

Foundation Model-Based Physiological Signal Representation for Generalizable Wearable Health Applications

Kasper Carlson

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.

contactkasper@oregonstate.edu

Keith Schmidt

Department of Computer Science, University of Houston, Houston, TX, USA.

contactkeith@uh.edu

Dylan C. Castro

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

castrodylan@unr.edu

Yogesh Siddiqui

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

siddiquiyogesh27@unh.edu

Abstract

Wearable health devices generate continuous, multimodal physiological signals that are noisy, context-dependent, and highly heterogeneous across individuals and sensing hardware. Foundation models offer a promising route toward generalizable physiological signal representation by learning transferable features from large-scale self-supervised pretraining. However, translating this promise into real-world wearable health applications requires careful attention to architecture, data governance, deployment infrastructure, robustness, fairness, sustainability, and regulatory oversight. This paper presents a system-level analysis of foundation model-based physiological signal representation. It examines the structural trade-offs between centralized and federated learning, the balance between representation capacity and on-device efficiency, and the challenges of distribution shift induced by motion, sensor placement, skin tone, and disease prevalence. The discussion integrates insights from language and vision foundation models, medical artificial intelligence, and wearable sensing. It emphasizes that physiological foundation models are not merely computational artifacts but socio-technical infrastructures that require lifecycle governance. The paper argues for a coordinated design perspective that couples representation learning with transparency mechanisms, energy-aware model updates, equity auditing, and clinical validation. Such a perspective is necessary to ensure that wearable health foundation models achieve generalizability without exacerbating existing disparities or creating unmanageable operational burdens.

Keywords

foundation models; physiological signal representation; wearable health; self-supervised learning; federated learning; fairness; sustainability.

1. Introduction

Foundation models have fundamentally altered the machine learning landscape by learning broad representations that can be adapted to a wide array of downstream tasks. The computational core of this shift is the attention mechanism, which allows models to capture long-range dependencies in sequential and spatial data [1]. Although attention was initially developed for natural language processing, its generality has encouraged the construction of large models that serve as reusable representation backbones across modalities. The resulting systems are often described as foundation models because they provide a base layer of learned structure that can be fine-tuned, prompted, or otherwise specialized [2]. In healthcare, this approach has generated considerable interest because clinical data are often scarce in individual domains, while shared physiological and anatomical structure exists across tasks. Generalist medical artificial intelligence systems have begun to show that pretraining on large and heterogeneous health data can support effective adaptation to multiple diagnostic and monitoring tasks [3]. At the same time, clinical machine learning has matured through narrow deep learning systems that achieve expert-level performance for particular tasks, such as imaging and electrocardiogram interpretation, but these systems often struggle when deployed outside the distributions represented in their training sets [4], [5].

Wearable health applications introduce a distinctive set of demands that differ from static clinical datasets. Wearable sensors collect photoplethysmography, electrocardiography, electrodermal activity, accelerometry, temperature, and other modalities under free-living conditions. The resulting signals are continuously changing, heavily affected by motion artifacts, sensor displacement, ambient conditions, and physiological variability across individuals. A foundation model for physiological signal representation must therefore internalize not only disease-related patterns but also nuisance variations related to activity, sensor configuration, and demographic context. This requires a system-level view that moves beyond isolated model architectures and toward the design of representational infrastructures that integrate data collection, governance, deployment, and updating. The present paper examines these broader infrastructures. It focuses on the structural trade-offs that shape whether physiological foundation models can become reliable, generalizable, and equitable components of wearable health systems. It does not propose a single architecture, but instead analyzes the architectural, governance, robustness, fairness, sustainability, and policy dimensions that jointly determine the success of these models in practice.

2. Background and Motivations

Representation learning in artificial intelligence has been strongly influenced by self-supervised objectives that derive training signals from the structure of the data itself. In language, bidirectional transformer models demonstrated that masked token prediction can produce representations that transfer to numerous downstream tasks, reducing the need for large task-specific datasets [6]. Subsequent work showed that scaling language models leads to emergent few-shot capabilities, in which a single pretrained model can perform tasks described in natural language without additional gradient updates [7]. This pattern motivated researchers to test whether similar scaling and transfer effects could be achieved for visual data. Vision transformers treat images as sequences of patches and apply attention mechanisms that were originally designed for text, allowing visual models to capture long-range spatial dependencies [8]. Contrastive learning methods further advanced this program

by aligning representations of different views of the same image, thereby learning invariances that support downstream classification with limited labels [9]. Multimodal models then connected image and text representations, enabling zero-shot transfer to new concepts by leveraging natural language supervision [10].

These successes have generated a strong pull toward applying foundation models to physiological signals. However, physiological data present different distributional and semantic properties. A single wearable record may contain multiple interacting processes, including cardiac rhythms, respiratory modulation, autonomic stress responses, and physical motion. The relationship between the observed signal and the underlying health state is indirect and can vary with sensor placement, skin properties, environmental conditions, and individual behavior. Benchmarks for in-the-wild distribution shifts have shown that machine learning models often degrade when the deployment environment differs from training, which is a central challenge for wearable systems that must operate across many contexts [11]. Privacy and data governance also differ from public language or vision corpora. Federated learning has been proposed as a strategy for training models across distributed medical institutions and wearable devices without centralizing raw physiological data [12]. These developments set the stage for foundation models in wearable health, but they also make clear that the representation problem cannot be separated from the infrastructure in which the model operates.

3. Architectural Considerations for Physiological Foundation Models

The architecture of a physiological foundation model must reconcile the need for expressive sequence modeling with the constraints of wearable deployments. Attention-based architectures are attractive because they can learn dependencies across time, sensor channels, and contextual variables. A recent development in opto-physiological monitoring uses joint attention mechanism learning to maintain signal quality during physical activity, which is one of the most demanding conditions for wearable sensing [13]. This line of work illustrates that attention can be directed toward multiple signal components, such as cardiac and motion features, to suppress artifacts and preserve physiologically meaningful information. The principle extends beyond photoplethysmography to electrocardiography, impedance cardiography, and multimodal sensor fusion. In each case, the model must determine which parts of the signal are informative under changing conditions rather than relying on fixed frequency bands or handcrafted features.

The clinical success of deep representation learning in related domains provides useful comparisons. In dermatology, a deep neural network trained on skin images achieved dermatologist-level classification for selected conditions, showing that learned visual representations can capture diagnostic patterns at scale [14]. In cardiology, a deep neural network trained on ambulatory electrocardiograms attained cardiologist-level arrhythmia detection across multiple rhythm classes [15]. Similarly, an artificial intelligence-enabled electrocardiogram was shown to screen for cardiac contractile dysfunction using standard ten-second recordings [16]. These achievements were largely task-specific, but they indicate that physiological signals contain rich features that can be exposed through deep networks. A foundation model for wearables would aim to unify many such tasks within a single representational space. The challenge is architectural: the model must handle variable sampling rates, irregular missingness, multimodal alignment, and low-power inference without sacrificing the representational breadth that makes foundation models useful.

Several structural trade-offs follow from this ambition. A fully centralized transformer with many layers may learn strong representations but is difficult to deploy on wearable processors and may incur high communication costs. Hierarchical architectures can separate a shared backbone from lightweight task heads, allowing the backbone to be compressed or distilled for edge devices. Temporal downsampling and channel grouping can reduce memory usage, but they may also obscure short-duration events such as arrhythmias or abrupt motion artifacts. Cross-modal attention can align physiological signals with activity and context, but it increases data requirements and may introduce additional privacy concerns if contextual labels are highly identifying. Therefore, architectural design cannot be evaluated solely by benchmark accuracy. It must also be evaluated by energy use, latency, memory footprint, robustness to missing channels, and the ease with which models can be updated across a fleet of heterogeneous devices.

4. Data Governance and Representational Quality

The quality of a physiological foundation model depends on the quality and governance of the data used for pretraining and adaptation. Physiological datasets vary widely in sampling rates, sensor types, clinical annotations, demographic composition, and collection protocols. A model trained predominantly on young, healthy, or institution-specific cohorts may learn representations that do not transfer to older adults, people with chronic conditions, or populations using different wearable platforms. Transparency about data collection and intended use is therefore essential. Datasheets for datasets provide a structured mechanism for documenting the motivation, composition, collection process, preprocessing, and limitations of a dataset [17]. When applied to wearable physiological data, such documentation can reveal whether a dataset contains sufficient variability in skin tone, body mass, medication use, activity patterns, and sensor placement. It can also expose gaps in disease prevalence or recording duration that would otherwise remain hidden in aggregate performance metrics.

Model documentation plays a complementary role. Model cards encourage developers to report intended use, performance across subgroups, evaluation datasets, and known limitations [18]. For wearable health models, model cards should report performance under motion, across sensor manufacturers, and across demographic strata rather than only in quiet laboratory conditions. Data governance must also address consent, secondary use, and device-level data rights. Federated learning can reduce the need to centralize raw physiological data, but it does not automatically solve governance problems because model updates can still leak information and because participation across institutions may be uneven [12]. A robust governance framework therefore links representational quality to downstream fairness, safety, and accountability. Without such linkage, a foundation model may appear generalizable in internal evaluation while embedding systematic blind spots that emerge only after deployment.

5. Infrastructure and Deployment Trade-offs

Deploying physiological foundation models in wearable health systems involves a hierarchy of computational environments. Cloud resources can host large pretraining runs and periodic fine-tuning, while edge devices such as smartwatches and chest patches require compact inference. The partition between cloud and edge is not simply an engineering detail; it determines how quickly a model can adapt to new users, how much raw data must be transmitted, and how vulnerable the system is to connectivity loss. A foundation model pretrained in the cloud may be distilled into a smaller edge model that retains a shared representational core. Alternatively, a federated architecture can allow models to be updated on devices without sending raw signals to a central server [12]. In either case, the

representation itself must be robust to compression because aggressive quantization and pruning can destroy the subtle temporal relationships that encode physiological events.

The choice of deployment architecture also affects update cadence and regulatory control. A centralized model can be patched and validated uniformly, but it may be slow to adapt to individual user drift. A personalized edge model can track individual baselines and respond to changes in activity or health status, but it raises questions about provenance and reproducibility when different devices run different model versions. Hybrid approaches are possible, in which a shared foundation model extracts general features and personalization layers adapt them on-device. This structure resembles a socio-technical assembly rather than a fixed product: the shared component must remain stable enough for clinical validation, while the personalized component must be allowed to evolve within bounded safety envelopes. For wearable health applications, the management of these boundaries is as important as the representational architecture itself.

Infrastructure decisions also determine resilience. Wearable health systems may operate during network outages, in low-resource settings, and across long battery lifetimes. If inference requires frequent cloud calls, the system becomes fragile in exactly the situations where continuous monitoring might be most valuable. Attention-based representations can be designed to support hierarchical inference, in which lightweight local models raise alerts and heavier cloud models provide detailed analysis only when needed [13]. This trade-off between immediate local awareness and deferred high-resolution analysis is central to sustainable deployment. It requires not only architectural choices but also organizational arrangements for monitoring performance, managing device heterogeneity, and ensuring that critical alerts are not lost due to infrastructure bottlenecks.

6. Robustness, Generalization, and Distribution Shift

Generalization in wearable health is best understood as robustness to distribution shift. Training distributions often contain clean laboratory recordings, whereas deployment involves free-living motion, variable sensor contact, and diverse ambient conditions. Benchmarks for in-the-wild distribution shifts have shown that models trained under one set of conditions frequently fail when tested under another [11]. Physiological foundation models can mitigate this problem by learning representations that are invariant to nuisance factors while preserving health-relevant features. Attention mechanisms are particularly suited to this goal because they can weight different temporal segments and channels according to current context by assigning higher relevance to signal segments and channel configurations that are most informative under a given activity state. From a systems perspective, this is not merely a filtering operation; it is a representational selection process that determines which physiological variations are retained in the learned embedding and which are suppressed as nuisance. The success of this process depends on the diversity of contexts seen during pretraining. If a model sees only gentle walking and resting states, it may learn to ignore motion frequencies that are essential for detecting effort-related cardiovascular changes during running or resistance exercise. Similarly, if the model is trained predominantly on wrist-worn devices, the representation may fail to transfer to chest patches or earbuds where the signal amplitude and noise structure differ. Generalization therefore requires evaluating the model under sensor-level and activity-level distribution shifts, not simply reporting average performance across a static held-out set.

Robust evaluation protocols for physiological foundation models should include prospective validation on unseen devices, controlled perturbation studies, and continuous monitoring

trials across multiple days. Out-of-distribution detection is equally important: the model should indicate when a signal falls outside its representational coverage rather than silently producing a confident but unreliable embedding. Such uncertainty-aware design is essential for safety-critical wearable applications because downstream classifiers may otherwise propagate errors into clinical alerts. Domain adaptation strategies can extend coverage, but they introduce additional governance questions about how much local personalization is permitted before a model loses the shared representational properties that made it manageable. Structural trade-offs between global stability and local adaptability must be resolved through explicit versioning, audit trails, and performance monitoring. In this respect, physiological foundation models resemble distributed software systems more than static diagnostic tools.

7. Fairness and Equity in Physiological Representation

The risk of representational bias is particularly acute in wearable health models because physiological signals are shaped by demographic, environmental, and social factors that are often correlated with protected attributes. A model pretrained on convenience samples may learn features that perform well for majority groups while failing for minority groups whose signal characteristics differ because of skin pigmentation, body composition, chronic conditions, or medication use. Fairness in machine learning is a multidimensional concept, and no single statistical criterion is sufficient for clinical deployment [19]. For wearable physiological models, equality of false negatives across groups may be more important than overall accuracy because a missed arrhythmia or apnea event has immediate safety implications. However, enforcing group fairness constraints can conflict with individual-level calibration, and the appropriate trade-off must be informed by clinical priorities and community engagement.

A cautionary example from health care resource allocation demonstrated that an algorithm using health costs as a proxy for health needs systematically underestimated illness severity among Black patients despite excluding race as an explicit input [20]. This illustrates how structural bias can enter through proxy variables and outcome labels even when raw demographics are not used. In wearable health, similar dynamics may arise if self-supervised objectives prioritize reconstruction of signals that are easier to measure in some skin tones or if annotations for stress, sleep, or arrhythmia are unevenly distributed across populations. The representational harm is compounded because the foundation model may absorb these biases into its embedding space, making them harder to detect and correct downstream. Datasheets and model cards remain important governance instruments, but they must be accompanied by regular subgroup audits and independent equity assessments that examine performance across intersecting axes of disadvantage. Participatory design with affected communities can also reveal representational requirements that are not visible in aggregated datasets, such as the need to account for cultural differences in activity patterns or the impact of chronic illness on baseline physiology. In this way, fairness is not a post hoc metric but a design constraint that shapes data collection, pretraining objectives, evaluation, and deployment.

8. Sustainability and Lifecycle Management

The environmental cost of foundation models has become a major concern in natural language processing, where training large transformer models can consume substantial electricity and emit significant carbon dioxide [21]. In response, the Green AI movement has called for evaluating machine learning systems not only by accuracy but also by energy, computational, and carbon efficiency [22]. Wearable health applications present a paradoxical situation. On one hand, pretraining a large physiological model in the cloud may be expensive;

on the other hand, once trained, the model may enable efficient on-device inference that reduces the need to transmit and process raw data continuously. This can lower the energy footprint of long-term monitoring, particularly if hierarchical inference sends only selected high-value segments to the cloud [13]. Still, the sustainability equation must account for the entire lifecycle: data collection, preprocessing, pretraining, fine-tuning, inference, periodic updates, and eventual model retirement. A foundation model that is updated frequently across millions of devices may incur substantial communication and computational overhead, even if individual inference steps are efficient.

Lifecycle management also includes the material and social sustainability of wearable hardware. If representational innovations require frequent replacement of sensors or processors, the environmental burden shifts from software energy use to electronic waste and manufacturing emissions. The design of physiological foundation models should therefore consider model distillation, dynamic channel selection, and adaptive sampling to extend the usable life of existing devices. Federated learning offers a potential path for reducing data center energy demand by keeping computation closer to the data source, but it also introduces transmission costs associated with aggregate model updates [12]. A comprehensive sustainability assessment must therefore examine trade-offs across compute, communication, data storage, and hardware longevity. Institutions that operate wearable health platforms can establish sustainability metrics that are reported alongside performance metrics, enabling more transparent comparisons between model architectures and update regimes. This shift from accuracy-only evaluation to multi-objective lifecycle evaluation is necessary for the responsible scaling of physiological foundation models.

9. Policy, Regulation, and Clinical Governance

Machine learning systems in medicine must navigate regulatory frameworks that were largely designed for static devices and discrete software versions [23]. A foundation model for wearable health challenges these frameworks because its representations may be continuously adapted through fine-tuning or federated updates. The Food and Drug Administration has proposed a regulatory framework for artificial intelligence and machine learning-based software as a medical device, emphasizing predetermined change control plans and post-market performance monitoring [24]. Such frameworks are relevant to physiological foundation models because they force developers to define which types of model updates are allowed, how safety is verified, and how adverse events are reported. However, the pretrained foundation layer may remain largely stable while downstream heads change frequently, raising the question of which components require full regulatory review and which can be updated under a less burdensome pathway.

Transparency and accountability are equally important. Responsible machine learning frameworks for healthcare emphasize the need for prospective validation, bias assessment, and clinician involvement before deployment [25]. For wearable health, these principles must extend beyond hospital-based evaluations to the nuanced settings where users live, exercise, sleep, and work. Governance structures should include independent audits of subgroup performance, incident reporting mechanisms, and clear procedures for model recall if safety thresholds are exceeded. Interoperability standards are also needed so that representations generated by different devices and vendors can be compared, transferred, and audited without proprietary lock-in. Without such standards, the generalizability of foundation models may be undermined by fragmentation in the wearable ecosystem. Policy makers, regulators, clinicians,

engineers, and patient communities must collaborate to ensure that representation learning advances do not outpace the governance mechanisms that make them trustworthy.

10. Conclusion

Foundation model-based physiological signal representation holds substantial promise for generalizable wearable health applications, but realizing this promise requires more than architectural innovation. It requires a coordinated design perspective that integrates representation learning with data governance, deployment infrastructure, robustness to distribution shift, fairness auditing, sustainability planning, and regulatory oversight. The structural trade-offs examined in this paper reveal that no single technical choice is sufficient. Attention-based backbones may improve robustness to motion, but they demand careful compression and hierarchical inference for edge deployment. Federated learning may protect data locality, but it complicates model versioning and auditability. Data documentation and model cards may improve transparency, but they must be enforced through institutional practices and regulatory requirements. Fairness and sustainability cannot be treated as secondary constraints because they shape the very representational space in which wearable health systems operate. As physiological foundation models move from research prototypes into everyday health infrastructures, interdisciplinary collaboration will be essential. The future of wearable health depends not only on the scale of pretraining or the expressiveness of attention mechanisms, but on the capacity of the broader system to remain safe, equitable, and sustainable across diverse populations and real-world conditions.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
2. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
3. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259-265.
4. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
5. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186.
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan,

- T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
8. 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*.
 9. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 1597-1607.
 10. 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. *Proceedings of the 38th International Conference on Machine Learning*, 8748-8763.
 11. Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I. S., Beery, S. M., Leskovec, J., Kundaje, A., ... Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of the 38th International Conference on Machine Learning*, 5637-5664.
 12. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 119.
 13. 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.
 14. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.
 15. 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.
 16. Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., Carter, R. E., & Friedman, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine*, 25(1), 70-74.
 17. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
19. 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.
20. 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.
21. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645-3650.
22. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54-63.
23. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
24. Food and Drug Administration. (2021). Proposed regulatory framework for modifications to artificial intelligence/machine learning based software as a medical device. U.S. Department of Health and Human Services.
25. 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., Thadane-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337-1340.