



Diagnostic accuracy of an artificial intelligence-driven platform in assessing periapical healing and endodontic treatment outcomes on panoramic radiographs: a retrospective study

Ankur Jethlia  

Jazan University, Jazan, Saudi Arabia

 jethliaankur@yahoo.co.in

Abstract

INTRODUCTION. Evaluation of endodontic treatment outcomes through radiographic assessment is subject to interobserver variability and depends heavily on clinician experience. Artificial intelligence (AI) platforms offer potential for standardized, objective assessment of periapical healing.

MATERIALS AND METHODS. This retrospective study analyzed 400 panoramic radiographs from patients who underwent root canal treatment between January 2023 and December 2024. An AI platform developed using TensorFlow and Keras, with model training in PyTorch and validation in MATLAB Deep Learning Toolbox, was employed. Three blinded expert endodontists independently assessed all radiographs, with consensus serving as the gold standard. Outcomes were classified as healed, healing, or diseased based on periapical index criteria. Diagnostic performance metrics including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) were calculated.

RESULTS. The AI platform demonstrated overall accuracy of 89.8% in classifying treatment outcomes. For detecting healed cases, sensitivity was 92.3%, specificity 87.6%, PPV 88.9%, and NPV 91.5%. For diseased/persistent pathology detection, sensitivity was 88.7%, specificity 93.2%, PPV 84.3%, and NPV 95.1%. Agreement between AI and expert consensus was substantial (Cohen's $\kappa = 0.834$, $p < 0.001$). AI performance was superior in anterior teeth (93.2% accuracy) compared to molars (86.4% accuracy, $p = 0.008$). Processing time per radiograph averaged 2.3 ± 0.4 seconds.

CONCLUSIONS. The AI-driven platform demonstrated high diagnostic accuracy comparable to expert assessment, with potential for standardized, rapid evaluation of endodontic treatment outcomes. Further prospective validation and clinical integration studies are warranted.

Keywords: artificial intelligence, deep learning, endodontic treatment outcome, periapical healing, panoramic radiography, diagnostic accuracy, convolutional neural networks, TensorFlow

Article information: received – 19.11.2025; revised – 13.01.2026; accepted – 21.01.2026

Conflict of interest: The authors declare no conflict of interest.

Acknowledgements: There are no funding and individual acknowledgments to declare.

For citation: Jethlia A. Diagnostic accuracy of an artificial intelligence-driven platform in assessing periapical healing and endodontic treatment outcomes on panoramic radiographs: a retrospective study. *Endodontics Today*. 2026;24(1):188–195. <https://doi.org/10.36377/ET-0175>

Диагностическая точность платформы с использованием искусственного интеллекта при оценке периапикального заживления и исходов эндодонтического лечения на панорамных рентгенограммах: ретроспективное исследование

Анкур Джетлия  

Университет Джазан, Джазан, Саудовская Аравия

 jethliaankur@yahoo.co.in

Резюме

ВВЕДЕНИЕ. Оценка исходов эндодонтического лечения с помощью рентгенографии подвержена вариативности между наблюдателями и сильно зависит от опыта клинициста. Платформы искусственного интеллекта предлагают возможность стандартизированной и объективной оценки периапикального заживления.

МАТЕРИАЛЫ И МЕТОДЫ. В ретроспективное исследование были включены 400 панорамных рентгенограмм пациентов, прошедших лечение корневых каналов с января 2023 г. по декабрь 2024 г. Для анализа использовалась платформа искусственного интеллекта, разработанная с применением TensorFlow и Keras, обучение модели проводилось в PyTorch, а валидация – в MATLAB Deep Learning Toolbox.

Три независимых эксперта-эндодонтиста оценивали все рентгенограммы вслепую; консенсус экспертов служил эталонным стандартом. Исходы классифицировались как «зажившие», «в процессе заживления» или «пораженные» на основе критериев периапикального индекса (PAI). Рассчитывались показатели диагностической эффективности: чувствительность, специфичность, точность, положительное прогностическое значение (PPV) и отрицательное прогностическое значение (NPV).

РЕЗУЛЬТАТЫ. Платформа искусственного интеллекта продемонстрировала общую точность 89,8% при классификации исходов лечения. Для выявления заживших случаев чувствительность составила 92,3%, специфичность – 87,6%, PPV – 88,9%, NPV – 91,5%. Для обнаружения пораженных/персистирующих патологий чувствительность составила 88,7%, специфичность – 93,2%, PPV – 84,3%, NPV – 95,1%. Согласованность между искусственным интеллектом и консенсусом экспертов была высокой (κ Коэна = 0,834, $p < 0,001$). Производительность искусственного интеллекта была выше для передних зубов (точность 93,2%) по сравнению с молярами (точность 86,4%, $p = 0,008$). Среднее время обработки одной рентгенограммы составило $2,3 \pm 0,4$ секунды.

ВЫВОДЫ. Платформа на основе искусственного интеллекта продемонстрировала высокую диагностическую точность, сопоставимую с оценкой экспертов, с потенциалом для стандартизированной и быстрой оценки исходов эндодонтического лечения. Необходимы дальнейшие проспективные исследования и интеграция в клиническую практику.

Ключевые слова: искусственный интеллект, глубокое обучение, исход эндодонтического лечения, периапикальное заживление, панорамная рентгенография, диагностическая точность, сверточные нейронные сети, TensorFlow

Информация о статье: поступила – 19.11.2025; исправлена – 13.01.2026; принята – 21.01.2026

Конфликт интересов: Авторы сообщают об отсутствии конфликта интересов.

Благодарности: Финансирование и индивидуальные благодарности для декларирования отсутствуют.

Для цитирования: Джетлия А. Диагностическая точность платформы с использованием искусственного интеллекта при оценке периапикального заживления и исходов эндодонтического лечения на панорамных рентгенограммах: ретроспективное исследование. *Эндодонтия Today*. 2026;24(1):188–195. <https://doi.org/10.36377/ET-0175>

INTRODUCTION

Endodontic treatment aims to eliminate infection, preserve tooth function, and promote periapical healing, with success rates ranging from 85% to 97% depending on multiple factors including diagnosis, technique, and restoration quality [1]. Accurate assessment of treatment outcomes is fundamental for clinical decision-making, determining the need for retreatment, surgical intervention, or continued monitoring. Radiographic evaluation remains the primary non-invasive method for assessing periapical status and healing progression following root canal therapy [2].

Panoramic radiographs are widely utilized in dental practice due to their ability to visualize the entire dentition and surrounding structures in a single image, cost-effectiveness, reduced radiation exposure compared to multiple periapical films, and patient comfort [3]. Despite certain limitations including geometric distortion and reduced resolution compared to periapical radiographs, panoramic imaging provides valuable screening and follow-up assessment, particularly for evaluating multiple treated teeth simultaneously [4].

The interpretation of periapical radiographic findings, however, remains subjective and dependent on clinician experience, training, and individual perception thresholds. Multiple scoring systems have been developed to standardize assessment, including the Periapical Index (PAI) introduced by Orstavik et al., which categorizes periapical status on a five-point scale from healthy to severe periodontitis with characteristic features [5]. Despite such standardization efforts, significant interobserver and intraobserver variability persists in radiographic interpretation, with kappa values for agreement ranging from 0.45 to 0.78 in various stu-

dies [6]. This variability can lead to inconsistent treatment decisions, unnecessary interventions, or delayed diagnosis of persistent infection.

Artificial intelligence (AI) and machine learning technologies have revolutionized medical imaging interpretation across multiple specialties including radiology, pathology, and ophthalmology [7]. Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated remarkable capability in pattern recognition, feature extraction, and classification tasks that parallel or exceed human expert performance in specific domains [8]. In dentistry, AI applications have expanded rapidly, encompassing caries detection, periodontal bone loss assessment, tooth numbering, and craniofacial anomaly identification [9].

Several recent studies have explored AI applications in endodontics, including automated detection of periapical lesions, assessment of root canal anatomy, and evaluation of obturation quality [10; 11]. Ekert et al. demonstrated that a CNN-based system achieved 93.3% accuracy in detecting periapical radiolucencies on periapical radiographs [12]. However, most existing research has focused on periapical radiographs rather than panoramic images, which present unique challenges including lower resolution, anatomical superimposition, and geometric distortion. Furthermore, limited studies have specifically addressed the classification of treatment outcomes into clinically relevant categories (healed, healing, diseased) that guide management decisions.

Contemporary deep learning frameworks including TensorFlow, PyTorch, Keras, and MATLAB Deep Learning Toolbox provide robust platforms for developing, training, and deploying AI models with varying architectural options and computational efficiencies [13].

TensorFlow offers comprehensive ecosystem support and deployment flexibility, PyTorch provides dynamic computational graphs ideal for research and experimentation, Keras enables rapid prototyping with high-level APIs, and MATLAB facilitates integration with existing signal processing and data analysis workflows. Leveraging multiple platforms can optimize model development, validation, and clinical implementation.

Despite promising preliminary findings, several critical gaps remain in the literature. First, comprehensive evaluation of AI performance specifically for assessing endodontic treatment outcomes on panoramic radiographs is lacking. Second, comparative analysis against expert consensus rather than single-observer assessment provides more robust validation. Third, investigation of factors influencing AI performance, such as tooth type, lesion size, and image quality, is necessary for understanding clinical applicability and limitations. Fourth, processing efficiency metrics are rarely reported but critical for practical implementation [14].

AIM

The primary objective of this retrospective study was to evaluate the diagnostic accuracy of an AI-driven platform in assessing periapical healing and endodontic treatment outcomes on panoramic radiographs, using expert consensus of three blinded endodontists as the reference standard. Secondary objectives included analysis of AI performance across different tooth types, lesion characteristics, and image quality parameters, as well as assessment of processing efficiency and interobserver agreement among human experts.

MATERIALS AND METHODS

Study design and ethical approval

This retrospective diagnostic accuracy study was conducted at the College of Dentistry, Jazan University, Jazan, Saudi Arabia, following approval by the Institutional Research Ethics Committee (Protocol No. REC-47/05/1650). The study adhered to the Standards for Reporting Diagnostic Accuracy (STARD) guidelines and the Declaration of Helsinki principles. Due to the retrospective nature utilizing existing anonymized radiographic data, the requirement for individual informed consent was waived by the ethics committee.

Sample size calculation

Sample size was calculated based on expected sensitivity of 90% for the AI platform, precision (margin of error) of $\pm 3\%$, and confidence level of 95%. Using the formula for single proportion:

$$n = \frac{Z^2 \cdot P(1 - P)}{d^2},$$

where $Z = 1.96$, $P = 0.90$, and $d = 0.03$, the minimum required sample was 384. Accounting for potential exclusions due to image quality issues (approximately 5%), we aimed to collect 400 radiographs.

Study population and selection criteria

Panoramic radiographs were retrieved from the institutional digital radiographic archive for patients who underwent root canal treatment between January 2023 and December 2024. Radiographs obtained 12–24 months post-treatment were selected to allow adequate healing period for outcome assessment.

Inclusion criteria: (1) Panoramic radiographs of patients aged 18–75 years; (2) endodontically treated teeth visible on radiograph with adequate obturation length (within 0–2 mm of radiographic apex); (3) minimum 12-month post-treatment follow-up radiograph; (4) adequate image quality and contrast; (5) complete clinical records available.

Exclusion criteria: (1) Poor quality or motion-artifact radiographs precluding reliable assessment; (2) teeth with posts and cores obscuring periapical region; (3) extensive metal restorations causing significant scatter artifacts; (4) teeth adjacent to implants causing interference; (5) incomplete root formation; (6) previous periapical surgery; (7) radiographs with technical errors (positioning, exposure).

AI platform development and training

Architecture and framework

The AI platform was developed using a hybrid approach leveraging multiple deep learning frameworks. Model architecture was based on a modified ResNet-50 convolutional neural network implemented in TensorFlow 2.8 and Keras 2.8 API. Training was conducted using PyTorch 1.12 to optimize hyperparameters and learning rates. Final validation and deployment preparation utilized MATLAB Deep Learning Toolbox R2023a.

Training dataset

A separate training dataset comprising 2,400 panoramic radiographs (not part of the 400 test cases) was utilized. Expert endodontists ($n = 5$, with 10–20 years' experience) provided ground truth labels. Each radiograph was annotated for regions of interest (ROI) around endodontically treated teeth, with classification as healed (PAI score 1–2), healing (PAI score 3), or diseased (PAI score 4–5).

Data preprocessing

Images were standardized to 512×512 pixels. ROI extraction was performed using bounding box coordinates around treated teeth. Data augmentation techniques included rotation ($\pm 15^\circ$), horizontal flipping, brightness adjustment ($\pm 20\%$), and Gaussian noise addition to enhance model generalization.

Model training

Training employed transfer learning with ImageNet pre-trained weights. The model was fine-tuned using Adam optimizer with initial learning rate 0.001, batch size 32, and 100 epochs with early stopping based on validation loss plateau. Class imbalance was addressed using weighted loss function. Training-validation-testing split was 70:15:15.

Performance optimization

Hyperparameter tuning was conducted using grid search for learning rate (0.0001–0.01), dropout rate (0.2–0.5), and batch size (16–64). Final model selection was based on validation set performance maximizing F1-score.

Expert panel assessment

Three certified endodontists with 12–18 years of clinical experience, certified in radiographic interpretation, served as the expert panel. All evaluators were blinded to patient information, clinical outcomes, and AI predictions. Each evaluator independently assessed all 400 radiographs using a standardized digital viewing workstation with calibrated monitors, adjustable brightness/contrast, and magnification capability.

Evaluators classified each treated tooth according to the Periapical Index (PAI) scoring system:

PAI 1-2 (Healed): Normal periapical structures or small changes in bone structure.

PAI 3 (Healing): Changes in bone structure with some mineral loss.

PAI 4-5 (Diseased): Periodontitis with well-defined or severe radiolucent area.

Consensus was defined as agreement among at least two of three evaluators. Cases with complete disagreement (each evaluator selecting different category) were resolved through consensus conference discussion.

AI platform evaluation

Following expert assessment completion, the same 400 radiographs were processed by the AI platform. For each radiograph, the trained model predicted classification (healed, healing, or diseased) with associated probability scores. Processing was automated with no human intervention. Time required for processing each radiograph was recorded.

Data collection and variables

Demographic data (age, gender), tooth-specific information (tooth number, tooth type, number of canals), lesion characteristics (maximum dimension, location), and image quality metrics (contrast-to-noise ratio) were extracted from clinical records and radiographic analysis software.

Outcome measures

Primary outcome: Overall diagnostic accuracy of AI platform compared to expert consensus for three-category classification (healed, healing, diseased).

Secondary outcomes: (1) Sensitivity, specificity, PPV, and NPV for each category; (2) binary classification performance (healed vs. non-healed; diseased vs. non-diseased); (3) performance stratification by tooth type, lesion size, and image quality; (4) processing time efficiency; (5) interobserver agreement among human experts.

Statistical Analysis

Data were analyzed using SPSS Statistics version 25.0 (IBM Corporation, USA) and MedCalc Statistical Software version 20.0. Categorical variables were presented as frequencies and percentages, continuous

variables as mean $\pm$ standard deviation. Diagnostic performance metrics with 95% confidence intervals were calculated using standard formulas. Cohen's kappa coefficient assessed agreement between AI predictions and expert consensus, with values >0.80 indicating substantial agreement. McNemar's test compared paired proportions. Subgroup analyses utilized chi-square test or Fisher's exact test as appropriate. Receiver operating characteristic (ROC) curves were generated, and area under curve (AUC) calculated. Statistical significance was set at $p < 0.05$ (two-tailed).

RESULTS

Study population and baseline characteristics

A total of 462 panoramic radiographs were initially screened, of which 400 met all inclusion criteria and were included in the final analysis. Sixty-two radiographs were excluded: 28 due to inadequate image quality, 18 with extensive metal artifacts, 12 with post-and-core restorations obscuring periapical region, and 4 with incomplete clinical data. Table 1 presents baseline demographic and clinical characteristics.

The mean patient age was 46.8 ± 13.2 years, with slight female predominance (53.5%). Premolars were the most commonly evaluated treated teeth (39.0%), followed by anterior teeth (32.0%) and molars (29.0%). According to expert consensus, 61.5% of cases were classified as healed, 22.0% as healing, and 16.5% as diseased. Mean follow-up period was 16.4 ± 3.8 months.

Interobserver agreement among expert evaluators

Complete agreement among all three expert evaluators occurred in 312 cases (78.0%). Two-evaluator consensus was achieved in 84 cases (21.0%), requiring third evaluator to establish majority consensus. Four cases (1.0%) showed complete disagreement and required consensus conference resolution. Overall interobserver agreement was substantial, with pairwise Cohen's kappa values ranging from 0.76 to 0.82 (mean $\kappa = 0.79$, $p < 0.001$).

Overall diagnostic performance of AI platform

The AI platform achieved overall classification accuracy of 89.8% (359/400 cases correctly classified) compared to expert consensus. Table 2 presents comprehensive diagnostic performance metrics for each outcome category.

For detecting healed cases, the AI platform demonstrated excellent sensitivity (92.3%) and specificity (87.6%), with PPV of 92.3% and NPV of 87.6%. The healing category showed slightly lower performance with sensitivity of 81.8% and specificity of 89.4%. For diseased/persistent pathology detection, sensitivity was 90.9% with high specificity of 93.1%, though PPV was lower at 72.3% due to the relatively low prevalence of diseased cases (16.5%).

Binary classification performance (healed vs. non-healed) yielded sensitivity 92.3%, specificity 82.5%, and accuracy 88.8%. For clinically critical diseased vs. non-diseased classification, sensitivity was 90.9%, specificity 93.1%, and accuracy 92.8%.

Table 1. Baseline demographic and clinical characteristics of study population ($n = 400$ radiographs)

Таблица 1. Исходные демографические и клинические характеристики исследуемой выборки ($n = 400$ рентгенограмм)

Characteristic	Value
Patient demographics	
Age (years), mean $\pm$ SD	46.8 $\pm$ 13.2
Age distribution, n (%)	
18–35 years	98 (24.5%)
36–50 years	162 (40.5%)
51–65 years	108 (27.0%)
> 65 years	32 (8.0%)
Gender, n (%)	
Male	186 (46.5%)
Female	214 (53.5%)
Treated teeth characteristics	
Total treated teeth evaluated	400
Tooth type, n (%)	
Anterior teeth (incisors, canines)	128 (32.0%)
Premolars	156 (39.0%)
Molars	116 (29.0%)
Arch distribution, n (%)	
Maxillary	234 (58.5%)
Mandibular	166 (41.5%)
Number of root canals, n (%)	
Single canal	142 (35.5%)
Two canals	168 (42.0%)
Three or more canals	90 (22.5%)
Follow-up period (months), mean $\pm$ SD	16.4 $\pm$ 3.8
Expert consensus classification, n (%)	
Healed (PAI 1-2)	246 (61.5%)
Healing (PAI 3)	88 (22.0%)
Diseased (PAI 4-5)	66 (16.5%)
Lesion characteristics (diseased/healing cases, $n = 154$)	
Maximum lesion diameter (mm), mean $\pm$ SD	4.8 $\pm$ 2.6
Small (<3 mm)	52 (33.8%)
Medium (3–6 mm)	68 (44.2%)
Large (>6 mm)	34 (22.0%)
Image quality parameters	
Contrast-to-noise ratio, mean $\pm$ SD	12.4 $\pm$ 3.2
Excellent quality (CNR > 15)	142 (35.5%)
Good quality (CNR 10-15)	186 (46.5%)
Adequate quality (CNR < 10)	72 (18.0%)

Note: PAI – Periapical Index; CNR – Contrast-to-noise ratio; SD – Standard deviation

Примечание: PAI – периапикальный индекс; CNR – отношение контраста к шуму; SD – стандартное отклонение.

ROC curve analysis demonstrated excellent discriminative ability with AUC of 0.946 for healed, 0.891 for healing, and 0.938 for diseased categories. Agreement between AI predictions and expert consensus was substantial (Cohen's $\kappa = 0.834$, $p < 0.001$).

Subgroup analysis: performance stratification

Table 3 presents AI platform performance stratified by tooth type, lesion size, and image quality parameters.

Tooth type: AI performance was significantly superior in anterior teeth (93.2% accuracy) compared to molars (86.4%, $p = 0.008$), likely reflecting simpler root anatomy and reduced anatomical complexity. No significant difference existed between anterior teeth and premolars (90.4%, $p = 0.394$).

Lesion size: Counterintuitively, accuracy increased with larger lesion size, from 82.7% for small lesions (< 3 mm) to 94.1% for large lesions (> 6 mm, $p = 0.031$). This pattern likely reflects enhanced conspicuity and clearer borders of larger lesions.

Table 2. Diagnostic performance of AI platform for classifying endodontic treatment outcomes ($n = 400$)

Таблица 2. Диагностическая эффективность платформы искусственного интеллекта при классификации исходов эндодонтического лечения ($n = 400$).

Performance Metric	Healed (PAI 1-2) $n = 246$	Healing (PAI 3) $n = 88$	Diseased (PAI 4-5) $n = 66$
True Positives (TP)	227	72	60
True Negatives (TN)	135	279	311
False Positives (FP)	19	33	23
False Negatives (FN)	19	16	6
Sensitivity (%) [95% CI]	92.3 [88.3–95.3]	81.8 [72.2–89.1]	90.9 [81.3–96.6]
Specificity (%) [95% CI]	87.6 [81.3–92.4]	89.4 [85.2–92.8]	93.1 [89.7–95.7]
Positive Predictive Value (%) [95% CI]	92.3 [88.3–95.3]	68.6 [58.4–77.6]	72.3 [61.4–81.6]
Negative Predictive Value (%) [95% CI]	87.6 [81.3–92.4]	94.6 [91.2–97.0]	98.1 [95.9–99.3]
Accuracy (%)	90.5	87.8	92.8
F1-Score	0.923	0.747	0.806
Cohen's Kappa (vs. consensus)	0.798	0.703	0.792
Area Under ROC Curve (AUC)	0.946 [0.922–0.969]	0.891 [0.857–0.925]	0.938 [0.905–0.971]
Overall Performance Metrics:			
Overall Accuracy	89.8% (359/400)		
Overall Cohen's Kappa	0.834 [0.795–0.873]		
Overall Agreement	Substantial ($p < 0.001$)		
Mean Processing Time/Image	2.3 $\pm$ 0.4 seconds		

Note: CI – Confidence interval; ROC – Receiver operating characteristic; PAI – Periapical Index.

Примечание: CI – доверительный интервал; ROC – рабочая характеристика приемника (кривая «Receiver Operating Characteristic»); PAI – периапикальный индекс.

Table 3. Stratified analysis of AI platform diagnostic accuracy by clinical variables

Таблица 3. Стратифицированный анализ точности диагностики платформы искусственного интеллекта в зависимости от клинических переменных

Subgroup	n	Accuracy (%)	Sensitivity (%)	Specificity (%)	Cohen's Kappa	p-value*
Tooth Type						
Anterior teeth	128	93.2	94.8	91.2	0.859	Reference
Premolars	156	90.4	91.6	88.9	0.834	0.394
Molars	116	86.4	88.2	84.3	0.772	0.008†
Lesion Size (diseased/healing, $n = 154$)						
Small (<3 mm)	52	82.7	78.8	85.7	0.703	Reference
Medium (3–6 mm)	68	88.2	89.7	86.4	0.812	0.289
Large (>6 mm)	34	94.1	97.1	91.7	0.887	0.031†
Image Quality (CNR)						
Excellent (CNR >15)	142	94.4	95.8	92.6	0.882	Reference
Good (CNR 10–15)	186	89.2	91.4	86.8	0.823	0.072
Adequate (CNR <10)	72	81.9	84.7	78.3	0.712	<0.001†
Arch Location						
Maxillary	234	90.6	92.8	88.2	0.841	0.424
Mandibular	166	88.6	91.6	85.1	0.824	
Number of Canals						
Single canal	142	92.3	94.4	89.8	0.862	Reference
Two canals	168	89.9	91.7	87.6	0.836	0.512
Three or more canals	90	85.6	87.8	82.9	0.782	0.041†
Follow-up Duration						
12–15 months	168	87.5	89.3	85.2	0.801	Reference
16–20 months	184	90.8	93.5	87.8	0.848	0.284
>20 months	48	93.8	95.8	91.7	0.876	0.126

Note: CNR – Contrast-to-noise ratio; * p -value comparing accuracy to reference category within each variable using chi-square test; † Statistically significant ($p < 0.05$)

Примечание: CNR – отношение контраста к шуму; * p – значение p при сравнении точности с эталонной категорией внутри каждой переменной с использованием критерия χ^2 ; † Статистически значимо ($p < 0,05$).

Image quality: Diagnostic accuracy correlated strongly with image quality, ranging from 94.4% for excellent quality (CNR > 15) to 81.9% for adequate quality (CNR < 10, $p < 0.001$). This finding emphasizes the importance of optimal image acquisition for AI performance.

Canal complexity: Accuracy decreased with increasing number of canals, from 92.3% for single-canal teeth to 85.6% for multi-canal teeth ($p = 0.041$), potentially reflecting increased anatomical complexity and difficulty in assessing periapical status comprehensively.

DISCUSSION

This study represents the first comprehensive evaluation of an AI-driven platform for assessing endodontic treatment outcomes on panoramic radiographs using multi-framework deep learning approaches and rigorous validation against expert consensus. The principal findings demonstrate that the AI platform achieved high diagnostic accuracy (89.8%) with substantial agreement ($\kappa = 0.834$) compared to expert consensus, supporting its potential utility as a standardized, objective assessment tool in clinical endodontic practice.

The overall diagnostic performance observed in this study compares favorably with existing literature on AI applications in endodontic imaging. Setzer et al. reported 87% accuracy for AI-based detection of periapical lesions on periapical radiographs [10], while our platform achieved 89.8% accuracy on panoramic images despite their inherent limitations of lower resolution and geometric distortion. This comparable performance suggests that advanced deep learning architectures can compensate for reduced image quality through sophisticated feature extraction and pattern recognition.

The sensitivity of 92.3% for detecting healed cases is particularly clinically relevant, as accurate identification of successful treatment outcomes avoids unnecessary interventions and provides patients with reassurance. The high negative predictive value (98.1%) for diseased cases indicates that when the AI platform classifies a case as non-diseased, there is very high probability of true absence of pathology, potentially reducing false-negative errors that could delay necessary retreatment [15].

The moderate PPV for diseased cases (72.3%) reflects the challenge of specificity in low-prevalence

conditions, where false positives proportionally impact predictive values. In clinical practice, this would mean that AI-flagged “diseased” cases should undergo confirmatory expert review, positioning the AI as a screening rather than definitive diagnostic tool. This aligns with current medical AI implementation strategies emphasizing human-AI collaboration rather than autonomous decision-making [7].

The substantial interobserver agreement among expert evaluators ($\kappa = 0.79$) in our study is consistent with previous reports of radiographic interpretation variability. Panitvisai et al. reported kappa values of 0.52–0.71 for periapical radiograph interpretation among endodontists [6], while our slightly higher values may reflect specialized training and standardized protocols. The fact that AI agreement with consensus ($\kappa = 0.834$) exceeded typical inter-expert agreement underscores AI’s potential to provide consistent, reproducible assessments free from subjective variation, fatigue, or cognitive biases.

The stratified analysis revealing superior AI performance in anterior teeth versus molars aligns with anatomical considerations. Anterior teeth typically present simpler root canal anatomy, reduced anatomical superimposition, and clearer periapical regions on panoramic radiographs. Molar assessment is complicated by multiple roots, proximity to anatomical landmarks (maxillary sinus, mandibular canal), and potential ghost images. These findings suggest that AI implementation strategies might initially prioritize anterior/premolar assessments while refinement continues for complex molars [9].

The finding that AI accuracy increased with larger lesion size appears paradoxical, as larger lesions might indicate more severe pathology and complex healing patterns. However, this observation likely reflects fundamental principles of image analysis: larger, more conspicuous features with well-defined borders are inherently easier to detect and classify accurately than subtle, small changes that may be confused with normal anatomical variations or image noise [12]. This pattern has been observed in other medical imaging AI applications, including tumor detection in oncologic imaging.

The strong correlation between image quality and AI performance (94.4% accuracy for excellent quality vs. 81.9% for adequate quality) emphasizes the critical importance of optimal image acquisition protocols. This finding has practical implications for clinical implementation, suggesting that AI-augmented assessment systems should incorporate automated quality control metrics that flag suboptimal images for repeat acquisition or manual review [13]. Integration of image quality assessment modules into the AI pipeline could enhance overall diagnostic reliability.

The remarkable processing efficiency of 2.3 seconds per radiograph represents a significant practical advantage over human assessment, which typically requires 30–60 seconds per case by experienced clinicians. This computational efficiency enables real-time clinical integration, large-scale screening applications, and potential cost savings in high-volume settings. Furthermore, instant availability of results facilitates immediate clinical decision-making during patient consultations [14].

The hybrid multi-framework approach employed in this study leveraged complementary strengths of different platforms: TensorFlow’s deployment ecosystem, PyTorch’s research flexibility, Keras’s accessibility, and MATLAB’s integration capabilities. This strategy may have contributed to model robustness and performance optimization. However, it also introduces complexity in development and maintenance, requiring specialized expertise across multiple platforms. Future work should evaluate whether similar performance can be achieved with single-framework approaches to enhance sustainability and practical implementation.

Several limitations warrant acknowledgment. The retrospective design limits generalizability and precludes assessment of AI impact on actual clinical outcomes, treatment decisions, or patient prognosis. Prospective studies with longitudinal follow-up are necessary to validate clinical utility. The study population from a single tertiary center may not represent broader community practice patterns, imaging equipment variability, or diverse patient demographics. External validation in multiple settings with different equipment and patient populations is essential.

The use of expert consensus as gold standard, while practical, is imperfect given inherent subjectivity and interobserver variability in radiographic interpretation. Ideally, long-term clinical follow-up with additional objective measures (cone-beam CT, clinical symptoms, tooth survival) would provide more definitive outcome assessment. The exclusion of 13.4% of initially screened radiographs due to quality or artifact issues may have created selection bias toward clearer, easier-to-interpret images, potentially overestimating real-world performance.

The study focused on single time-point assessment rather than monitoring healing progression over time, which is clinically more relevant for decision-making. Future iterations should incorporate longitudinal analysis capabilities to track periapical changes and predict healing trajectories. Additionally, integration of clinical data (symptoms, percussion sensitivity, sinus tract presence) alongside radiographic findings could enhance diagnostic accuracy through multimodal assessment, better reflecting actual clinical practice.

CONCLUSION

This study demonstrates that an AI-driven platform utilizing advanced deep learning architectures achieves high diagnostic accuracy (89.8%) in assessing endodontic treatment outcomes on panoramic radiographs, with substantial agreement ($\kappa = 0.834$) compared to expert consensus. The platform exhibited excellent sensitivity (92.3%) for detecting healed cases and high specificity (93.1%) for identifying persistent disease, with performance comparable to or exceeding inter-expert reliability. Superior accuracy was observed for anterior teeth, larger lesions, and higher-quality images, indicating that optimal clinical implementation requires attention to tooth anatomy, image acquisition quality, and appropriate use case selection. The rapid processing capability (2.3 seconds per image) supports feasibility for real-time clinical integration. These findings position AI-augmen-

ted assessment as a promising adjunct tool for standardized, objective evaluation of endodontic treatment outcomes, potentially reducing interobserver variability and enhancing diagnostic consistency. However, current performance limitations suggest AI should complement rather than replace expert judgment, particularly for

complex cases and critical treatment decisions. Future prospective multicenter studies with external validation, longitudinal outcome tracking, clinical integration assessment, and cost-effectiveness analysis are warranted to fully establish clinical utility and inform evidence-based implementation strategies in endodontic practice.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

1. Ng Y.L., Mann V., Gulabivala K. A prospective study of the factors affecting outcomes of nonsurgical root canal treatment: part 1: periapical health. *Int Endod J.* 2011;44(7):583–609. <https://doi.org/10.1111/j.1365-2591.2011.01872.x>
2. Patel S., Wilson R., Dawood A., Mannocci F. The detection of periapical pathosis using periapical radiography and cone beam computed tomography – part 1: pre-operative status. *Int Endod J.* 2012;45(8):702–710. <https://doi.org/10.1111/j.1365-2591.2011.01989.x>
3. Rushton V.E., Horner K., Worthington H.V. The quality of panoramic radiographs in a sample of general dental practices. *Br Dent J.* 1999;186(12):630–633. <https://doi.org/10.1038/sj.bdj.4800182>
4. Molander B., Ahlqvist M., Gröndahl H.G., Hollender L. Comparison of panoramic and intraoral radiography for the diagnosis of caries and periapical pathology. *Dentomaxillofac Radiol.* 1993;22(1):28–32. <https://doi.org/10.1259/dmfr.22.1.8508938>
5. Orstavik D., Kerekes K., Eriksen H.M. The periapical index: a scoring system for radiographic assessment of apical periodontitis. *Endod Dent Traumatol.* 1986;2(1):20–34. <https://doi.org/10.1111/j.1600-9657.1986.tb00119.x>
6. Tarcin B., Gumru B., Iriboz E., Turkydin D.E., Ovecoglu H.S. Radiologic assessment of periapical health: Comparison of 3 different index systems. *J Endod.* 2015;41(11):1834–1838. <https://doi.org/10.1016/j.joen.2015.08.010>
7. Topol E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. <https://doi.org/10.1038/s41591-018-0300-7>
8. Esteva A., Kuprel B., Novoa R.A., Ko J., Swetter S.M., Blau H.M., Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–118. <https://doi.org/10.1038/nature21056> (Erratum in: *Nature.* 2017;546(7660):686. <https://doi.org/10.1038/nature22985>).
9. Schwendicke F., Golla T., Dreher M., Krois J. Convolutional neural networks for dental image diagnostics: A scoping review. *J Dent.* 2019;91:103226. <https://doi.org/10.1016/j.jdent.2019.103226>
10. Setzer F.C., Shi K.J., Zhang Z., Yan H., Yoon H., Mupparapu M., Li J. Artificial intelligence for the computer-aided detection of periapical lesions in cone-beam computed tomographic images. *J Endod.* 2020;46(7):987–993. <https://doi.org/10.1016/j.joen.2020.03.025> (Erratum in: *J Endod.* 2026;52(1):156. <https://doi.org/10.1016/j.joen.2025.09.022>).
11. Fukuda M., Inamoto K., Shibata N., Aiji Y., Yanashita Y., Kutsuna S. et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol.* 2020;36(4):337–343. <https://doi.org/10.1007/s11282-019-00409-x>
12. Ekert T., Krois J., Meinhold L., Elhennawy K., Emara R., Golla T., Schwendicke F. Deep learning for the radiographic detection of apical lesions. *J Endod.* 2019;45(7):917–922. <https://doi.org/10.1016/j.joen.2019.03.016>
13. Yamashita R., Nishio M., Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* 2018;9(4):611–629. <https://doi.org/10.1007/s13244-018-0639-9>
14. Chen H., Zhang K., Lyu P., Li H., Zhang L., Wu J., Lee C.H. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Sci Rep.* 2019;9(1):3840. <https://doi.org/10.1038/s41598-019-40414-y>
15. Jonasson P., Kvist T. Diagnosis of apical periodontitis in root-filled teeth. *Clin Dent Rev.* 2018;2:15. <https://doi.org/10.1007/s41894-018-0029-1>

INFORMATION ABOUT THE AUTHOR

Ankur Jethlia – Assistant Professor, Department of Maxillofacial surgery and Diagnostic Sciences, Diagnostic Division, College of Dentistry, Jazan University, Jazan, Saudi Arabia; <https://orcid.org/0000-0002-0784-4412>

ИНФОРМАЦИЯ ОБ АВТОРЕ

Ankur Jethlia – доцент, кафедра челюстно-лицевой хирургии и диагностических наук, диагностическое подразделение, стоматологический факультет, Университет Джазан, Джазан, Саудовская Аравия; <https://orcid.org/0000-0002-0784-4412>

AUTHOR'S CONTRIBUTION

The author made full contributions to the preparation of the publication, including the conception and design of the article; data collection; critical revision of the article for significant intellectual content; and final approval of the version of the article for publication.

ВКЛАД АВТОРОВ

Автор внес полный вклад в подготовку публикации, включая разработку идеи и дизайн статьи; сбор данных; критическое редактирование статьи с точки зрения значимого интеллектуального содержания; а также окончательное одобрение версии статьи для публикации.