ISSN :2584-2706

Algorithmic Bias and "Black Box" Accountability in Dental Artificial Intelligence: Beyond Diagnostic Accuracy Metrics toward Clinical Equity and Medico-Legal Governance

Dr. Althaaf Khan, BDS (SMU), Assoc Fcgdent (UK)
Private Practice, South Africa

Corresponding Author

Abstract

Objective: Artificial intelligence (AI), particularly deep learning applied to dental imaging, routinely reports high sensitivity and specificity in controlled evaluations. However, rapid commercial adoption has outpaced critical inquiry into population equity, therapeutic over-intervention, and legal accountability. This review synthesizes evidence on dataset bias, radiographic overdiagnosis of incipient caries, and medico-legal liability in AI-assisted dental practice.

Methodology: A structured narrative literature synthesis was conducted across peer-reviewed dental, biomedical informatics, bioethics, and health-law publications. Literature was analyzed across three domains: training dataset representativeness and subgroup performance, diagnostic threshold effects on clinical decision-making, and liability attribution under opaque ("black box") model architectures.

Major Findings: Dental AI models are trained predominantly on retrospective, single-institution datasets from high-income settings, introducing representation and spectrum biases that compromise generalizability across diverse demographic groups, imaging hardware, and restorative profiles. Furthermore, elevated algorithmic sensitivity for enamel-confined proximal radiolucencies interacts with clinician automation bias to encourage premature operative restoration rather than minimally invasive management, reducing cost-effectiveness and accelerating iatrogenic tooth-tissue loss. Concurrently, post-hoc explainability tools cannot fully resolve neural network opacity, creating unresolved tension around informed consent and negligence allocation when AI-guided treatment fails. Conclusion: Safe translation of dental AI requires moving beyond isolated accuracy benchmarks toward multi-center external validation across underrepresented populations, integration of risk-adaptive minimally invasive clinical thresholds, structured clinician AI literacy, and transparent medico-legal governance.

Keywords: *Artificial Intelligence; Algorithmic Bias; Dental Radiography; Overdiagnosis; Medical Liability;Explainable Artificial Intelligence.*

I. INTRODUCTION

Deep learning and computer vision have rapidly transformed dental diagnostics over the past decade, particularly in the automated interpretation of bitewing, periapical, and panoramic radiographs as well as cone-beam computed tomography **[1]**. Convolutional neural networks (CNNs) trained to detect proximal caries, periapical radiolucencies, and periodontal bone loss frequently report sensitivity and specificity values matching or exceeding those of experienced clinicians in retrospective evaluations **[2]**, **[3]**.

Despite expanding regulatory clearances and commercial deployment in routine dental workflows, the academic literature remains disproportionately fixated on internal discrimination metrics such as sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and pixel-level Intersection-over-Union **[3]**, **[4]**. These metrics evaluate mathematical agreement against curated reference labels under controlled conditions, yet they reveal little about downstream clinical utility, demographic fairness, or patient safety in heterogeneous practice environments.

Three critical translational vulnerabilities warrant urgent scrutiny. First, training corpora derived from homogeneous, high-income clinical populations risk encoding dataset bias that degrades diagnostic equity when deployed in diverse global settings **[4]**, **[5]**. Second, heightened algorithmic sensitivity for early enamel demineralization can trigger clinical overdiagnosis and premature operative intervention **[2]**. Third, the opaque ("black box") architecture of deep neural networks complicates shared decision-making and blurs medico-legal accountability when AI-informed treatment plans fail **[4]**, **[5]**. The objective of this review is to critically evaluate these three interrelated challenges and establish an evidence-informed governance framework for responsible clinical adoption.

ISSN :2584-2706

II. MATERIALS AND METHODS

This study was designed as a structured narrative review synthesizing interdisciplinary literature across clinical dentistry, medical computer vision, bioethics, health economics, and medical jurisprudence. Targeted literature searches were executed across PubMed/MEDLINE, Scopus, Web of Science, and policy repositories of the World Health Organization (WHO) and International Telecommunication Union (ITU) Focus Group on Artificial Intelligence for Health.

Search terms combined controlled vocabulary and free-text descriptors spanning *artificial intelligence*, *deep learning*, *dentistry*, *dental radiography*, *algorithmic bias*, *dataset representativeness*, *overdiagnosis*, *incipient caries*, *automation bias*, *explainable artificial intelligence*, and *medical liability*. Peer-reviewed systematic reviews, randomized controlled diagnostic trials, health-economic microsimulations, international reporting checklists **[6]**, and ethical governance frameworks **[5]** were prioritized for inclusion.

Retrieved literature was synthesized thematically across four analytical dimensions: (i) methodological limitations of benchmark diagnostic metrics; (ii) sources and clinical manifestations of dataset bias across diverse populations; (iii) behavioral and economic consequences of incipient lesion overdetection; and (iv) explainability, informed consent, and tort liability in AI-augmented care. Because this investigation exclusively analyzed previously published literature without human participants, clinical interventions, or identifiable patient records, formal institutional ethics committee approval and informed consent were not required.

III. RESULTS AND DISCUSSION

3.1.The Diagnostic Accuracy Paradigm and Dataset Homogeneity

Systematic evaluations of dental deep learning models demonstrate substantial heterogeneity in dataset construction, lesion definitions, and validation rigor **[3]**, **[6]**. Most published algorithms are developed using retrospective convenience samples sourced from single university hospitals or private imaging centers, split internally into training, validation, and test partitions without independent external validation **[1]**, **[3]**.

Internal validation inflates apparent diagnostic performance through spectrum bias and technical overfitting. Models learn scanner-specific noise profiles, exposure geometries, sensor resolutions, and institutional patient case-mix characteristics rather than generalizable pathophysiological features **[6]**. Furthermore, because histological ground truth is unavailable for routine clinical radiographs, reference labels rely on consensus annotations by small panels of local clinicians—a "fuzzy gold standard" that embeds institutional diagnostic habits directly into the algorithm's objective function **[6]**.

3.2.Algorithmic Bias and Population Disparities in Dental Imaging

In broader medical AI, landmark investigations have established that algorithms trained on unrepresentative cohorts or flawed clinical proxies systematically disadvantage underserved populations. Obermeyer et al. **[7]** demonstrated that commercial healthcare algorithms using expenditure proxies underestimated illness severity in Black patients relative to White patients with identical comorbidity burdens. Similarly, Seyyed-Kalantari et al. **[8]** showed that state-of-the-art chest radiograph classifiers exhibited significantly higher false-negative underdiagnosis rates among female, Black, Hispanic, and lower-socioeconomic patients.

Comparable structural mechanisms threaten dental AI when models trained predominantly on European, North American, or East Asian cohorts are deployed in under-resourced regions such as Sub-Saharan Africa **[4]**, **[5]**. Population-level differences in craniofacial morphology, root canal anatomy, enamel developmental defects, fluorosis prevalence, and restorative material distributions (such as extensive metallic restorations versus tooth-colored composites) alter radiographic contrast and artifact patterns. When algorithms lack training exposure to these epidemiological profiles, calibration drift occurs, yielding asymmetrical false-positive or false-negative error rates across patient subgroups.

3.3.Overdiagnosis of Incipient Lesions and Automation Bias

A critical clinical paradox arises when AI achieves superior sensitivity for early, non-cavitated enamel lesions. In a diagnostic study on bitewing radiographs, Cantu et al. **[2]** reported that a U-Net deep learning model detected initial enamel-restricted caries lesions with markedly higher sensitivity than dentists, who missed a substantial proportion of early radiolucencies. In a subsequent randomized controlled diagnostic trial, Mertens et al. **[9]** confirmed that AI support significantly increased dentists' sensitivity for enamel caries detection, yet also elevated invasive restorative treatment decisions.

From a cariology standpoint, outer-enamel radiolucencies (E1/E2) rarely require surgical excavation and are appropriately managed via non-invasive remineralization, biofilm control, or micro-invasive resin infiltration. However, visual bounding boxes and high-confidence probability overlays generated by commercial software induce *automation bias*—the cognitive tendency of clinicians to defer to automated cues rather than conducting independent clinical synthesis **[12]**. Health-economic microsimulations by Schwendicke et al. **[10]** and Gomez Rossi et al. **[11]** demonstrate that AI-assisted caries detection is cost-effective only when positive detections are paired with non-operative or micro-invasive care; when AI triggers unnecessary composite restorations, lifetime tooth retention declines and healthcare expenditures rise.

3.4. "Black Box" Opacity and Medico-Legal Accountability

Modern deep neural networks comprise millions of non-linear parameters whose internal feature representations are opaque to human inspection **[1]**, **[13]**. While post-hoc explainable AI (XAI) techniques such as saliency heatmaps and Layer-wise Relevance Propagation highlight image regions influencing model output, Ma et al. **[13]** note that heatmaps indicate *where* a network attends rather than providing causal clinical reasoning, and can

highlight spurious radiographic artifacts.

This epistemic opacity creates acute medico-legal ambiguity when AI-guided care causes injury—such as missed periapical pathology, iatrogenic pulp exposure during unnecessary caries excavation, or compromised implant placement. Under prevailing tort doctrines analyzed by Price et al. **[14]**, clinicians currently serve as "learned intermediaries" who bear primary liability whenever they follow an erroneous algorithmic recommendation that deviates from the accepted standard of care. Paradoxically, practitioners face a legal double bind: accepting an incorrect AI flag risks liability for iatrogenic overtreatment, whereas rejecting an accurate AI flag risks liability for diagnostic omission **[14]**, **[15]**. Furthermore, failing to disclose AI involvement during treatment planning undermines patient autonomy and valid informed consent **[4]**, **[5]**.

Table 1. Synthesis of Translational Vulnerabilities in Dental Artificial Intelligence and Governance Mitigations

Vulnerability Domain	Underlying Mechanism	Clinical & Legal Consequence	Recommended Mitigation
Dataset & Representation Bias	Single-center training cohorts lacking geographic, demographic, and imaging hardware diversity **[3]**, **[6]**.	Subgroup calibration drift; inequitable false-negative or false-positive rates in underserved populations **[7]**, **[8]**.	Multi-center external validation; mandatory subgroup performance stratification and dataset nutrition labels **[5]**, **[16]**.
Incipient Lesion Overdiagnosis	High sensitivity for E1 /E2 enamel radiolucencies interacting with clinician automation bias **[2]**, **[9]**, **[12]**.	Unnecessary surgical restoration, premature restorative cycle entry, and loss of cost-effectiveness **[10]**, **[11]**.	Pair AI detection overlays with lesion-depth staging and explicit non-operative/micro-invasive care protocols **[9]**, **[10]**.
"Black Box" Opacity	Non-linear deep network reasoning poorly captured by post-hoc pixel saliency heatmaps **[1]**, **[13]**.	Inability to audit spurious correlations; impaired clinician verification of algorithmic confidence **[13]**.	Report calibrated uncertainty intervals alongside localized explanations; mandate clinical-radiographic correlation **[6]**, **[13]**.
Medico-Legal & Consent Gaps	Unclear liability allocation between clinician, vendor, and institution; omission of AI disclosure **[4]**, **[14]**.	Disproportionate practitioner tort exposure; erosion of shared decision-making and patient trust **[5]**, **[14]**.	Standardized informed consent disclosures, auditable clinical override logs, and regulatory post-market surveillance **[5]**, **[16]**.

3.5. Pathways toward Equitable Validation and Clinical Governance

Bridging the gap between computational promise and safe clinical deployment requires coordinated technical, educational, and regulatory reform. In alignment with the WHO guidance on AI ethics in health **[16]** and the ITU/WHO dental AI ethical checklist developed by Rokhshad et al. **[5]**, regulatory clearance should require prospective, multi-center external validation demonstrating consistent performance across diverse age, sex, ethnic, and socioeconomic strata.

At the chairside level, decision-support interfaces must evolve from binary pathology flags toward calibrated probabilistic outputs integrated with caries risk assessment and minimally invasive treatment thresholds **[10]**, **[13]**. Simultaneously, undergraduate and continuing dental education must cultivate critical AI literacy so practitioners understand dataset limitations, recognize automation bias, document clinical reasoning when accepting or overriding software outputs, and transparently communicate the role of AI during patient consent **[1]**, **[5]**.

IV. CONCLUSION

While deep learning systems demonstrate compelling diagnostic accuracy on curated dental radiograph datasets, sensitivity and specificity alone are insufficient benchmarks for clinical readiness. Unexamined dataset homogeneity threatens diagnostic equity across diverse populations, high sensitivity for incipient enamel lesions risks catalyzing iatrogenic overtreatment through automation bias, and algorithmic opacity complicates informed consent and medico-legal accountability. Sustainable integration of AI into dental practice depends on rigorous multi-population external validation, alignment of diagnostic outputs with minimally invasive clinical workflows, and clear governance frameworks that safeguard patient welfare and practitioner accountability.

V. ACKNOWLEDGEMENTS

The author gratefully acknowledges the international dental informatics and bioethics research community whose foundational scholarship informed this synthesis.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares no financial, professional, institutional, or personal conflicts of interest relevant to the content of this manuscript.

REFERENCES

[1]. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res.
[2]. 2020;99(7):769-774.
[3]. Cantu AG, Gehrung S, Krois J, Chaurasia A, Gomez Rossi J, Gaudin R, et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent. 2020;100:103425.
[4]. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E, et al. Deep learning for caries detection: a systematic review. J Dent. 2022;122:104115.
[5]. Mörch CM, Atsu S, Cai W, Li X, Madathil SA, Liu X, et al. Artificial intelligence and ethics in dentistry: a scoping review. J Dent Res. 2021;100(13):1452-1460.
[6]. Rokhshad R, Ducret M, Chaurasia A, Karteva T, Radenkovic M, Roganovic J, et al. Ethical considerations on artificial intelligence in dentistry: a framework and checklist. J Dent. 2023;135:104593.
[7]. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent. 2021;107:103610.
[8]. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.
[9]. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-2182.
[10]. Mertens S, Krois J, Cantu AG, Arsiwala LT, Schwendicke F. Artificial intelligence for caries detection: randomized trial. J Dent. 2021;115:103849.
[11]. Schwendicke F, Rossi JG, Göstemeyer G, Elhennawy K, Cantu AG, Gaudin R, et al. Cost-effectiveness of artificial intelligence for proximal caries detection. J Dent Res. 2021;100(4):369-376.
[12]. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. JAMA Netw Open. 2022;5(3):e220269.
[13]. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-127.

ISSN :2584-2706

[14]. Ma J, Schneider L, Lapuschkin S, Achtibat R, Duchrau M, Krois J, et al. Towards trustworthy AI in dentistry. J Dent Res. 2022;101(11):1263-1268.
[15]. Price WN 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA.
[16]. 2019;322(18):1765-1766.
[17]. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med.
[18]. 2019;25(1):44-56.
[19]. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.