Artificial Intelligence for Difficulty Prediction in Impacted Mandibular Third Molar Surgery: A Scoping Review

Trong Tinh Nguyen
Faculty of Dentistry
Van Lang University
Ho Chi Minh City, Vietnam
tinh.nt@vlu.edu.vn

Cao My Ngan Nguyen
Faculty of Odonto-Stomatology
Can Tho University of Medicine and Pharmacy
Can Tho, Vietnam
25350112026@student.ctump.edu.vn

Huu Trieu Truong
Faculty of Odonto-Stomatology
Can Tho University of Medicine and Pharmacy
Can Tho, Vietnam
2253020066@student.ctump.edu.vn

Trang Thao Nguyen
Laboratory for Artificial Intelligence, Institute for Computational Science and Artificial Intelligence
Faculty of Information Technology, Van Lang School of Technology
Van Lang University
Ho Chi Minh City, Vietnam
thao.nguyentrang@vlu.edu.vn

Nhut Khue Truong*
Faculty of Odonto-Stomatology
Can Tho University of Medicine and Pharmacy
Can Tho, Vietnam
*Corresponding author: tnkhue@ctump.edu.vn

***Abstract*—Surgical removal of impacted mandibular third molars (M3M) is a common procedure, yet preoperative difficulty assessment remains heterogeneous and operator-dependent. While artificial intelligence (AI) has been increasingly investigated to automate radiographic interpretation and risk prediction, the literature lacks a comprehensive synthesis specifically focusing on how AI operationalizes surgical difficulty. Following PRISMA-ScR guidelines, this scoping review identified 12 original studies from PubMed, ScienceDirect, and Google Scholar up to December 31, 2025. Results show that most models utilize panoramic radiographs as primary input, with architectures evolving from traditional CNNs to advanced Transformers and YOLO-based detectors. We found that surgical difficulty is operationalized through non-equivalent endpoints, ranging from subjective radiographic indices to objective intraoperative time. Our analysis further highlights a paradigm shift where Transformer-based models outperform CNNs by capturing the long-range spatial dependencies essential for complex tooth-nerve risk assessment. Critical appraisal via PROBAST and CLAIM reveals that while models are highly applicable, 100% carry a high risk of bias due to insufficient external validation and reporting deficits. We propose the M3M-AI Framework to harmonize input fidelity, core outcome sets, and model explainability, providing a roadmap for reliable clinical translation and robust cross-study comparison.**

***Keywords*—artificial intelligence; deep learning; mandibular third molar; panoramic radiography; CBCT; difficulty prediction; scoping review**

I. INTRODUCTION

Surgical extraction of impacted mandibular third molars is routinely performed worldwide. Operative complexity varies markedly and may be associated with postoperative morbidity and complications, particularly sensory disturbance related to inferior alveolar nerve (IAN) injury [4], [15]. Robust preoperative assessment is therefore essential for surgical planning, referral decisions, and patient counseling.

Conventional evaluation is commonly based on panoramic radiography combined with classification systems such as Winter's angulation and the Pell and Gregory relationship to the ramus and occlusal plane [1], [4]. Composite indices, including the Pederson difficulty index, were developed to support standardization but demonstrate inconsistent predictive validity across settings and operators [1], [5]. When panoramic signs suggest proximity between the third molar and the IAN canal, cone-beam computed tomography (CBCT) may be requested for three-dimensional assessment, although access, radiation dose, and workflow constraints limit routine use [8], [15].

Artificial intelligence (AI), particularly deep learning, has rapidly advanced in dento-maxillofacial radiology and can automate detection, segmentation, and risk stratification tasks [4], [10]. For mandibular third molars, published systems range from automated localization and impaction classification to prediction of surgical difficulty, extraction time, and IAN-

related risk [1–5], [7–12], [15]. However, the literature remains heterogeneous in target definitions, reference standards, datasets, and validation strategies, making it challenging to compare results and to judge clinical readiness.

While previous systematic reviews have evaluated the broader potential of AI in dento-maxillofacial radiology [4], [10] there is a distinct lack of synthesis focusing specifically on the prediction of surgical difficulty for impacted mandibular third molars. Existing reviews often prioritize diagnostic accuracy and exclude studies with heterogeneous outcome measures. Because current AI models utilize highly diverse endpoints—ranging from classic radiographic indices to operative time and surgeon-estimated difficulty grades—these studies cannot be readily combined via meta-analysis. Therefore, a scoping review is necessary to systematically map these diverse methodologies, model architectures, and input-output designs [6].

Beyond dentistry, these models exemplify AI-driven risk assessment and decision-support workflows that can be embedded into healthcare infrastructure to standardize triage, optimize resource allocation, and improve procedural safety.

This scoping review maps current evidence on AI models for preoperative assessment and prognosis in impacted mandibular third molar surgery, characterizes model families and input-output designs, and identifies gaps that should be addressed before clinical implementation.

II. Methods

A. Study Design and Reporting

A scoping review was conducted in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [6]. PRISMA-ScR is a reporting guideline designed for scoping reviews, which map the breadth, characteristics, and knowledge gaps of a body of evidence, rather than estimating a pooled effect size as in systematic reviews or meta-analysis. In contrast to narrative reviews, scoping reviews use a transparent, predefined search and selection process with explicit eligibility criteria and systematic data charting.

B. Eligibility Criteria

We included original research that developed or evaluated AI models applied to impacted mandibular third molars and explicitly described model architecture with clearly defined inputs and outputs relevant to preoperative assessment or prognosis (e.g., detection/localization, impaction classification, difficulty scoring, extraction time prediction, IAN proximity/contact assessment, or complication risk estimation). We excluded non-original publications, studies without sufficient technical detail, non-English manuscripts, and studies unrelated to mandibular third molar surgery.

The specificity of these criteria was maintained to ensure that the mapped evidence directly informs the surgical decision-making process for third molars, as general dental AI models often lack the specialized features (e.g., root curvature or IAN proximity) necessary for surgical difficulty estimation [6].

C. Information Sources and Search Strategy

PubMed, ScienceDirect, and Google Scholar were searched from inception to December 31, 2025. The following Boolean search strings were used and adapted to each database syntax. PubMed: (("Molar, Third"[Mesh] OR "Tooth, Impacted"[Mesh] OR "impacted mandibular third molar"[tiab] OR "mandibular third molar impaction"[tiab] OR "wisdom tooth"[tiab] OR "wisdom teeth"[tiab])) AND (("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Neural Networks, Computer"[Mesh] OR "Artificial Intelligence"[tiab] OR "Deep learning"[tiab] OR "Machine learning"[tiab] OR "Neural network"[tiab])) AND (("Risk Assessment"[Mesh] OR "Forecasting"[Mesh] OR "Difficulty"[tiab] OR "Prediction"[tiab] OR "Assessment"[tiab] OR "Classification"[tiab])). ScienceDirect: ("mandibular third molar" OR "wisdom tooth") AND ("artificial intelligence" OR "machine learning" OR "deep learning") AND (difficulty OR prediction OR assessment). Google Scholar: (intitle:"mandibular third molar" OR intitle:"wisdom tooth" OR intitle:"wisdom teeth") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network") AND (difficulty OR prediction). Quotation marks were used for multi-word phrases, and AND/OR operators were applied as shown.

D. Selection Process and Data Charting

Records were screened by title and abstract, followed by full-text assessment. At the title/abstract stage, records were excluded if they (i) were not original research, (ii) did not involve impacted mandibular third molars or surgical assessment, (iii) did not use an AI-based method, or (iv) did not report a clinically relevant outcome for preoperative assessment or prognosis. A standardized charting form extracted imaging modality, dataset characteristics, labeling and reference standards, model family, target outputs, and evaluation metrics (e.g., accuracy/AUC for classification, mean average precision for detection, Dice/IoU for segmentation). Findings were synthesized descriptively and grouped by task category.

E. Quality Assessment and Risk of Bias

The scientific validity and reporting transparency of the included studies were evaluated using two standardized frameworks: PROBAST (Prediction model Risk Of Bias Assessment Tool) for clinical prediction models and the CLAIM (Checklist for Artificial Intelligence in Medical Imaging) guidelines [13], [14]. The appraisal was performed independently by two reviewers across four PROBAST domains (Participants, Predictors, Outcome, and Analysis). Any discrepancies between the two primary reviewers were resolved through discussion or, if necessary, by consultation with a third senior reviewer to reach a final consensus. This dual-framework approach, combined with a multi-reviewer verification process, ensures both clinical rigor and technical transparency in our critical analysis.

III. RESULTS

A. *Study Selection*

The initial search yielded 399 records; 370 remained for screening after removing 29 duplicates. Exclusions at the title and abstract stage (n = 354) were primarily due to a lack of original AI methodology or clinical relevance to third molar surgery. Of the 16 articles undergoing full-text evaluation, 4 were excluded for insufficient data or incorrect interventions, resulting in 12 studies for final synthesis (Fig. 1) [1–5], [7–12], [15].

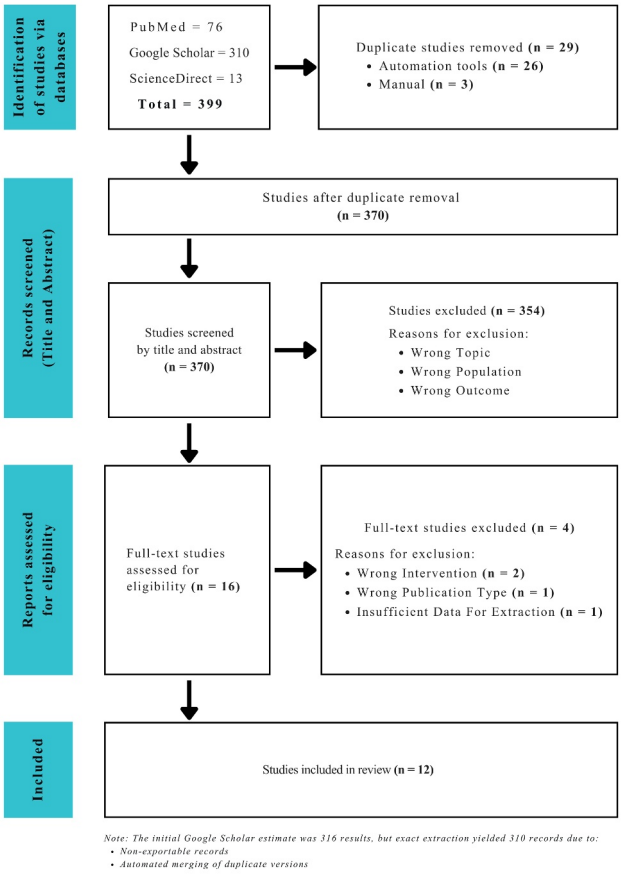

Figure 1. PRISMA-ScR flow diagram of study selection.

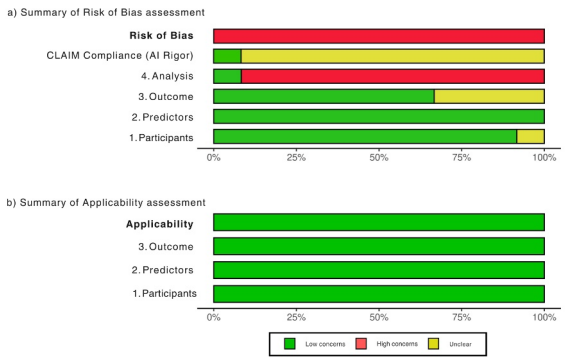

Figure 2. Quality appraisal summary of included studies based on PROBAST and CLAIM criteria.

B. *Synthesis of Study Characteristics*

Detailed in Table II, datasets ranged from 419 surgical cases to 4,903 panoramic radiographs. OPG was the primary input modality (n = 11), with architectures evolving from standard CNNs (e.g., ResNet, VGG) toward advanced multi-stage pipelines and object detection frameworks like YOLO11 and Vision Transformers.

C. *Heterogeneity in Outcome Definitions and Ground Truth*

A critical analysis of the evidence revealed substantial heterogeneity in how ground truth for surgical difficulty was operationalized. The majority of studies (n = 7) relied on subjective radiographic indices—primarily the Pederson or Pell & Gregory scales, with labels established by the consensus of dental experts or radiologists. Conversely, an emerging trend prioritized objective hard endpoints, such as actual intraoperative time or the specific procedural requirements (e.g., need for tooth sectioning or ostectomy) recorded in surgical logs.

D. *Model Performance and Task-Specific Challenges*

Model performance varied significantly across the taxonomy of AI tasks (Table I). While detection and localization tasks consistently achieved near-perfect accuracy (mAP > 99%) [2], [10], [15], high-level difficulty prediction remained a technical challenge. Accuracy for complex cases, such as those requiring expert-level intervention or involving intricate root curvature, ranged from 61% to 89% [3]. Notably, multimodal fusion models that integrated OPG features with clinical variables demonstrated superior predictive power compared to unimodal image-based models (correlation R increasing from 0.39 to 0.83; R^2 = 0.68) [11]. Representative studies include automated detection and segmentation of third molars [3], [12], [15], difficulty prediction from panoramic images [4], [10], comparison of AI performance against human specialists [5], and IAN-related positional assessment on panoramic radiography [15].

E. *Quality Assessment and Critical Appraisal*

The critical appraisal results are visualized in Fig. 2. Analysis using PROBAST revealed that 100% of the studies (n = 12) had an overall high risk of bias, primarily due to a lack of external validation in the Analysis domain. Conversely, 100% showed low concern for applicability, confirming high clinical relevance to third molar surgery. Regarding technical transparency, 11 studies (91.7%) achieved medium CLAIM compliance, while Li et al (2023) [9] demonstrated high compliance, setting a benchmark for reporting rigor.

TABLE I. AI TASK CATEGORIES AND TYPICAL EVALUATION IN MANDIBULAR THIRD MOLAR STUDIES

AI task category	Operational definition (what the AI does)	Typical input(s)	Target output(s)	Common evaluation metrics	No. of studies (n, %)
Detection / localization	Detects mandibular third molars and/or generates a bounding box/ROI for downstream prediction	Panoramic radiograph (OPG)	Bounding box/ROI crop (often includes adjacent teeth ± IAN region)	mAP (e.g., IoU 0.5), precision/recall, F1	6 (50.0%)
Radiographic attribute / index-component classification	Classifies radiographic attributes used to construct a difficulty index (e.g., Winter, Pell & Gregory components, root-related criteria), sometimes as multi-label outputs	OPG (or ROI) or CBCT report text (NLP feature extraction)	Winter / Pell & Gregory attributes; radiographic criteria used to compute DI; structured feature set from reports (angulation, roots, curvature, root-canal relationship)	Accuracy, AUC, F1, kappa; confusion matrix	6 (50.0%)
Difficulty prediction	Predicts extraction difficulty level (e.g., 3–4 classes) or difficulty index category/score	OPG/ROI ± clinical variables or CBCT report text or surgical records	Difficulty level (easy→very difficult); Pederson DI level/score; surgeon-specific difficulty class	Accuracy, sensitivity/specificity, AUC, F1, kappa	12 (100%)
Operative time prediction	Predicts time to extract (continuous) using imaging ± clinical variables	OPG + clinical variables	Predicted extraction time (minutes)	MAE/RMSE, correlation (R), R^2	1 (8.3%)
Procedure-step prediction	Predicts the need for specific surgical steps (e.g., crown/root separation, bone removal), used as a procedural proxy for difficulty	OPG/ROI	Need for crown/root separation; bone removal/ostectomy requirement	Sensitivity/specificity, accuracy, AUC	1 (8.3%)
Risk/complication prediction	Predicts likelihood of IAN injury (or similar complication risk) as a clinically important difficulty-adjacent endpoint	OPG/ROI	IAN injury likelihood (class/probability)	Accuracy, AUC, F1	1 (8.3%)
Tool integration / decision-support implementation	Implements a pipeline into a usable tool (e.g., GUI) for clinical use (not just a standalone model)	OPG (pipeline-based)	Automated output display (e.g., DI printed/GUI workflow)	Usability/expert feedback (when reported); sometimes performance metrics	1 (8.3%)

TABLE II. CHARACTERISTICS OF INCLUDED STUDIES (N = 12)

Study (Author, Year)	Dataset Size	Input	Model Architecture	Reference Standard & Ground Truth	Key Performance	XAI	GUI
Yoo (2021)	600 OPGs (1,053 M3Ms)	OPG	ResNet-34 + SSD	PDS; GT: 3 observers	Acc: 78.9% (D), 82.0% (R), 90.2% (A)	No	No
Kwon (2022)	724 OPGs	OPG	CNN + MLP (Concat)	Time; GT: Records	R = 0.83, MAE = 2.95 min	Grad-CAM SHAP	No
Lee (2022)	4,903 OPGs	OPG	RetinaNet (Detection) + ViT (Classification)	Winter/P&G; GT: 2 OMFS	DI Acc: 83.5%, IAN Acc: 81.1%	No	No
Li (2023)	N = 608 OPGs	OPG	ResNet-152 + GAN	IAN contact; GT: 2 OMFS	Acc: 88.7%	No	No
Danjo (2024)	1,095 OPGs	OPG	Faster R-CNN (VGG16)	Winter's; GT: 2 OMFS	Acc: 57.7% (3-class task)	No	No
Khorshidi (2024)	738 CBCT reports	Clinical Text	NLP + MLP	PDS; GT: 2 experts	Acc = 95%, F1 = 0.96 (Train); Acc = 95%, F1 = 0.92 (Val).	No	No
Trachoo (2025)	1,367 M3Ms	OPG	ResNet101V2 +	PDS; GT: 3 OMFS	Acc: 89.8%–95.6%	No	No
Akdoğan (2025)	2,000 OPGs	OPG	YOLO11 (nano-xl)	PDS; GT: 1 radiologist	P: 97.0%, R: 94.6%, F1: 95.8%	No	Yes
Achararit (2025)	1,200 OPGs	OPG (ROI)	CNN (Custom)	PDS; GT: 2 experts	DI Acc: 96.0%	Score-CAM heatmaps	No
Chindanuruks (2025)	1,730 OPGs	OPG	YOLOv5x (CNN object detection)	PDS; GT: Consensus	Acc: 61.0%–89.0%	No	No
Kang (2025)	419 cases	Records	SVM + Lasso	Time; GT: Records	Acc: 80.0%, AUC: 0.85	LASSO trajectory	No
Balel (2025)	3,830 OPGs	OPG	YOLO11	Pederson DI; GT: 1 OMFS	P: 98.0%, R: 94.8%, F1: 97.4%	No	No

Notes: (1) Studies are ordered chronologically by publication year. (2) Performance metrics reflect the results of the best-performing model on the test or validation set as reported in the original studies. (3) External validation indicates evaluation using a dataset from a distinct geographic or institutional source. (4) XAI (Explainable AI) denotes the application of interpretability techniques (e.g., SHAP, Grad-CAM, Score-CAM); GUI indicates the implementation of a clinical decision-support interface.
Abbreviations: OPG = orthopantomogram; CBCT = cone-beam computed tomography; ROI = region of interest; M3M = mandibular third molar; IAN = inferior alveolar nerve; DI = difficulty index; PDS = Pederson difficulty score; GT = ground truth; OMFS = oral and maxillofacial surgeon; CNN = convolutional neural network; YOLO = You Only Look Once; ViT = Vision Transformer; SSD = Single Shot MultiBox Detector; SVM = support vector machine; MLP = multilayer perceptron; GAN = generative adversarial network; SR = super-resolution; NLP = natural language processing; LASSO = least absolute shrinkage and selection operator; LR = logistic regression; KNN = k-nearest neighbors; CV = cross-validation; Acc = accuracy; P = precision; R = recall; AUC = area under the curve; mAP = mean average precision; F1 = F1-score; MAE = mean absolute error; r = Pearson correlation; R^2 = coefficient of determination; ICC = intraclass correlation coefficient; CI = confidence interval; SD = standard deviation; Sens = sensitivity; Spec = specificity; D = depth; R = ramus relationship; A = angulation; VE = vertical eruption; STI = soft tissue impaction; PBI = partial bony impaction; CBI = complete bony impaction.

F. Performance Reporting and Validation Patterns

Performance reporting and validation strategies were heterogeneous. Most studies relied on internal validation (hold-out split or cross-validation), whereas external or multi-center validation was uncommon. In addition, the definition of difficulty varied across studies (indices/grades, operative time, or procedure-step surrogates), limiting direct comparability of model performance. These findings highlight methodological gaps relevant to future translation rather than enabling a pooled estimate of effect. To strengthen reproducibility and clinical translation, future studies should align reporting and appraisal with established guidance for clinical AI and prediction models [12-14].

IV. DISCUSSION

This scoping review highlights the dominance of OPG-based inputs for AI-driven M3M preoperative planning. Our analysis reveals a critical architectural transition: while CNNs are optimal for local feature extraction and segmentation, Transformer-based models leverage self-attention to capture long-range spatial dependencies across the entire radiograph. This global geometric context is essential for assessing complex root-IAN relationships, where surgical risk is determined by overarching anatomical geometry rather than isolated pixel intensities.

Surgical difficulty is operationalized through non-equivalent endpoints, composite indices, operative time, and procedural surrogates that are not interchangeable [6], [13], [14]. PROBAST and CLAIM appraisals reveal that, despite

high clinical applicability, 100% of models face a high risk of bias due to insufficient multi-center validation. A 91.7% medium CLAIM compliance underscores persistent reporting deficits. Following the transparency benchmark of Li et al. (2023) [9], future research must prioritize clearer endpoint definitions and alignment with clinically actionable decisions.

Key research gaps hindering clinical implementation include: limited multi-center external validation; significant outcome heterogeneity; insufficient transparency in reference standard reporting; and a lack of clinical utility metrics (e.g., calibration) or deployment-oriented evaluations. Resolving these deficiencies is essential for achieving reliable, reproducible, and comparable AI-driven surgical decision support [6], [13], [14].

Advancing predictive accuracy requires a transition toward multimodal learning, integrating 2D panoramic radiographs (for broad screening) with 3D CBCT and EHR data to overcome single-modality information loss and enable robust risk stratification. Consequently, we propose the M3M-AI (Mandibular Third Molar AI) Framework to harmonize future research and bridge identified validity gaps.

To bridge identified validity gaps, the M3M-AI Framework mandates: (i) Input Fidelity (reporting resolution, preprocessing, and hardware); (ii) a Core Outcome Set (COS) including radiographic metrics, surgical surrogates (e.g., operative time), and procedural requirements (e.g., ostectomy) [1], [4]; (iii) Reference Rigor (documenting expertise, kappa reliability, and adjudication); (iv) Explainable AI (XAI) using saliency maps or attention weights to align model reasoning with surgical intuition and anatomical landmarks; and (v) Validation Benchmarks via mandatory multi-center external testing.

Focusing on direct clinical utility, the 12 identified studies represent the AI frontier for complex mandibular third molar extractions. While feasibility is established, persistent methodological diversity precludes pooled effectiveness claims. Instead, this review clarifies operational definitions and prioritizes research limitations to guide future standardization.

V. Conclusions

This scoping review of 12 studies identifies OPG as the primary AI input for M3M difficulty prediction, noting a heavy focus on radiographic scores over operative time or IAN-related risk. Critical gaps persist in external validation, endpoint consistency, and reporting transparency [12–14]. To enable clinical translation, future research must prioritize standardized outcomes, rigorous labeling, and multi-center validation to ensure safe and interpretable preoperative decision support.

Acknowledgment

The authors thank the Faculty of Odonto-Stomatology, Can Tho University of Medicine and Pharmacy and Faculty of Dentistry, Van Lang University for institutional support in literature retrieval and academic discussion. No external funding was received for this work.

References

[1] P. Achararit et al., "Impacted lower third molar classification and difficulty index assessment: Comparisons among dental students, general practitioners and deep learning model assistance," BMC Oral Health, vol. 25, no. 1, Art. no. 152, Jan. 2025.

[2] Y. Balel and K. Sağtaş, "Deep learning-based approach to third molar impaction analysis with clinical classifications," Sci. Rep., vol. 15, no. 1, Art. no. 23688, Jul. 2025.

[3] T. Chindanuruks et al., "Development and validation of a deep learning algorithm for the classification of the level of surgical difficulty in impacted mandibular third molar surgery," Int. J. Oral Maxillofac. Surg., vol. 54, no. 5, pp. 452–460, May 2025.

[4] J. H. Yoo et al., "Deep learning based prediction of extraction difficulty for mandibular third molars," Sci. Rep., vol. 11, Art. no. 1954, 2021.

[5] A. Danjo et al., "Limitations of panoramic radiographs in predicting mandibular wisdom tooth extraction and the potential of deep learning models to overcome them," Sci. Rep., vol. 14, no. 1, Art. no. 30806, Dec. 2024.

[6] A. C. Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," Ann. Intern. Med., vol. 169, no. 7, pp. 467–473, 2018.

[7] C. Kang et al., "A data-driven method for surgeon-specific difficulty assessment in third molar extraction," Front. Med. (Lausanne), vol. 7, Art. no. 1654727, Nov. 2025.

[8] F. Khorshidi, R. Esmaeilyfard, and M. Paknahad, "Enhancing predictive analytics in mandibular third molar extraction using artificial intelligence: A CBCT-based study," Saudi Dent. J., vol. 36, no. 12, pp. 1582–1587, Dec. 2024.

[9] W. Li et al., "Transfer learning-based super-resolution in panoramic models for predicting mandibular third molar extraction difficulty: A multi-center study," Medical Data Mining, 2023.

[10] V. Trachoo et al., "Deep learning for predicting the difficulty level of removing the impacted mandibular third molar," Int. Dent. J., vol. 75, no. 1, pp. 144–150, Feb. 2025.

[11] D. Kwon, J. Ahn, C. S. Kim, D. O. Kang, and J. Y. Paeng, "A deep learning model based on concatenation approach to predict the time to extract a mandibular third molar tooth," BMC Oral Health, vol. 22, no. 1, Art. no. 571, Dec. 2022.

[12] S. Akdoğan, M. U. Öziç, and M. Tassoker, "Development of an AI-supported clinical tool for assessing mandibular third molar tooth extraction difficulty using panoramic radiographs and YOLO11 sub-models," Diagnostics, vol. 15, no. 4, Art. no. 462, 2025.

[13] J. Mongan, L. Moy, and C. E. Kahn Jr., "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," Radiology, vol. 295, no. 1, pp. 202–208, 2020.

[14] R. F. Wolff et al., "PROBAST: A tool to assess risk of bias and applicability of prediction model studies," BMJ, vol. 364, Art. no. l1327, 2019.

[15] J. Lee, J. Park, S. Y. Moon, and K. Lee, "Automated prediction of extraction difficulty and inferior alveolar nerve injury for mandibular third molar using a deep neural network," Appl. Sci., vol. 12, no. 1, Art. no. 475, 2022.