Transformer-Based Physiological Signal Analysis for Continuous Health Monitoring

Slava Bergman
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
slava.online@uc.edu

Henrik Harrison
Department of Computer Science, University of Houston, Houston, TX, USA.
harrisonhenrik252@uh.edu

Abstract

Continuous health monitoring systems increasingly depend on high-frequency physiological signals collected from wearable, clinical, and ambient sensing infrastructures. The ability to extract clinically meaningful patterns from electrocardiograms, photoplethysmograms, electroencephalograms, and respiratory signals is central to early detection, remote care, and long-term disease management. Transformer-based architectures have attracted substantial attention in this domain because they can model long-range dependencies, integrate heterogeneous channels, and support transferable representations across different physiological modalities. However, translating these modeling advances into reliable continuous health monitoring requires careful consideration of system architecture, data governance, deployment constraints, robustness, fairness, and sustainability. This paper presents a system-level analysis of transformer-based physiological signal analysis. It examines architectural trade-offs between representational capacity and computational efficiency, discusses the role of data infrastructure and signal acquisition, and evaluates deployment patterns from cloud-centric to edge-assisted architectures. It further considers regulatory, ethical, and operational challenges that arise when machine learning systems are embedded in clinical and home settings. The discussion emphasizes that technical performance alone is insufficient for effective continuous health monitoring. Instead, sustainable deployment depends on integrated design choices that balance accuracy, latency, privacy, energy consumption, interpretability, and equity. The paper concludes by identifying directions for future research that align transformer-based physiological analysis with the broader requirements of health systems and long-term sociotechnical infrastructure.

Keywords

transformer architectures; physiological signal analysis; continuous health monitoring; system deployment; data governance; fairness; robustness; clinical infrastructure.

1. Introduction

Contemporary continuous health monitoring systems are built around the acquisition, transmission, and interpretation of physiological signals that reflect cardiovascular, respiratory, neurological, and metabolic states. These signals are often high-dimensional, nonstationary, and contaminated by motion artifacts, sensor drift, and environmental variability. Traditional signal processing pipelines rely on feature engineering, wavelet transforms, and domain-specific filters to reduce complexity before classification or regression. While such approaches have produced many clinically validated algorithms, they

often struggle with cross-patient variability, multimodal integration, and long-range temporal dependencies. The emergence of deep learning has shifted attention toward learnable representations, and the transformer architecture in particular has introduced a powerful set of mechanisms for modeling sequential and multivariate data. The core innovation of the transformer is its use of multi-head self-attention, which allows the model to attend to relevant time points and channels without relying on recurrence or convolution [1]. This design has been generalized through bidirectional pretraining strategies that learn contextual representations from large-scale unlabeled corpora [2]. The same attention mechanism has been successfully applied to visual data by dividing inputs into patches and treating them as tokens [3]. These developments have encouraged researchers to adapt transformer models to multivariate physiological time series, where patient states evolve over multiple interacting channels and require representations that can capture both local morphology and global temporal structure [4]. The ability to model long sequences has also been demonstrated in forecasting tasks, including influenza prevalence, and has motivated interest in applying such models to continuous physiological prediction [5].

Yet continuous health monitoring is not solely a modeling problem. It is a systems challenge that spans edge sensing, data transmission, cloud computation, clinical decision support, and long-term maintenance. The choices made at each level influence whether a high-performing model can be deployed safely and equitably in real-world settings. For example, a transformer model that achieves excellent arrhythmia detection in a curated dataset may be too computationally demanding for wearable hardware, may degrade under missing sensor data, or may produce systematically different performance across demographic groups. The integration of transformer-based physiological signal analysis into continuous monitoring therefore requires an interdisciplinary perspective that connects machine learning with biomedical engineering, software architecture, governance, and health policy. This paper examines these dimensions from a systems perspective. It discusses structural trade-offs in transformer architectures, data infrastructure requirements, deployment strategies, robustness and fairness concerns, and sustainability implications. The goal is to provide an integrated analysis that can guide researchers, engineers, and policymakers in designing systems that are not only accurate but also dependable, interpretable, and aligned with the operational constraints of healthcare infrastructures.

2. Background and System Context

The adoption of transformer models for physiological signal analysis occurs within a broader movement toward learning-based health monitoring. The technical motivation for this movement is rooted in the limitations of recurrence and convolution for very long sequences. Recurrent networks suffer from vanishing gradients and sequential computation bottlenecks, while convolutional networks require careful receptive field design to capture distant dependencies. Transformer variants have sought to address these limitations while controlling memory and computational cost. For instance, locality-aware transformations and log-sparse attention patterns have been proposed to improve efficiency for time series forecasting [6]. The Informer architecture further reduces complexity through probsparse self-attention and a generative decoder that directly produces long sequence outputs [7]. These efficiency-oriented designs demonstrate that the representational benefits of attention can be partially preserved without imposing prohibitive resource demands. Other architectural strategies have focused on memory compression and local attention windows. Reformer uses locality-sensitive hashing and reversible layers to reduce memory consumption in large-scale

sequence models [8]. Longformer similarly introduces a combination of sliding window attention and global attention to handle long documents and extended sequences [9]. These advances are relevant to continuous health monitoring because physiological records can span hours, days, or weeks, and the cost of processing such records must remain manageable.

From a systems perspective, these architectural improvements shape the viability of transformer models in clinical and home monitoring contexts. A model that requires high-end graphics processing units may be suitable for retrospective analysis in a centralized data center but unsuitable for ambulatory settings where computational resources are limited and real-time response is expected. At the same time, the selection of an efficient attention variant is not merely a technical optimization. It has implications for model interpretability, error propagation, and the ability to comply with regulatory requirements that demand auditable decision processes. The system context also includes the need to integrate multiple physiological modalities, each with distinct sampling rates, noise profiles, and clinical meanings. A transformer model that can attend across modalities may provide a unified representation of patient state, but this unification introduces new dependencies that must be carefully validated. The use of pretraining and transfer learning further complicates governance because representations learned from general unlabeled physiological data may carry hidden biases or may not align with the clinical constructs of interest. Thus, the background of transformer-based physiological analysis is characterized by a tension between representational power and operational accountability.

3. Transformer Architectures for Physiological Signal Analysis

The application of transformer models to physiological time series has evolved alongside broader research on deep time series classification. Early studies showed that relatively simple deep neural networks can serve as strong baselines for time series classification tasks when sufficient labeled data are available [10]. More comprehensive reviews have organized the landscape of deep learning methods for time series classification and identified important considerations such as data augmentation, model selection, and evaluation protocols [11]. In the healthcare domain, reviews of deep learning for physiological signals have highlighted the unique challenges posed by inter-subject variability, class imbalance, and the high cost of expert annotation [12]. These reviews emphasize that physiological signal analysis is not a homogeneous problem but rather a family of tasks with different sensing modalities, clinical objectives, and tolerance for error. The broader adoption of artificial intelligence in healthcare has been influenced by the availability of large datasets, improvements in computing infrastructure, and the development of algorithms that can learn directly from raw or minimally processed signals [13]. Transformer architectures fit within this trajectory because they can operate on raw tokenized signal segments and can be adapted to diverse tasks through self-supervised pretraining and fine-tuning.

The role of public datasets has been particularly important for advancing transformer-based physiological analysis. The MIT-BIH Arrhythmia Database, for example, has served as a benchmark for electrocardiogram classification for decades and has shaped the development of many algorithmic approaches [14]. The PhysioNet research resource has further expanded access to complex physiologic signals and has enabled reproducible studies across institutions [15]. These data resources provide the foundation for training and evaluating transformer models, but they also introduce questions about representativeness and generalization. Many public datasets are collected under controlled conditions and may not capture the variability encountered in continuous home monitoring. Transformer models trained on such datasets

may learn features that are highly predictive within the benchmark but less robust in the presence of motion artifacts, changing electrode placement, or heterogeneous skin types. The architectural challenge is therefore not only to model long-range dependencies but also to ensure that these dependencies reflect genuine physiological structure rather than dataset-specific artifacts. Attention mechanisms can potentially learn spurious correlations, especially when the training data contain subtle confounders. From a systems perspective, this creates a need for evaluation frameworks that move beyond aggregate accuracy and consider calibration, subgroup performance, and robustness under distribution shift.

4. Data Infrastructure and Signal Acquisition

Continuous health monitoring depends on a data infrastructure that extends from sensors on the body to storage and processing systems in hospitals, clinics, or cloud environments. The signal acquisition layer determines the quality, continuity, and context of the data that transformer models ultimately process. Recent work in opto-physiological monitoring has demonstrated rapid vital sign extraction under varying physical activity intensity levels, indicating that wearable optical sensors can provide clinically useful information outside controlled settings [16]. Such methods illustrate the potential for unobtrusive monitoring, but they also raise questions about signal quality during motion, ambient light variation, and differences in skin tone. Gradient boosting models have been widely used as scalable baselines for structured health data and provide an important comparison point for deep learning systems when tabular features are extracted from physiological signals [17]. Deep learning approaches have achieved high accuracy in cardiology tasks, as shown by convolutional neural networks that classify arrhythmias from ambulatory electrocardiograms at a level comparable to cardiologists [18]. Similarly, artificial intelligence models applied to electrocardiograms have been used to identify patients with atrial fibrillation during normal sinus rhythm, demonstrating the potential to detect latent disease before it becomes symptomatic [19]. These successes highlight the clinical value of learning-based physiological analysis, but they also depend on high-quality electrocardiographic data and substantial computational resources for training.

For continuous monitoring, the data infrastructure must address more than model training. It must support real-time ingestion, missing data handling, versioning, consent management, and secure transmission. Transformer models are particularly sensitive to input tokenization, sequence length, and missing values, which means that the signal acquisition pipeline must be designed with model requirements in mind. For example, electroencephalogram decoding with convolutional neural networks has shown that deep architectures can learn representations directly from raw signals, but the performance depends on careful preprocessing and artifact rejection [20]. In transformer-based systems, similar preprocessing steps must be transformed into tokenization strategies that preserve phase, amplitude, and temporal relationships. The infrastructure must also support multimodal alignment when electrocardiogram, photoplethysmogram, accelerometry, and respiratory signals are collected simultaneously. This alignment is nontrivial because different sensors have different sampling rates and time delays. A well-designed data layer can provide synchronized windows, quality flags, and contextual metadata that improve model robustness. Without such infrastructure, transformer models may learn from misaligned or artifact-heavy data and produce predictions that are difficult to interpret clinically.

5. Deployment Architectures for Continuous Monitoring

The deployment of transformer-based physiological signal analysis must account for the constraints of clinical workflows, home monitoring programs, and remote patient management. One important consideration is the use of transfer learning, which allows models pretrained on large datasets to be adapted to smaller clinical cohorts. Transfusion-style transfer learning has shown that pretraining on general data can improve performance on medical tasks when the source and target domains are carefully matched [21]. However, the generalizability of clinical machine learning models is often limited by differences in patient populations, equipment, and data collection protocols [22]. Continuous health monitoring amplifies these concerns because the model must operate over time and across changing environments. A deployment architecture that relies on frequent retraining may be unsustainable, while a static model may become increasingly inaccurate as the patient population or device characteristics evolve. The convergence of human and artificial intelligence in high-performance medicine suggests that algorithms should be integrated into decision support systems rather than treated as autonomous diagnostic tools [23]. This perspective has direct implications for deployment, because clinical users need transparent indications of model confidence, data quality, and the presence of distribution shift. A transformer model that produces a continuous risk score without contextual information may be less useful than a system that also communicates uncertainty and explains which input segments contributed to the prediction.

Deployment architectures range from cloud-centric designs to edge-native and hybrid approaches. Cloud-centric designs benefit from scalable computation and central model management, but they raise concerns about data privacy, network reliability, and latency. Edge-assisted architectures can perform initial preprocessing, artifact detection, and lightweight inference on the device, while more complex transformer layers run in the cloud. This hybrid pattern reduces the amount of raw physiological data that must be transmitted and can improve responsiveness in critical scenarios. However, it also introduces engineering complexity, because the model may need to be split across heterogeneous hardware with different memory and energy budgets. The choice of deployment architecture is therefore not merely a technical decision but a governance decision that affects the distribution of risk among patients, clinicians, device manufacturers, and cloud providers. For example, if a cloud service fails during a cardiac monitoring session, the responsibility for missed detections must be clearly defined. In addition, regulatory frameworks may require that the system be validated as a whole, including sensor hardware, preprocessing software, transformer model, and clinical interface. These system-level dependencies mean that model developers must work closely with hardware engineers, clinical informaticians, and regulatory specialists from the earliest stages of design.

6. Robustness, Fairness, and Clinical Governance

Robustness and fairness are central to the responsible deployment of transformer-based physiological signal analysis in continuous health monitoring. Machine learning models can inadvertently learn biases from historical data, and these biases may lead to unequal performance across demographic groups, clinical conditions, or device types. The healthcare community has articulated a roadmap for responsible machine learning that emphasizes the need for external validation, proactive bias assessment, and transparent reporting of limitations [24]. Transformer models, with their large parameter counts and complex attention distributions, present specific challenges for interpretability and auditability. Although attention weights can provide some insight into which parts of an input are influential, they do

not constitute a complete explanation of model behavior. Clinical governance structures must therefore require more rigorous evaluation, including subgroup analyses, calibration checks, and stress tests under missing data and sensor degradation. The governance of machine learning in medicine also requires clarity about the role of the algorithm in clinical decision making, the criteria for overriding model recommendations, and the procedures for monitoring performance after deployment [25]. Continuous health monitoring systems add a further layer of governance complexity because they generate longitudinal predictions that may influence patient behavior and clinical decisions over extended periods. A model that produces frequent false alarms may lead to alarm fatigue, while a model that misses subtle deterioration may create false reassurance. The design of alarm thresholds, notification workflows, and clinician review mechanisms is as important as the underlying model architecture.

Fairness considerations are also shaped by the data infrastructure and deployment context. Physiological signals are influenced by age, sex, skin pigmentation, body habitus, and comorbidities, all of which can affect signal quality and model performance. If transformer models are trained predominantly on data from one population, they may not generalize to other populations, even when the architecture is highly expressive. System designers must therefore invest in representative data collection, transparent documentation of dataset characteristics, and continuous monitoring for performance disparities. This is not a one-time task but an ongoing process that requires collaboration among data scientists, clinicians, ethicists, and patient representatives. The regulatory environment is also evolving, with increasing attention to algorithmic accountability in medical devices and clinical decision support systems. Transformer-based physiological monitoring systems may be subject to different rules depending on whether they are classified as diagnostic devices, wellness applications, or clinical decision support tools. The distinction between these categories can be ambiguous for continuous monitoring systems that provide early warnings rather than formal diagnoses. Clear governance frameworks are needed to define the legal responsibilities of developers and operators, particularly when models are updated continuously or personalized to individual patients.

7. Sustainability and Resource Management

The sustainability of transformer-based continuous health monitoring depends on managing computational, financial, and environmental resources over the full lifecycle of the system. Large transformer models consume significant energy during training, and the carbon footprint of repeated retraining can be substantial. Although the operating phase of a deployed model may be less energy intensive, continuous monitoring involves frequent inference over long periods, and the cumulative cost may be considerable. Efficient attention mechanisms such as those introduced in Informer, Reformer, and Longformer reduce memory and computational requirements [7, 8, 9]. These optimizations are not only technical conveniences; they are essential for enabling transformer models to run on edge devices with limited batteries and for reducing the environmental impact of large-scale health monitoring programs. From a systems perspective, sustainability also includes the human and institutional capacity to maintain the monitoring infrastructure. A health system may lack the data engineering staff, computational resources, or clinical informatics expertise needed to support a sophisticated transformer-based system over many years. The choice of model complexity should therefore be aligned with the operational maturity of the deploying organization.

Resource management also intersects with data governance and privacy. Federated learning and distributed training can reduce the need to centralize sensitive physiological data, but they introduce their own computational and coordination costs. The use of synthetic data and data augmentation can improve model robustness, but synthetic signals must be carefully validated to avoid introducing artifacts. Another sustainability concern is model drift, which occurs when the relationship between input signals and clinical outcomes changes over time due to shifts in sensor hardware, patient behavior, or clinical protocols. Continuous monitoring systems must include mechanisms for detecting drift and deciding when retraining is warranted. However, frequent retraining may not be feasible in low-resource settings, and a poorly timed update may disrupt clinical workflows. System designers can address these issues by building modular architectures that separate stable feature extractors from task-specific heads, thereby reducing the scope of updates. They can also adopt model compression techniques that lower the computational burden of inference while preserving clinically relevant accuracy. Ultimately, sustainability requires a planning process that considers not only the accuracy of the transformer model but also the long-term reliability, maintainability, and equity of the entire monitoring system. A technically impressive model that cannot be maintained or afforded by the health system is unlikely to produce sustained clinical benefit.

8. Conclusion

Transformer-based physiological signal analysis offers substantial promise for continuous health monitoring by enabling the integration of long temporal contexts, multiple physiological modalities, and transferable representations. However, the successful translation of these models into health systems depends on a broad set of system-level considerations that extend beyond architectural innovation. This paper has examined the structural trade-offs among expressive attention mechanisms and computational efficiency, the foundational role of signal acquisition and data infrastructure, the implications of deployment architectures, and the governance requirements associated with robustness, fairness, and clinical accountability. It has further considered sustainability as a first-order design constraint rather than an afterthought. The analysis indicates that future research should move toward integrated system design in which transformer models are developed alongside the sensing, computing, and governance layers that define their operational context. Such an approach requires interdisciplinary collaboration across machine learning, biomedical engineering, clinical informatics, regulatory science, and health policy. It also requires evaluation frameworks that capture not only aggregate predictive performance but also subgroup equity, calibration under distribution shift, energy consumption, latency, and long-term reliability. By adopting this broader perspective, researchers and practitioners can build transformer-based monitoring systems that are not only accurate but also trustworthy, equitable, and sustainable within the complex sociotechnical infrastructure of healthcare.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

2. 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.

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

4. Zerveas, G., Jayaraman, S., Patel, D., Bhamidipati, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2114-2124.

5. Wu, N., Green, B., Ben, X., & O'Banion, S. (2020). Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv:2001.08317.

6. Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Advances in Neural Information Processing Systems, 32, 5243-5253.

7. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106-11115.

8. Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. International Conference on Learning Representations.

9. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

10. Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. 2017 International Joint Conference on Neural Networks, 1578-1585.

11. Ismail Fawaz, H., 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. Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine, 161, 1-13.

13. Rong, G., Mendez, A., Bou Assi, E., Zhao, B., & Sawan, M. (2020). Artificial intelligence in healthcare: Review and prediction case studies. Engineering, 6(3), 291-301.

14. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45-50.

15. Goldberger, A. L., Amaral, L. A., 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.

16. 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.

17. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

18. Hannun, A. Y., Rajpurkar, P., Haghpanahi, 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.

19. Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S., & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm. The Lancet, 394(10201), 861-867.

20. Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping, 38(11), 5391-5420.

21. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems, 32, 3347-3357.

22. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health, 2(9), e489-e492.

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

24. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

25. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.