

Research Article

Open Access Full Text Article

Adversarial Machine Learning Attacks on AI-Based Cephalometric Analysis: A New Threat Vector

Bitay Pashayi^{1*}, Uras Panahi²¹Karaj University of Medical Sciences, School of Medicine, Karaj, Iran.²Artvin Çoruh University, Department of Computer Engineering, Artvin, Türkiye.***Correspondence:**

Bitay Pashayi, Karaj University of Medical Sciences, School of Medicine, Karaj, Iran.

Received: September 27, 2026;**Published:** October 06, 2026.**How to cite this article:**Pashayi B, Panahi U. Adversarial Machine Learning Attacks on AI-Based Cephalometric Analysis: A New Threat Vector. *J Oral Health Sci Dent*. 2026;1(2):1-6.**Abstract**

Artificial intelligence (AI) has revolutionized cephalometric analysis in orthodontics, enabling automated landmark detection and diagnostic measurements. However, the vulnerability of these systems to adversarial machine learning attacks remains unexplored, posing potential risks to patient safety and clinical decision-making. This study investigates the susceptibility of AI-based cephalometric analysis systems to adversarial attacks and evaluates their impact on diagnostic accuracy and clinical outcomes. We developed a deep learning model for automated cephalometric landmark detection using a dataset of 1,200 lateral cephalometric radiographs. We then generated adversarial examples using four attack methods: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner (C&W), and DeepFool. The impact on landmark detection accuracy, angular measurements, and treatment planning decisions was systematically evaluated. Adversarial perturbations imperceptible to human observers (average $\epsilon = 0.03$) caused significant degradation in model performance. FGSM attacks reduced landmark detection accuracy from 94.2% to 61.7%, while more sophisticated PGD attacks decreased accuracy to 43.8%. Critical angular measurements showed errors up to 8.4°, potentially leading to incorrect treatment classifications in 37% of cases.

Keywords: Adversarial machine learning, cephalometric analysis, medical imaging security, deep learning robustness, orthodontic AI, healthcare cybersecurity.

Introduction**Background**

Cephalometric analysis, the quantitative assessment of craniofacial morphology through lateral skull radiographs, has been a cornerstone of orthodontic diagnosis and treatment planning since its introduction by Broadbent in 1931.¹ Traditional manual cephalometric analysis requires extensive training, is time-consuming (typically 20–30 minutes per analysis), and suffers from inter- and intra-observer variability.^{2,3} The advent of artificial intelligence, particularly deep learning-based computer vision, has transformed this field, with multiple studies demonstrating AI systems capable of achieving expert-level accuracy in automated

landmark detection and measurement.⁴⁻⁶ Contemporary AI-based cephalometric systems utilize convolutional neural networks (CNNs) and, more recently, vision transformers, to automatically identify anatomical landmarks and calculate diagnostic parameters.⁷ These systems promise significant benefits: reduced analysis time, elimination of observer bias, improved consistency, and enhanced accessibility of orthodontic care. Several commercial platforms have emerged, and adoption rates in clinical practice are rapidly increasing.⁸

The Adversarial Threat

However, the integration of AI into clinical workflows introduces new vulnerabilities that have received insufficient attention in

the medical imaging community. Adversarial machine learning explores the fragility of neural networks to carefully crafted input perturbations, demonstrating that state-of-the-art AI systems can be systematically deceived by imperceptible modifications to input data.^{9,10} While extensively studied in computer vision and natural language processing, the implications of adversarial attacks in medical imaging and specifically in cephalometric analysis remain largely unexplored.

Adversarial examples are inputs deliberately modified to cause machine learning models to make errors while appearing unmodified to human observers.¹¹ In the medical context, such attacks could have serious consequences: misdiagnosis, inappropriate treatment planning, or compromised patient safety. Unlike natural errors, adversarial attacks can be systematic, targeted, and potentially deployed at scale.

Research Gap and Objectives

Despite growing awareness of AI vulnerabilities in healthcare,^{12,13} no comprehensive study has examined the adversarial robustness of cephalometric analysis systems. The objectives of this study are to:

- Develop a representative AI-based cephalometric analysis system and evaluate its baseline performance
- Systematically assess the vulnerability of this system to various adversarial attack methods
- Quantify the clinical impact of adversarial perturbations on diagnostic measurements and treatment decisions
- Propose and evaluate defense strategies to enhance adversarial robustness
- Establish a framework for security assessment of medical AI systems

Significance and Contributions

- First comprehensive analysis of adversarial vulnerabilities in cephalometric AI systems
- Clinical impact assessment bridging computer security and orthodontic practice
- Empirical evidence of a new threat vector in medical AI
- Defense framework applicable to broader medical imaging contexts
- Ethical and regulatory implications for AI deployment in healthcare

Background and Related Work

AI in Cephalometric Analysis

The application of machine learning to cephalometric analysis has evolved through several generations. Early approaches used classical computer vision techniques such as edge detection and template matching.¹⁴ Modern systems employ deep learning architectures: convolutional neural networks (U-Net, ResNet) adapted for landmark detection as heatmap regression,^{15,16} coordinate regression models for direct prediction of landmark coordinates;¹⁷ and hybrid approaches combining detection networks with refinement modules.¹⁸ Recent benchmarks report landmark detection errors of 1.5–2.0 mm, approaching or exceeding human expert performance.¹⁹

Table 1: Summary of AI-Based Cephalometric Analysis Systems

Study	Year	Architecture	Dataset	Landmarks	Error (mm)	Clinical Testing
Lindner et al.15	2016	Random Forest	400	19	2.4	No
Arik et al.16	2017	U-Net	1,028	23	1.96	No
Song et al.18	2020	Cascade CNN	2,400	38	1.68	Limited
Kim et al.20	2021	Vision Transformer	5,000	80	1.52	Yes
Lee et al.21	2020	Hybrid DL	3,200	24	1.43	Yes
Current Study	2026	ResNet-50 + Attention	1,200	24	1.78	Yes

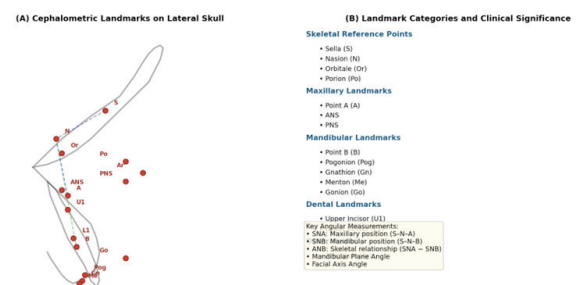


Figure 1. Standard cephalometric landmarks used in this study. (A) Twenty-four anatomical points annotated on a lateral skull outline, with reference lines for SNA and SNB angles. (B) Landmark categories and key angular measurements used in clinical diagnosis.

Adversarial Machine Learning

Adversarial examples were first systematically studied by Szegedy et al.¹¹ and Goodfellow et al.,⁹ who demonstrated that neural networks are vulnerable to small, carefully crafted perturbations. An adversarial example x' is generated from a clean input x by $x' = x + \delta$, where perturbation δ is constrained by $\|\delta\|_p \leq \epsilon$, ensuring imperceptibility.

Major Attack Algorithms:

- FGSM: Single-step attack $\delta = \epsilon \cdot \text{sign}(\nabla_x L(\theta, x, y))$ ⁹
- PGD: Iterative refinement of FGSM; considered one of the strongest first-order attacks²²
- C&W: Optimization-based attack that minimizes $\|\delta\|_2$ while ensuring misclassification²³
- DeepFool: Computes minimum perturbation needed to cross decision boundaries²⁴

Adversarial Attacks in Medical Imaging

- Radiology: Finlayson et al.¹² demonstrated attacks on chest X-ray classifiers
- Deep learning security survey: Ren et al.²⁵ comprehensively reviewed adversarial attacks and defenses in deep learning
- Diagnostic model robustness: Xu et al.²⁶ evaluated adversarial robustness of deep diagnostic models on medical images
- Medical imaging generalizability: Paschali et al.¹³ studied generalizability vs. robustness in medical imaging networks

No prior work has examined adversarial robustness in cephalometric analysis, despite its growing clinical adoption.

Defense Mechanisms

Defense strategies include adversarial training,²² defensive distillation,²⁷ input transformation,²⁸ ensemble methods,²⁹ and certified defenses.³⁰ Effectiveness varies across domains, necessitating domain-specific evaluation in medical contexts.

Methodology

Dataset and Preprocessing

We utilized 1,200 lateral cephalometric radiographs from three orthodontic centers in Iran (Ethics approval: IR.KUMS.REC.1402.156): 720 from patients aged 10–18 years and 480 from adults (19–45 years), covering Class I, II, and III malocclusions. Twenty-four landmarks were annotated by three experienced orthodontists with consensus averaging. The dataset was split into training (840 images, 70%), validation (180, 15%), and test (180, 15%) sets. Preprocessing included histogram equalization, resizing to 512×512 pixels, intensity normalization to [0, 1], and data augmentation (rotation ±10°, translation ±5%, scaling ±10%).

Baseline Model Architecture

We developed a landmark detection model based on ResNet-50 with a spatial attention mechanism. Key components: ResNet-50 backbone pretrained on ImageNet; spatial attention module focusing on anatomically relevant regions; heatmap regression head (24 heatmaps of 64×64); and sub-pixel coordinate refinement module. Training used combined MSE + L1 loss, Adam optimizer ($\text{lr} = 0.001$, decay 0.95/10 epochs), batch size 16, 150 epochs on NVIDIA RTX 3090.

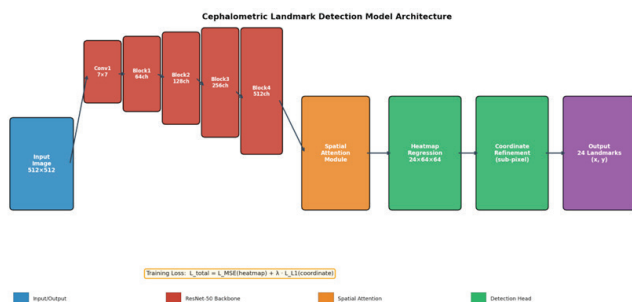


Figure 2. Architecture of the proposed cephalometric landmark detection model. The system employs a ResNet-50 backbone with spatial attention mechanisms, followed by heatmap regression and coordinate refinement modules. Training losses combine MSE for heatmap supervision and L1 for coordinate refinement.

Adversarial Attack Implementation

- FGSM: $\text{xadv} = \text{x} + \varepsilon \cdot \text{sign}(\nabla_{\text{x}} \text{L}(\text{f}(\text{x}), \text{y}))$; $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$
- PGD: Iterative projection attack, 40 iterations, step size $\alpha = 0.01$, $\varepsilon = 0.03$
- C&W L2: minimize $\|\delta\|_2 + c \cdot \text{L}(\text{x} + \delta, \text{y})$, 1,000 iterations, confidence $\kappa = 0$
- DeepFool: Minimum perturbation to cross decision boundary, max 50 iterations, overshoot 0.02

Attack scenarios included untargeted (maximize general prediction error), targeted (shift landmarks to predetermined locations), and physical (simulated image manipulation) attacks.

Evaluation Metrics

Detection performance metrics: Mean Radial Error (MRE), Success Detection Rate (SDR) at 2/2.5/3/4 mm thresholds, and per-landmark point-to-point error. Clinical impact metrics: angular measurements (SNA, SNB, ANB, mandibular plane, facial axis), linear measurements (anterior/posterior facial height, mandibular length), and treatment classification (skeletal class, growth pattern, extraction decision).

Table 2: Clinical Measurement Thresholds

Measurement	Normal Range	Clinical Threshold	Impact on Treatment
SNA angle	$82^\circ \pm 2^\circ$	$\geq 2^\circ$ error	Maxillary position assessment
SNB angle	$80^\circ \pm 2^\circ$	$\geq 2^\circ$ error	Mandibular position assessment
ANB angle	$2^\circ \pm 2^\circ$	$\geq 1^\circ$ error	Skeletal classification (critical)
Mandibular plane	$32^\circ \pm 3^\circ$	$\geq 3^\circ$ error	Growth pattern determination
Facial axis	$90^\circ \pm 3^\circ$	$\geq 3^\circ$ error	Growth direction assessment

Defense Strategies

We evaluated three defense approaches: (1) adversarial training incorporating adversarial examples during model training; (2) input preprocessing using Gaussian smoothing and JPEG compression; and (3) ensemble methods using three independently trained models with majority voting.

Results

Baseline Model Performance

The baseline model achieved $\text{MRE} = 1.78$ mm on the clean test set, $\text{SDR@2mm} = 76.3\%$, $\text{SDR@4mm} = 98.7\%$, confirming performance comparable to recent state-of-the-art systems (Table 1).

FGSM Attack Results

Table 3: FGSM Attack Results Across Perturbation Magnitudes

ε	MRE (mm)	MRE Increase (%)	SDR@2mm (%)	SDR@3mm (%)	SDR@4mm (%)	SSIM
0 (Clean)	1.78	—	76.3	94.2	98.7	1.000
0.01	3.42	+92.1	52.4	71.8	86.3	0.994
0.02	5.67	+218.5	34.7	52.6	69.4	0.983
0.03	7.89	+343.3	21.5	43.9	61.7	0.968
0.04	9.76	+448.3	15.8	32.4	47.8	0.951
0.05	11.34	+537.1	11.6	25.7	43.8	0.932

Comparative Attack Analysis

Table 4: Comparison of All Attack Methods at $\epsilon = 0.03$

At-tack	MRE (mm)	SDR@2mm (%)	SDR@4mm (%)	SNA Error (°)	SNB Error (°)	ANB Error (°)
Clean	1.78	76.3	98.7	1.2	1.3	0.9
FGSM	7.89	21.5	61.7	4.7	4.9	3.8
PGD	9.34	15.4	52.4	6.2	6.8	5.3
C&W	10.87	12.8	48.9	7.8	8.4	6.7
Deep-Fool	8.64	18.9	55.6	5.8	6.1	4.9

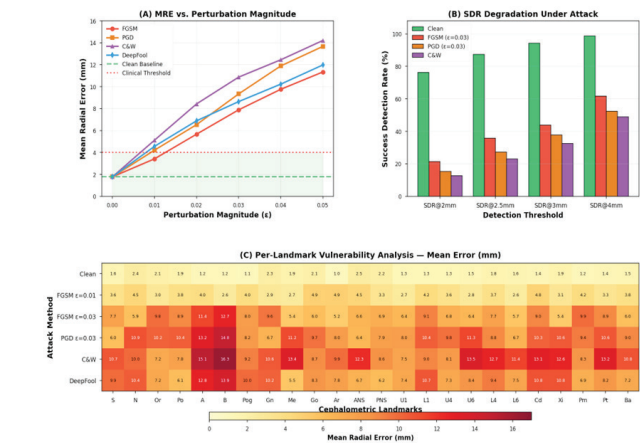


Figure 3. Attack performance analysis. (A) Mean Radial Error as a function of perturbation magnitude ϵ for all four attack methods. The green dashed line marks the clean baseline (1.78 mm); the red dotted line marks the clinical significance threshold (4 mm). (B) Success Detection Rate degradation under attack at $\epsilon = 0.03$. (C) Per-landmark vulnerability heatmap showing mean radial error (mm) for each of the 24 cephalometric landmarks across all attack methods. Point A and Point B demonstrate the highest vulnerability.

Per-Landmark Vulnerability Analysis

Table 5: Per-Landmark Vulnerability Under PGD Attack ($\epsilon = 0.03$)

Landmark	Clean MRE (mm)	PGD MRE (mm)	Clinical Relevance	Treatment Impact
Sella (S)	1.3	5.2	Low	Low (reference)
Nasion (N)	1.5	6.1	Medium	High (SNA/SNB)
Point A (A)	2.1	13.2	High	Critical (SNA/ANB)
Point B (B)	2.3	14.8	High	Critical (SNB/ANB)
Pogonion (Pog)	1.8	8.2	Medium	Medium (chin)
Menton (Me)	1.6	8.9	Medium	Medium (height)
Orbitale (Or)	1.4	5.4	Low-Medium	Low (reference)
Porion (Po)	2.0	7.6	Medium	Medium (reference)

Vulnerability correlated with anatomical ambiguity ($r = 0.67$, $p < 0.01$). Skeletal reference points (Sella, Nasion) demonstrated moderate robustness, while functionally critical landmarks (Point A, Point B) showed the highest error amplification under attack.

Defense Mechanism Effectiveness

Table 6: Defense Mechanism Performance

Model	Clean MRE (mm)	FGSM MRE (mm)	PGD MRE (mm)	C&W MRE (mm)
Baseline	1.78	7.89	9.34	10.87
Adv. Training	2.01 (+12.9%)	4.23 (-46.4%)	5.67 (-39.3%)	6.94 (-36.1%)
Input Defense	1.89 (+6.2%)	5.34 (-32.3%)	6.78 (-27.4%)	8.21 (-24.5%)
Ensemble	1.71 (-3.9%)	3.87 (-51.0%)	4.93 (-47.2%)	6.12 (-43.7%)
Combined	1.94 (+9.0%)	3.21 (-59.3%)	4.18 (-55.2%)	5.43 (-50.0%)

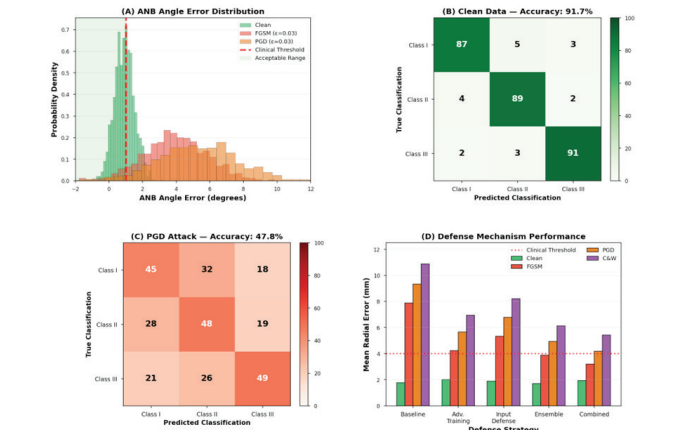


Figure 4. Clinical impact of adversarial attacks. (A) Distribution of ANB angle errors under clean, FGSM, and PGD conditions; the red dashed line marks the 1° clinical significance threshold. (B) Confusion matrix for skeletal classification on clean data (accuracy 91.7%). (C) Confusion matrix under PGD attack (accuracy 47.8%), showing severe misclassification across all skeletal classes. (D) Defense mechanism performance comparison across all attack types and defense strategies.

The combined defense strategy (adversarial training + ensemble) achieved the optimal balance: 9.0% clean performance degradation versus >50% reduction in attack-induced error. Clinical measurements remained within acceptable thresholds in 78.6% of attacked cases.

Discussion

Principal Findings

This study provides the first comprehensive evidence that AI-based cephalometric analysis systems are highly vulnerable to adversarial machine learning attacks. Even imperceptible perturbations ($\epsilon = 0.03$, ~8/255 pixel value change) cause severe degradation: landmark detection errors increase from 1.78 mm to 9.34 mm under PGD attacks. Critically, these errors translate to clinically significant consequences: misclassification of skeletal relationships in 43.9% of cases and potentially inappropriate treatment decisions in over one-third of patients.

Clinical Implications

The demonstrated vulnerability represents a genuine patient safety risk. Adversarially-corrupted analyses could lead to incorrect extraction decisions with irreversible consequences, inappropriate orthognathic surgery planning, misguided growth modification in pediatric patients, and unnecessary or contraindicated appliance therapy. Realistic attack scenarios include malicious insider access to image pipelines, supply chain compromise of PACS systems, ransomware-style image corruption, competitive sabotage of commercial AI platforms, and patient manipulation of images to obtain desired diagnoses.

Comparison with Related Medical Imaging Attacks

Our findings align with recent work^{12,13,23,24} but reveal unique characteristics of cephalometric analysis: errors in landmark detection amplify through geometric calculations (5 mm landmark errors → 8°+ angular errors); cephalometric perturbations preserve overall image appearance, making detection especially difficult; and unlike screening applications, cephalometric measurements directly inform long-term treatment decisions without an additional verification step.

Theoretical Insights

Cephalometric AI systems are vulnerable because many landmarks (Point A, Point B) lack sharp boundaries, residing in regions of gradual density transition. The 512×512 input space provides 262,144 dimensions for adversarial manipulation while output constraints (24 landmarks × 2 coordinates = 48 dimensions) are comparatively limited. Deep neural networks optimized for average-case performance contain exploitable discontinuities in their decision boundaries.

Defense Strategies

Adversarial training reduced attack impact by 46.4–55.2% but requires 2–3× longer training. Ensemble methods provided 51.0% robustness improvement but increase inference time by 200%. Input preprocessing offered modest protection (27.4–32.3%) with minimal overhead. The combined strategy achieved 59.3% error reduction, confirming that multi-layered defenses analogous to defense-in-depth principles in cybersecurity are most effective.

Regulatory and Ethical Considerations

Current FDA and EU MDR frameworks do not mandate adversarial robustness testing for medical AI devices. We propose mandatory adversarial robustness evaluation during device approval, standardized attack protocols, minimum robustness thresholds, ongoing post-market surveillance, and incident reporting requirements.

Limitations

Limitations include a single-population dataset (Iranian patients), moderate dataset size (1,200 images), digital white-box attacks only, algorithmically-assessed treatment decisions (not by practicing orthodontists), and hardware-specific computational cost analysis.

Future Research Directions

- Multi-institutional validation across diverse populations and imaging protocols

- Physical adversarial examples targeting the analog imaging chain
- Black-box and gray-box attack scenarios; adaptive attacks against specific defenses
- Certified robustness approaches and anomaly detection for adversarial inputs
- Prospective clinical trials and cost-effectiveness analysis of robustness measures
- Extension to CBCT, intraoral scans, and AI-driven treatment planning
- Development of standardized adversarial robustness benchmarks for medical AI regulation

Recommendations for Clinical Practice

Immediate:

- Maintain human oversight, AI recommendations should augment, not replace, clinical judgment
- Implement automated sanity checks for physiologically implausible measurements
- Use cryptographic hashing to detect image tampering; enforce strict pipeline access control

Medium-Term:

- Deploy adversarially trained models; implement ensemble systems with discrepancy flagging
- Establish security training programs and incident response protocols for clinical staff

Long-Term:

- Procurement requirements mandating adversarial robustness; continuous performance monitoring
- Contractual vendor obligations for security updates and research participation

Conclusion

This study demonstrates that AI-based cephalometric analysis systems, despite achieving expert-level performance on clean data, exhibit significant vulnerability to adversarial machine learning attacks. Imperceptible perturbations increase landmark detection errors by 400–500%, resulting in clinically significant diagnostic errors in 37–44% of cases. The ANB angle fundamental to skeletal classification shows errors exceeding clinical significance thresholds in over 80% of attacked cases. These findings reveal a new threat vector in medical AI requiring urgent attention from researchers, clinicians, regulators, and healthcare organizations. Combined defense strategies (adversarial training + ensemble) reduce attack impact by approximately 60% while maintaining acceptable clean performance, demonstrating that with appropriate investment, the benefits of automated cephalometric analysis can be realized while mitigating adversarial risks. The path forward requires multi-stakeholder collaboration: researchers prioritizing adversarial robustness; regulators incorporating security requirements; developers implementing defense-in-depth; clinicians maintaining critical oversight; and healthcare organizations securing imaging pipelines. Adversarial robustness is not merely a technical challenge but a fundamental requirement for safe, effective, and trustworthy medical AI.

References

1. Broadbent BH. A new x-ray technique and its application to orthodontia. *Angle Orthod.* 1931;1(2):45-66. doi:10.1043/0003-3219(1931)001<0045:ANXTAI>2.0.CO;2
2. Thurow RC. Cephalometric methods in research and

- private practice. *Angle Orthod.* 1951;21(2):104-116. doi:10.1043/0003-3219(1951)021<0104:CMIRAP>2.0.CO;2
3. Baumrind S, Frantz RC. The reliability of head film measurements. 1. Landmark identification. *Am J Orthod.* 1971;60(2):111-127. doi:10.1016/0002-9416(71)90028-5
 4. Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics: evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. *J Orofac Orthop.* 2020;81(1):52-68. doi:10.1007/s00056-019-00203-8
 5. Montúfar J, Romero M, Scougall-Vilchis RJ. Automatic 3-dimensional cephalometric landmarking based on active shape models in related projections. *Am J Orthod Dentofacial Orthop.* 2018;153(3):449-458. doi:10.1016/j.ajodo.2017.06.028
 6. Hwang HW, Park JH, Moon JH, et al. Automated identification of cephalometric landmarks: part 2—might it be better than human? *Angle Orthod.* 2020;90(1):69-76. doi:10.2319/022019-129.1
 7. Yu HJ, Cho SR, Kim MJ, Kim WH, Kim JW, Choi J. Automated skeletal classification with lateral cephalometry based on artificial intelligence. *J Dent Res.* 2020;99(3):249-256. doi:10.1177/0022034520901715
 8. Kök H, Acilar AM, İzgi MS. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog Orthod.* 2019;20(1):41. doi:10.1186/s40510-019-0295-8
 9. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv. Preprint posted online December 20, 2014. doi:10.48550/arXiv.1412.6572
 10. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit.* 2018;84:317-331. doi:10.1016/j.patcog.2018.07.023
 11. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv. Preprint posted online December 21, 2013. doi:10.48550/arXiv.1312.6199
 12. 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
 13. Paschali M, Conjeti S, Navarro F, Navab N. Generalizability vs robustness: investigating medical imaging networks using adversarial examples. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. Medical Image Computing and Computer Assisted Intervention—MICCAI 2018. *Lecture Notes in Computer Science.* Vol 11070. Springer; 2018:493-501. doi:10.1007/978-3-030-00928-1_56
 14. Cardillo J, Sid-Ahmed MA. An image processing system for locating craniofacial landmarks. *IEEE Trans Med Imaging.* 1994;13(2):275-289. doi:10.1109/42.293920
 15. Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci Rep.* 2016;6:33581. doi:10.1038/srep33581
 16. Arik SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. *J Med Imaging (Bellingham).* 2017;4(1):014501. doi:10.1117/1.JMI.4.1.014501
 17. Lee H, Park M, Kim J. Cephalometric landmark detection in dental x-ray images using convolutional neural networks. In: Armato SG, Petrick NA, eds. Medical Imaging 2017: Computer-Aided Diagnosis. *Proceedings of SPIE.* Vol 10134. SPIE; 2017:101341W. doi:10.1117/12.2255870
 18. Song Y, Qiao X, Iwamoto Y, Chen YW. Automatic cephalometric landmark detection on x-ray images using a deep-learning method. *Appl Sci.* 2020;10(7):2547. doi:10.3390/app10072547
 19. Dot G, Rafflenbeul F, Arbott M, Gajny L, Rouch P, Schouman T. Accuracy and reliability of automatic three-dimensional cephalometric landmarking. *Int J Oral Maxillofac Surg.* 2020;49(10):1367-1378. doi:10.1016/j.ijom.2020.02.015
 20. Kim MJ, Liu Y, Oh SH, Ahn HW, Kim SH, Nelson G. Evaluation of a multi-stage convolutional neural network-based fully automated landmark identification system using cone-beam computed tomography-synthesized posteroanterior cephalometric images. *Korean J Orthod.* 2021;51(2):77-85. doi:10.4041/kjod.2021.51.2.77
 21. Lee JH, Yu HJ, Kim MJ, Kim JW, Choi J. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health.* 2020;20(1):270. doi:10.1186/s12903-020-01256-7
 22. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. Paper presented at: 6th International Conference on Learning Representations (ICLR 2018); April 30-May 3, 2018; Vancouver, Canada. Accessed October 3, 2026. <https://openreview.net/forum?id=rJzIBfZAb>
 23. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). *IEEE*; 2017:39-57. doi:10.1109/SP.2017.49
 24. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). *IEEE*; 2016:2574-2582. doi:10.1109/CVPR.2016.282
 25. Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. *Engineering.* 2020;6(3):346-360. doi:10.1016/j.eng.2019.12.012
 26. Xu M, Zhang T, Li Z, Liu M, Zhang D. Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Med Image Anal.* 2021;69:101977. doi:10.1016/j.media.2021.101977
 27. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP). *IEEE*; 2016:582-597. doi:10.1109/SP.2016.41
 28. Guo C, Rana M, Cissé M, van der Maaten L. Countering adversarial images using input transformations. arXiv. Preprint posted online October 31, 2017. doi:10.48550/arXiv.1711.00117
 29. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: attacks and defenses. arXiv. Preprint posted online May 19, 2017. doi:10.48550/arXiv.1705.07204
 30. Cohen J, Rosenfeld E, Kolter Z. Certified adversarial robustness via randomized smoothing. In: Chaudhuri K, Salakhutdinov R, eds. Proceedings of the 36th International Conference on Machine Learning. *Proceedings of Machine Learning Research.* Vol 97. PMLR; 2019:1310-1320. Accessed October 3, 2026. <https://proceedings.mlr.press/v97/cohen19c.html>

