Benchmarking Adversarial Robustness of AI Medical Assistants in Emergency Triage Scenarios

Otis Reed
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
contactotis@buffalo.edu

Nikhil Ganguly
Department of Computer Science, University of North Texas, Denton, TX, USA.
gangulynikhil@unt.edu

Landon R. Gustafsson
Department of Computer Science, University of Houston, Houston, TX, USA.
landonrgustafsson368@uh.edu

Abstract

The increasing deployment of artificial intelligence in emergency medicine, particularly for triage decision support, raises critical questions about system resilience under adversarial manipulation. While AI medical assistants promise to reduce diagnostic delays and improve resource allocation, their reliance on deep learning models makes them vulnerable to crafted perturbations that can alter clinical recommendations. This paper proposes a comprehensive benchmarking framework for evaluating the adversarial robustness of AI medical assistants in emergency triage scenarios. We examine the structural trade-offs inherent in current system architectures, including the tension between model accuracy and robustness, the role of input modality heterogeneity, and the deployment constraints of real-time clinical environments. The analysis extends beyond technical metrics to encompass governance challenges, fairness implications, and policy requirements for trustworthy deployment. By integrating insights from adversarial machine learning, human factors engineering, and socio-technical systems theory, we identify critical failure modes that transcend conventional evaluation practices. The benchmarking framework incorporates multi-level stress testing, adaptive attack simulations, and clinical utility-preserving measures. We argue that robustness cannot be divorced from operational context and that standards must evolve to account for adversarial dynamics in triage workflows. This work contributes a systems-oriented perspective to the growing literature on adversarial machine learning in healthcare and provides concrete guidance for researchers, regulators, and clinical administrators seeking to responsibly integrate AI into emergency care.

Keywords

Adversarial robustness, AI medical assistants, emergency triage, benchmarking, system architecture, fairness, governance, socio-technical systems.

1. Introduction

Artificial intelligence systems are increasingly being integrated into emergency departments to support triage decisions, prioritizing patients based on severity of condition and available resources [1,2]. These AI medical assistants leverage large-scale neural networks trained on historical clinical data to predict outcomes such as hospital admission, mortality risk, or need

for intensive care [3]. However, the safety-critical nature of emergency triage demands that these systems maintain reliable performance even under unexpected or malicious input perturbations. Adversarial attacks, originally demonstrated in image classification by Szegedy et al. [4], have been shown to transfer to medical domains, where imperceptible modifications to radiology images or clinical text can flip a model's prediction [5,6]. This vulnerability is particularly dangerous in triage because a misclassification could delay care for a high-acuity patient or overwhelm resources for a low-acuity case. Existing research on adversarial robustness in medical AI has largely focused on individual model-level defenses, such as adversarial training and certified robustness [7,8], but has not sufficiently addressed the system-level integration challenges specific to emergency triage workflows. There is a pressing need for benchmarking methodologies that capture the unique constraints of clinical deployment: real-time latency requirements, multi-modal inputs (vital signs, imaging, free-text notes), and the interplay between human oversight and automated recommendations. This paper develops a benchmarking framework for adversarial robustness tailored to AI medical assistants in emergency triage. We examine structural trade-offs in system architecture, including the balance between model complexity and robustness, the role of input preprocessing pipelines, and the governance mechanisms needed to ensure fairness across demographic groups. By situating robustness within the broader socio-technical infrastructure of emergency care, we aim to shift the discourse from isolated model-level defenses to holistic system resilience.

2. Background and Related Work

Adversarial examples are inputs that have been intentionally perturbed to cause a machine learning model to produce an incorrect output [4]. In healthcare, such attacks have been demonstrated on medical imaging classifiers [5,9], clinical NLP systems [6,10], and risk prediction models [11]. The broader machine learning community has developed a range of defensive techniques, including adversarial training, wherein models are trained on perturbed examples [7,12]; gradient masking [13]; and certified defenses such as randomized smoothing [14]. However, these defenses often come at the cost of standard accuracy or increased computational overhead [7]. In emergency triage, the stakes are high, and the operational environment introduces additional complexities. Triage systems must process heterogeneous inputs: numerical vital signs, categorical complaint codes, free-text clinical narratives, and sometimes medical images [3,15]. Each modality has its own vulnerability surface. For example, an attacker could manipulate a patient's recorded heart rate by a small amount to shift a triage score, or inject misleading keywords into a clinical note to downgrade severity. Moreover, the triage recommendation is typically one of several possible outputs (e.g., Emergency Severity Index level 1 through 5), each associated with a specific resource allocation protocol. A successful attack that alters the recommendation by one level can have cascading effects on patient flow and staff workload. Research on adversarial robustness in this domain must therefore move beyond accuracy metrics and consider clinical utility preservation, human-machine teaming, and system-level failure propagation. The work by Hu [16] provides a foundational discussion on security enhancement methods for adversarial robust large language models in medical decision-making, but a comprehensive benchmarking approach for triage-specific scenarios remains absent. Similarly, prior studies on medical AI benchmarking have focused on natural distribution shifts rather than adversarial perturbations [17]. The present paper builds on these foundations by proposing a structured evaluation framework that accounts for the multi-layered nature of adversarial threats in emergency triage.

3. System Architecture and Deployment Considerations

An AI medical assistant for emergency triage typically comprises several interconnected components: a data ingestion layer that collects real-time patient information from electronic health records, wearable devices, and clinician inputs; a preprocessing pipeline that normalizes and encodes the data; a prediction engine that runs the trained model; and a decision support interface that presents the recommendation to the triage nurse or physician [2,3]. This architecture introduces multiple points of potential adversarial interference. An attacker could inject malicious inputs at the data ingestion stage, compromise the preprocessing logic (for example, by exploiting a vulnerability in a natural language parser), or directly manipulate the model weights or inference environment [16]. From a system perspective, the robustness of the entire pipeline is only as strong as its weakest component. Most existing defenses assume that the attacker has direct access to the model input, but in a clinical setting, an adversary may instead target the data source, such as by tampering with a sensor or altering a patient record. Therefore, a realistic benchmarking framework must consider a threat model that spans the full system stack. Furthermore, deployment constraints such as low latency (triage decisions must be made within minutes) and limited computational resources (edge devices in some emergency departments) constrain the types of defenses that can be applied. For example, iterative adversarial training or expensive certified defenses may be infeasible in real time [14]. Trade-offs between robustness and speed must be explicitly evaluated. Another architectural consideration is the use of ensemble methods or human-in-the-loop designs. A system that requires clinician confirmation of the AI recommendation may be more resilient to adversarial attacks, but also introduces cognitive load and potential automation bias [18]. Benchmarking should therefore include not only model-level robustness metrics but also human-centered measures such as time to decision, override rates, and diagnostic accuracy under adversarial conditions. The governance of such systems also involves audit trails, logging mechanisms, and post-hoc explanation capabilities to detect and attribute failures [19]. These components must be tested against adversarial scenarios to ensure that the system as a whole can maintain safe operation.

4. Adversarial Threat Models in Emergency Triage

The design of a benchmarking framework begins with a clear specification of the adversarial threat model. In emergency triage, the attacker's goals can be categorized as targeted or untargeted, white-box or black-box, and opportunistic or strategic [4,5]. A targeted attack might aim to downgrade a high-acuity patient to a lower triage level, thereby delaying care and causing harm. An untargeted attack may simply cause any misclassification, leading to resource misallocation and system chaos. The attacker's capability is also critical: a black-box attacker may have no knowledge of the model internals and rely on transferable adversarial examples generated using surrogate models [7,20]. A white-box attacker can exploit full gradient information to craft optimized perturbations. In a clinical setting, the adversary could be an external entity (e.g., a hacker compromising an interface) or an insider (e.g., a disgruntled staff member manipulating input data). The benchmarking framework must include all these scenarios. Additionally, domain-specific constraints shape the feasibility of attacks. For instance, adversarial perturbations to vital signs must remain within physiologically plausible ranges to avoid immediate detection by human reviewers [11]. Similarly, modifications to free-text clinical notes must preserve grammatical coherence and clinical plausibility to evade manual inspection [6,10]. This introduces a realistic boundary for adversarial perturbation magnitude, which differs from the typical Lp-norm constraints used

in computer vision. The benchmark should therefore define modality-specific perturbation budgets based on clinical expert input. Another critical dimension is the timing of the attack. Triage decisions are made under time pressure, and an adversary could delay or manipulate data in real-time streaming scenarios. For example, a sensor feeding heart rate data could be spoofed to slowly drift values over several minutes, making the attack harder to detect than a sudden spike. Adversarial robustness evaluation must incorporate temporal dynamics and continuous monitoring periods, not just single-shot input perturbations. Lastly, the attack surface includes not only the input but also the output: an attacker might manipulate the AI assistant's interface to display misleading recommendations, or intercept the communication channel between the model and the clinician. These system-level attacks require a broader security perspective than typical adversarial example research.

5. Benchmarking Framework: Metrics and Methodologies

A robust benchmarking framework for AI medical assistants in emergency triage should encompass multiple evaluation dimensions. First, we propose a set of robustness metrics that go beyond classification accuracy under attack. Specifically, for each triage level, we compute the worst-case misclassification rate under a given threat model, as well as the expected clinical cost of misclassification, where costs can be derived from expert-defined harm scores (e.g., a false negative for a high-acuity patient carries higher cost than a false positive for a low-acuity patient). This translates adversarial robustness into clinically meaningful terms. Second, framework should measure the efficacy of deployed defenses, including adversarial training, input sanitization, and detection-based defenses (e.g., anomaly detection on input statistics). For each defense, we assess not only the reduction in attack success but also the impact on standard accuracy, inference latency, and computational resource usage. Third, we introduce scenario-based stress tests that simulate realistic attack sequences. For example, one scenario might involve an adversary manipulating vital sign streams over a ten-minute window while also inserting plausible but misleading text in the patient's chief complaint. The benchmark should record the system's overall triage output, clinician override behavior (if human-in-the-loop is present), and the time to detect the anomaly. Fourth, the framework must evaluate fairness under adversarial conditions. Adversarial vulnerability can be unevenly distributed across demographic groups due to differences in data representation or model sensitivity [21]. A robust system should maintain equitable performance across race, gender, age, and socioeconomic status, even when under attack. We propose group-wise metrics such as differential robustness, defined as the maximum difference in worst-case error rates between groups. Fifth, the framework should incorporate reproducibility and transparency requirements. All evaluation datasets, attack implementations, and defense configurations should be documented systematically to enable independent verification [17]. Given the scarcity of public triage datasets with adversarial perturbations, we also recommend a methodology for generating synthetic but clinically plausible adversarial examples using generative models, with validation by domain experts. Finally, the framework should allow for incremental updates as new attack techniques and defenses emerge, similar to the model of continual learning benchmarks.

6. Structural Trade-offs and Robustness

The pursuit of adversarial robustness in emergency triage introduces fundamental trade-offs that must be carefully managed. The most well-known trade-off is between standard accuracy and adversarial robustness: models trained to be robust often exhibit lower performance on clean, unperturbed data [7,22]. In a triage setting, reducing accuracy on routine cases may

erode clinician trust and reduce the utility of the AI assistant. Yet accepting vulnerability to adversarial perturbations poses unacceptable safety risks. One way to address this tension is to adopt a hybrid architecture that uses a robust model for high-stakes decisions (e.g., high-acuity suspicious cases) and a more accurate but less robust model for low-stakes cases, with a decision rule that switches between them based on an input uncertainty estimate. Another trade-off involves interpretability: simpler, more interpretable models (e.g., decision trees) may be inherently more robust to small perturbations but lack the predictive power of deep neural networks [23]. Conversely, deep models offer high accuracy but are notoriously brittle. A third trade-off is between robustness and latency: defenses that involve multiple forward passes (e.g., randomized smoothing [14]) or iterative optimization at inference time may violate the strict time constraints of emergency triage. In some cases, edge deployment with limited compute may force a compromise that favors speed over robustness. The benchmarking framework should explicitly measure these trade-offs and provide decision-makers with Pareto frontier curves showing achievable combinations of accuracy, robustness, latency, and cost. A further structural consideration is the integration of human oversight. While a fully automated system may be faster, it also centralizes risk. A triage system that requires clinician confirmation can mitigate some adversarial failures, but introduces the possibility of automation bias, where clinicians over-rely on the AI recommendation even when it is obviously wrong [18]. The distribution of decision authority between human and machine is itself a robustness parameter. For example, under high-uncertainty conditions, the system might default to a rule-based triage protocol that is immune to adversarial perturbations but less adaptive. The benchmarking framework must evaluate these governance choices as part of the system's overall resilience.

7. Governance, Fairness, and Policy Implications

Deploying an AI medical assistant in emergency triage without robust adversarial safeguards raises serious ethical and legal concerns. The notion of informed consent is complicated when patients are unaware that their clinical data may be manipulated by an adversary [19]. Furthermore, if an attack leads to patient harm, determining liability between the system developer, the hospital, and the attacker is legally ambiguous. Regulatory bodies such as the U.S. Food and Drug Administration have begun to issue guidance on AI/ML-based medical devices, but adversarial robustness testing is not yet a standard requirement for premarket approval [24]. The benchmarking framework proposed here could serve as a basis for developing regulatory standards, analogous to the concept of adversarial validation sets used in other safety-critical domains. Fairness is another central governance concern. Studies have shown that adversarial examples can disproportionately affect underrepresented groups due to biases in training data [21]. In emergency triage, if an AI assistant is less robust for certain racial or socioeconomic groups, those patients could experience higher rates of undertriage or overtriage under attack. The framework's differential robustness metric allows regulators to set acceptable thresholds for group disparities. Moreover, the process of designing adversarial defenses itself may introduce biases. For example, a robust model trained primarily on data from a single hospital may fail to generalize to diverse populations, and the adversarial examples used in training may encode latent demographics [25]. Therefore, the benchmarking process must include transparency requirements for training data composition and attack generation methodology. Policy implications also extend to the procurement and lifecycle management of these systems. Hospitals should mandate periodic adversarial stress testing as part of their AI governance protocols, with results reported to a centralized repository for independent auditing. The costs of such testing (computational, financial, and labor) must be

weighed against the potential harm from adversarial failures. Given the fast-evolving nature of attack techniques, the framework should incorporate a mechanism for continuous re-evaluation and updating of threat models. Finally, international collaboration is needed to harmonize benchmarking standards, as emergency triage AI systems are increasingly deployed across borders with different regulatory environments.

8. Conclusion

This paper has presented a comprehensive framework for benchmarking the adversarial robustness of AI medical assistants in emergency triage scenarios. We have argued that current evaluation practices, which focus narrowly on model-level accuracy under perturbation, are insufficient for the complex, multi-modal, and time-critical context of emergency care. The proposed framework integrates system-level threat models, clinical utility metrics, human-in-the-loop considerations, fairness auditing, and governance requirements. Structural trade-offs between accuracy, robustness, latency, interpretability, and cost must be systematically evaluated using Pareto analysis. The work by Hu [16] underscores the importance of security enhancement methods for large language models in medical decision-making, and our framework extends this perspective to the triage domain. Future research should focus on empirical deployment studies that apply the framework across multiple hospital systems and attack scenarios. Additionally, the development of open-source benchmark datasets with clinically validated adversarial examples would accelerate progress. As AI systems become embedded in emergency care, adversarial robustness must be treated not as an optional feature but as a foundational safety requirement. The framework proposed here offers a roadmap for achieving that goal while maintaining the trust and equity that every patient deserves.

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. Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning applications. NPJ Digital Medicine, 4(1), 110.

4. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

5. 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.

6. Tschiatschek, S., Patil, K., & Ghahramani, Z. (2019). Adversarial attacks on clinical NLP models. Proceedings of the Machine Learning for Health Workshop at NeurIPS, 85, 95-110.

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. Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. International Conference on Machine Learning, 97, 7472-7482.

9. 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, 107575.

10. Xu, H., Li, Y., Huang, K., & Zhu, X. (2020). Adversarial attacks on clinical language models. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1639-1642.

11. Goetz, L. H., & Lehmann, H. P. (2021). Adversarial vulnerabilities in clinical risk prediction models. Journal of the American Medical Informatics Association, 28(6), 1195-1202.

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

13. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2017). Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy, 582-597.

14. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. International Conference on Machine Learning, 97, 1310-1320.

15. Madjarov, M., Radev, D., & Mihaylova, T. (2021). Machine learning for emergency triage: A systematic review. Artificial Intelligence in Medicine, 117, 102087.

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. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Steinhardt, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. International Conference on Computer Vision, 8340-8349.

18. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121-127.

19. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689.

20. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. International Conference on Learning Representations.

21. Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2019). Use privacy in data-driven systems: Theory and experiments. Proceedings of the 2019 ACM Conference on Computer and Communications Security, 253-270.

22. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. International Conference on Learning Representations.

23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

24. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.

25. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.