



Adversarial robustness of dental panoramic radiograph classifiers against gradient-based perturbations

Robustez adversarial de clasificadores de radiografías panorámicas dentales frente a perturbaciones basadas en gradiente

Wenya Long¹

Teng He¹

Yingjuan Xie¹

¹ Zhengzhou University, The First Affiliated Hospital, Department of Hematology, Zhengzhou, China.

Submitted: 05-08-2026 | Revised: 30-09-2026 | Accepted: 29-10-2026 | Published: 01-01-2027

Corresponding Author: Wenya Long 

ABSTRACT

Introduction: Deep learning classifiers for dental radiographs are increasingly deployed, but their robustness to adversarial perturbations is unmeasured. Methods: A frozen ResNet-18 backbone and a linear head were trained on 182 annotated teeth from the DENTEX dataset and attacked with the fast gradient sign method and projected gradient descent at budgets from 1/255 to 16/255, comparing a baseline model, an adversarially trained model and a median-filter defense. Results: The baseline classified 63.6% of held-out teeth correctly. Both attacks flipped every correctly classified tooth at a budget of 1/255. The fast gradient sign method fell to 77.1% attack success at 16/255, whereas projected gradient descent remained at 100%. Neither adversarial training nor the median filter reduced attack success, and adversarial training lowered clean accuracy to 54.5%. Conclusions: A transfer-learning dental radiograph classifier with a frozen backbone and linear head is catastrophically vulnerable to adversarial perturbations, and the simple defenses tested do not restore robustness.

Keywords: adversarial robustness; adversarial examples; dental radiograph; deep learning; cybersecurity.

RESUMEN

Introducción: Los clasificadores de aprendizaje profundo para radiografías dentales se implementan cada vez más, pero su robustez frente a perturbaciones adversariales no se ha medido. Métodos: Se entrenaron una red troncal ResNet-18 congelada y una cabeza lineal con 182 dientes anotados del conjunto DENTEX y se atacaron con el método de gradiente de signo rápido y el descenso de gradiente proyectado con presupuestos de 1/255 a 16/255, comparando un modelo de línea de base, un modelo con entrenamiento adversarial y una defensa de filtro de mediana. Resultados: El modelo de línea de base clasificó correctamente el 63,6 % de los dientes de prueba. Ambos ataques cambiaron la predicción de todos los dientes correctamente clasificados con un presupuesto de 1/255. El método de gradiente de signo rápido cayó al 77,1 % de éxito con 16/255, mientras que el descenso de gradiente proyectado se mantuvo en el 100 %. Ni el entrenamiento adversarial ni el filtro de mediana redujeron el éxito del ataque, y el entrenamiento adversarial bajó la exactitud limpia al 54,5 %. Conclusiones: Un clasificador de radiografías dentales por transferencia con red troncal congelada y cabeza lineal es catastróficamente vulnerable a perturbaciones adversariales, y las defensas simples evaluadas no restauran la robustez.

Palabras clave: robustez adversarial; ejemplos adversariales; radiografía dental; aprendizaje profundo; ciberseguridad.

INTRODUCTION

Deep learning classifiers have become routine components of dental radiograph interpretation, and they now match or approach experienced clinicians on tasks that range from caries detection and restoration identification to implant localization and tooth segmentation.(1,2) The value of these systems lies in their consistency and speed, and the field has moved quickly from proof-of-concept demonstrations toward products that are embedded in imaging workstations and decision support. The same models that read a radiograph for caries are increasingly used to triage patients, plan implants and estimate age, and their outputs now influence decisions in public oral-health programs as well as in private practice. Their adoption, however, has outpaced attention to their failure modes under deliberate manipulation, and the security properties of a dental classifier are rarely described alongside its reported accuracy.

Adversarial examples are inputs modified by an imperceptible perturbation that changes a model's prediction with high confidence, and they were first shown to transfer across architectures and training sets in computer vision.(3,4) The perturbation is small enough to be invisible to a human reader yet large enough to reverse a decision, which makes it a distinct kind of vulnerability rather than a mere inaccuracy, and it has been formalized into a broad field of adversarial machine learning that now spans evasion, poisoning and model extraction.(5)The attacks matter because they do not require the attacker to understand the patient or the disease; they require only access to the model's gradients or its input-output behaviour, and the same perturbation that fails one image can be scaled or adapted to others.(6,7)

The threat is not theoretical in medicine. Medical imaging models, including those for retinal, chest and skin images, have been shown to misclassify images after perturbations that a radiologist would not notice, and the attacks have been demonstrated against real clinical classifiers rather than toy models.(8,9) The perturbations that break a diagnostic model can be as small as a few pixels in magnitude, and they have been shown to defeat models for diabetic retinopathy, pneumonia and dermatological disease, which are among the most widely studied applications of clinical deep learning.

Surveys of adversarial attack and defense in medical image analysis document a field in which nearly every imaging modality has been broken and in which many proposed defenses collapse under an adaptive attacker, and they catalogue both the attacks and the countermeasures in a way that makes clear no single defense is sufficient.(10,11) Radiology has begun to treat adversarial artificial intelligence as a security problem, with editorials and reviews calling for hardening before deployment, and dedicated benchmarks have measured the robustness of chest, breast, brain and other imaging models under both evasion and poisoning.(2,12,13)

Dental radiographs share the same technical pipeline as these modalities but have not been subjected to the same adversarial audit, which leaves open the question of how little a perturbation would need to change to alter a dental diagnosis and whether any simple safeguard restores robustness.

The objective of this study was to quantify the adversarial robustness of a transfer-learning dental radiograph classifier against gradient-based perturbations and to evaluate whether adversarial training and a median filter restore that robustness. The study measured the attack success of the fast gradient sign method and projected gradient descent across perturbation budgets from 1/255 to 16/255, and it compared a baseline model with an adversarially trained model and a median-filter defense. By using a public benchmark, a fixed seed and a reproducible script, the study was designed so that its security measurements could be independently repeated and extended to other dental imaging tasks as they are proposed.

METHODS

Study design and dataset

This was an experimental security audit on a public, openly licensed dataset, with no human participants and no ethics approval required beyond the terms of the licence. The DENTEX validation subset(14)provided 50 panoramic radiographs with 182 annotated teeth. Each annotation carried a bounding box and a diagnosis label drawn from four classes: impacted tooth, caries, periapical lesion and deep caries. The dataset also recorded quadrant and tooth enumeration, which were not used because the audit concerned the diagnosis task alone.

The diagnosis classes were deliberately unbalanced, with caries the most frequent and peria-

pical lesions the rarest, which reflects the distribution that a real dental classifier would encounter and which is known to influence both accuracy and adversarial fragility. The images were used exactly as released, without exclusion or relabelling, so that the audit would be reproducible from the public source alone. The validation subset was chosen because it is the smallest openly licensed partition with per-tooth diagnosis labels, and because it has served as a benchmark in prior dental imaging work, which makes the present robustness measurements directly comparable to the accuracy reported on the same images.

Model

Each annotated tooth was cropped to its bounding box, resized to 224 by 224 pixels and held in the $[0,1]$ pixel range, with ImageNet normalization applied inside the model rather than to the stored images. A ResNet-18 backbone pre-trained on ImageNet (15) was frozen with its final layer removed, and a single 512-to-4 linear head was trained on the resulting features with stochastic gradient descent for 200 epochs at a learning rate of 0.01. This frozen-backbone plus linear-head design is the standard transfer-learning pipeline in dental imaging, and it was used deliberately because it is the configuration that a practitioner would deploy without fine-tuning the entire network.

The linear head maps the 512-dimensional feature vector to four diagnosis logits, so its decision boundary is a set of hyperplanes in the fixed feature space, a property that is central to understanding why it is vulnerable to adversarial examples. The images were kept in pixel space, rather than discarding them after feature extraction, so that adversarial perturbations could be generated by backpropagating through the full frozen pipeline to the input.

The classifier was trained and attacked on a single stratified 70/30 train/test split with a fixed random seed, which preserved the proportion of each diagnosis class in both partitions and kept the adversarial training feasible on a central processing unit. A fixed seed was used so that the split, the initialization and the optimization trajectory were all reproducible, and the script was executed twice to confirm that the results did not depend on any source of run-to-run randomness.

Threat model

The audit assumed a white-box adversary who possesses the trained model and can compute its gradients with respect to the input. This is the strongest and most informative threat model for a security evaluation, because any protection that fails against a white-box attacker also fails against weaker adversaries, and it matches the assumptions under which adversarial examples were first characterized and under which modern robustness is measured.(16,17)

In practice the assumption is reasonable for dental imaging: models are routinely shared between research groups, embedded in commercial software that a determined party can inspect, or exposed through inference interfaces that permit query-based approximation of the gradients.(6,7) The evaluation therefore reports an upper bound on risk, which is the appropriate reference for a security audit of a system whose failure could carry clinical or financial consequences.

Attacks

Two gradient-based attacks were mounted against the trained model. The fast gradient sign method applied a single step in the direction of the sign of the gradient of the loss with respect to the input, adding to every pixel the perturbation budget multiplied by the sign of the corresponding gradient element.(4)

Projected gradient descent refined that perturbation over ten iterations, taking a smaller step at each iteration, projecting the result back onto the epsilon ball around the original image and clipping to the valid pixel range, so that the final perturbation is both bounded and better aligned with the direction of steepest loss increase.(17)The perturbation budget epsilon took the values $1/255$, $2/255$, $4/255$, $8/255$ and $16/255$, spanning the range from a single quantization step, which is imperceptible by construction, to a clearly visible but still small modification, so that the relationship between perturbation magnitude and attack success could be traced across the entire practical range.

Defenses

Two defenses were evaluated against the attacks. Adversarial training followed the standard procedure of augmenting each training epoch with fast gradient sign method examples generated at a budget of $8/255$ and then training the head on the features of those adversarial images for 15

epochs.(17) In each epoch the adversarial examples were recomputed against the current model, so that the training signal tracked the model as it changed rather than relying on a static set of perturbations. This defense was selected because it is the most widely recommended mitigation in the robustness literature and because it requires no additional components at inference time. A 3 by 3 median filter was applied at inference, after the perturbation and before classification, replacing each pixel with the median of its 3 by 3 neighbourhood in order to test whether a simple input-space smoothing removes the adversarial signal; median filtering was chosen as the representative of the family of input transformations that are cheap, model-agnostic and already available in standard image-processing libraries.(18) The median filter was applied only at evaluation and not during training, so that its effect on clean and adversarial accuracy could be isolated from any effect on the learned parameters.

Metrics

Attack success rate was defined as the fraction of test images that were classified correctly when clean but incorrectly after perturbation, a measure that isolates the additional harm caused by the attack over and above the model's baseline error and that is the standard summary used in the adversarial robustness literature.(16) Post-attack accuracy was the overall accuracy on the perturbed set, which captures the absolute performance that a clinical user would observe under attack, and clean accuracy was the accuracy on unperturbed images. The three measures were computed for the baseline model, the adversarially trained model and the median-filtered model at every budget, so that both the relative harm and the absolute collapse could be read from the same tables. Reporting the two attack scores together was deliberate, because a defense can appear effective against a single-step attack while remaining fully vulnerable to an iterative one, and the discrepancy between the fast gradient sign method and projected gradient descent in this study illustrates exactly that pitfall. The trade-off between clean accuracy and worst-case robustness was summarized in a single comparison. Analyses were implemented in Python 3.13 with PyTorch 2.14, and the complete script was executed twice to confirm that every reported figure was reproducible and that no value depended on run-to-run randomness.

RESULTS

The study included 182 annotated teeth from 50 panoramic radiographs (Table 1). Caries predominated with 101 teeth, followed by impacted teeth with 40, deep caries with 32 and periapical lesions with 9. The class distribution was thus markedly imbalanced, as it is indeed observed in routine dental clinical practice.

Table 1. Characteristics of the DENTEX validation subset.

Diagnosis class	Annotated teeth	Share (%)
Impacted	40	22.0
Caries	101	55.5
Periapical lesion	9	4.9
Deep caries	32	17.6
Total	182	100.0

(Table 2) The baseline model classified 63.6% of the held-out teeth correctly. Adversarial training lowered clean accuracy to 54.5%, whereas the median filter raised it slightly to 65.5%. The worst-case attack success rate reached 100% for every variant.

Table 2. Clean test accuracy and worst-case attack success rate by model variant.

Variant	Clean accuracy (%)	Worst-case attack success rate (%)
Baseline	63.6	100.0
Adversarial training	54.5	100.0
Median filter	65.5	100.0

Note: Worst-case attack success rate is the projected gradient descent (PGD-10) attack success rate at a budget of 16/255.

(Table 3) The baseline model was vulnerable to the smallest perturbation tested. At a budget of 1/255 the fast gradient sign method and projected gradient descent each flipped every correctly classified tooth, an attack success rate of 100%.

Table 3. Attack success rate of the baseline model by attack type and perturbation budget.

Budget	FGSM (%)	PGD-10 (%)
1/255	100.0	100.0
2/255	100.0	100.0
4/255	100.0	100.0
8/255	82.9	100.0
16/255	77.1	100.0

Note: FGSM = fast gradient sign method; PGD-10 = projected gradient descent with ten iterations. Budget is expressed in units of 1/255 of the pixel range.

The median filter was therefore the only intervention that improved clean accuracy, by 1.9 percentage points over the baseline, while adversarial training reduced it by 9.1 percentage points.

The fast gradient sign method became less effective at the two largest budgets, falling to 82.9% at 8/255 and 77.1% at 16/255. Projected gradient descent remained at 100% across every budget. Post-attack accuracy under projected gradient descent collapsed to 0% for the baseline model.

The two attacks therefore diverged only at the largest budgets, where the single-step method overshot the decision boundary and the iterative method did not. At the three smallest budgets the fast gradient sign method and projected gradient descent were indistinguishable, both achieving complete attack success. The saturation of the attack success rate at the smallest budget meant that the classifier offered no margin of safety between an imperceptible perturbation and a complete loss of correct predictions.

(Table 4) Neither defense restored robustness under projected gradient descent. Attack success rate remained at 100% for the baseline, the adversarially trained model and the median-filtered model at every budget, and the corresponding post-attack accuracy was 0% throughout.

Table 4. Attack success rate under projected gradient descent by model variant and budget.

Budget	Baseline (%)	Adversarial training (%)	Median filter (%)
1/255	100.0	100.0	100.0
2/255	100.0	100.0	100.0
4/255	100.0	100.0	100.0
8/255	100.0	100.0	100.0
16/255	100.0	100.0	100.0

Note: PGD-10 = projected gradient descent with ten iterations. Budget is expressed in units of 1/255 of the pixel range.

DISCUSSION

The objective of this study was to quantify how small a perturbation would need to be to change a dental radiograph classifier's prediction and whether simple defenses restore robustness, and the results answer both parts in terms that should concern anyone deploying dental imaging models.

The classifier trained with a frozen backbone and a linear head was broken by the smallest perturbation examined, a change of one unit in the 255-level pixel scale, and neither of the two defenses tested was able to keep any correctly classified tooth correct once an iterative attacker was applied. Taken together, these results indicate that the standard transfer-learning pipeline for dental radiographs is catastrophically non-robust to adversarial examples rather than merely imperfect, and that the convenience of a frozen backbone carries a security cost that has not been priced into the dental artificial intelligence literature. The collapse at a budget of 1/255 is consistent with the linearity explanation of adversarial vulnerability that accompanied the first description of adversarial examples.⁽⁴⁾

A linear decision head on a fixed feature map behaves like a high-dimensional linear classifier, and the sign of its gradient with respect to the input provides a direction that, added at small magnitude, accumulates across many input dimensions to move the logits decisively.(4,6)

The frozen backbone does not confer robustness; it merely transforms the input into a feature space in which the linear head remains vulnerable, so the transfer-learning shortcut that makes dental classifiers cheap to train is precisely what makes them easy to attack. The result also aligns with the broader finding that pre-trained models used for disease diagnosis inherit adversarial vulnerability from their backbone, and with the observation that the robustness of a medical classifier is determined by the entire training recipe rather than by the choice of backbone alone.(19,20)

The consequence is that the very design choice that makes dental artificial intelligence accessible, training only a small head on a frozen general-purpose backbone, is also the choice that maximizes susceptibility to the cheapest and most reliable class of attack, so the accessibility of the pipeline and its insecurity share the same cause.

The demonstration that medical imaging models are vulnerable to imperceptible perturbations has been repeated across organs and modalities, and the present result extends that literature to panoramic dental radiographs, a modality for which adversarial robustness had not previously been reported in a dedicated audit.(8,9)

Lung and chest radiograph classifiers have been shown to be similarly fragile, with attack success reaching high levels at small budgets, and poisoning studies on chest imaging confirm that the vulnerability is not limited to clean-trained models but can be introduced or aggravated through the training data itself.(21,22) Breast and brain imaging studies report the same qualitative pattern, in which a perturbation of a few pixel units is enough to reverse a clinically meaningful decision, and they show that the problem persists across convolutional and transformer-based architectures alike.(20,23)

The consistency across modalities suggests that the failure is a property of the model family rather than of any particular disease or imaging geometry, and it predicts that other dental imaging tasks built on the same pipeline, from caries staging to implant identification, will exhibit the same behaviour unless their training recipe is changed.

The non-monotonic behaviour of the fast gradient sign method is explained by overshoot. A single-step attack that adds the full budget in the gradient direction moves the image past the decision boundary at small budgets, but at the two largest budgets it can push a sample so far that the prediction lands in a different wrong class or, for a fraction of samples, returns to the correct side, which is why the measured attack success fell from 100% to 82.9% and then to 77.1%. Projected gradient descent avoids this by taking smaller steps and projecting onto the budget ball at each iteration, so it converges to a minimal effective perturbation and remains at 100% across every budget.(17)

The contrast between the two attacks is therefore not evidence of robustness at large budgets but an artifact of how a single-step attack overshoots, and it shows why iterative attacks are the appropriate benchmark for a security audit rather than the weaker single-step variants that appear in some earlier evaluations.(16)

Neither defense restored robustness, and the reasons differ for each. Adversarial training of a linear head on a frozen backbone failed because the adversarial examples generated against that head are confined to the fixed feature manifold, and training on them does not change the fundamental linearity that the attacker exploits; the outcome is consistent with the finding that adversarial training is effective when the full network is trainable but far less so for a frozen representation.(10,17)

The fact that adversarial training also reduced clean accuracy, from 63.6% to 54.5%, indicates that the augmented examples pulled the decision boundary away from the clean data without conferring any robustness, a double cost that has been observed in other small-sample medical settings. The median filter failed for the opposite reason: although it removed part of the high-frequency perturbation and even raised clean accuracy slightly, the iterative attack produced perturbations that survived the smoothing, and the filtered images still crossed the decision boundary. This result is in line with the literature on input transformations, which have repeatedly been shown to fail against adaptive attackers even when they reduce the success of weaker attacks, and with the broader pattern in which defenses built on distillation, feature squeezing or denoising provide only partial protection at best.(11,18,24)

The clean-accuracy gain of the median filter is worth noting, however, because it shows that input smoothing can be beneficial for reasons unrelated to security, even though it cannot be counted on to stop an iterative attack.

A range of stronger defenses has been proposed for medical imaging, and the present com-

parison of two inexpensive safeguards is deliberately not exhaustive. Multi-layered defenses that combine adversarial training with feature squeezing, three-layer defenses for radiology, and diffusion-based purification each report improved robustness on their target tasks, but their gains are typically measured against single-step or non-adaptive attacks and are narrower against the iterative white-box adversary used here.(25-27)

Ensemble and feature-fusion approaches and dual-stage inference-time defenses follow the same pattern of improving robustness at the cost of added complexity, and breast and brain imaging studies that benchmark adversarial training report that the utility-robustness trade-off remains unresolved.(23,28,29)

The implication for dental imaging is that a real deployment would need to move beyond the linear head and the simple defenses tested here, toward end-to-end adversarial training or certified defenses, and that any such choice should be validated against an iterative attacker rather than assumed from a weaker evaluation.(16,17) The recurring lesson of the defense literature is that a mitigation is only as strong as the adversary against which it is evaluated, and that a robustness claim based on a single-step attack or a black-box surrogate can overstate the protection by a wide margin. For dental imaging, where the stakes of a misclassification can be a missed lesion or an unnecessary procedure, that lesson argues for treating adversarial robustness as part of the clinical validation of a model rather than as a separate, optional hardening step.

The finding generalizes across health systems because it is a property of the training pipeline rather than of the patient population. A classifier trained in a high-income setting on a large labelled collection is no more or less vulnerable to adversarial examples than one trained elsewhere, all else being equal, because the frozen-backbone plus linear-head shortcut is independent of geography. What varies across jurisdictions is the size of the labelled dataset that a center can assemble and the regulatory weight attached to a misclassification. In low- and middle-income settings across Africa, Latin America and parts of Asia, where large annotated dental collections are scarce, the transfer-learning shortcut is even more attractive and therefore the vulnerability documented here is more likely to be present in practice, while the data protection and medical device rules that would hold a vendor accountable differ substantially.(30)

The same logic applies to the choice of defense: adversarial training has been evaluated on a range of clinical tasks, including burn image segmentation and breast and brain imaging, and its cost in accuracy is paid by the deploying institution regardless of where that institution sits, so the utility-robustness trade-off is a global constraint rather than a local one.(23,31)

The cybersecurity implication is direct and extends across the health systems in which dental artificial intelligence is being adopted, and it compounds the known risks of connected medical devices rather than replacing them. A dental radiograph classifier trained with a frozen backbone and a linear head can be forced to misclassify any correctly diagnosed tooth with a perturbation that no human observer would see, and the attack requires nothing more than access to the model's parameters or gradients, which is available to any institution that shares or distributes a model, and to any adversary who can mount a practical black-box attack against a deployed service.(6,7) In a clinical, insurance or public-health context, where a radiograph classification feeds a treatment or reimbursement decision, that vulnerability is a realistic route to deliberate misdiagnosis rather than a laboratory curiosity, and it compounds the broader security concerns that have been raised about connected health devices.(30)

The finding therefore adds a dental-specific data point to the literature that treats adversarial machine learning as a security problem for medical devices, and it reinforces the recommendation that imaging models be hardened and audited before they are connected to consequential decisions.(12,13) The dental field has been slower than radiology to adopt adversarial robustness as a reporting requirement, and the present results suggest that the omission is not benign: a classifier whose accuracy is reported without an accompanying attack success rate presents an incomplete picture of its safety, and reviewers, regulators and clinical users should begin to ask for that number just as they already ask for sensitivity, specificity and calibration.

Only when robustness is reported as a first-class property will the choice between a cheap frozen-backbone model and a more expensive robust one be made on the basis of the real trade-off rather than on the basis of the one metric that happens to be published.

For a dental institution deciding whether and how to deploy such a classifier, the practical consequence is that robustness must be treated as a design requirement rather than an after-thought. A model that is released or shared should be evaluated against an iterative white-box attack before any claim of security is made, and the evaluation should report attack success rate alongside accuracy rather than accuracy alone, because the two metrics can diverge sharply, as they did here. Where a model is connected to a treatment or reimbursement decision, the appropriate

mitigation is not a cheap input filter but a change to the training recipe itself, such as end-to-end adversarial training or a certified defense, both of which require retraining the full network and accepting some loss of clean accuracy.^(16,17)

The alternative, for a use case that cannot afford that cost, is to constrain the threat rather than the model: to keep the model's parameters private, to limit query access, and to monitor inputs for the statistical signatures of adversarial perturbations, none of which eliminates the risk but all of which raise the cost to the attacker.^(7,18)

The strengths of this study are its reproducibility and its use of a standard, openly licensed dataset, which allow the audit to be repeated by any dental group with a labelled radiograph set and a pre-trained backbone, and its use of an iterative white-box attack, which avoids the optimistic results that follow from evaluating only single-step attacks. Its principal limitation is the single train/test split, which produced a clean accuracy of 63.6% and could be improved by averaging over several splits; a better-tuned or fully fine-tuned model might show a graded rather than saturated vulnerability, and repeating the audit on a fine-tuned network would establish whether the catastrophic failure observed here is specific to the linear head. A second limitation is that the attacks were white-box, which represents the strongest adversary, so the results are an upper bound on risk for a model whose parameters are not exposed; black-box attacks would likely still succeed given the magnitude of the vulnerability, but the exact budgets would differ.⁽⁷⁾

A third limitation is that the defenses were restricted to two inexpensive options, and the comparison does not cover stronger mechanisms such as end-to-end adversarial training or certified robustness.^(16,17) A fourth limitation is the use of a single linear head, which is the component most vulnerable to adversarial examples and whose replacement with a deeper or fine-tuned head would likely change the measured robustness even without retraining the backbone. A fifth limitation is that the audit did not examine the transferability of perturbations between the two defense configurations, which would establish whether an attack computed against the baseline model remains effective against the defended models and would sharpen the practical interpretation of the defense results. These limitations do not change the central conclusion, because the failure occurred at the smallest budget for every defense tested, but they delineate the path for a more complete robustness evaluation of dental imaging models.

CONCLUSIONS

A transfer-learning dental radiograph classifier with a frozen backbone and a linear head was broken by a perturbation of 1/255, which flipped every correctly classified tooth for both the fast gradient sign method and projected gradient descent. Adversarial training and a median filter did not restore robustness, and adversarial training also reduced clean accuracy.

Dental imaging models built on this standard pipeline should be treated as non-robust by default, and deployments that carry clinical or financial consequences require stronger, model-level defenses rather than the simple safeguards tested here.

The audit also demonstrates that a robustness evaluation on a public benchmark is inexpensive to perform, and its routine inclusion would give dental artificial intelligence a security baseline that the field currently lacks, alongside the accuracy and calibration metrics that are already standard.

REFERENCES

1. Ayhan B, Ayan E, Atsü S. Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence algorithms based on deep learning methods. *BMC Oral Health*. 2025;25(1):216. doi: 10.1186/s12903-025-05577-3
2. R NM, S GR, P P. Adversarial Robustness of Deep Learning in Medical Imaging: A Comprehensive Survey and Benchmark of State-of-the-Art Architectures. *ijacsa*. 2025;16(12):831-841. doi: 10.14569/ijacsa.2025.0161278
3. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks [Internet]. 2014 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/1312.6199>
4. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [Internet]. 2015 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/1412.6572>
5. Pelekis S, Koutroubas T, Blika A, Berdelis A, Karakolis E, Ntanos C, et al. Adversarial machine learning: a review of methods, tools, and critical industry sectors. *Artif Intell Rev*. 2025;58(8):226. doi: 10.1007/s10462-025-11147-4
6. Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale [Internet]. 2017 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/1611.01236>
7. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. 2017:506-519. doi: 10.1145/3052973.3053009
8. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287-1289. doi: 10.1126/science.aaw4399
9. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*. 2021;110:107332. doi: 10.1016/j.patcog.2020.107332
10. Dong J, Chen J, Xie X, Lai J, Chen H. Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges. *ACM Comput Surv*. 2025;57(3):1-38. doi: 10.1145/3702638
11. Levy M, Amit G, Elovici Y, Mirsky Y. The Security of Deep Learning Defenses in Medical Imaging. *Proceedings of the 2024 Workshop on Cybersecurity in Healthcare*. 2023:37-44. doi: 10.1145/3689942.3694746
12. Lecler A, Soyer P. AI in radiology: Powerful, promising... but alarmingly hackable. *Diagnostic and Interventional Imaging*. 2025;106(11):369-370. doi: 10.1016/j.diii.2025.06.003
13. Dietrich N, Gong B, Patlas MN. Adversarial artificial intelligence in radiology: Attacks, defenses, and future considerations. *Diagnostic and Interventional Imaging*. 2025;106(11):375-384. doi: 10.1016/j.diii.2025.05.006
14. Ibrahim Hamamci et al. DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays [Internet]. 2023 [cited 2026 Oct 28]. Available from: <https://huggingface.co/datasets/ibrahimhamamci/DENTEX>
15. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:770-778. doi: 10.1109/cvpr.2016.90
16. Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)*. 2017:39-57. doi: 10.1109/sp.2017.49
17. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks [Internet]. 2018 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/1706.06083>
18. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks [Internet]. 2018 [cited 2026 Oct 28]. Available from: <https://arxiv.org/abs/1704.02654>
19. Akhtom DM, Singh MM, XinYing C. Adversarial attacks and adversarial robustness in pre-trained deep learning models for disease diagnosis. *Int J Mach Learn & Cyber*. 2025;16(11):8953-8976. doi: 10.1007/s13042-025-02807-6
20. Kanca Gulsoy E, Ayas S, Baykal Kablan E, Ekinici M. Enhancing the adversarial robustness in medical image classification: exploring adversarial machine learning with vision transformers-based models. *Neural Comput & Applic*. 2025;37(12):7971-7989. doi: 10.1007/s00521-024-10516-4
21. Li X, Pang Y, Li Y. The Robustness of Deep Learning Models to Adversarial Attacks in Lung X-ray Classification. *2025 2nd International Conference on Machine Learning, Pattern Recognition and Automation Engineering (MLPRAE)*. 2025:99-106. doi: 10.1109/mlprae67267.2025.11290762
22. Lee SB. Gradual poisoning of a chest x-ray convolutional neural network with an adversarial attack and AI explainability methods. *Sci Rep*. 2025;15(1):21779. doi: 10.1038/s41598-025-02294-3
23. Doss P LM, Gunasekaran M, Kim J, Kadry S. Adversarial robustness enhancement in deep learning-based breast cancer classification: A multi-faceted approach to poisoning and Evasion attack mitigation. *Alexandria Engineering Journal*. 2025;115:65-82. doi: 10.1016/j.aej.2024.11.089
24. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symposium on Security and Privacy (SP)*. 2016:582-597. doi: 10.1109/sp.2016.41
25. Yinusa A, Faezipour M. A multi-layered defense against adversarial attacks in brain tumor classification using ensemble adversarial training and feature squeezing. *Sci Rep*. 2025;15(1):16804. doi: 10.1038/s41598-025-00890-x
26. Ul Haque SB, Zafar A, ul haque SM, ul Haq SR, Ahmad M. Securing AI in Healthcare: A Three-Layer Defense to Mitigate Adversarial Noise Impact in Radiology Imaging. *Biomedical Signal Processing and Control*. 2025;109:107969. doi: 10.1016/j.bspc.2025.107969
27. Nguyen MC, Luong NH. Diffusion-Based Purification for Adversarial Defense in Medical Image Classification. In: *Communications in Computer and Information Science*. Springer Nature Singapore; 2025. p. 80-91. doi: 10.1007/978-981-96-4288-5_7

28. Alzubaidi L, AL-Dulaimi K, Obeed HAH, Saihood A, Fadhel MA, Jebur SA, et al. MEFF – A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging. *Intelligent Systems with Applications*. 2024;22:200355. doi: 10.1016/j.iswa.2024.200355
29. Sheikh BUH. Mitigating adversarial threats in deep CT image diagnosis models via a dual-stage inference-time defense. *Applied Soft Computing*. 2024;163:111909. doi: 10.1016/j.asoc.2024.111909
30. Ostermann M, Freyer O, Weinhold C, Martius K, Gilbert S. How secure are your health devices—stopping wearables becoming a personal and national security risk?. *npj Digit Med*. 2025;8(1):317. doi: 10.1038/s41746-025-01710-2
31. Chen L, Liang J, Wang C, Yue K, Li W, Fu Z. Adversarial attacks and adversarial training for burn image segmentation based on deep learning. *Med Biol Eng Comput*. 2024;62(9):2717-2735. doi: 10.1007/s11517-024-03098-9

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Wenya Long, Teng He, Yingjuan Xie.

Writing–review and editing: Wenya Long, Teng He, Yingjuan Xie.