Semantic-Aware Approximate Nearest Neighbor Search for Personalized Cardiovascular Monitoring Using PPG Foundation Models

Abhishek M. Srinivasan
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
abhishekwork@colostate.edu

Ganjemin Rrawford
Department of Computer Science, University of Houston, Houston, TX, USA.
grawford1981@uh.edu

Vikram R. Batra
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
vikram.work@missouri.edu

Abstract

Photoplethysmography has emerged as a ubiquitous modality for ambulatory cardiovascular assessment, but its full clinical potential remains constrained by the heterogeneity of signal morphology across diverse populations, sensor configurations, and pathophysiological manifestations. Foundation models pre-trained on large-scale photoplethysmographic repositories offer a unified representational space that can capture subtle physiological signatures, yet the challenge of efficiently retrieving clinically relevant exemplars from massive embedding databases for personalized inference has not been addressed. This paper introduces a semantic-aware approximate nearest neighbor search framework tailored to cardiovascular monitoring tasks that leverages photoplethysmography foundation model embeddings. The architectural design couples deep semantic hashing with graph-based indexing to enable millisecond-latency retrieval of diagnostically similar physiological states while preserving clinically meaningful semantics. We systematically analyze the trade-offs between retrieval accuracy, computational efficiency, and interpretability through the lens of multi-level infrastructure spanning edge wearable devices, fog gateways, and cloud-based model repositories. Critical considerations surrounding robustness to distributional shift, demographic fairness in semantic similarity spaces, differential privacy during query execution, and sustainable model lifecycle management are examined. The paper further explores the governance structures required to maintain trustworthiness when personalized retrieval systems operate across health system boundaries, and proposes a policy-oriented architecture that embeds auditability and federated accountability mechanisms directly into the retrieval pipeline. Our analysis suggests that semantic-aware approximate nearest neighbor search, when integrated with structured fine-grained access controls and continuous monitoring for concept drift, can serve as a pivotal enabling technology for the next generation of equitable and efficient cardiovascular digital twins.

Keywords

Approximate nearest neighbor search; semantic hashing; photoplethysmography; foundation models; personalized cardiovascular monitoring; fairness-aware retrieval; edge-cloud infrastructure; health data governance.

1. Introduction

Cardiovascular diseases remain the leading cause of global mortality, driving an urgent need for continuous, non-invasive monitoring technologies that can detect early signs of decompensation and guide personalized interventions. Photoplethysmography (PPG) has become the modality of choice for consumer and clinical wearables due to its low cost, ease of acquisition, and rich information content related to cardiac timing, vascular tone, and autonomic regulation [1, 2]. However, the raw photoplethysmographic signal is highly sensitive to skin pigmentation, motion artifacts, sensor placement, and inter-individual hemodynamic differences, making population-level diagnostic models brittle and prone to failure in underrepresented subgroups. In recent years, foundation models have begun to reshape biomedical time-series analysis by learning universal representations from vast and diverse datasets, enabling downstream tasks to be addressed with limited labeled data and improved generalization [3]. When applied to PPG, these models promise to extract robust latent features that abstract away nuisance variations while preserving clinically actionable physiological signatures [4, 5]. Yet a fundamental system-level challenge remains largely unaddressed: how to efficiently retrieve the most relevant historical physiological states from a continuously growing, multi-institutional embedding database to inform real-time, patient-specific cardiovascular decisions.

Semantic-aware approximate nearest neighbor (ANN) search provides a natural computational framework for this retrieval problem. Departing from traditional distance-based similarity search, semantic-aware approaches seek to align the notion of proximity in the embedding space with clinically meaningful concepts such as arrhythmia subtype, hemodynamic compensation status, or chronotropic response. This alignment is crucial for building trust in automated retrieval systems that operate in safety-critical domains. The deployment of such systems, however, cuts across a complex socio-technical fabric involving edge accelerators on wearables, hospital edge-cloud gateways, and centralized model registries, each introducing distinct constraints on latency, energy consumption, and data residency [6, 7]. Semantic hashing and graph-based indexing have independently demonstrated their effectiveness in accelerating retrieval while respecting semantic boundaries, and their combination with foundation model representations opens a design space that has not been systematically explored for cardiovascular monitoring [8]. Furthermore, the introduction of statistical-prior informed generative masking architectures for PPG foundation models has recently expanded the representational fidelity achievable in this domain, creating a renewed impetus for building retrieval infrastructure that can fully exploit such representations [9].

This paper offers a comprehensive systems-oriented examination of semantic-aware ANN search in the context of personalized cardiovascular monitoring using PPG foundation models. We do not propose a new algorithmic method; rather, we interrogate the architectural, infrastructural, and governance dimensions of integrating semantic retrieval into real-world health ecosystems. The discussion is organized around a conceptual platform that pipelines PPG signals through a foundation model encoder, projects the resulting embeddings into a semantic hashing space optimized for clinical relevance, indexes them with a navigable graph structure, and exposes a query interface that supports personalized retrieval under differential

privacy guarantees. Throughout this analysis, we emphasize the structural trade-offs that arise when balancing retrieval recall against computational cost; fairness across demographic strata; robustness under wearable noise and temporal drift; and the policy frameworks needed for cross-jurisdictional deployment. By treating semantic-aware ANN not as an isolated algorithmic component but as an infrastructural capability, we aim to bridge the gap between deep learning research and the operational realities of equitable cardiovascular care.

2. Background and Related Work

The evolution of PPG analysis has progressed from hand-engineered features and classical signal processing to deep learning architectures capable of end-to-end extraction of heart rate variability, respiratory rate, and blood pressure trends [2]. The limitations of fully supervised models trained on homogeneous datasets have motivated a shift toward self-supervised representation learning, where models are pre-trained on unlabeled PPG data using contrastive objectives that encourage invariance to domain-specific distortions while preserving temporal coherence. Contrastive learning of cardiac signals across space, time, and patients, as well as general temporal contrasting frameworks, have demonstrated that features learned without manual annotation can rival or exceed the performance of traditional features on downstream tasks including atrial fibrillation detection and sleep staging [4, 5]. Building on these foundations, the broader adoption of foundation models across multiple modalities has inspired efforts to pre-train large-scale PPG encoders that can serve as a universal backbone for cardiovascular applications [3]. The SIGMA-PPG architecture, which incorporates statistical priors from physiological modeling into a generative masking pre-training scheme, represents a notable advance by explicitly encoding domain knowledge about pulse wave propagation and sensor frequency responses into the representation learning process [9].

Parallel to these advances in representation learning, the field of approximate nearest neighbor search has matured significantly. Graph-based indices, most notably the Hierarchical Navigable Small World method, offer logarithmic scaling of query time with database size while maintaining high recall through multi-scale navigability [7]. For billion-scale datasets, GPU-accelerated approaches that combine vector quantization with product quantization have become standard [6]. However, these methods operate on generic distance metrics that do not inherently reflect the semantic distinctions that matter in clinical contexts. Deep semantic hashing bridges this gap by learning compact binary codes that maximize the preservation of semantic similarity as defined by clinical labels or physiological meta-data [8]. Self-supervised asymmetric semantic excavation has shown that even without explicit class labels, the semantic structure of the data can be recovered and embedded into hash codes with margin-scalable constraints that adapt to the desired granularity of retrieval. This body of work forms the technical substrate upon which our architectural discussion builds, yet its integration with PPG foundation models and the attendant deployment challenges have not been systematically articulated.

In the healthcare AI ecosystem, fairness and robustness concerns have come to the forefront. Groundbreaking work dissecting racial bias in widely used population health algorithms has illustrated how reliance on proxies like healthcare cost can systematically disadvantage minority populations [10]. In PPG-based monitoring, skin pigmentation, sensor calibration, and motion patterns can introduce similar biases that propagate into the embedding space. The design of semantic-aware retrieval systems must therefore account for both representational biases in the foundation model and retrieval biases introduced by the indexing structure itself.

Federated learning frameworks offer one path toward collaborative model improvement without centralizing sensitive physiological data, and have been proposed as a foundational architecture for digital health [11, 12]. Differential privacy mechanisms further enable the quantification and bounding of information leakage during query execution [13, 14]. The present work extends these perspectives into the retrieval layer, where the interaction between privacy budgets, semantic fidelity, and latency introduces new design dilemmas.

3. System Architecture and Semantic Retrieval Design

Our architectural framework posits a modular pipeline comprising four principal stages: signal acquisition and pre-processing on wearable or point-of-care devices; encoding via a PPG foundation model; indexing into a semantic-aware ANN structure; and personalized retrieval for downstream clinical decision support. This section analyzes the structural trade-offs inherent in each stage and their interdependencies, emphasizing the design rationale behind coupling semantic hashing with graph-based navigation.

The foundation model encoder serves as the representational anchor of the entire system. By mapping variable-length PPG windows into a fixed-dimensional embedding space, it enables subsequent retrieval to operate on a condensed yet information-rich summary. The choice of encoder architecture, pre-training objective, and the granularity of temporal segments directly impacts the semantic quality of retrieval. A model that is overly attuned to low-level signal noise will produce embeddings where distance correlates with artifact rather than physiology, undermining downstream clinical utility. Conversely, a representation that aggressively discards patient-specific variability may homogenize embeddings to the point where personalized retrieval becomes infeasible. The system architect must therefore navigate a tension between domain invariance and individual discriminability. Foundation models pre-trained with multi-task objectives that include reconstruction, contrastive, and physiological forecasting components can partially resolve this by encoding information at multiple levels of abstraction, but they also increase encoder complexity and inference cost on resource-constrained edge devices.

After encoding, the embedding vectors are projected into a semantic hashing space. Unlike traditional hashing that focuses solely on compression and collision probability, semantic-aware hashing integrates a learned mapping that realigns the distance metric with clinical semantics. This mapping can be trained using triplet losses, asymmetric pairwise constraints, or self-supervised objectives that leverage metadata such as diagnostic codes, medication classes, or laboratory values co-occurring with PPG segments. The resulting binary codes dramatically reduce storage requirements and accelerate distance computations via Hamming distance, enabling ANN queries to be performed with minimal arithmetic overhead. A critical design decision is the length of the hash code: shorter codes enable faster scanning but reduce the capacity to preserve fine-grained semantic differences, while longer codes retain more information at the expense of larger indices and increased collision rates under noise. This trade-off must be calibrated against the expected volume of the embedding database, the diversity of cardiovascular conditions represented, and the acceptable false positive rate in clinical retrieval tasks.

The graph-based indexing layer organizes the hash codes into a navigable graph structure where edges connect semantically proximal states. The high-dimensional embedding space is first reduced via the semantic hashing layer, and the resulting binarized representations are used as node identifiers in a hierarchical graph. This two-tier design leverages the compression and semantic alignment of hashing with the logarithmic search properties of

navigable graphs, providing sublinear query times even as the corpus grows to contain millions of patient-specific state representations. An important architectural nuance concerns the selectivity of graph edges: edges can be constructed purely based on Hamming distance, or they can incorporate additional constraints such as temporal adjacency, demographic similarity, or clinical outcome similarity. The latter approach transforms the graph into a multi-relational structure that supports richer query semantics—for instance, retrieving physiological states that are not only similar in waveform morphology but also observed in patients with similar therapeutic trajectories. However, maintaining such a graph under continuous data ingestion and concept drift introduces substantial operational overhead that must be weighed against the incremental clinical value.

Personalization of retrieval is accomplished by maintaining a lightweight user-specific context vector that modulates the similarity function. This context vector encodes an individual's baseline hemodynamic profile, common artifact patterns, and condition-specific prior distributions, allowing the retrieval engine to upweight exemplars from demographic and phenotypic peers. The context can be seamlessly integrated into the semantic hashing function by applying an affine transformation before binarization, or it can be realized as a re-ranking step that adjusts the raw ANN candidate list using a personalized scorer. The former approach embeds personalization into the index itself but requires re-indexing when the context changes; the latter is more flexible but incurs additional latency at query time. Determining the appropriate balance hinges on the expected rate of context drift and the computational envelope of the target deployment tier.

4. Infrastructure and Deployment Considerations

The deployment of semantic-aware ANN search for cardiovascular monitoring spans a heterogeneous computing continuum. At the edge, wearable devices with embedded accelerators can perform signal pre-processing and lightweight encoder inference, generating embeddings that are streamed to a nearby fog node or smartphone gateway. The choice of where to place the indexing and retrieval logic has profound implications for latency, privacy, and energy consumption. If the full ANN index is pushed to the edge, retrieval latency can be minimized and physiological data need not leave the personal device, but the index size is severely constrained by storage and memory limitations. Conversely, a cloud-hosted index can scale to billions of embeddings and support graph updates from multiple institutions, but incurs network latency and raises data governance concerns.

Splitting the retrieval pipeline across tiers offers a pragmatic middle path. A small, frequently updated personalized cache index can be maintained on-device or at the gateway, storing embeddings most relevant to the user's recent physiological history. A larger, less frequently updated cluster index resides in the cloud and services queries that fall outside the cache's semantic reach. This tiered design mirrors classical web caching architectures but must incorporate additional logic to ensure that cached embeddings do not become stale due to changes in the underlying foundation model or in the user's cardiovascular state. Cache invalidation strategies must be tightly coupled with model versioning and drift detection mechanisms, which we discuss in the robustness section.

Energy efficiency is a critical metric for continuous monitoring systems that must operate for days or weeks without recharging. Semantic hashing significantly reduces the energy footprint of retrieval by replacing costly floating-point vector operations with integer-level Hamming distance computations. Moreover, graph navigation based on binary codes can be implemented using memory-efficient data structures that exploit bit-level parallelism on

modern microcontrollers. Still, the energy cost of the encoder forward pass remains dominant. Techniques such as knowledge distillation, quantization-aware training, and neural architecture search for edge-friendly PPG encoders become essential complements to efficient retrieval. The interplay between encoder sparsity and retrieval accuracy is a multi-objective optimization problem that has not been fully characterized: aggressive compression may degrade embedding quality to the point where semantic hashing collisions mask clinically relevant distinctions, resulting in misdirected retrieval that could propagate to erroneous clinical decisions.

Data residency and regulatory compliance add further constraints. In jurisdictions governed by strict health data localization laws, the cloud-hosted index may not store or process identifiable biomedical signals across borders. This necessitates a federated retrieval infrastructure where each jurisdictional zone maintains its own index and meta-index that routes inter-zone queries without exposing raw embeddings. Federated graph indexing protocols, where subgraphs are updated locally and only structural summaries are exchanged, present an open systems challenge. Ensuring that the global semantic alignment of the hashing function is preserved across zones requires periodic synchronization of the hash projection layers, which can be achieved through federated hashing schemes that share model parameters but not data. The design of the synchronization frequency and the communication protocols must balance consistency against bandwidth and latency constraints.

5. Robustness and Fairness in Personalized Monitoring

The reliability of semantic-aware ANN retrieval in cardiovascular applications is contingent on the system's ability to maintain accurate and fair nearest-neighbor relationships under real-world perturbations. Distributional shifts occur along multiple axes: sensor drift over time, changes in daily activity patterns, the onset of new cardiovascular conditions, and the inclusion of data from previously unseen demographic groups. The embedding space produced by a foundation model trained on static snapshots may fail to correctly position novel physiological states, leading to retrieval of clinically inappropriate exemplars. Continuous monitoring of embedding-space topology through techniques such as persistent homology can detect emerging clusters or fragmentation that signal a degradation in semantic coherence. When such drift is detected, the system should trigger fine-tuning of the semantic hashing layer or, in more severe cases, a re-indexing operation that restructures the graph to reflect the new distribution.

Robustness is intricately linked to fairness. If the foundation model is pre-trained on datasets that overrepresent certain demographic or phenotypic groups, the embedding space will exhibit anisotropic density, with higher granularity and tighter clustering for majority groups and sparser, more ambiguous representations for minority groups. This unevenness directly translates into retrieval bias: queries from underrepresented groups may have fewer semantically proximal neighbors, leading to lower-quality personalized recommendations or, worse, misclassification driven by spurious proximity to physiologically distinct but embedding-space-adjacent states. Addressing this requires fairness interventions at both the representation and retrieval levels. At the representation level, data augmentation strategies that synthetically generate diverse PPG morphologies informed by physiological models can help densify underrepresented regions [9]. At the retrieval level, fairness-aware graph construction can explicitly enforce diversity constraints during edge formation, ensuring that every node connects to exemplars from multiple demographic strata. Moreover, the

personalization context vector can be reweighted to compensate for representation disparities, although this must be done transparently to maintain clinical interpretability.

Adversarial robustness is another concern. Maliciously crafted perturbations to PPG signals, achievable through electromagnetic interference or deliberate manipulation of sensor contact, could cause the encoder to produce embeddings that are drastically different from the true physiological state. If an attacker can drive the embedding toward a desired pathological cluster, they could potentially trigger false clinical alerts or mask genuine anomalies. Defensive strategies include adversarial training of the encoder, input-level anomaly detectors that reject signals with unrealistic frequency spectra, and graph-based consistency checks that flag queries whose nearest neighbors exhibit suspiciously high variance in clinical outcomes. The last approach leverages the very structure of the retrieval system itself as a defense, turning the neighborhood's diagnostic heterogeneity into a signal of anomalous input. These robustness dimensions underscore the necessity of viewing the retrieval system not as a passive lookup mechanism but as an active participant in the system's safety assurance.

6. Governance, Policy, and Ethical Implications

The deployment of semantic-aware ANN retrieval for cardiovascular monitoring engages a web of governance and policy questions that extend well beyond technical performance metrics. The core capability—finding similar physiological states across a diverse population—can simultaneously empower precision medicine and enable new forms of sensitive inference about individuals' health trajectories, genetic predispositions, or lifestyle patterns. Establishing trust requires embedding ethical principles into the retrieval architecture itself rather than treating them as external compliance checkboxes.

Data governance in a multi-tenant retrieval system must address several interrelated concerns: consent for secondary use of embeddings, the right to withdraw one's physiological signatures from the index, and the delineation of responsibilities when retrieval results influence clinical decisions across institutional boundaries. Traditional data access control models based on role-based permissions struggle to capture the nuanced, context-dependent nature of similarity queries. A patient may consent to their embeddings being used for retrieval in a diabetes management program but not for life insurance risk assessment, yet these downstream uses are not always separable at the query level. Attribute-based access control integrated with the retrieval pipeline can enforce such constraints by checking the query's purpose metadata against the data subject's consent policy before allowing the graph traversal to access certain nodes. Furthermore, immutable audit logs built on distributed ledger technology can provide a tamper-evident record of which embeddings were accessed, for what purpose, and with what outcome, enabling post-hoc accountability.

The dynamic nature of personalized retrieval introduces additional regulatory challenges under frameworks such as the EU Medical Device Regulation and the FDA's evolving stance on adaptive AI in healthcare. If the semantic hashing function or the graph index is continuously updated based on new data, the system may be classified as a learning medical device subject to stricter post-market surveillance requirements. The architecture must therefore include mechanisms for version pinning and rollback, where a clinician can fall back to a previously validated index snapshot if an update introduces unanticipated behavior. The safety case for such a system should include not only algorithmic performance metrics but also evidence from structured human-factors studies examining how clinicians interpret and act upon retrieval results in time-pressured scenarios. Designing for appropriate reliance—where the retrieval output is treated as a consultative second opinion rather than a

prescriptive recommendation—is as much a user-interface and training challenge as it is a technical one.

Cross-border deployment of federated indices further complicates the policy landscape. While federated retrieval prevents direct transfer of raw PPG data, the queries themselves, when combined with side information, can leak sensitive information about the population structure of the index. Differential privacy mechanisms can provide formal guarantees on the amount of information revealed per query, but these guarantees must be carefully calibrated against the clinical tolerance for false neighbors. This calibration involves engaging diverse stakeholders—regulators, clinicians, patient advocacy groups, and ethicists—to define acceptable risk thresholds and to translate those thresholds into epsilon budgets for the privacy-preserving retrieval interface. A participatory design process, institutionalized through ongoing multi-stakeholder governance boards, is therefore an essential non-technical component of the overall system architecture.

7. Sustainability and Long-term Maintenance

The long-term viability of semantic-aware ANN retrieval systems depends on their ability to adapt to evolving clinical knowledge, population demographics, and sensor technology without necessitating prohibitive retraining costs or environmental impact. Foundation model updates, driven by the availability of new pre-training corpora or improved self-supervised objectives, can render existing embedding collections semantically stale. A naive re-encoding of the entire historical corpus with the new model not only entails substantial computational expenditure but also disrupts the continuity of personalized contexts that have been refined over time. A more sustainable approach involves maintaining backward-compatible embeddings through constrained fine-tuning objectives that penalize large deviations in pairwise similarity structures, or through learned embedding transformation modules that map old embeddings into the new space without full recomputation.

Concept drift in the underlying PPG signal characteristics, induced by transitions from one device generation to another or by changes in population-level health patterns, similarly demands continuous lifecycle management. Automated monitoring of index quality via proxy metrics—such as the temporal stability of nearest-neighbor sets for a fixed set of anchor queries—can trigger alerts when re-indexing is warranted. However, the energy and carbon footprint of re-indexing operations at scale must be explicitly accounted for. Strategies such as incremental graph updates that only rewire neighborhoods affected by drift, rather than rebuilding the entire hierarchy, offer substantial savings. Additionally, the choice of hardware for index maintenance, including the potential use of carbon-aware scheduling for cloud-based re-indexing jobs, aligns the system with broader sustainability goals in digital health infrastructure.

An often-overlooked sustainability challenge is the human capital required to curate, validate, and govern the retrieval system over decades. Semantic alignment depends critically on the quality of the clinical metadata used to supervise hashing. As diagnostic standards evolve and new cardiovascular sub-phenotypes are discovered, the semantic labels attached to historical PPG segments may become outdated. Building institutional memory through version-controlled semantic ontologies and semi-automated label curation pipelines that combine domain expert input with weak supervision can mitigate this brittleness. Equally important is the cultivation of a multidisciplinary workforce that spans systems engineering, clinical informatics, and medical ethics, capable of steering the system through the inevitable socio-

technical shifts. Long-term funding models that decouple system maintenance from grant-based project cycles should be considered a policy prerequisite for responsible deployment.

8. Conclusion

Semantic-aware approximate nearest neighbor search over PPG foundation model embeddings offers a powerful paradigm for personalized cardiovascular monitoring, bridging the gap between the richness of pre-trained representations and the operational need for fast, clinically meaningful retrieval. This paper has examined the architectural, infrastructural, and governance dimensions of such a system, highlighting that its success depends as much on the deliberate design of fairness, robustness, and sustainability mechanisms as on the algorithmic efficiency of the retrieval core. By coupling deep semantic hashing with hierarchical graph indexing, the retrieval pipeline can achieve the latency and energy profiles needed for edge deployment while preserving the semantic granularity required for clinical personalization. The structural analysis revealed that key trade-offs—between personalization granularity and index update frequency, between semantic hash length and fairness across demographic subgroups, and between privacy budgets and clinical recall—must be navigated through a systems perspective that integrates technical and policy choices.

Looking forward, several open challenges demand interdisciplinary attention. The standardization of semantic similarity metrics for cardiovascular states, informed by domain ontologies and clinician consensus, would facilitate interoperability across different PPG foundation models and institutional indices. Methods for continuous and verifiable fairness auditing of retrieval outcomes in production settings remain nascent, particularly in federated deployments where access to raw demographic attributes is restricted. The integration of causal reasoning into the retrieval process, enabling the system to distinguish between states that merely co-occur and those that share etiological mechanisms, could further elevate clinical trust and utility. Ultimately, the development of semantic-aware retrieval infrastructure for cardiovascular monitoring constitutes a socio-technical endeavor that requires sustained collaboration among machine learning researchers, healthcare systems engineers, clinicians, regulators, and patient communities to ensure that the promise of personalized digital health is realized equitably and sustainably.

References

1. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3), R1–R39.

2. Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. Current Cardiology Reviews, 8(1), 14–25.

3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

4. Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., Li, X., & Guan, C. (2021). Time-series representation learning via temporal and contextual contrasting. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 2352–2359.

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

6. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.

7. Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836.

8. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

9. Guo, Z., Chen, T., Jiao, Y., Pan, Y., Hu, X., & Ferrario, M. (2026). SIGMA-PPG: Statistical-prior Informed Generative Masking Architecture for PPG Foundation Model. arXiv preprint arXiv:2601.21031.

10. 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.

11. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Maier-Hein, K. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1–7.

12. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273–1282.

13. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

14. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 308–318.

15. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

16. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

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

18. Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. Health Affairs, 39(3), 359–365.

19. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, 2(1), 1–10.

20. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1–37.

21. Yue, Y., Khanal, A., Lyu, T., Weissman, S., & Liang, C. (2025, May). EHR Phenotyping Methods for Measuring Treatment Adherence Among People Living With HIV in All of Us: Towards Disparities and Inequalities in HIV Care Continuum. In AMIA Annual Symposium Proceedings (Vol. 2024, p. 1294).