Continual Learning for Adversarially Robust Medical AI Agents in Evolving Disease Landscapes

Albert Adams
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
albertwork@unh.edu

Lin Ren
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
lin525@uab.edu

Ross Lindberg
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
ross.lindberg112@missouri.edu

Abstract

The deployment of artificial intelligence agents in clinical settings demands not only high diagnostic accuracy but also sustained robustness against adversarial perturbations and the capacity to adapt to constantly shifting disease landscapes. Medical AI systems, from diagnostic imaging classifiers to clinical decision support tools, are currently trained on static datasets that quickly become outdated as pathogens mutate, treatment protocols change, and population demographics evolve. This paper presents a comprehensive system-level analysis of continual learning frameworks designed to maintain adversarial robustness in medical AI agents operating under real-world constraints. We examine architectural trade-offs between plasticity for new knowledge acquisition and stability for retaining previously learned representations, particularly when those representations are vulnerable to adversarial attacks that can degrade patient safety. The discussion spans infrastructure requirements for online model updates, governance mechanisms for certifying deployed models after each retraining cycle, and the ethical implications of adaptive systems that must balance fairness across subpopulations while defending against malicious inputs. Through cross-domain comparisons with autonomous driving and cybersecurity, we illustrate how medical AI faces unique challenges due to high stakes, heterogeneous data distributions, and regulatory oversight. We propose a conceptual framework that integrates continual learning with adversarial training, uncertainty quantification, and human-in-the-loop validation. The paper further addresses sustainability concerns, including computational cost, energy consumption, and the need for decentralized data governance in federated learning topologies. Finally, we outline policy recommendations for regulatory bodies to ensure that adaptive medical AI agents remain both safe and equitable as disease landscapes continue to evolve.

Keywords

continual learning, adversarial robustness, medical artificial intelligence, disease evolution, system architecture, clinical decision support, model governance, fairness.

1. Introduction

The integration of artificial intelligence into healthcare has advanced rapidly over the past decade, with deep learning models achieving expert-level performance in tasks ranging from dermatological classification to radiological interpretation [1,2]. However, the static nature of most deployed models presents a fundamental limitation: diseases are not static. Pathogens mutate, environmental factors shift, and clinical practices evolve, rendering models trained on historical data progressively less accurate over time [3]. This temporal drift is compounded by the growing threat of adversarial attacks, where carefully crafted perturbations can cause models to misdiagnose conditions with potentially catastrophic consequences [4]. For medical AI agents to remain safe and effective, they must be capable of continual learning—updating their knowledge incrementally as new data arrives—while simultaneously defending against adversaries who may exploit the model's plasticity.

Continual learning, also known as lifelong learning, has been extensively studied in machine learning research as a means to overcome catastrophic forgetting [5,6]. When applied to medical AI, continual learning must reconcile the need for rapid adaptation to new disease patterns with the imperative to preserve existing knowledge about rare conditions and established pathologies. Simultaneously, adversarial robustness requires that the model maintain its performance under worst-case perturbations, a property that is notoriously difficult to guarantee when the model's own weights are being updated online [7]. The intersection of these two demands creates a complex design space with significant implications for system architecture, deployment infrastructure, and regulatory oversight.

This paper adopts a systems perspective to examine how continual learning can be designed to support adversarially robust medical AI agents in evolving disease landscapes. Rather than focusing on a specific algorithmic innovation, we analyze the structural trade-offs that arise when these capabilities are integrated into real-world clinical workflows. We consider the entire pipeline from data acquisition and model training to deployment monitoring and governance. The discussion is organized as follows. Section 2 provides background on adversarial threats in medical AI and the fundamental challenges of continual learning. Section 3 explores architectural considerations for building continual learning systems that preserve robustness. Section 4 delves into specific mechanisms for adversarial defense under non-stationary distributions. Section 5 addresses governance, fairness, and ethical implications, emphasizing the need for transparent accountability frameworks. Section 6 discusses deployment sustainability, including computational costs and federated learning configurations. Section 7 offers future research directions and policy recommendations. Section 8 concludes.

2. Background and Related Work

Adversarial attacks on deep learning models have been documented across multiple domains, but their potential impact in medicine is uniquely severe. A small perturbation imperceptible to human experts can cause a model to misclassify a malignant lesion as benign or to overlook a life-threatening anomaly in a chest radiograph [4,8]. Finlayson and colleagues demonstrated that adversarial examples could be crafted to fool medical image classifiers without altering the clinical diagnosis made by human radiologists, highlighting the insidious nature of such attacks [9]. Defenses, including adversarial training, have been shown to improve robustness but at the cost of reduced accuracy on clean data and increased computational burden [10]. Moreover, adversarial training typically assumes a stationary data distribution, an assumption that is violated in evolving disease landscapes.

Continual learning research has produced several families of algorithms to mitigate catastrophic forgetting. Regularization-based approaches, such as elastic weight consolidation, penalize changes to parameters deemed important for previous tasks [5]. Memory replay methods store representative samples from past distributions and interleave them with new data during training [6]. Dynamic architecture methods allocate new parameters for new tasks while freezing older ones [11]. Each approach involves trade-offs between memory, computational overhead, and the degree of forgetting. In medical applications, where data privacy regulations like HIPAA restrict data sharing, memory replay that requires storing patient-level examples raises ethical and legal concerns. Alternative approaches using generative replay—where a model generates synthetic samples from past distributions—offer a promising path, but the fidelity of generative models themselves can degrade over time [12].

Adversarial robustness in a continual learning setting is less explored but gaining attention. Recent work has shown that naively applying adversarial training during each incremental learning step can lead to catastrophic forgetting of robust features [13]. Conversely, models that are robust to attacks on one distribution may be vulnerable to new types of perturbations introduced by distribution shift [14]. This dynamic is particularly relevant for medical AI because adversaries may deliberately craft attacks that target the model's weak spots during transition periods between disease outbreaks. The required reference [18] provides insights into security enhancement methods specifically for large language model agents used in medical decision-making, emphasizing the need for robust adversarial defenses that can be updated as new tasks emerge.

3. Architectural Considerations for Continual Learning in Medical AI

Designing a continual learning system for medical AI agents requires careful architectural decisions that balance plasticity, stability, and robustness. At the highest level, the architecture must support online or batch-wise model updates without interrupting clinical service. A common approach is to maintain a primary inference model and a separate shadow model that is periodically updated; after validation, the shadow model replaces the primary model in a controlled cutover [15]. This architecture introduces latency between data availability and deployment, which may be unacceptable during rapidly evolving pandemics. Alternatively, meta-learning frameworks can enable the model to adapt with very few examples, but they often require careful regularization to avoid overfitting to new tasks at the expense of prior knowledge [16].

The choice of representation learning strategy is critical. Medical data is inherently multimodal, comprising imaging, genomics, electronic health records, and clinical notes. Continual learning across modalities introduces the challenge of aligning representations learned at different times. One solution is to maintain a shared encoder that is fine-tuned with a small learning rate while task-specific heads are appended for each new modality or disease class [17]. However, sharing parameters across all tasks increases vulnerability to adversarial attacks because a perturbation that affects the shared encoder can propagate to all downstream tasks. The system must therefore incorporate defensive mechanisms at multiple levels, such as input sanitization, feature squeezing, and ensemble methods.

The architecture must also account for the heterogeneity of clinical environments. A model deployed in a tertiary referral hospital may encounter a different distribution of diseases than a model in a rural clinic. Continual learning should be personalized to local data distributions while still benefiting from global knowledge. Federated learning offers a framework where models are trained across decentralized sites without sharing raw data [18], but coordinating

robust updates across nodes with different disease prevalences is non-trivial. Adversarial attacks could be mounted by a compromised node, injecting malicious updates that degrade the global model's performance on specific subpopulations. Robust aggregation techniques, such as trimmed mean or Krum, can mitigate Byzantine failures, but they add computational overhead and may reduce efficiency [19].

4. Adversarial Robustness Mechanisms in Dynamic Clinical Environments

Traditional adversarial defenses assume a stationary threat model, but in medical AI the adversary's capabilities may evolve as the model's knowledge expands. For example, an attacker who knows that a model has been retrained on recent influenza strains may craft perturbations that mimic those strains in a way that triggers misclassification of bacterial pneumonia. Continuous monitoring of the adversarial landscape is essential. One approach is to maintain an adversarial detector that flags inputs likely to be perturbed, but detectors themselves can be fooled by adaptive adversaries [20]. A more robust strategy is to combine adversarial training with domain randomization, exposing the model to a wide variety of potential perturbations during each training epoch regardless of the current disease distribution.

In the context of continual learning, adversarial training must be integrated with memory mechanisms. For instance, during each learning episode, the model can be trained on both current data and replayed adversarial examples from past tasks. This ensures that the model retains robustness against previously seen attack types while acquiring defenses against new ones. The computational cost, however, scales with the number of tasks and the diversity of attacks. System-level trade-offs arise between the frequency of retraining, the number of adversarial examples generated, and the inference latency for clinical decisions. Scheduling policies can prioritize retraining during periods of low clinical demand, such as overnight, but this delays the deployment of updated models in fast-moving outbreaks.

Uncertainty quantification is another layer of defense. A model that can express high uncertainty on out-of-distribution or adversarial inputs allows human clinicians to intervene before a potentially harmful decision is executed [21]. Continual learning models that maintain a probabilistic representation of their knowledge, such as Bayesian neural networks or deep ensembles, can provide calibrated uncertainty estimates even after multiple updates. However, maintaining a full Bayesian posterior over time is computationally prohibitive. Approximate methods, such as Monte Carlo dropout with temperature scaling, offer a practical compromise but require careful tuning to avoid overconfident predictions on adversarially perturbed inputs [22].

5. Governance, Fairness, and Ethical Implications

The deployment of continual learning medical AI agents raises profound governance questions. Regulatory bodies such as the U.S. Food and Drug Administration have established frameworks for approving locked algorithms, but adaptive algorithms that change over time challenge existing premarket review processes [23]. A system that updates itself weekly could drift into a state that is less accurate or less fair for certain demographic groups. Fairness metrics must be monitored continuously across all protected attributes, and any degradation must trigger a rollback or human review. This requires the infrastructure to log every model version, the data used for its update, and the performance metrics on holdout validation sets stratified by subpopulation.

Adversarial attacks can be targeted to exacerbate fairness issues. An adversary could craft perturbations that cause a model to misdiagnose individuals from a minority group more frequently, eroding trust and potentially causing harm. Continual learning systems that incorporate fairness constraints into their optimization objective, such as adversarial debiasing, must ensure that these constraints remain effective after each update. However, fairness definitions themselves may need to evolve as disease landscapes change; a model that is fair with respect to a historical population may become unfair when a new disease disproportionately affects a previously underrepresented group [24].

Ethical considerations extend to the autonomy of clinicians. If a continual learning agent is perceived as always updating itself, clinicians may lose confidence in its stability. Transparency mechanisms, such as explainable AI outputs that highlight which features drove a decision, become even more critical when the model's internal representations are in flux. Furthermore, the consent process for using patient data in continual learning loops must be carefully designed. While de-identified data may be permissible for model updates, the risk of re-identification through adversarial reconstruction attacks requires robust privacy-preserving techniques like differential privacy [25]. The trade-off between privacy and model utility is a persistent tension that must be managed at the system level.

6. Deployment Sustainability and System-Level Trade-Offs

Sustained operation of continual learning medical AI agents imposes significant computational and energetic demands. Each retraining cycle consumes GPU hours, and adversarial training further multiplies this cost. In resource-constrained settings, such as low-income countries or rural healthcare facilities, the infrastructure required to support frequent model updates may be prohibitive. Cloud-based solutions can offload computation, but they introduce latency and dependence on reliable internet connectivity. Edge computing with on-device learning offers lower latency and privacy benefits, but the limited compute and memory on edge devices restrict the complexity of models and the frequency of updates [26]. A distributed architecture that performs lightweight updates at the edge and more comprehensive retraining in the cloud can balance these constraints, but it requires careful orchestration.

Federated learning configurations introduce additional sustainability challenges. Communication costs between clients and a central server can be high, especially if the model is large or the number of clients is large. Techniques like gradient compression and asynchronous updates reduce overhead but may compromise convergence or robustness. The energy consumption of a federated learning system that spans hundreds of hospitals worldwide is non-negligible and must be weighed against the clinical benefits of improved model accuracy. Carbon footprint considerations are increasingly relevant for institutional review boards and funding agencies.

The trade-off between model performance and system complexity is perhaps the most salient. Adding continual learning and adversarial robustness layers increases the number of hyperparameters, the risk of implementation bugs, and the difficulty of debugging failures. A simpler static model with periodic manual updates may be more reliable in some contexts, even if it is less adaptive. The decision to adopt a continual learning architecture should be driven by the rate of disease evolution and the anticipated adversarial threat. For slow-moving diseases like certain cancers, a manually updated model with regular retraining every six months may suffice. For pandemic response, an automated continual learning system with robust defenses is indispensable.

7. Future Directions and Policy Recommendations

Future research should focus on developing theoretical guarantees for continual learning under adversarial conditions specific to medical domains. Current guarantees from robust optimization assume i.i.d. data, but continual learning introduces temporal dependencies that break these assumptions. New frameworks that bound the cumulative regret or the worst-case degradation in robust accuracy over a sequence of tasks would provide a foundation for certification. Practical benchmarks that simulate evolving disease landscapes with realistic attack models are needed to evaluate systems before deployment.

Policymakers should consider adaptive regulatory pathways that allow for ongoing model changes under strict monitoring. The concept of a total product lifecycle regulatory framework, as proposed for software as a medical device, can be extended to include continual learning with mandatory reporting of performance drifts [23]. Standards organizations should develop protocols for adversarial testing during each model update cycle, possibly using third-party red teams. Additionally, liability frameworks must be clarified: if a continually learning model causes harm because of an update that introduced a new vulnerability, who is responsible? The developer, the deploying institution, or the regulator who approved the adaptive algorithm?

Interdisciplinary collaboration between computer scientists, clinicians, ethicists, and legal scholars is essential to ensure that continual learning medical AI agents serve the public good. The integration of these technologies must be guided by principles of beneficence, non-maleficence, autonomy, and justice. As the required reference [18] underscores, security enhancement for large language models in medical decision-making is a pressing research area that cannot be divorced from the system-level considerations discussed here. Only by addressing architectural, governance, and sustainability challenges holistically can we realize the promise of truly adaptive and robust medical AI.

8. Conclusion

Continual learning for adversarially robust medical AI agents represents a confluence of multiple demanding requirements: adaptation to evolving disease landscapes, defense against malicious perturbations, adherence to fairness and privacy, and sustainable deployment in heterogeneous clinical environments. This paper has examined the system-level trade-offs inherent in such an endeavor, arguing that no single algorithmic solution suffices. Instead, a holistic architecture integrating memory mechanisms, adversarial training, uncertainty quantification, federated learning, and continuous monitoring is necessary. Governance structures must be redesigned to accommodate adaptive algorithms while preserving safety and equity. The path forward demands rigorous research, thoughtful regulation, and collaborative engagement across disciplines. With careful design, continual learning can empower medical AI to not only keep pace with disease evolution but also withstand the adversarial pressures that threaten its integrity.

References

1. 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.

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

3. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.

4. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

5. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.

6. Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. Advances in Neural Information Processing Systems, 30, 6467–6476.

7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.

8. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition, 110, 107641.

9. Finlayson, S. G., Beam, A. L., & van der Schaar, M. (2021). Adversarial attacks on medical machine learning: A visual explanation. The New England Journal of Medicine, 384(24), 2333–2335.

10. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. International Conference on Machine Learning, 80, 274–283.

11. Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... & Hadsell, R. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.

12. Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. Advances in Neural Information Processing Systems, 30, 2990–2999.

13. French, R. M. (1999). Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4), 128–135.

14. Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. Advances in Neural Information Processing Systems, 33, 1633–1645.

15. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.

16. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. International Conference on Machine Learning, 70, 1126–1135.

17. Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). iCaRL: Incremental classifier and representation learning. IEEE Conference on Computer Vision and Pattern Recognition, 2001–2010.

18. 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.

19. Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 30, 119–129.

20. Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. ACM Workshop on Artificial Intelligence and Security, 3–14.

21. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 48, 1050–1059.

22. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

23. US Food and Drug Administration. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). FDA Discussion Paper.

24. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

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

26. Wang, J., Cao, Z., Parada, P., & Song, Y. (2020). Lifelong learning with a mobile edge computing framework for medical image analysis. IEEE Transactions on Mobile Computing, 21(2), 630–643.