Explainable Medical Decision-Making through Adversarially Hardened LLM Agents and Semantic-Aware Multi-Label Image Hash Retrieval

Micolas Males
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
hellonicolas@unr.edu

Makle Lease
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
lase530@colostate.edu

Kaeaeth Ghodes
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
kenneth.rhodes691@oregonstate.edu

Wishal Gain
Department of Computer Science, University of North Texas, Denton, TX, USA.
gain902@unt.edu

Abstract

The rapid adoption of large language models (LLMs) in clinical environments has introduced unprecedented opportunities for decision support, yet it has simultaneously magnified concerns surrounding adversarial vulnerability, explainability, and the scalable retrieval of multimodal evidence. This paper presents a comprehensive systems-level analysis of an integrated framework that combines adversarially hardened LLM agents with semantic-aware multi-label image hash retrieval to deliver robust and interpretable medical decision-making. The discussion moves beyond algorithmic novelty to examine the structural trade-offs inherent in deploying such a hybrid system within real-world healthcare infrastructures. We explore how adversarial hardening techniques, including input purification, representation smoothing, and constrained decoding, can be embedded into the agent architecture without undermining clinical fluency or diagnostic accuracy. In parallel, we investigate the role of deep semantic hashing that preserves multi-label diagnostic relationships, enabling efficient similarity-preserving retrieval of medical images from large-scale repositories while offering traceable evidence paths for model recommendations. The interplay between the LLM agent and the retrieval engine is analyzed through the lenses of latency, memory footprint, trust calibration, and failure mode containment. Special attention is devoted to the governance challenges that arise when explainability mechanisms must satisfy both algorithmic transparency requirements and the epistemic needs of heterogeneous clinical stakeholders. The paper further interrogates fairness considerations in multi-label retrieval when disease prevalence distributions are imbalanced, and outlines deployment pathways that account for data sovereignty, model updating cadence, and energy sustainability. By synthesizing perspectives from adversarial machine learning, information retrieval, human-computer

interaction, and health informatics, the work articulates a design philosophy in which explainability is not retrofitted but engineered as a first-class property of a secure, retrieval-augmented agentic system. The analysis concludes with policy-oriented reflections on certification, auditability, and cross-institutional governance for AI-enabled clinical decision support.

Keywords

explainable AI, large language models, adversarial robustness, medical image retrieval, semantic hashing, multi-label learning, clinical decision support, agentic systems, health informatics.

1. Introduction

The infusion of artificial intelligence into clinical workflows has evolved from isolated pattern recognition tasks toward comprehensive agentic systems capable of synthesizing heterogeneous data streams, engaging in multi-turn diagnostic dialogue, and recommending evidence-grounded interventions. Large language models (LLMs) are at the forefront of this shift, demonstrating substantial proficiency in interpreting unstructured clinical narratives, summarizing patient histories, and suggesting differential diagnoses [1,2]. Yet their deployment in safety-critical medical settings remains circumscribed by two deeply entangled problems: an acute susceptibility to adversarial perturbations and a persistent opacity that undermines clinician trust and institutional accountability [3,4]. While much research has focused on either improving model interpretability or hardening models against adversarial inputs in isolation, real-world clinical decision-making demands a unified architecture in which robustness and explainability are co-designed rather than independently optimized. This work examines a systems-level integration of adversarially hardened LLM agents with semantic-aware multi-label image hash retrieval, constructing a decision-support pipeline that simultaneously resists malicious manipulation and provides verifiable evidentiary grounding through retrieved visual analogues.

The motivation for binding these two components stems from the observation that many high-stakes medical errors arise not from a single misclassification but from a cascade of poorly calibrated interactions between reasoning modules, retrieval backends, and human interpreters. Adversarial inputs, whether introduced via poisoned electronic health records, manipulated imaging metadata, or crafted dialogue prompts, can propagate through an LLM agent to produce confident yet clinically dangerous recommendations. At the same time, a retrieval system that searches large-scale medical image archives without preserving the nuanced multi-label structure of radiological findings can return superficially similar cases that mislead rather than inform [5]. By deploying a retrieval mechanism whose hash codes are trained under asymmetric semantic excavation constraints, the system can maintain high recall for rare disease co-occurrences while bounding the influence of spurious correlations. The LLM agent, in turn, consumes these retrieved exemplars along with a calibrated confidence decomposition, producing explanations that link its reasoning traces to concrete, inspectable evidence rather than opaque latent representations. The present paper articulates the architectural principles, trade-off spaces, and governance frameworks that govern such a system, paying sustained attention to the infrastructure demands, sustainability metrics, and fairness implications that separate laboratory prototypes from clinically viable deployments.

2. Foundational Concepts and Cross-Disciplinary Landscape

Understanding the proposed integration requires situating it at the confluence of four distinct research streams: adversarial robustness in language models, semantic hashing for large-scale image retrieval, multi-label medical image analysis, and explainable agent design. Adversarial vulnerability in LLMs has been documented across a spectrum of attack vectors, including synonym substitution, prompt injection, and latent-space perturbations that preserve surface semantics while radically altering model outputs [6,7]. In clinical contexts, these vulnerabilities translate into concrete patient harms when an agent mischaracterizes a radiology report or omits a critical contraindication due to a subtly altered input. Defensive strategies have ranged from adversarial training with clinical corpora to input sanitization pipelines that project embeddings onto plausibility manifolds, yet each strategy introduces its own trade-offs between robustness and linguistic flexibility.

Semantic hashing, originally conceived as a method for mapping semantically similar documents to compact binary codes, has been extended to high-resolution medical images through deep convolutional architectures that jointly optimize representation learning and quantization [8,9]. The shift from single-label to multi-label retrieval is critical for medical imaging, where chest radiographs routinely exhibit multiple concurrent findings such as cardiomegaly, pleural effusion, and nodular opacities. Hash functions that collapse multi-label combinations into a single code must carefully balance inter-class separation with intra-code coherence, a challenge addressed by self-supervised semantic excavation methods that mine asymmetric similarity structures from label co-occurrence statistics [10]. When integrated into a clinical decision-support system, the retrieval module transforms from a generic search engine into a context-aware assistant that surfaces cases with pertinent combinatorial evidence.

Explainability in agentic systems extends beyond saliency maps or feature attribution scores to encompass the broader communicative act of rendering a decision trajectory intelligible to a human practitioner. Clinicians require not only a justification of why a particular diagnosis was favored but also an account of how alternative hypotheses were excluded, what evidence was retrieved, and how the system's confidence should be calibrated given known training-data limitations [11,12]. An LLM agent that can reference specific retrieved cases, articulate the relevance of their multi-label profiles, and signal uncertainty intervals embodies a form of interactive explainability that aligns with the cognitive workflow of differential diagnosis more faithfully than static explanation interfaces.

3. System Architecture and Integration Paradigm

The architecture under examination consists of three principal subsystems: a retrieval-augmented LLM agent, a semantic-aware multi-label image hashing engine, and an orchestration layer that manages synchronous and asynchronous information flows among components while enforcing adversarial defense checkpoints. The LLM agent is responsible for natural language understanding of clinical queries, iterative reasoning over patient context, and generation of diagnostic hypotheses with associated explanations. Its adversarial hardening is realized through a layered defense strategy: an input sanitizer that detects and neutralizes known perturbation patterns, a representation regularizer that constrains hidden-state activations to stay within empirical bounds observed during training on clean clinical data, and a constrained decoding module that suppresses output distributions inconsistent with retrieved evidence.

The image hash retrieval engine operates on a continuously updated repository of anonymized medical images, each indexed by a compact binary code produced by a deep hashing network

trained with self-supervised asymmetric semantic excavation and margin-scalable constraints. This training paradigm, which does not rely on exhaustively curated similarity annotations, allows the system to scale to institutional repositories containing millions of studies while preserving fine-grained multi-label relationships. Upon receiving a retrieval request from the LLM agent—formulated as a structured representation of the current diagnostic hypothesis—the hashing engine returns a ranked list of cases whose hash codes exhibit maximal semantic alignment under a task-specific distance metric. Crucially, the retrieval is not a one-shot query; the agent can refine its request based on initial results, engaging in a sequential hypothesis-evidence refinement loop that mirrors the clinical reasoning process.

The orchestration layer embeds several non-functional mechanisms that are indispensable for deployment. A latency budget controller ensures that the combined LLM inference and retrieval latency remains within clinically acceptable bounds, dynamically trading off retrieval depth against generation speed. A model versioning registry tracks the exact configuration of both the LLM checkpoint and the hashing network used for each clinical encounter, enabling retrospective audits. An adversarial intrusion monitor continuously evaluates incoming requests and retrieved images for distributional anomalies that could indicate poisoning attacks or model inversion attempts. These infrastructural components collectively transform a conceptual algorithm into a system that satisfies the reliability and traceability requirements of healthcare IT governance.

4. Adversarial Hardening of LLM Agents for Medical Contexts

Designing adversarially robust LLM agents for medicine demands strategies that go beyond generic robustness benchmarks and engage with the specificity of clinical language and the severity of downstream consequences. One foundational insight is that clinical adversarial threats are not limited to direct prompt manipulation; they include indirect attacks that poison the retrieval index, corrupt the lab values fed into the agent, or exploit temporal inconsistencies in longitudinal patient records. The hardening architecture adopted here consequently embeds defenses at multiple abstraction levels. At the input level, a learned perturbation detector distinguishes between plausible clinical paraphrasing and adversarially crafted rephrasings by analyzing distributional properties of token sequences against a corpus of authentic clinical notes. Representations that fall outside the support of the training distribution are mapped back onto the nearest semantically equivalent safe region using a denoising autoencoder trained on clean clinical text, a technique that has shown promise in preserving medical named entity integrity while neutralizing adversarial noise [13].

At the model level, the LLM agent employs representation smoothing through localized Lipschitz constraints that limit the sensitivity of hidden-state trajectories to small input variations. This approach is particularly suited to clinical dialogue because it prevents a single perturbed token from cascading into a radically different diagnostic trajectory while still allowing the model to express genuine medical uncertainty. Moreover, the agent leverages a retrieval-conditioned confidence estimator that explicitly decomposes predictive uncertainty into epistemic and aleatoric components, reporting to the clinician not only its top hypotheses but also the extent to which those hypotheses depend on evidence that could be undermined by an adversary. This decomposition is critical for enabling the clinician to gauge when to override the system's suggestion—a capability that constitutes a practical instantiation of meaningful human control.

The adversarial hardening extends to the interaction between the LLM agent and the hashing retrieval engine. An adversary who cannot directly attack the LLM may instead inject subtly

manipulated images into the retrieval index such that hash codes for certain diagnostic categories become systematically distorted. To counter this, the system incorporates cross-modal consistency checks: the LLM agent compares the textual description of a retrieved image against its known clinical metadata and flags discrepancies for human review. This cross-modal verification loop leverages the complementary failure modes of vision and language modules, making it substantially harder for an adversary to compromise the integrated system without triggering multiple independent alarms. Recent work has underscored the importance of such layered defenses in medical agent architectures, demonstrating that even modest adversarial hardening at the agent level can significantly reduce the rate of clinically consequential errors under targeted attack scenarios [16].

5. Semantic-Aware Multi-Label Image Hash Retrieval

The image retrieval component must reconcile three competing demands: retrieval speed suitable for interactive clinical workflows, preservation of fine-grained multi-label semantics, and robustness to natural and adversarial distribution shifts. Deep hashing methods satisfy the speed requirement by mapping high-dimensional image features into compact binary codes that enable fast Hamming-distance comparisons, yet the mapping can erase subtle diagnostic distinctions essential for multi-label retrieval. The semantic-aware hashing framework adopted here addresses this tension through an asymmetric training objective that treats a given image and its multi-label annotation as two views of the same underlying clinical state, excavating semantic relationships by maximizing the agreement between hash codes of images that share label subsets while applying a margin-scalable constraint that gently pushes apart codes for dissimilar combinations [10]. This objective yields hash spaces in which images exhibiting co-occurring findings such as consolidation and pleural effusion are located in proximity, while cases where these findings appear in isolation occupy distinct but neighboring regions, preserving the topology of diagnostic co-expression.

Maintaining such a structured hash space across large-scale, continually updated institutional repositories introduces substantial engineering challenges. The hashing network must be periodically retrained or fine-tuned as new imaging modalities, annotation standards, and disease taxonomies emerge, yet the binary codes of legacy images must remain stable to avoid invalidating downstream clinical audit trails. Incremental hashing strategies that append new bits to the code length while keeping existing bits frozen offer a pragmatic compromise, though they introduce a trade-off between storage overhead and retrieval precision that must be tuned according to the specific epidemiology of the target clinical domain. For instance, in thoracic imaging where the co-occurrence patterns of common findings like cardiomegaly and pulmonary edema are relatively stable, a smaller code length with periodic full re-indexing may be acceptable, whereas in dermatology where novel lesion taxonomies evolve rapidly, incremental schemes may be preferred.

The multi-label nature of medical images further complicates the attribution of retrieval evidence in explainability workflows. When a retrieved image exhibits three concurrent findings and the clinician’s query focuses on only one of them, the system must articulate why that image remains relevant without implying that all three findings are present in the current patient. This requires the hash retrieval engine to expose not only a global similarity score but also per-label relevance gradients computed via backpropagation through the hashing network, which can then be verbalized by the LLM agent. The computational overhead of such per-query gradient computation is non-trivial and necessitates caching strategies that exploit the hash space geometry to approximate relevance scores for frequently

retrieved clusters. These design choices illustrate how the pursuit of explainability in retrieval-augmented systems propagates into low-level indexing infrastructure decisions.

6. Explainability Mechanisms and Interpretability Design

Explainability in the proposed system is conceptualized as a multi-tiered communicative process rather than a single algorithmic output. At the first tier, the LLM agent provides a natural language summary of its diagnostic hypothesis, explicitly linking each component of the differential to specific retrieved cases and to snippets of the clinical guidelines encoded in its parametric knowledge. At the second tier, the interface surfaces a relevance-annotated gallery of retrieved images, where each image is accompanied by a visually highlighted overlay indicating the regions most influential for the multi-label hash similarity, along with a quantitative decomposition of how each label contributed to the retrieval score. At the third tier, the system exposes an interactive uncertainty dashboard that displays the confidence decomposition discussed earlier, allowing the clinician to drill down into which evidence sources most strongly affect the recommendation and to simulate counterfactual scenarios by virtually removing certain findings or retrieved cases.

This layered explainability design recognizes that different clinical stakeholders possess distinct informational needs and cognitive bandwidths. A radiologist verifying a finding may predominantly engage with the image gallery and region highlights, while an attending physician weighing a treatment decision may focus on the narrative summary and confidence decomposition, and a hospital administrator auditing for quality assurance may examine aggregate statistics of retrieval-evidence concordance across cases. By decoupling the generation of the evidence base from its presentation, the architecture supports customizable explainability views without altering the underlying decision logic. This decoupling also facilitates compliance with emerging regulatory frameworks such as the European Union's Artificial Intelligence Act, which mandates varying levels of transparency depending on the risk classification of the AI system and the role of the human overseer [14,15].

A persistent tension in designing such explainability interfaces lies in balancing completeness against comprehensibility. Exhaustively enumerating every piece of retrieved evidence and its associated confidence interval can overwhelm the clinician, paradoxically reducing decision quality by increasing cognitive load. The system addresses this through an evidence triage mechanism that prioritizes the subset of retrieved cases that are both highly influential on the recommendation and sufficiently dissimilar to one another in their multi-label profiles, thereby maximizing the informational diversity of the presented evidence while keeping the volume manageable. This triage itself must be explainable: the agent briefly articulates why certain cases were selected as exemplars and how they differ from others that were withheld, fostering a meta-level transparency that bolsters trust without overwhelming the user.

7. Deployment, Infrastructure, and Sustainability Considerations

Transitioning the integrated system from a research prototype to a clinically sustained deployment requires confronting a cluster of infrastructural realities that are often elided in algorithmic publications. The first is computational heterogeneity across healthcare institutions: a large academic medical center may operate on-premise GPU clusters capable of hosting full-scale LLM inference and hash index serving, whereas a rural clinic may rely on a thin-client model with inference offloaded to a regional cloud hub. The orchestration layer must therefore support tiered deployment profiles, dynamically resolving whether retrieval and inference occur locally or remotely based on latency constraints, data sovereignty

regulations, and available bandwidth. Hybrid deployment models that keep the hash index and a compressed LLM variant on-site while synchronizing with a more capable cloud backend for complex cases can balance these competing pressures, but they introduce consistency challenges that demand careful state management and failover protocols.

Data sovereignty and patient privacy impose additional architectural constraints. The retrieval index must be constructed without exposing raw patient images across institutional boundaries, which can be achieved through federated hashing frameworks where each institution contributes locally trained hash functions to a global consensus code space without sharing image data. The LLM agent, when operating on patient-specific data, must execute in environments compliant with the Health Insurance Portability and Accountability Act or equivalent regulations, employing techniques such as differential privacy during fine-tuning and secure multi-party computation during collaborative inference. These privacy-preserving mechanisms inevitably degrade retrieval precision and generation fluency to some degree, and quantifying this degradation as a function of privacy budget is essential for enabling institutional risk-benefit assessments.

Sustainability concerns extend beyond privacy to encompass the energy footprint and lifecycle carbon cost of continuously operating large models in clinical settings. The hashing retrieval engine contributes relatively modestly to overall energy consumption because binary code comparisons are extremely lightweight; however, the periodic retraining of deep hashing networks and LLM checkpoints can incur substantial computational cost. Strategies such as sparse model updating, where only a subset of parameters relevant to newly encountered disease patterns are fine-tuned, and knowledge distillation from large teacher models into smaller student agents that are sufficient for routine queries, can reduce both energy use and inference latency. A comprehensive sustainability assessment should also account for the avoided environmental costs when accurate, explainable decision support reduces unnecessary repeat imaging or prevents diagnostic errors that lead to prolonged treatment cascades, framing the system's environmental impact in the broader context of healthcare resource optimization.

8. Fairness, Governance, and Policy Implications

A system that retrieves medical images to ground clinical decisions inherits and can amplify the biases present in its training data and institutional repositories. Multi-label retrieval is particularly susceptible to fairness degradations when certain disease combinations are underrepresented in the source population, causing the hash space to sparsely cover those regions and thereby reducing retrieval quality for patients with rare intersectional profiles. Mitigating such disparities requires proactive measures, including stratified hash code calibration that re-weights training objectives to upweight underrepresented label combinations, and post-retrieval fairness auditing that monitors differential retrieval precision across demographic and clinical subgroups. These fairness interventions must be documented in model cards and made accessible to clinical oversight committees, enabling ongoing accountability that extends beyond initial deployment.

Governance frameworks for adversarially hardened, retrieval-augmented LLM agents must address the challenge of continuous learning in a safety-critical environment. If the hashing index is updated weekly and the LLM agent is fine-tuned quarterly, the behavior of the integrated system can drift in ways that are difficult to anticipate through pre-deployment testing alone. Shadow deployment, wherein the updated system operates silently alongside the incumbent version and discrepancies are logged and reviewed, provides a mechanism for

detecting regressions before they affect patient care. The governance structure should also define clear thresholds for when an observed performance drop in a particular subpopulation or clinical condition triggers a rollback or a targeted remediation sprint, embedding algorithmic accountability into operational workflows.

The policy implications of such systems extend to the certification and liability frameworks that govern medical software. Current regulatory pathways, such as the U.S. Food and Drug Administration's Software as a Medical Device framework, were designed for static or periodically updated algorithms and are strained by the continuous learning and retrieval dynamics of agentic systems. A shift toward a total product lifecycle approach, where manufacturers submit not a frozen model but a validated change management protocol that specifies how updates are monitored, tested, and communicated, would align regulatory oversight with the technical reality of adaptive clinical AI. Cross-institutional data sharing agreements that enable federated hashing while respecting jurisdictional privacy laws represent another policy frontier, as they require harmonizing the data protection principles of multiple legal regimes without compromising the clinical utility of the shared representation space.

9. Conclusion

This paper has advanced a system-level perspective on merging adversarially hardened LLM agents with semantic-aware multi-label image hash retrieval to deliver explainable medical decision-making that is simultaneously robust, traceable, and cognizant of real-world deployment constraints. The architecture is characterized by a layered defense strategy that spans input sanitization, representation regularization, and cross-modal consistency checks, integrated with a hashing engine trained under asymmetric semantic excavation objectives that preserve nuanced multi-label relationships. Explainability is engineered as a multi-tiered facility that adapts to the informational needs of diverse clinical stakeholders, linking natural language narratives to retrievable visual evidence through interactive uncertainty decompositions. The analysis has further illuminated the infrastructure trade-offs inherent in tiered deployment, federated indexing, and privacy-preserving inference, and has foregrounded the sustainability and fairness considerations that must accompany clinical AI from conception through continuous governance. While substantial engineering and regulatory work remains before such systems can be broadly entrusted with life-critical decisions, the conceptual integration articulated here offers a principled template for building AI agents whose recommendations are not only accurate but also defensible, auditable, and aligned with the ethical imperatives of modern medicine.

References

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

2. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

3. Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44-58.

4. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

5. Quellec, G., Cazuguel, G., Cochener, B., & Lamard, M. (2017). Multiple-instance learning for medical image retrieval and classification. IEEE Transactions on Medical Imaging, 36(5), 1087-1096.

6. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 31-36).

7. Li, J., Ji, S., Du, T., Li, B., & Wang, T. (2020). TextBugger: Generating adversarial text against real-world applications. In 26th Annual Network and Distributed System Security Symposium.

8. Salakhutdinov, R., & Hinton, G. (2009). Semantic hashing. International Journal of Approximate Reasoning, 50(7), 969-978.

9. Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 769-790.

10. 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.

11. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310.

12. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

13. Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2021-2031).

14. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

15. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. AI Magazine, 38(3), 50-57.

16. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

17. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2023). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633-2650).

18. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 590-597).

19. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (pp. 1597-1607).

20. Lai, H., Pan, Y., Liu, Y., & Yan, S. (2015). Simultaneous feature learning and hash coding with deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3270-3278).

21. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In International Conference on Machine Learning (pp. 1885-1894).

22. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

23. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119.

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

25. U.S. Food and Drug Administration. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning-based software as a medical device. Discussion paper.