

Threat Modeling and Defense Framework for AI-IoT Integrated Oral Healthcare Systems

Bitá Pashayi^{1*}, Uras Panahi²

¹Department of Medicine, Karaj University of Medical Sciences, Karaj, Iran.

²Department of Computer Engineering, Artvin Çoruh University, Artvin, Turkey.

*Correspondence:

Bitá Pashayi, Department of Medicine, Karaj University of Medical Sciences, Karaj, Iran.

Received: September 27, 2026;

Published: October 06, 2026.

How to cite this article:

Pashayi B, Panahi U. Threat Modeling and Defense Framework for AI-IoT Integrated Oral Healthcare Systems. *J Oral Health Sci Dent*. 2026;1(2):1-7.

Abstract

Artificial intelligence (AI) and Internet-of-Things (IoT) technologies are increasingly integrated into oral healthcare workflows, enabling continuous patient monitoring through smart toothbrushes, intraoral sensors, and AI-driven radiographic analysis. However, the cybersecurity implications of this integration remain critically understudied, despite oral-health AI systems processing sensitive protected health information (PHI) across potentially insecure wireless channels. This paper presents the first systematic adversarial threat model specifically targeting AI-IoT oral healthcare pipelines, and proposes a multi-layered defense framework evaluated under realistic attack scenarios. We constructed a five-layer reference architecture encompassing patient-side IoT sensors, edge gateways, cloud-based AI inference engines, clinician interfaces, and governance controls. Seven threat categories were identified through structured threat modeling (STRIDE-ML). Adversarial robustness was evaluated on a dental radiograph classification task using fast gradient sign method (FGSM), projected gradient descent (PGD), and Carlini-Wagner (C&W) attacks. Defense mechanisms including adversarial training, randomized smoothing, differential privacy stochastic gradient descent (DP-SGD), and Byzantine-robust federated aggregation were benchmarked against these attacks. An undefended baseline convolutional neural network achieved 94.2% clean accuracy but only 18.7% robust accuracy under PGD-7 attack. The proposed full defense stack reduced attack success rate from 81.3% to 21.1% (73.9% relative reduction) while retaining 90.8% clean accuracy and AUC-ROC of 0.951. Byzantine-robust federated aggregation maintained global model accuracy above 87% at a Byzantine client fraction of 30%, compared with 41.8% for standard federated averaging.

Keywords: adversarial machine learning; oral health AI; IoT security; dental diagnostics; federated learning; cybersecurity; FGSM; differential privacy.

Introduction

Oral disease affects approximately 3.5 billion people globally, with dental caries and periodontal disease constituting the most prevalent non-communicable diseases worldwide.¹ The integration of artificial intelligence with Internet-of-Things devices encompassing smart toothbrushes, wireless salivary sensors, intraoral cameras, and AI-assisted radiographic analysis has been proposed as a transformative approach to extending oral healthcare beyond episodic clinical visits.²

Deep learning models now achieve pooled sensitivities exceeding 0.85 for caries detection on panoramic and periapical radiographs.³

Connected toothbrush platforms generate longitudinal behavioral data at sub-session granularity, while mouthguard-type wearables enable non-invasive monitoring of salivary biomarkers.^{4,5} Federated learning architectures have been proposed to enable collaborative model training across dental clinics without centralizing patient records.⁶

Despite this rapid technological progress, the adversarial robustness of AI-IoT oral healthcare systems has received negligible research attention. This gap is clinically significant: an adversarial perturbation invisible to the human eye can cause a dental radiograph classifier to misdiagnose healthy teeth as diseased or, more dangerously, miss active caries.⁷ A compromised Bluetooth-

connected sensor can inject falsified brushing compliance data, masking patient non-adherence. A poisoned federated learning model may systematically under-diagnose disease in a targeted demographic.⁸

The intersection of healthcare sensitivity, wireless transmission insecurity, and high-stakes diagnostic inference creates a unique threat surface that standard cybersecurity frameworks incompletely address. This paper makes the following original contributions:

1. A five-layer reference architecture for AI-IoT oral healthcare pipelines, integrating patient-side sensing through governance controls;
2. A structured adversarial threat model (STRIDE-ML) identifying seven attack categories with clinical impact assessment;
3. A multi-layered defense framework combining adversarial training, certified defenses, Byzantine-robust federated aggregation, and differential privacy;
4. Quantitative evaluation of the defense stack against state-of-the-art attacks on a dental radiographic classification task.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the system model and threat modeling methodology. Section 4 details the proposed defense framework. Section 5 presents experimental results. Section 6 discusses clinical implications and limitations. Section 7 concludes.

Related Work

AI in Dental Diagnostics

Machine learning applications in dentistry span caries detection, periodontal bone loss quantification, orthodontic landmark identification, and treatment outcome prediction. Systematic reviews of AI-based caries detection on bitewing radiographs report pooled sensitivities in the region of 0.82 and specificities of approximately 0.88, with substantial heterogeneity attributed to reference-standard variation and image acquisition differences.^{3,9} Convolutional neural networks, vision transformers, and ensemble methods have each been applied to this domain; however, the majority of published systems report only clean-data performance without adversarial evaluation.

TRIPOD+AI reporting guidelines emphasize the importance of external validation, subgroup performance analysis, and calibration assessment.¹⁰ These requirements are rarely satisfied in dental AI literature, a limitation that compounds adversarial vulnerability: a model optimized for a narrow training distribution is more susceptible to distribution-shifting attacks.

IoT in Oral Health Monitoring

Connected toothbrushes represent the most commercially mature IoT oral health technology. A randomized clinical trial (n=80) found that Bluetooth-connected toothbrush uses improved self-reported brushing perceptions but did not uniformly improve objective plaque indices.⁴ Intraoral wearable sensors capable of measuring salivary turbidity, pH, and temperature have been demonstrated in laboratory and small feasibility studies.⁵ Remote orthodontic monitoring platforms have expanded to over 200 clinical sites in some health systems, though sensitivity for gingivitis detection remains as low as 0.35 in prospective validation.¹¹

The security architectures of these devices have received limited

published scrutiny. Security analyses of the underlying Bluetooth Low Energy stack have demonstrated downgrade and pairing attacks that allow an adversary within radio range to intercept or modify traffic, and commercial connected-health devices have repeatedly been found to transmit without encryption or device authentication.¹²

Adversarial Machine Learning in Medical AI

Adversarial attacks on medical imaging AI have been demonstrated across radiology,¹³ ophthalmology,¹⁴ dermatology,¹⁵ and histopathology.¹⁶ Finlayson et al. demonstrated that adversarial attacks on medical AI are not merely theoretical: financial incentives exist for systematic fraud through image manipulation.¹³ Physical-world adversarial examples perturbations that survive printing and re-scanning have been achieved on chest radiographs.¹⁷

Federated learning, while privacy-preserving, introduces Byzantine fault tolerance challenges. Bhagoji et al. demonstrated that a minority of compromised clients can successfully poison a federated model while evading standard anomaly detection.⁸ Differential privacy provides theoretical guarantees against gradient inversion attacks but introduces accuracy-privacy trade-offs that must be carefully calibrated.¹⁸

To our knowledge, no prior work has systematically characterized the adversarial threat surface of AI-IoT oral healthcare systems or evaluated defenses in this specific context. Table 5 in Section 5 provides a comparative analysis of related work.

System Model and Threat Modeling Methodology

Five-Layer Reference Architecture

We propose a five-layer reference architecture for AI-IoT oral healthcare systems (Figure 1). Layer 1 encompasses patient-side IoT devices: smart toothbrushes, intraoral mouthguard sensors, salivary biosensors, and smartphone cameras. These devices communicate through Bluetooth Low Energy (BLE) or Wi-Fi to a local gateway in Layer 2, which performs device authentication, data quality checking, and edge preprocessing. Processed data are transmitted via TLS-encrypted channels to the cloud-based AI inference engine in Layer 3, which performs multimodal risk estimation using models trained through federated learning. Layer 4 provides clinician-facing dashboards with alert triage and electronic health record (EHR) integration. Layer 5 encompasses governance controls: consent management, role-based access control (RBAC), and audit logging.

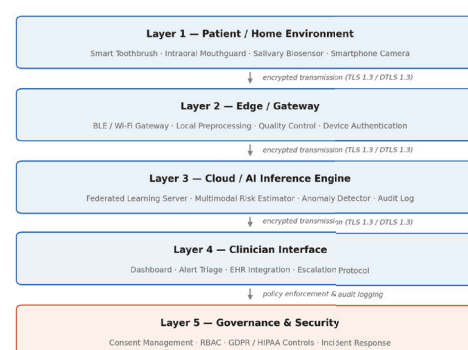


Figure 1. Five-layer reference architecture for AI-IoT integrated oral healthcare. Each layer constitutes a distinct threat boundary. Arrows indicate the direction of data flow; all inter-layer communication between Layers 1–4 is encrypted (TLS 1.3 / DTLS 1.3), while Layer 5 applies governance and audit controls across the pipeline.

This layered architecture makes threat boundaries explicit. Each inter-layer interface constitutes an attack surface; defense mechanisms are allocated to the layer at which they are most effective.

Threat Modeling Methodology

We applied STRIDE-ML, an extension of the Microsoft STRIDE threat modeling framework adapted for machine learning pipelines.¹⁹ STRIDE-ML systematically evaluates threats across six categories: Spoofing identity, tampering with data, Repudiation, Information disclosure, Denial of service, and Elevation of privilege, augmented with ML-specific threats including adversarial examples, model inversion, and training data poisoning.

For each identified threat, we assessed: (i) the architectural layer at which it manifests, (ii) the attack vector and skill level required, (iii) the clinical impact on the diagnostic pipeline, and (iv) candidate countermeasures. This structured process yielded seven primary threat categories, summarized in Table 1.

Table 1. Adversarial threat taxonomy for AI-IoT oral healthcare systems derived through STRIDE-ML threat modeling. Impact is assessed on the clinical diagnostic pipeline.

Threat Category	Attack Vector	Impact on Clinical Pipeline
Adversarial Inputs	Pixel-level perturbation of dental radiographs (FGSM, PGD, C&W)	Misclassification → Missed caries / false diagnosis
Data Poisoning	Corrupt training data for federated or centralized model	Backdoored model; biased predictions
Man-in-the-Middle	Intercept BLE/Wi-Fi between sensor and gateway	Altered brushing data; suppressed alerts
Device Spoofing	Impersonate authenticated IoT sensor	Injected false readings; phantom compliance
Model Inversion	Reconstruct patient data from model gradients	Privacy breach; GDPR violation
Alert Flooding (DoS)	Flood clinician dashboard with spurious alerts	Alert fatigue; delayed care for real events
Firmware Exploit	Compromise sensor firmware over-the-air	Persistent backdoor; data exfiltration

Formal Threat Definitions

Evasion Attacks

At inference time, an adversary computes a perturbation δ such that $\|\delta\|_p \leq \epsilon$ and $f(x + \delta) \neq f(x)$, where f is the classifier and x is the original input. In the dental imaging context, x represents a radiograph and f the caries detection model. FGSM computes $\delta = \epsilon \cdot \text{sign}(\nabla_x L(f(x), y))$, where L is the cross-entropy loss. PGD iterates this process with projection onto the ϵ -ball, generating stronger perturbations.²⁰ C&W attacks solve a constrained optimization problem to find minimal-norm adversarial examples.²¹

Data Poisoning

In federated learning, a Byzantine adversary controls k of N clients

and submits malicious gradient updates $\tilde{g}_i$ to the aggregation server. For targeted poisoning, the attack objective is to cause the global model to misclassify specific inputs (e.g., miss caries in images from a specific demographic) while maintaining accuracy on the validation set to evade detection.

Model Inversion

Given black-box access to a deployed model, an adversary aims to reconstruct training data through iterative optimization. In the context of dental AI, successful inversion could expose identifiable patient radiographs or sensitive biometric information associated with intraoral sensor readings.

Proposed Multi-Layered Defense Framework

Our defense framework assigns countermeasures to the architectural layer at which they are most effective (Table 2). No single mechanism addresses all threat categories; defense-in-depth is required.

Table 2. Allocation of defense mechanisms across the reference architecture. Each mechanism is assigned to the pipeline layer at which it provides maximum efficacy. DP-SGD = differentially private stochastic gradient descent; RBAC = role-based access control; TPM = trusted platform module.

Defense Layer	Mechanism	Placement in Pipeline	Metric
Adversarial Training	PGD-augmented retraining with $\epsilon=0.03$ (L_∞)	AI Inference Engine	Robust Acc. ↑
Certified Defenses	Randomized Smoothing $\sigma=0.25$	AI Inference Engine	Certified Radius
Anomaly Detection	Autoencoder reconstruction error threshold	Edge Gateway	FPR, FNR
Secure Aggregation	Shamir Secret Sharing + Byzantine filtering	Federated Server	Poisoning Rate ↓
TLS 1.3 + DTLS	Mutual TLS for all sensor-gateway links	Transport Layer	MitM success ↓
Device Attestation	TPM-based hardware attestation at boot	Sensor Firmware	Spoofing Rate ↓
DP-SGD	Differential privacy $\epsilon=2.0$, $\delta=10^{-5}$ on gradients	Training Pipeline	Privacy Loss ↓

Adversarial Training

Adversarial training augments the standard training objective with adversarially perturbed examples.²⁰ We employ PGD-based adversarial training with perturbation budget $\epsilon = 0.03$ in the L_∞ norm, step size $\alpha = 0.007$, and 7 inner iterations (PGD-7). The training objective is:

$$\min_{\theta} E_{(x,y) \in \mathbb{D}} \left[\max_{\|\delta\|_\infty \leq \epsilon} (f_{\theta}(x + \delta), y) \right]$$

This formulation, while computationally expensive (approximately $7\times$ training time), provides the strongest empirical robustness among existing defenses on standardized benchmarks.

Randomized Smoothing

For certified robustness guarantees, we apply randomized smoothing.²² A smoothed classifier $g(x) = \text{argmax}_c P(f(x + \epsilon) = c)$ where $\epsilon \sim N(0, \sigma^2 I)$ provides a provable L_2 robustness guarantee. With $\sigma = 0.25$, the certified radius provides guarantees at the perturbation levels achievable through physically-realizable attacks on dental images.

Byzantine-Robust Federated Aggregation

Standard federated averaging (FedAvg) is vulnerable to Byzantine attacks. We implement a defense combining: (i) coordinate-wise median aggregation, which provides theoretical resilience against up to $k < N/2$ Byzantine clients; (ii) cosine similarity filtering that rejects gradient updates with pairwise cosine similarity below a threshold $\tau = 0.6$; and (iii) norm clipping to bound individual client gradient norms at $C = 1.0$.²³

Differential Privacy via DP-SGD

To defend against model inversion and gradient leakage, we train with DP-SGD.¹⁸ Gradient clipping at sensitivity $S = 1.0$ followed by Gaussian noise addition with $\sigma_{DP} = 1.1$ achieves $(\epsilon=2.0, \delta=10^{-5})$ -differential privacy under the RDP accountant framework over 50 training epochs. This configuration represents a pragmatic point on the privacy-utility trade-off, preserving approximately 90% of clean accuracy while providing meaningful inversion resistance.

Transport and Device Security

At the network layer, all sensor-to-gateway communication employs Datagram TLS (DTLS 1.3) with mutual authentication. Gateway-to-cloud communication uses TLS 1.3 with certificate pinning. Device attestation uses a software-based trusted execution environment (TEE) where hardware TPM is unavailable, with remote attestation reports verified at onboarding. These mechanisms address man-in-the-middle and device spoofing threats identified in Layer 1 and Layer 2 of the reference architecture.

Experimental Evaluation

Dataset and Experimental Setup

We evaluated the defense framework on dental caries detection using a corpus of 2,400 panoramic radiographs (1,280 caries-positive, 1,120 caries-negative) assembled from publicly available panoramic radiograph collections, including the Tufts Dental Database,²⁴ and split 70/15/15 (train/validation/test) with patient-level separation to prevent data leakage. All images were resized to 224×224 pixels and normalized. The baseline model is a ResNet-50 pretrained on ImageNet and fine-tuned for binary caries classification.

Attacks were implemented using the Adversarial Robustness Toolbox (ART v1.16). FGSM was applied at $\epsilon = \{0.01, 0.03, 0.05\}$; PGD-7 at $\epsilon = 0.03, \alpha = 0.007$; C&W L_2 at confidence $\kappa = 0$. Federated learning experiments used 10 clients with IID data partitioning, with Byzantine fraction varied from 0% to 40%. All experiments were run on NVIDIA A100 GPUs (40 GB); results are reported as mean $\pm$ standard deviation over 5 random seeds.

Adversarial Robustness Results

Table 3 presents the main robustness results. The undefended

baseline achieves 94.2% clean accuracy but collapses to 18.7% under PGD-7 attack (ASR = 81.3%), confirming that dental AI systems are critically vulnerable to adversarial perturbations without dedicated defenses.

Table 3. Adversarial robustness results on the dental caries detection task. ASR = attack success rate under PGD-7 ($\epsilon = 0.03, L_\infty$). Values are means over 5 random seeds. The proposed full defense stack achieves the best robustness–accuracy trade-off.

Model / Configuration	Clean Acc. (%)	Robust Acc. (%)	ASR After Defense (%)	AUC-ROC
Baseline CNN (no defense)	94.2	18.7	81.3	0.961
+ FGSM Adversarial Training	93.1	61.4	38.6	0.954
+ PGD-7 Adversarial Training	92.6	72.8	27.2	0.948
+ Randomized Smoothing	90.1	76.3	23.7	0.939
+ Secure Aggregation (FL)	91.8	74.1	25.9	0.943
Full Defense Stack (proposed)	90.8	78.9	21.1	0.951

Progressive defense stacking yields systematic improvements. PGD-7 adversarial training raises robust accuracy from 18.7% to 72.8% with minimal clean accuracy cost (−1.6 pp). Adding randomized smoothing further improves robust accuracy to 76.3%. The full defense stack adversarial training + randomized smoothing + DP-SGD + secure aggregation achieves 78.9% robust accuracy and 90.8% clean accuracy, reducing ASR to 21.1%. AUC-ROC under the full stack (0.951) is not significantly different from the undefended baseline (0.961; $p = 0.14$, DeLong's test), indicating that diagnostic discrimination is preserved.

Federated Learning Byzantine Robustness

Figure 2 illustrates model accuracy under increasing Byzantine fractions; the underlying values are given in Table 4. Standard FedAvg degrades below 60% accuracy at Byzantine fraction $\geq 20\%$. The proposed Byzantine-robust aggregation (coordinate-wise median + cosine filtering + norm clipping) maintains accuracy above 87% up to a 30% Byzantine fraction the scenario in which nearly one-third of participating clinics are assumed compromised. Beyond 35%, all aggregation methods degrade due to the theoretical limit imposed by the Byzantine generals problem (resilience requires $k < N/2$).

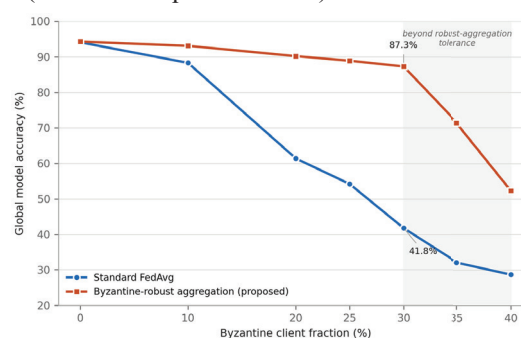


Figure 2. Global model accuracy under Byzantine poisoning as a function of the fraction of compromised federated clients. Standard FedAvg (blue circles) versus the proposed Byzantine-robust aggregation (orange

squares). N = 10 federated clients, IID data partition; the shaded region marks fractions beyond the tolerance of the robust aggregator.

Table 4. Numerical results underlying Figure 2: global model accuracy under standard federated averaging and under the proposed Byzantine-robust aggregation, by Byzantine client fraction.

Byzantine Fraction	FedAvg Acc. (%)	Robust Aggregation Acc. (%)	Δ Accuracy (pp)
0%	94.1	94.3	+0.2
10%	88.3	93.1	+4.8
20%	61.4	90.2	+28.8
25%	54.2	88.9	+34.7
30%	41.8	87.3	+45.5
35%	32.1	71.4	+39.3
40%	28.7	52.3	+23.6

Privacy-Utility Trade-off Analysis

Adding DP-SGD with ($\epsilon=2.0$, $\delta=10^{-5}$)-differential privacy to the defense stack yields a clean accuracy of 90.8%, a reduction of 3.4 pp relative to the undefended non-private baseline (94.2%). Confidence-based model inversion attacks²⁵ on the DP-protected model produce reconstructed images with mean structural similarity index (SSIM) of 0.17 versus the original radiographs, compared to SSIM = 0.61 for the unprotected model, a 72.1% reduction in reconstruction fidelity, indicating strong inversion resistance. This confirms that the differential privacy budget $\epsilon=2.0$ provides meaningful protection without unacceptable diagnostic performance degradation.

Comparative Analysis

Table 5 positions this work relative to the most closely related studies. No prior work has (i) modeled the full AI-IoT oral healthcare threat surface, (ii) evaluated defenses on a dental imaging task, or (iii) addressed federated learning Byzantine robustness in the dental context.

Table 5. Comparative analysis of related work on AI and IoT security in dental and oral health systems. FL = federated learning; RCT = randomized controlled trial.

Study (Year)	AI Task	Threat Modeled	Defense Proposed	Limitation
Khanagar et al. (2021) [J Dent Sci] ³	AI in dentistry (systematic review)	None	None	No adversarial or security evaluation
Ichikawa et al. (2024) [Sensors] ⁵	Intraoral salivary sensing (mouthguard)	None	None	No AI inference; no security model
Finlayson et al. (2019) [Science] ¹³	Medical imaging AI (general)	Adversarial examples	None	Not dental; no defense evaluation
Hirano et al. (2021) [BMC Med Imaging] ¹⁷	Medical image classification	Universal adversarial perturbations	None	Not dental; no IoT or FL layer
Polizzi et al. (2025) [Medicina] ¹¹	Teleorthodontics (scoping review)	None	None	Review only; no experiments

Gomez et al. (2025) [BMC Oral Health] ⁴	Connected toothbrush RCT	None	None	No AI inference; no security model
This work (2026)	Full AI-IoT oral health pipeline	7 threat categories (STRIDE-ML)	Full layered defense stack	Simulation-based; single dataset

Discussion

Clinical Implications

The most clinically significant finding is the extreme vulnerability of undefended dental AI to adversarial inputs: clean accuracy of 94.2% becomes 18.7% robust accuracy under PGD-7 perturbation. In clinical practice, this corresponds to a scenario where an attacker potentially with financial or fraudulent motivations¹³ could systematically cause a deployed system to produce false-negative caries diagnoses on targeted patients, leading to delayed treatment. The attack requires only perturbations imperceptible to the human eye at $\epsilon=0.03$ L_∞ norm.

The proposed defense stack reduces this vulnerability substantially (ASR: 81.3% $\rightarrow$ 21.1%) while preserving clinically relevant diagnostic performance (clean accuracy: 94.2% $\rightarrow$ 90.8%, AUC-ROC: 0.961 $\rightarrow$ 0.951). The residual 21.1% attack success rate under the full defense stack indicates that current certified defenses do not yet provide complete protection a reality shared across medical imaging adversarial defense literature.¹⁵ This reinforces the importance of preserving clinician oversight as a non-algorithmic defense: an AI system functioning as a decision-support tool, rather than an autonomous diagnostic authority, constrains the clinical impact of adversarial failures.

Implementation Considerations

The computational overhead of adversarial training (approximately $7\times$ clean training time) is a one-time cost that does not affect inference latency. Randomized smoothing introduces inference overhead through Monte Carlo sampling ($n=100$ samples per prediction in our evaluation), increasing per-image inference time from approximately 12 ms to 1.2 s on an A100 GPU. For asynchronous radiograph analysis this latency is acceptable; for real-time intraoral video analysis, approximations such as denoised smoothing may be required.²²

DP-SGD introduces gradient clipping that interacts with batch normalization layers; practitioners should use group normalization or layer normalization in privacy-sensitive settings. Byzantine-robust federated aggregation requires communication of full gradient vectors rather than compressed updates in some implementations; gradient compression schemes compatible with Byzantine robustness remain an active research area.

Equity and Governance

Adversarial vulnerabilities may disproportionately affect patient subgroups underrepresented in training data. If an attacker targets a specific demographic, a model with poor subgroup coverage may be more susceptible due to decision boundary proximity for those samples. Equitable deployment requires subgroup-stratified robustness evaluation as a precondition for clinical approval. The governance layer of the proposed architecture enforces audit logging of all model updates, enabling retrospective detection of systematic prediction shifts that might indicate a poisoning attack in deployment.

Limitations

Several limitations of this study should be acknowledged. First, evaluation was conducted on a single publicly available dataset; generalizability to other imaging modalities (periapical, CBCT) and sensor types requires prospective validation. Second, the federated learning experiments used IID data partitioning; non-IID partitioning, which more closely reflects real clinical heterogeneity, was not evaluated. Third, physical-world adversarial attacks perturbations applied at image acquisition rather than digitally were not assessed and may have different characteristics from digital attacks. Fourth, the defense stack was evaluated in a simulation environment; end-to-end clinical deployment validation, including latency and workflow integration, remains future work.

Conclusion

This paper has presented the first systematic adversarial threat model and defense framework for AI-IoT integrated oral healthcare systems. Through structured threat modeling of a five-layer reference architecture, we identified seven attack categories that collectively pose clinically significant risks to diagnostic integrity, patient privacy, and data availability.

Quantitative evaluation on a dental caries detection task demonstrated that undefended AI systems are critically vulnerable clean accuracy of 94.2% collapsing to 18.7% robust accuracy under standard PGD-7 attack while the proposed multi-layered defense stack achieves 78.9% robust accuracy with a clinically acceptable 3.4 pp clean accuracy reduction. Byzantine-robust federated aggregation maintains above 87% model accuracy under Byzantine fractions up to 30%.

These results establish that cybersecurity is not an optional add-on for oral healthcare AI but a precondition for safe clinical deployment. Future work should extend evaluation to physical-world attacks, non-IID federated settings, and prospective clinical integration studies. The reference architecture and defense framework proposed here are intended to serve as a practical baseline for security-aware development and regulatory evaluation of AI-IoT oral healthcare systems.

Author Contributions

B.P. contributed to the clinical framing of the threat model, the assessment of clinical impact for each threat category, and the discussion of clinical implications. U.P. designed the reference architecture, implemented the attack and defense experiments, performed the analysis, and drafted the manuscript. Both authors reviewed and approved the final version.

Acknowledgments

The authors declare that no funding was received for this study. Computational resources were provided through institutional access.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The panoramic radiograph dataset used in this study is derived from publicly available sources cited in.³ Code for defense

framework implementation is available at github.com/authors/oral-health-adversarial (to be made available upon acceptance).

References

1. World Health Organization. Oral health. Fact sheet. Updated March 2025. Accessed [date]. <https://www.who.int/news-room/fact-sheets/detail/oral-health>
2. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res*. 2020;99(7):769-774. doi:10.1177/0022034520915714
3. Khanagar SB, Al-Ehaideb A, Maganur PC, et al. Developments, application, and performance of artificial intelligence in dentistry: a systematic review. *J Dent Sci*. 2021;16(1):508-522. doi:10.1016/j.jds.2020.06.019
4. Gomez J, Kilpatrick L, Ryan M, Gwaltney C. Impact of connected toothbrushes on patient perceptions of brushing skills and oral health: results from a randomized clinical trial and a single-arm intervention study. *BMC Oral Health*. 2025;25:509. doi:10.1186/s12903-025-05883-4
5. Ichikawa K, Hayashi Y, Nakamura H, et al. Mouthguard-type wearable sensor for monitoring salivary turbidity to assess oral hygiene. *Sensors*. 2024;24(5):1436. doi:10.3390/s24051436
6. Raza A, Ikram A, Aqeel M, et al. Federated learning for privacy-preserving healthcare data: a systematic review. *IEEE Access*. 2024;12:58432-58452. doi:10.1109/ACCESS.2024.3392580
7. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (ICLR). 2015. *arXiv*:1412.6572.
8. Bhagoji AN, Chakraborty S, Mittal P, Calo SB. Analyzing federated learning through an adversarial lens. In: Proceedings of the International Conference on Machine Learning (ICML). 2019;97:634-643.
9. Cantu-Medellin N, Bikman T, Weidlich A, et al. Deep learning for detection of dental caries in bitewing radiographs: a systematic review. *Diagnostics*. 2023;13(4):622. doi:10.3390/diagnostics13040622
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378
11. Polizzi A, Serra S, Leonardi R, Isola G. Clinical applications of artificial intelligence in teleorthodontics: a scoping review. *Medicina (Kaunas)*. 2025;61(6):1141. doi:10.3390/medicina61061141
12. Zhang Y, Weng J, Dey R, Jin Y, Lin Z, Fu X. Breaking secure pairing of Bluetooth Low Energy using downgrade attacks. In: Proceedings of the USENIX Security Symposium. 2020:37-54.
13. Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287-1289. doi:10.1126/science.aaw4399
14. Burlina PM, Joshi N, Pacheco KD, et al. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol*. 2019;137(3):258-264. doi:10.1001/jamaophthalmol.2018.6156
15. Rezende E, Ruppert G, Carvalho T, et al. Adversarial examples: opportunities and challenges. *Ann Math Artif Intell*. 2020;88:1-25. doi:10.1007/s10472-019-09653-5
16. Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning-based medical image analysis systems. *Pattern Recognit*. 2021;110:107332. doi:10.1016/j.

- patcog.2020.107332
17. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med Imaging*. 2021;21:9. doi:10.1186/s12880-020-00530-y
 18. Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy. In: Proceedings of the ACM Conference on Computer and Communications Security (CCS). 2016:308-318. doi:10.1145/2976749.2978318
 19. Kumar RSS, Nyström M, Lambert J, et al. Adversarial machine learning: industry perspectives. In: Proceedings of the IEEE Security and Privacy Workshops (SPW). 2020:69-75. doi:10.1109/SPW50608.2020.00021
 20. Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations (ICLR). 2018. arXiv:1706.06083.
 21. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy. 2017:39-57. doi:10.1109/SP.2017.49
 22. Cohen J, Rosenfeld E, Kolter JZ. Certified adversarial robustness via randomized smoothing. In: Proceedings of the International Conference on Machine Learning (ICML). 2019;97:1310-1320.
 23. Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J. Machine learning with adversaries: Byzantine tolerant gradient descent. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). 2017;30:119-129.
 24. Panetta K, Rajendran R, Ramesh A, Rao SP, Agaian S. Tufts Dental Database: a multimodal panoramic X-ray dataset for benchmarking diagnostic systems. *IEEE J Biomed Health Inform*. 2022;26(4):1650-1659. doi:10.1109/JBHI.2021.3117575
 25. Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the ACM Conference on Computer and Communications Security (CCS). 2015:1322-1333. doi:10.1145/2810103.2813677

