South American Publishing

SAP Artificial Intelligence in Dentistry. 2025; 1:18
doi: 10.62486/aid202518

BRIEF COMMUNICATION

Artificial intelligence for dental diagnosis and decision support in Latin America

Inteligencia artificial para el diagnóstico dental y el apoyo a la toma de decisiones en América Latina

Jadier Wong-Silva[1], Arian González-Cruz[2], Jhossmar Cristians Auza-Santivañez[3]

[1]Universidad de Ciencias Médicas de Pinar del Río, Hospital Pediátrico Provincial Docente "Pepe Portilla". Pinar del Río, Cuba.
[2]Hospital General Docente "Abel Santamaría Cuadrado". Pinar del Río, Cuba.
[3]Hospital de Tercer Nivel Fray Quebracho, Instituto Académico Científico Quispe-Cornejo. La Paz, Bolivia.

Cite as
Wong-Silva J, González-Cruz A, Auza-Santivañez JC. Artificial intelligence for dental diagnosis and decision support in Latin America. SAP Artificial Intelligence in Dentistry. 2025; 1:18. https://doi.org/10.62486/aid202518

Publication History
Submitted: 20-03-2025
Revised: 17-05-2025
Accepted: 28-06-2025
Published: 29-06-2025

Corresponding author
Jhossmar Cristians Auza-Santivañez

ABSTRACT

Artificial intelligence has been applied in dentistry in Latin America, primarily in imaging-based diagnosis, risk prediction, and clinical decision support. Available studies, largely concentrated in Brazil, employ architectures such as convolutional neural networks, YOLO-based detectors, and machine learning models including XGBoost. Reported performance varies across tasks, with high sensitivity in specific applications but frequent false positives and limited specificity in some approaches, particularly in large language model-based systems. Relevant methodological limitations are also observed, including retrospective designs, small sample sizes, lack of external validation, and heterogeneity in evaluation metrics. The evidence suggests that, despite significant technical advances, widespread clinical implementation in the region remains constrained by validation, infrastructure, and regulatory challenges.

Keywords: Artificial Intelligence; Dentistry; Dental Diagnosis; Machine Learning; Latin America.

RESUMEN

La inteligencia artificial ha sido aplicada en odontología en América Latina, principalmente en diagnóstico por imágenes, predicción de riesgo y apoyo a la toma de decisiones clínicas. Los estudios disponibles, concentrados en su mayoría en Brasil, emplean arquitecturas como redes neuronales convolucionales, modelos de detección tipo YOLO y algoritmos de aprendizaje automático como XGBoost. Los resultados reportan desempeños variables, con sensibilidades elevadas en tareas específicas pero con frecuencias altas de falsos positivos y limitada especificidad en algunos enfoques, especialmente en modelos basados en grandes modelos de lenguaje. Asimismo, se observan limitaciones metodológicas relevantes, incluyendo diseños retrospectivos, tamaños muestrales reducidos, falta de validación externa y heterogeneidad en métricas de evaluación. La evidencia sugiere que, si bien existen avances técnicos significativos, la implementación clínica generalizada de estos sistemas en la región se encuentra restringida por desafíos de validación, infraestructura y regulación.

Palabras clave: Inteligencia Artificial; Odontología; Diagnóstico Dental; Aprendizaje Automático; América Latina

© 2025; The Authors. This Open access article is distributed under the Creative Commons Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Artificial intelligence (AI) has increasingly been incorporated into dental research and clinical practice, particularly in diagnostic imaging, risk prediction, and decision support systems. In Latin America, the adoption of these technologies has accelerated in recent years, driven by the availability of digital radiographic datasets and advances in machine learning architectures. However, the extent to which these models achieve clinically meaningful performance and generalizability within regional healthcare contexts remains uncertain. This communication synthesizes recent evidence from Latin American studies, focusing on applications, model architectures, diagnostic performance, and methodological limitations.

DEVELOPMENT

Empirical Latin American studies report diagnostic performance across oral photography, dental radiography, and cytopathology, using architectures spanning EfficientNet-based multi-task CNNs, YOLO detectors, MobileNet classifiers, and multimodal large language models (LLMs).[1,2,3] In Brazilian panoramic radiographs, an EfficientNet-B3 multi-task model with attention (CBAM) was trained retrospectively on 2,200 images split into train/validation/test (1 320/440/440) and reported best-in-set performance for age prediction (lowest MAE, highest) and sex classification (highest accuracy and AUC), with Grad-CAM supporting anatomically plausible attention; ablations removing CBAM or altering task weights reduced performance.[1] For interproximal caries detection, a Brazilian systematic review/meta-analysis of six studies (from nine eligible publications; 5 930 radiographs) estimated pooled sensitivity 0,90 and specificity 0,25, yielding LR− 0,21 and LR+ 1,38, a profile consistent with high false-positive burden at the tested thresholds.[4]

Oral lesion detection and classification

A Brazilian two-step pipeline for oral potentially malignant disorders and oral squamous cell carcinoma combined YOLOv11 lesion localization with MobileNetV2 classification of cropped regions.[2] The best YOLOv11 detector reached mAP50 0,820 with precision 0,897, recall 0,744, and F1-score 0,813, indicating non-trivial miss rates at the detection stage even when precision is high.[2] The best MobileNetV2 classifier achieved accuracy 0,846 with AUC-ROC 0,852 on internal testing and accuracy 0,850 with AUC-ROC 0,935 on external validation, suggesting that classification performance can improve when evaluated on a distinct held-out set with compatible labeling and image characteristics.[2] When the detector and classifier were chained, external validation reached accuracy 0,863 and AUC-ROC 0,934, but 23/132 images (17,4 %) produced no detection box and were excluded from classification, introducing pipeline attrition that is clinically material because non-detections are implicitly treated as non-actionable rather than as uncertain outputs needing adjudication.[2]

Multimodal LLMs evaluated as image readers produced a markedly different operating point on periapical radiographs of anterior teeth: ChatGPT 5.0 achieved sensitivity 0,97 with specificity 0,11 (balanced accuracy 0,54), while Gemini Flash 2.0 achieved sensitivity 0,84 with specificity 0,11 (balanced accuracy 0,48).[3] Both systems were described as having high false-positive rates and frequent tooth localization errors, consistent with a positive classification bias that limits stand-alone diagnostic value despite high sensitivity in this constrained setting.[3]

Radiograph-based models for population and disease inference

For sex estimation from panoramic radiographs, a large Brazilian retrospective dataset from 15 clinical centers (207 946 images) yielded 95,02 % accuracy for a CNN approach, with image resolution identified as a significant factor influencing performance.[5] A ResNet variant exceeded 86 % accuracy in subjects older than 6 years and exceeded 96 % in those older than 16 years, while AUC exceeded 96 % for most age groups but not for the youngest group; performance was also reported to be higher on female images, implying sensitivity to demographic composition and acquisition characteristics even at very large sample size.[5]

For periapical bone rarefaction assessment in panoramic radiographs, Xception and DenseNet121 (with or without attention) achieved accuracy above 68 % and sensitivity up to 76 % for binary classification of healthy versus diffuse cases, indicating moderate discrimination that may be insufficient for high-stakes triage without additional calibration and validation across clinical settings.[6] For caries risk prediction using non-image epidemiological features, XGBoost achieved AUC 0,84 versus 0,81 for a decision tree in 615 Brazilian adolescents, showing incremental gains from gradient boosting in a tabular prediction task that is closer to risk stratification than to image diagnosis.[7]

Evidence syntheses on caries imaging emphasize a recurrent sensitivity–specificity asymmetry. In a meta-analysis of deep learning CNNs on interproximal radiographs, pooled sensitivity was 0,90 while pooled specificity was 0,25, with very high heterogeneity (reported values 99,3 % for sensitivity and 99,9 % for specificity), which constrains transferability of a single operating threshold across studies and suggests that apparent "accuracy" claims are likely to be threshold- and prevalence-dependent.[4] Complementary Brazilian literature reviews identify CNN-based approaches for interproximal caries (including U-Net, YOLO, and DenseNet121 in the reviewed literature) and explicitly flag false positives and false negatives as reliability challenges alongside ethical and regulatory issues.[8]

Cytopathology and histopathology applications

In oral exfoliative cytology for cancer screening, a Brazilian AI system for AgNOR quantification used a cut-off of 3,69 AgNORs/nucleus and achieved sensitivity 86 %, specificity 93 %, and accuracy 90 % in phase 1, indicating clinically plausible separation at the smear level for the studied cohorts.[9] In phase 2, agreement between the CNN system and human counting was high (ICC 0,896, 95 % CI 0,875–0,915), and analysis time differed markedly (20 minutes with the CNN vs 67 hours for human analysis as reported),

https://doi.org/10.62486/aid202518

framing efficiency gains as an operational advantage contingent on stable staining quality and consistent smear preparation across settings.[9]

For histopathology of oral epithelial dysplasia, a Brazilian multi-center dataset study reported that VGG16 performed best during development but exhibited "massive overfitting," and that EfficientNetB0 had comparable metrics with the lowest loss, while the broader conclusion emphasized high learning rates with very low generalization potential across most trained models.[10] Interobserver agreement for region overlap was summarized as mean IoU 0,675 (±0,030), underscoring that model training targets inherit annotation variability that can cap achievable performance even before external validation is attempted.[10]

Methodological limitations

Across diagnostic domains, limitations arise from study design, reporting, and the structure of multi-stage pipelines. Retrospective assembly and random splits can overstate generalization when acquisition conditions are homogeneous, as in the panoramic radiograph age/sex estimation study that explicitly describes retrospective collection and a random train/validation/test partitioning scheme.[1] In oral lesion pipelines, the dependence of downstream classification on detector outputs produces measurable attrition: 17,4 % of external-validation images had no detection box and were excluded, so reported classifier metrics apply to a selected subset rather than to the full intended-use population of images encountered in practice.[2] In multimodal LLM image reading, poor specificity (0,11 for both tested models) coupled with reported frequent localization errors indicates that high sensitivity can coexist with systematic error modes that are not well captured by a single aggregate metric such as balanced accuracy.[3]

Reporting and governance limitations also constrain comparability. A systematic review of deep learning for teeth, caries, and restoration detection noted that 41,66 % of analyzed papers lacked clear information on ethics committee approval for data use, and that limited public datasets and insufficient methodological detail hinder fair cross-study comparison, limiting reproducibility and the ability to attribute performance differences to architecture rather than to dataset curation or labeling practices.[11] A Brazilian methodological review similarly argues that study design and reporting guidance should be rigorously followed, including improvement of datasets, and it identifies limited interaction between dentistry and technical disciplines as a barrier to applicable dental AI, which can manifest as opaque preprocessing choices and incomplete documentation of training regimes.[12] Evidence syntheses also identify broader bias risks: a scoping review on ML for oral cavity disease diagnostic prediction reports that sensitivity, accuracy, and specificity are the most common metrics and that SVM, Random Forest, and logistic regression are frequently used, pointing to heterogeneity in both algorithms and evaluation conventions that complicates synthesis without standardized endpoints and datasets.[13] In a related systematic review embedded within the oral dysplasia paper, high risk of bias was identified in 23 studies, aligning with concerns that model evaluation quality is a limiting factor even when architectures are conventional.[10]

Ethical and regulatory constraints are repeatedly raised in Latin American commentaries, focusing on privacy, informed consent, and accountability for automated decisions, framing clinical deployment as a socio-technical problem rather than a pure performance optimization exercise.[14] Mexican editorials and overviews also emphasize dependence on large datasets, the need for transparency and explainability to gain trust, and the risk posed by unclear data regulations to patient privacy, themes that intersect with the empirical studies' reliance on proprietary or non-public clinical image collections.[15]

Analytical perspective

The retrieved Latin American evidence base is uneven across tasks: Brazil contributes multiple primary diagnostic evaluations spanning oral photography (YOLOv11/MobileNetV2), radiographic inference (panoramic sex estimation; periapical lesion screening by LLMs), and cytology quantification, while Mexico and Peru are represented mainly by ethical or narrative syntheses rather than by model validation studies with reportable sensitivity/specificity/AUC estimates.[2,3,9,15,16] This distribution matters because the strongest reported metrics tend to occur in bounded problems with constrained input variation (AgNOR counting with ICC 0,896; panoramic sex classification with accuracy 95,02 %), while tasks closer to routine diagnosis and triage exhibit either moderate discrimination (periapical bone rarefaction sensitivity up to 76 % with accuracy above 68 %) or sensitivity-dominant trade-offs that shift burden to downstream clinical confirmation (caries meta-analysis specificity 0,25; LLM specificity 0,11).[3,4,5,6,9]

A second pattern is that methodological maturity correlates more with evaluation design than with model novelty. The oral lesion pipeline reports external validation metrics and reveals pipeline attrition explicitly (17,4 % non-detections), which supplies clinically relevant failure-mode information that single-stage classifiers and many reviews do not report.[2] By contrast, the oral dysplasia histopathology work documents "massive overfitting" and low generalization potential, illustrating that high-capacity CNNs can fail under class overlap and annotation variability even when trained on large patch counts, and that without external testing the direction and magnitude of such failures can remain unknown.[10] In caries detection, pooled sensitivity and specificity coexist with extreme heterogeneity (I^2 near 100 %), which implies that performance is context-sensitive and that cross-site transfer in Latin America will likely require recalibration, harmonized labeling, and explicit prevalence-aware threshold selection rather than model replacement alone.[4]

A third pattern is that governance and infrastructural constraints appear directly in the literature's quality signals. The report that 41,66 % of papers omit clear ethics approval information indicates that data governance practices are not consistently documented, which can slow multi-institutional pooling and reduce confidence in downstream deployment in public systems that require auditable data provenance.[11] Ethical and regulatory concerns around privacy, consent, and accountability recur in Brazilian and Mexican sources, suggesting that implementation barriers in Latin America are shaped by legal fragmentation and limited institutional capacity for continuous model monitoring, not only by model accuracy metrics.[14,15] Taken together, the empirical studies support a view of Latin American dental AI as technically diverse but evaluation-limited: strong point estimates exist for specific pipelines

and measurement problems, while generalizable diagnostic deployment remains constrained by heterogeneity, attrition in multi-step systems, low specificity in some paradigms, and incomplete reporting of data governance and methods across the wider literature. (2,3,4,11)

FUNDINGS

None.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Jadier Wong-Silva, Arian González-Cruz, Jhossmar Cristians Auza-Santivañez.
Research: Jadier Wong-Silva, Arian González-Cruz, Jhossmar Cristians Auza-Santivañez.
Writing – original draft: Jadier Wong-Silva, Arian González-Cruz, Jhossmar Cristians Auza-Santivañez.
Writing – proofreading and editing: Jadier Wong-Silva, Arian González-Cruz, Jhossmar Cristians Auza-Santivañez.

REFERENCES

1. Oliveira ML, Yang S, Sampaio-Oliveira M, Freitas DQ, Haiter-Neto F, Lim SH, et al. Multi-Task Deep Learning for Sex and Age Estimation from Panoramic Radiographs in a Brazilian Young Population. International Dental Journal 2026; 76:109381. https://doi.org/10.1016/j.identj.2025.109381.

2. Araújo ALD, Silva AVB da, Gonçalves ARM, Saldivia-Siracusa C, Ferraz DLF, Calderipe CB, et al. Two-step pipeline for oral diseases detection and classification: a deep learning approach. Front Oral Health 2025;6. https://doi.org/10.3389/froh.2025.1659323.

3. Bubna DP, Mattos NHR, Luiz LB das P, Baratto-Filho F, Mattos-Calil MT de, Silva-Sousa YTC, et al. Can Large Language Models Detect Periapical Lesions in Anterior Teeth? A Comparative Study. Braz Dent J 2025;36:e25. https://doi.org/10.1590/0103-644020256861.

4. Alves VP, Souza MRF, Gomes IP, Falci SGM, Souza GM de. Efetividade da inteligência artificial em detectar cárie dentária em exames interproximais: uma revisão sistemática com metanálise. REVISTA DO CROMG 2023;22. https://doi.org/10.61217/rcromg.v22.490.

5. Ciconelle ACM, da Silva RLB, Kim JH, Rocha BA, Dos Santos DG, Vianna LGR, et al. Deep learning for sex determination: Analyzing over 200,000 panoramic radiographs. J Forensic Sci 2023;68:2057–64. https://doi.org/10.1111/1556-4029.15376.

6. Filho JE da S, Aguiar AWO de, Silva CM, Albuquerque DF de, Filho EDG. Automated detection and classification of periapical bone rare factions in panoramic radiographs: a deep learning-based approach. Brazilian Journal of Dentistry Oral Radiology 2025;4:xxviiijao17–xxviiijao17. https://doi.org/10.52600/2965-8837.bjdor.2025.4.xxviiijao17.

7. Bomfim RA. Machine learning to predict untreated dental caries in adolescents. BMC Oral Health 2024;24:316. https://doi.org/10.1186/s12903-024-04073-4.

8. Martins C de A, Rodrigues JHM, Soares LFF, Pires LGS, Nogueira BR, Lopes MABS, et al. Análise da inteligência artificial no diagnóstico de lesões cariosas em radiografia interproximal: Uma Revisão de literatura. Brazilian Journal of Implantology and Health Sciences 2025;7:751–69. https://doi.org/10.36557/2674-8169.2025v7n10p751-769.

9. Lepper TW, Amaral LN do, Espinosa ALF, Guedes IC, Rönnau MM, Daroit NB, et al. Cytopathological quantification of NORs using artificial intelligence to oral cancer screening. Braz Oral Res 2025;39:e056. https://doi.org/10.1590/1807-3107bor-2025.vol39.056.

10. Araujo ALD. Machine Learning aplicada às investigações nos campos da Estomatologia e da Patologia Oral e Maxilofacial= Machine Learning applied to research in Stomatology and Oral and Maxillofacial Pathology fields. PhD Thesis. 2022.

11. Carneiro JA, Zancan BAG, Tirapelli C, Macedo AA. Deep learning to support the detection and classification of teeth, dental caries and restorations: a review. Brazilian Journal of Health Review 2024;7:e68218–e68218. https://doi.org/10.34119/bjhrv7n2-156.

12. Rodrigues JA, Krois J, Schwendicke F. Demystifying artificial intelligence and deep learning in dentistry. Braz Oral Res 2021;35:e094. https://doi.org/10.1590/1807-3107bor-2021.vol35.0094.

13. Lopes IM. Aprendizado de máquina para prever o diagnóstico de doenças da cavidade oral : uma revisão de escopo. Machine learning in predicting diagnosis of oral cavity diseases: a scoping review 2025.

14. Neto RN da SN, Camardela TWM, Carvalho HD de. Inteligência Artificial na Odontologia: avanços tecnológicos, desafios éticos e diretrizes para uma aplicação responsável no Brasil. Brazilian Journal of Health Review 2025;8:e79411–e79411. https://doi.org/10.34119/bjhrv8n2-406.

15. Chanes-Cuevas OA. Impacto de la Inteligencia Artificial (IA) en la odontología. Revista odontológica mexicana 2023;27:1–2. https://doi.org/10.22201/fo.1870199xp.2023.27.3.89461.

16. Rojas Ortega RA, Barzola Loayza MG, Gómez Carrión CE, Saravia Alviar RA, Rojas Ortega RA, Barzola Loayza MG, et al. El rol de la inteligencia artificial en técnicas no invasivas para el diagnóstico del cáncer oral. Revista Minerva 2025;6:19–29. https://doi.org/10.47460/minerva.v6i18.220.