Multi-Agent Collaboration with Adversarial Filtering for Reliable Medical Diagnosis Support

Nathan Hawkins
School of Computing, Clemson University, Clemson, SC, USA.
hawkins146@clemson.edu

Keith Bell
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
keith.work@unr.edu

Abstract

The integration of artificial intelligence into clinical decision support systems promises to augment diagnostic accuracy and operational efficiency, yet the adoption of such systems remains constrained by concerns over reliability, robustness, and adversarial vulnerability. This paper proposes a novel architectural framework that combines multi-agent collaboration with adversarial filtering to enhance the trustworthiness of medical diagnosis support. The system comprises a suite of specialized diagnostic agents, each trained on distinct data modalities or clinical subdomains, whose outputs are aggregated through a central arbitrator. Prior to aggregation, an adversarial filtering module screens each agent contribution for potential manipulated or out-of-distribution inputs, thereby mitigating the impact of adversarial perturbations and improving overall system resilience. We examine the structural trade-offs inherent in such a distributed architecture, including the balance between agent specialization and coordination overhead, the computational cost of real-time filtering, and the implications for governance and fairness. Drawing on perspectives from large-scale socio-technical systems, we discuss deployment considerations such as interoperability with existing health information infrastructures, regulatory compliance, and sustainability of model maintenance. Ethical dimensions around algorithmic bias, accountability, and patient safety are analyzed within the context of adversarial filtering mechanisms. The paper further explores policy implications for certification and continuous monitoring of medical AI systems. By situating the proposed framework within the broader landscape of robust and trustworthy AI, we argue that multi-agent architectures enhanced with adversarial defenses offer a viable path toward reliable clinical decision support. The work contributes a systems-level blueprint for future research and practical implementation in high-stakes medical environments.

Keywords

multi-agent systems, adversarial filtering, medical diagnosis support, robust AI, clinical decision support, socio-technical infrastructure, governance, algorithmic fairness.

1. Introduction

Artificial intelligence has demonstrated remarkable potential in medical diagnosis, achieving performance comparable to or exceeding human experts in tasks such as dermatology classification, radiology interpretation, and pathology pattern recognition [1][2]. However, the translation of these capabilities into routine clinical practice has been slow, hampered by

concerns about the reliability of AI outputs under real-world conditions. One critical vulnerability is the susceptibility of deep learning models to adversarial perturbations—small, often imperceptible changes to input data that can cause a model to produce incorrect and potentially harmful outputs [4]. In medical contexts, adversarial attacks could be directed against imaging data, electronic health records, or even natural language inputs from clinical notes, with severe consequences for patient safety. The problem is compounded by the fragmentation of medical AI systems into siloed, single-purpose models that lack the redundancy and cross-checking capabilities inherent in human diagnostic teams.

Multi-agent systems offer a promising alternative by distributing diagnostic reasoning across multiple specialized agents that can collaborate, critique, and converge on a consensus diagnosis [3]. Such architectures mimic the collaborative nature of multidisciplinary medical teams, where radiologists, pathologists, and clinicians integrate their expertise. However, multi-agent systems themselves are not immune to adversarial threats; an adversary could compromise a subset of agents or inject malicious inputs into the communication channel. To address this, we introduce a framework that incorporates an adversarial filtering module operating as a gatekeeper between the agent ensemble and the final decision arbitrator. This module evaluates each agent's output for consistency, plausibility, and adherence to learned distributions of trustworthy responses, effectively screening out contributions that are likely adversarial or anomalous.

The present paper provides a comprehensive system-level analysis of this architecture. We examine the design choices that govern the trade-off between diagnostic accuracy and robustness, the infrastructural requirements for deployment in healthcare settings, and the governance mechanisms needed to ensure fairness and accountability. The analysis draws on recent advances in adversarial robustness research, particularly the work of Hu [5] on security enhancement methods for adversarial robust large language model intelligent agents in medical decision-making tasks, which underscores the importance of filtering mechanisms in high-stakes natural language processing applications. By integrating multi-agent collaboration with adversarial filtering, we aim to contribute a foundational framework that addresses both the technical and socio-technical dimensions of reliable medical diagnosis support.

2. Background and Related Work

The development of AI for medical diagnosis has evolved from single-model classifiers to more sophisticated ensembles and collaborative frameworks. Early deep learning systems demonstrated high accuracy on benchmark datasets but often failed when deployed in heterogeneous clinical environments due to distribution shift, label noise, and adversarial exploitation [1][2]. Concurrently, the field of adversarial machine learning revealed that even state-of-the-art models could be deceived with high confidence [4]. Finlayson et al. [10] specifically demonstrated the feasibility of adversarial attacks on medical machine learning systems, including the modification of medical images to cause misdiagnosis. These findings motivated the exploration of defensive mechanisms such as adversarial training, input sanitization, and ensemble methods.

Multi-agent systems have been studied extensively in distributed AI, with applications ranging from robotics to information retrieval [3]. In healthcare, multi-agent architectures have been proposed for collaborative diagnosis, treatment planning, and clinical workflow management [23]. The key advantage of such systems is their ability to combine diverse perspectives and mitigate individual model weaknesses through consensus. However, the security of the multi-agent communication and decision process remains underexplored.

Adversarial filtering, as a specific defense, involves detecting and removing manipulated or out-of-distribution inputs before they influence the system's final output. This concept draws on anomaly detection, adversarial detection, and robust statistics. Hu [5] proposed security enhancement methods specifically for large language model agents in medical decision-making, employing filtering layers that assess the semantic coherence and factual consistency of agent outputs. This approach aligns with the broader goal of building trustworthy AI systems that can withstand both intentional attacks and accidental data corruption.

Beyond technical defenses, the reliability of medical AI requires attention to governance, fairness, and regulatory compliance. Obermeyer and Emanuel [19] highlighted the risk of algorithmic bias in clinical predictions, while Char et al. [7] discussed ethical challenges in implementing machine learning in healthcare. The National Academy of Medicine has called for continuous monitoring and auditing of AI systems [17]. Multi-agent architectures, with their modular structure, facilitate such auditing by isolating the performance of individual agents and enabling independent validation. Adversarial filtering adds an extra layer of accountability by providing a mechanism to log and review screened contributions.

3. Multi-Agent Collaborative Architecture for Medical Diagnosis

The proposed architecture consists of three primary components: a set of diagnostic agents, an adversarial filter, and an arbitrator. Each agent is trained on a specific aspect of the clinical problem—for example, one agent may specialize in chest radiograph interpretation, another in laboratory value trend analysis, a third in natural language processing of patient history, and a fourth in genomic markers. This specialization allows each agent to achieve high accuracy within its domain while also providing redundancy; if one agent is compromised or encounters an out-of-distribution input, the others may still produce reliable outputs.

The agents operate in parallel, receiving either the same input (e.g., a patient case summary) or different modalities of the same case. They produce diagnostic outputs that include a predicted diagnosis, confidence scores, and supporting evidence. These outputs are sent to the adversarial filter before reaching the arbitrator. The filter's role is to evaluate each agent's contribution for signs of adversarial manipulation or internal failure. The filter may employ techniques such as consistency checks between agents (e.g., cross-validation of a finding present in both imaging and laboratory data), distributional anomaly detection (e.g., using density estimators trained on normal agent output distributions), and adversarial detection networks that flag inputs with high uncertainty or unusual gradient patterns. Importantly, the filter does not simply reject outliers; it provides a graded risk score for each agent contribution, which the arbitrator can then weight accordingly.

The arbitrator aggregates the filtered contributions to produce a final diagnosis. Aggregation methods can range from simple voting or averaging to more complex Bayesian fusion that accounts for agent reliability and filter confidence. The arbitrator also has the authority to request additional information or a second opinion from the human clinician. This human-in-the-loop component is critical for high-stakes decisions where confidence is low. The overall architecture is modular and extensible: new agents can be added or retired without disrupting the entire system, and the filter can be updated independently as adversarial threats evolve.

4. Adversarial Filtering: Design and Mechanisms

Adversarial filtering in our framework serves as a prophylactic defense that operates in real time, screening agent outputs before they are used for decision making. The design of the filter must balance sensitivity and specificity: too aggressive filtering may reject legitimate

but unusual cases (e.g., rare diseases), while too permissive filtering may allow adversarial attacks to propagate. Several mechanisms are employed to achieve this balance.

First, the filter uses a set of statistical and learning-based detectors trained on a corpus of agent outputs from both benign and adversarially manipulated scenarios. These detectors include density-based approaches that model the typical output distribution of each agent using techniques like kernel density estimation or variational autoencoders. An output that lies in a low-density region is flagged as suspicious. Second, the filter performs cross-agent consistency checks: if one agent's output contradicts the majority of other agents on a factual assertion, that output is given a high risk score. This leverages the diversity of the agent ensemble, as an adversary would need to compromise multiple agents simultaneously to maintain consistency. Third, the filter incorporates a meta-learning component that adapts to emerging attack strategies over time, updating its detection boundaries based on a continuous stream of labeled and unlabeled data from deployment.

A crucial consideration is the computational overhead of filtering. In a clinical setting where decisions may need to be made within minutes, the filter must operate with low latency. This can be achieved through parallel processing and hardware acceleration, as well as by using lightweight detector architectures for initial triage, with more expensive detectors applied only to borderline cases. The trade-off between detection accuracy and computational cost is a key engineering challenge that we address by proposing a hierarchical filter structure: a fast, approximate detector first filters clear benign outputs; ambiguous outputs are then passed to a more sophisticated detector; and only those that remain suspicious are flagged for human review.

The adversarial filter also plays a role in maintaining the long-term robustness of the system. As medical knowledge evolves and new adversarial techniques emerge, the filter must be updated periodically. This requires a governance framework that includes regular retraining of detectors, validation against an independent test set, and documentation of filter performance metrics. Hu [5] emphasizes that such dynamic updating is essential for maintaining defense effectiveness against adaptive adversaries in medical language modeling tasks.

5. System-Level Trade-Offs and Robustness

The integration of multi-agent collaboration and adversarial filtering introduces several structural trade-offs that must be carefully managed. One major trade-off is between agent specialization and communication overhead. Highly specialized agents can achieve superior performance on their subdomain, but the arbitration process becomes more complex as the number of agents grows. The adversarial filter must process more inputs, increasing latency and the risk of false positives. Conversely, reducing the number of agents may simplify the system but diminish the benefits of diverse perspectives. Optimal agent granularity depends on the clinical application: for a broad diagnostic support system, perhaps five to ten agents covering major modalities may be appropriate, while for a narrow task such as skin lesion classification, two or three agents with different training algorithms might suffice.

Another trade-off involves the stringency of the adversarial filter. A strict filter that rejects a high fraction of outputs may reduce the risk of successful attacks but also increase the rate of false alarms, leading to unnecessary delays and clinician frustration. In medical settings, a false alarm that causes a clinician to order additional tests may be acceptable if it prevents a misdiagnosis, but excessive false alarms can erode trust and lead to filter abandonment. Therefore, the filter's threshold should be calibrated using cost-sensitive metrics that weigh

the harm of a missed attack against the cost of a false positive. This calibration should be informed by the specific diagnostic context and the prevalence of adversarial threats.

Robustness also extends to the system's ability to cope with failures in individual agents or the filter itself. The architecture should be designed with graceful degradation: if the filter fails, the arbitrator can fall back to a simple majority vote among agents, albeit with reduced security. If an agent fails (e.g., due to data drift), the other agents can still produce a diagnosis, possibly with lower confidence. Redundancy in the filter, such as using multiple independent detectors with distinct training data, can protect against a single point of failure. These design principles align with best practices in large-scale distributed systems and are essential for deployment in critical infrastructure.

6. Governance, Fairness, and Policy Implications

The deployment of a multi-agent diagnostic support system with adversarial filtering raises significant governance questions. Who is responsible when the system makes an error? Is it the developer of the adversarial filter, the trainer of the agent that produced the flawed output, the clinician who overruled the system, or the hospital that deployed it? Clear lines of accountability must be established, and the system should maintain audit trails that record which agent contributed what output, how the filter scored it, and what final decision was made. These records are crucial for post-hoc analysis and legal liability.

Fairness is another critical dimension. Multi-agent systems may exacerbate existing biases if the agents are trained on data that underrepresents certain demographic groups. The adversarial filter, if not carefully designed, could also introduce bias by disproportionately flagging outputs from agents that serve minority populations—for example, if the density model was trained predominantly on majority group data. To mitigate this, the filter must be evaluated for fairness across demographic subgroups, and its thresholds should be adjusted to ensure equal false positive and false negative rates. Raji et al. [22] argue that end-to-end algorithmic auditing is necessary to uncover such biases, and we advocate for incorporating fairness auditing into the continuous monitoring framework.

Policy implications include the need for regulatory standards for multi-agent medical AI systems. Current frameworks, such as those from the U.S. Food and Drug Administration, primarily address single-model software as a medical device. The modular nature of multi-agent systems with an intermediate filter challenges the traditional approval process. Regulators must consider whether each agent should be certified independently, or whether the entire ensemble should be approved as a system. Additionally, the adversarial filter itself must meet performance criteria. Hu [5]'s work on security enhancement methods suggests that certification bodies could require evidence of robustness against a common set of adversarial benchmarks. International harmonization of such standards would facilitate global deployment while ensuring patient safety.

7. Deployment and Sustainability Considerations

Deploying a multi-agent adversarial filtering system in a real healthcare environment requires overcoming infrastructural and organizational barriers. The system must interface with hospital information systems, electronic health records, and picture archiving and communication systems. Data interoperability standards such as HL7 FHIR must be supported to allow agents to access relevant patient data. The adversarial filter and arbitrator must be integrated as middleware, possibly running on a secure cloud platform or on-premises cluster depending on privacy regulations.

Sustainability involves both financial and environmental costs. Training and maintaining multiple specialized agents is resource-intensive. However, transfer learning and pre-trained foundation models can reduce the need for extensive retraining. The adversarial filter, particularly if it uses deep learning detectors, also incurs computational costs. Over time, model drift and emerging adversarial threats necessitate periodic updates, which require sustained investment in data curation and model retraining. Organizations should plan for a lifecycle that includes continuous monitoring, performance evaluation, and retirement of obsolete agents.

From a human factors perspective, the system must be designed to support clinician trust. Explainability of agent outputs and filter decisions is essential. Clinicians need to understand why a particular agent's output was flagged or why the arbitrator reached a certain conclusion. Techniques from explainable AI, such as attention maps, counterfactual explanations, and natural language justifications, can be integrated into the user interface. As Ghassemi et al. [11] caution, current explainability methods may provide false reassurance, but careful design and validation can mitigate this risk. Training programs for clinicians on how to interpret system outputs and when to override them are equally important.

8. Conclusion

This paper has presented a comprehensive system-level framework for multi-agent collaboration with adversarial filtering to support reliable medical diagnosis. The architecture leverages the strengths of specialized diagnostic agents while defending against adversarial attacks through a dedicated filtering module. We have analyzed the structural trade-offs, including agent granularity, filter stringency, and computational overhead, and discussed the governance, fairness, and policy implications of deploying such systems in clinical environments. The incorporation of adversarial filtering, informed by recent advances such as those of Hu [5], provides a defense-in-depth approach that can enhance the robustness of medical AI. Future work should focus on empirical evaluation of the framework on real-world clinical datasets, the development of standardized benchmarks for adversarial robustness in multi-agent medical systems, and the design of adaptive filters that can learn from deployment feedback. As AI becomes increasingly embedded in healthcare, architectures that combine collaboration with security will be essential to realizing the promise of safe, equitable, and effective diagnostic support.

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. Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 38(2), 156–172.

4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).

5. 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.

6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

7. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.

8. Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In International Joint Conference on Artificial Intelligence (pp. 4691–4697).

9. 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.

10. 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.

11. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.

12. Horvitz, E. (2022). AI in healthcare: Balancing opportunities and risks. Journal of the American Medical Association, 328(12), 1211–1212.

13. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., … & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), 230–243.

14. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM Computing Surveys, 55(2), 1–38.

15. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 1–9.

16. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: users' guides to the medical literature. JAMA, 322(18), 1806–1816.

17. Matheny, M. E., Whicher, D., & Thadaney Israni, S. (2020). Artificial intelligence in health care: a report from the National Academy of Medicine. Journal of the American Medical Association, 323(6), 509–510.

18. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21.

19. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.

20. Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of Global Health, 8(2), 020303.

21. Park, Y., & Ho, J. C. (2021). Artificial intelligence in healthcare: A systematic review of applications and challenges. Healthcare Informatics Research, 27(1), 3–16.

22. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for

internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44).

23. Saria, S., Butte, A., & Saria, S. (2021). The case for multi-agent systems in clinical decision support. Journal of Biomedical Informatics, 114, 103674.

24. Stoyanovich, J., Van Bavel, J. J., & West, I. V. (2020). The imperative of computational transparency in AI. Nature Machine Intelligence, 2(10), 590–592.

25. Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731.