



Deep learning for tooth segmentation in cone-beam computed tomography: a systematic review

Aprendizaje profundo para la segmentación de dientes en tomografía computarizada de haz cónico: revisión sistemática

Maria Fernanda Escalante Freitez¹  

¹ Florida International University, USA.

Submitted: 24-01-2025 | Revised: 26-03-2025 | Accepted: 30-06-2025 | Published: 10-07-2025

Corresponding Author: Maria Fernanda Escalante Freitez 

ABSTRACT

Introduction: Tooth segmentation in cone-beam computed tomography is a prerequisite for digital treatment planning, but manual segmentation is slow and subjective. Deep learning has been proposed to automate it. Methods: Following PRISMA 2020, PubMed, OpenAlex, and Crossref were searched for studies published from 2018 to 2025 that evaluated deep learning models for tooth segmentation in cone-beam computed tomography against a reference standard. Risk of bias was assessed with QUADAS-2. Results: Twenty-two studies were included, most using U-Net or nnU-Net architectures. The Dice similarity coefficient ranged from 0.84 for impacted canines to 0.99, with most studies reporting values above 0.90. External validation was reported in a minority of studies. Conclusions: Deep learning segments teeth in cone-beam computed tomography with high Dice coefficients in most studies, but performance varies with tooth type and dataset, and external validation remains uncommon.

Keywords: deep learning; cone-beam computed tomography; tooth; image segmentation; artificial intelligence; systematic review.

RESUMEN

Introducción: La segmentación de dientes en tomografía computarizada de haz cónico es un requisito para la planificación digital del tratamiento, pero la segmentación manual es lenta y subjetiva. Se ha propuesto el aprendizaje profundo para automatizarla. Métodos: Siguiendo PRISMA 2020, se buscó en PubMed, OpenAlex y Crossref estudios publicados entre 2018 y 2025 que evaluaran modelos de aprendizaje profundo para la segmentación de dientes en tomografía computarizada de haz cónico frente a un estándar de referencia. El riesgo de sesgo se evaluó con QUADAS-2. Resultados: Se incluyeron veintidós estudios, la mayoría con arquitecturas U-Net o nnU-Net. El coeficiente de Dice osciló entre 0,84 para caninos impactados y 0,99, y la mayoría de los estudios reportó valores superiores a 0,90. Solo una minoría reportó validación externa. Conclusiones: El aprendizaje profundo segmenta los dientes en tomografía computarizada de haz cónico con coeficientes de Dice altos en la mayoría de los estudios, pero el rendimiento varía según el tipo de diente y el conjunto de datos, y la validación externa sigue siendo poco común.

Palabras clave: aprendizaje profundo; tomografía computarizada de haz cónico; diente; segmentación de imágenes; inteligencia artificial; revisión sistemática.

INTRODUCTION

Cone-beam computed tomography has become a standard imaging modality in implant dentistry, orthodontics, endodontics, and oral surgery, and its three-dimensional data support the digital workflows on which modern treatment planning depends.⁽¹⁻³⁾ The first step of these workflows, the segmentation of individual teeth from the volumetric image, remains a bottleneck, because manual segmentation is slow, operator-dependent, and difficult to reproduce. A single full-arch segmentation can take an experienced operator an hour or more, and the resulting surfaces vary between operators.⁽⁴⁾ The transition to lower-dose and differently configured scanners has added further variability to the input data, which a robust automatic method must accommodate.⁽⁵⁾

The demand for accurate tooth surfaces is growing, not shrinking. Digital workflows for implant planning, aligner design, and orthognathic simulation all consume the tooth surface as their primary input, and the quality of the plan is bounded by the quality of the segmentation that feeds it. The clinical and economic cost of a poor segmentation is therefore realized downstream, in a plan that does not fit, and this downstream dependence is what gives tooth segmentation its outsized importance relative to its apparent simplicity as a task. As these digital workflows spread beyond specialized centers into routine practice, the need for an automatic, reproducible segmentation becomes a prerequisite for scaling, rather than a convenience for the few operators who already have the skill to trace manually.

The appeal of automating tooth segmentation is therefore concrete and direct. A model that delineates each tooth accurately and repeatably would remove the bottleneck, standardize the input to downstream planning, and enable large-scale retrospective analyses that are impractical with manual tracing. Deep learning, and convolutional networks in particular, has been applied to this task with increasing frequency, and the reported performance has risen steadily as architectures and datasets have matured.^(6,7)

The task is also well posed for supervised learning in a way that many clinical imaging tasks are not. The target of segmentation, the tooth surface, is anatomically defined and largely consistent across patients, the input is a standardized volumetric acquisition, and the ground truth, a manual segmentation, can be produced by trained operators with reasonable consistency. These favorable properties lower the barrier to entry, and they help to explain both the number of primary studies and the relatively high and consistent performance they report. These properties make tooth segmentation a natural first application of deep learning in dental imaging, and they explain why it has attracted a comparatively large and mature literature relative to other dental tasks.

There is nonetheless a difficulty that is easy to underestimate. The boundary between the root and the surrounding alveolar bone is radiographically subtle, because the two tissues have similar density and the boundary can be obscured by beam hardening and by the thinness of the periodontal ligament. It is at this root-to-bone interface, and at the contact points between adjacent teeth, that the segmentation error concentrates, and it is the same places that a human operator finds hardest to trace. The models and the humans therefore fail in the same places, which is why a high average Dice can coexist with a clinically meaningful residual error at the boundaries that matter most. This shared failure mode is also why the models are unlikely to dramatically exceed human performance on this task in the near term: the information that would be needed to resolve the ambiguous boundary is, in part, simply absent from the image.

The segmentation task differs in kind from the detection and classification tasks that dominate much of dental artificial intelligence. Segmentation requires the model to assign a label to every voxel of the volume, and its accuracy is summarized by the Dice similarity coefficient and the intersection over union, which measure the overlap between the predicted and reference surfaces. These metrics are well suited to aggregation, which makes tooth segmentation one of the most quantitatively comparable tasks in dental imaging.⁽⁴⁾ The metric is also directly interpretable: a Dice of 0.95 means that the predicted and reference surfaces overlap almost completely, whereas a Dice of 0.80 means that a fifth of the volume is mislabeled, which is enough to matter for a surface used in surgical planning.

The aim of this systematic review was to synthesize the segmentation performance, expressed as the Dice similarity coefficient, of deep learning models for tooth segmentation in cone-beam computed tomography scans, and to characterize the architectures, datasets, and reference standards of the included studies. The review focuses on the Dice because it is the metric that the field reports most consistently and the one that is most directly interpretable, but it also records the intersection over union and the surface distances where reported, so that the overlap and the boundary error can be considered together.

METHODS

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement.⁽⁸⁾

Studies were eligible when they were original investigations that developed or evaluated a deep learning or machine learning model for segmenting individual teeth or the full dentition in cone-beam computed tomography and reported at least one quantitative overlap metric, defined as the Dice similarity coefficient, the intersection over union, a surface distance, or a segmentation accuracy, against a reference standard. Studies whose primary input was a two-dimensional radiograph, an intraoral scan, or a different volumetric modality were excluded, as were reviews, editorials, and studies reporting no quantitative metric. The publication window was 2018 to 2025.

PubMed, OpenAlex, and Crossref were searched for records published between 2018 and 2025. Queries combined terms for deep learning, tooth segmentation, and cone-beam computed tomography, with several synonymous strings run in each source to maximize recall. The PubMed query set comprised three strings: ("deep learning" OR "convolutional neural network") AND ("tooth segmentation" OR "teeth segmentation") AND ("cone-beam computed tomography" OR CBCT OR "cone beam"); "deep learning" AND tooth AND segmentation AND CBCT AND (Dice OR IoU OR Jaccard OR "segmentation performance"); and "automatic" AND "tooth segmentation" AND ("cone beam" OR CBCT) AND "deep learning". Equivalent free-text queries were adapted to the OpenAlex and Crossref syntax, and each source returned up to 50 records per query. Records were deduplicated by digital object identifier and by title before screening.

Screening proceeded in two stages. Title and abstract screening excluded clearly irrelevant records, and uncertain records were carried forward rather than discarded. The title and abstract screen applied the eligibility criteria conservatively, so that any record whose relevance could not be determined was retained for full-text assessment. Full text was available for a subset of reports; for the remainder, eligibility and data extraction relied on the abstract, which reports the key overlap metrics in most segmentation studies.

For each included study the following items were extracted: first author and year, country, model architecture, sample size, reference standard, and reported segmentation performance, including the Dice similarity coefficient, the intersection over union, and, where reported, the Hausdorff distance and the average symmetric surface distance. Values were recorded as reported, without recalculation, and when a study reported separate values for internal and external test sets or for different tooth types, all values were retained. When a study reported a segmentation accuracy or a surface deviation rather than an overlap coefficient, the value was recorded with its unit, and when a study reported no extractable numeric metric, it was excluded at the eligibility stage rather than included with missing data.

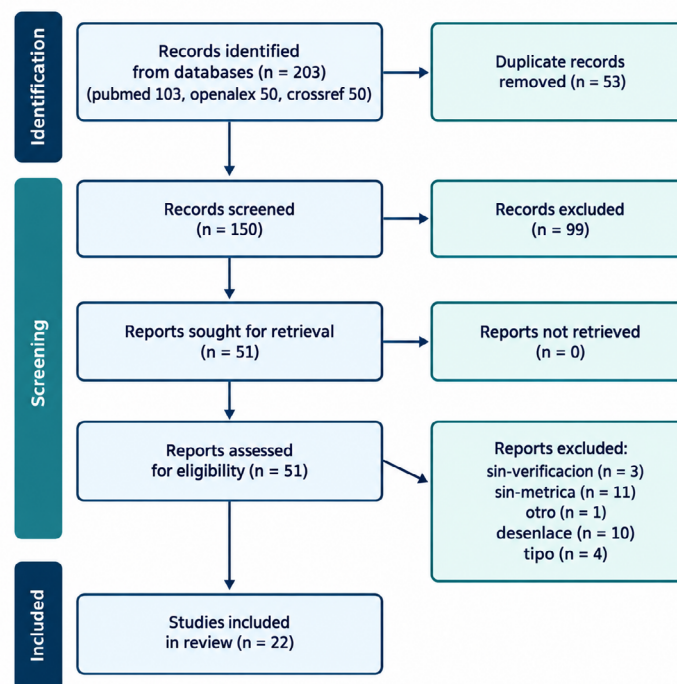
Risk of bias and applicability were assessed with QUADAS-2 across patient selection, the index test, the reference standard, and flow and timing.⁽⁹⁾ Because most studies reported no prospective registration and no independent external validation, risk of bias was summarized narratively by domain. In the patient selection domain, the concern is that most studies used a single-institution retrospective cohort, which limits applicability to other scanners and populations. In the index test domain, the concern is minimal, because the model is deterministic and the preprocessing is usually described, but augmentation and hyperparameter choices were reported inconsistently. In the reference standard domain, the concern is that manual segmentation is imperfect and its variability was not quantified. In the flow and timing domain, the concern is that the test set was not always demonstrably independent of the training distribution.

RESULTS

Study selection

The searches retrieved 203 records, of which 53 duplicates were removed and 150 unique records were screened (Figure 1). Title and abstract screening excluded 99 records, leaving 51 reports assessed for eligibility. Of these, 29 were excluded because the outcome was not tooth segmentation or because no quantitative metric was reported. Twenty-two studies were included⁽²²⁾.

Figure 1. PRISMA 2020 flow diagram of study selection.



Note: Records were excluded at eligibility for a non-tooth-segmentation outcome, no extractable overlap metric, a non-CBCT modality, a non-original design, a retraction, a duplicate, or an unverifiable reference. Source: Elaborated from the systematic search of PubMed, OpenAlex, and Crossref.

Characteristics of included studies

The included studies spanned 2021 to 2025, with the largest number published in 2025 (7 studies) and 2024 (5 studies) (Table 1). The absence of studies before 2021 reflects the fact that the deep learning era for this task began only after the first large annotated datasets became available, and the rapid growth since then tracks the broader expansion of dental artificial intelligence. The geographic distribution was concentrated in East Asia and Europe, with a smaller number of studies from elsewhere, and many studies did not report the country of origin, which limits any assessment of how the models transfer across populations and scanner conventions. The U-Net family and its self-configuring variant nnU-Net were the dominant architectures, appearing in a large proportion of the studies.(10-14)Transformer-based and hybrid architectures appeared in the most recent reports,(15-17)and a smaller group used multi-stage or hierarchical networks.(18-21) Several studies evaluated commercially available or cloud-based software, including Diagnocat and DentalSegmentator,(11,22-24)which is notable because it means the results are not limited to research prototypes but reflect tools a clinician could already use.

The reference standard was manual segmentation by expert operators in nearly all studies, and a few used a physical reference from an intraoral scan of extracted teeth, which is the strongest available criterion because it does not depend on another human operator.(25)Sample sizes ranged from 20 CBCT scans(22)to several hundred,(10,13,15)with one study using a large set of virtually generated training volumes, which addresses the scarcity of annotated data without requiring manual labeling.(26)Only a minority reported validation on an independent external cohort.(10,11)

The studies also differed in the scope of the segmentation. Some segmented the full dentition, some a single arch, and some a single tooth or a single tooth type such as the impacted canine.(12,23) A few extended the segmentation beyond the teeth to the alveolar bone, the maxillary sinus, and the mandibular canal, which places the tooth segmentation within a broader digital planning pipeline.(19,27) The scope of the task, whether full-arch, single-tooth, or multi-structure, affects the difficulty and the reported metric, and it is a further source of the variation observed across studies.

Table 1. Characteristics of the 22 included studies evaluating deep learning for tooth segmentation in cone-beam computed tomography.

First author, year	Country	Architecture	Sample	Reference standard	Notes
ElShebiny 2025	NR	multiclass deep learning	210 CBCT (140/40/30)	ground truth labels	segmentation accuracy 93.16 +/- 1.18%
Elsonbaty 2025	NR	CNN cloud platform	37 CBCT (402 primary teeth)	manual segmentation	accuracy 98%; ICC 0.99
Ntovas 2025	NR	AI tools (CDX, BSB)	67 extracted teeth (14 CBCT)	intraoral scanner	RMS 0.27 mm
Sinard 2025	NR	4 DL solutions (Diagnocat, DentalSegmentator, others)	20 CBCT	expert semi-automatic	
Unal 2025	NR	nnU-Net v2	159 CBCT	manual	impacted canines; 95% HD 7.07 mm
Zhou 2025	China	3D TransUNet + nnU-Netv2	450 CBCT	manual annotation	ASSD 0.501 mm
Zhou 2025	Hong Kong	UNetR (Swin + Mamba)	NR	NR	instance segmentation
Dot 2024	NR	nnU-Net (DentalSegmentator)	470 train; 133 + 123 test	manual	
Hu 2024	NR	nnU-Net and U-Net	336 train / 120 test	manual	mixed dentition; ASSD 0.091 mm
Liu 2024	NR	multi-stage CNN	NR	manual	also bone, sinus, canal
Swaity 2024	Belgium	CNN cloud platform	100 CBCT	manual	impacted maxillary canine
Takeya 2024	NR	deep learning	>170,000 virtual CBCT	manual contouring	
Alqahtani 2023	NR	multiclass CNN	215 CBCT (1780 teeth)	manual	teeth with brackets
Chen 2023	NR	CTA-UNet (CNN-Transformer)	200 patients (45 annotated)	manual	HD95 0.525 mm; ASSD 0.199 mm
Deng 2023	NR	SkullEngine (deep learning)	61 CBCT	manual	ASSD 0.1 mm
Fontenele 2022	NR	CNN (AI-driven)	175 train / 74 test CBCT	manual	effect of restorations
Jang 2022	South Korea	hierarchical multi-step CNN	NR	NR	instance segmentation + FDI
Nogueira-Reis 2022	NR	three integrated CNNs	40 CBCT	manual	HD 0.045 mm
Duan 2021	NR	U-Net (two-phase)	NR	manual	single-rooted / multi-rooted
Lahoud 2021	Belgium	feature pyramid network	433 DICOM (314 CBCT)	manual	
Shaheen 2021	NR	3D U-Net	186 CBCT	manual	precision 0.98; recall 0.83; 95% HD 0.56 mm
Wang 2021	NR	mixed-scale dense CNN	30 CBCT	4 dentists	

Note: NR = not reported; CBCT = cone-beam computed tomography; CNN = convolutional neural network. Source: Compiled from data extracted from the included studies.

A subset of studies addressed instance segmentation, in which each tooth is separated from its neighbors and individually identified, which is a harder problem than binary tooth-versus-background segmentation and a prerequisite for the tooth-level planning that most downstream workflows require.(16,18) Instance segmentation adds the problem of distinguishing adjacent teeth that touch at the contact points, and the models that solved it did so with a hierarchical approach that first separates the teeth and then refines each boundary.(18)The contact points are difficult because the interproximal space can be sub-voxel in width, so that a single voxel must be assigned to one tooth or the other with no visual evidence to decide the boundary. The distinction between semantic and instance segmentation is important, because a high Dice on the tooth region as a whole does not imply that the individual teeth are correctly separated, and the studies that reported instance-level metrics are the ones most directly relevant to clinical use.

The instance segmentation studies also reported tooth identification, in which the model assigns the correct tooth number, and this identification is what connects the segmented surface to the clinical record and to the treatment plan.(18,28) A segmentation that is geometrically accurate but assigned to the wrong tooth is clinically worse than a slightly inaccurate one assigned correctly, because it would direct the plan to the wrong site. The studies that reported both the segmentation accuracy and the identification accuracy therefore provide the most complete picture of clinical readiness, and the high identification accuracies reported, near or above 95 percent in several studies, are as important as the high Dice coefficients.

Risk of bias

Across QUADAS-2 domains, the reference standard was the principal source of concern because it was manual segmentation in nearly all studies, and the inter-operator variability of the reference was not quantified in the included studies, which means the model is evaluated against an imperfect target. Patient and image selection raised further concerns because cohorts were assembled retrospectively from single institutions in most studies, and few described consecutive or random sampling. Flow and timing were generally acceptable, but the independence of the test set from the training distribution was not demonstrated in most studies, and only a minority reported external validation. The overall risk of bias was therefore moderate, and the applicability of the benchmark-based estimates to routine clinical images is uncertain.

Segmentation performance

Eighteen studies reported a Dice similarity coefficient (Table 2). The reported values ranged from 0.84 for the segmentation of impacted canines(12) to 0.99 for a multiclass network applied to teeth with brackets(28) and for a system of three integrated networks.(21) Most studies reported Dice coefficients above 0.90.(10,11,14-16,19,20,23,24,27) One study reported a segmentation accuracy of 93.16% rather than an overlap coefficient.(29) The intersection over union, reported in six studies, ranged from 0.74 for impacted canines(12) to 0.99 for teeth with brackets,(28) and a feature pyramid network reported intersection over union of 0.87 to 0.88 across two evaluation settings.(30)

The tooth type conditioned the performance. Impacted canines, which are surrounded by bone and overlap adjacent teeth, were segmented less accurately than erupted teeth, with a Dice of 0.84(12) and a Dice of 0.99 in a study that focused specifically on the impacted maxillary canine but with a larger and more curated dataset.(23) Teeth with restorations showed slightly lower Dice than unrestored teeth, with values of 0.95 to 0.97 for restored teeth compared with 0.98 to 0.99 for control teeth, a difference that the authors attributed to the radiopaque artifact introduced by the filling material.(31) Single-rooted and multi-rooted teeth were segmented with similar accuracy in one study.(14)

The dataset and the evaluation setting also conditioned the performance. A study that reported both internal and external validation found a Dice of 0.964 internally and 0.951 externally, a measurable but modest drop,(10) and a benchmark tool evaluated on both internal and external datasets reported a Dice of 92.2% internally and 94.2% externally,(11) which suggests that the external gap is not always in the same direction. The two external-validation studies are too few to generalize from, but together they indicate that the models are robust to a change of institution and scanner, which is the property that matters most for clinical adoption. The studies that evaluated commercial or cloud software reported Dice coefficients from 88.5% to 95.6% across four solutions,(22) which is within the range of the research models but not at the top of it. A cloud platform applied to primary teeth reported a Dice of 95% with an accuracy of 98% and an intra-class correlation of 0.99 against the manual reference,(24) which indicates that the tools transfer to the pediatric dentition, where the smaller tooth size and the presence of developing roots add difficulty. Several studies reported surface distances rather than overlap, with average symmetric surface distances from 0.045 mm(21) to 0.5 mm,(15,17) and Hausdorff distances that ranged from 0.56 mm(13) to 7.07 mm for the difficult impacted canine.(12) The Hausdorff distance is a worst-case measure that is dominated by a single outlier point, so its wide spread is expected even when the average surface distance is small, and the two measures together indicate that the typical error is small but that occasional large deviations occur.

Table 2. Reported segmentation performance of deep learning models for tooth segmentation in cone-beam computed tomography.

First author, year	Dice	IoU
ElShebiny 2025	NR	NR
Elsonbaty 2025	95 +/- 2%	NR
Ntovas 2025	NR	NR
Sinard 2025	88.5-95.6%	NR
Unal 2025	0.84	0.74
Zhou 2025	0.941	NR
Zhou 2025	0.832 +/- 0.031	NR
Dot 2024	92.2% / 94.2%	NR
Hu 2024	0.964 internal / 0.951 external	0.931 / 0.921
Liu 2024	96.5%	NR
Swaity 2024	0.99 +/- 0.02	NR
Takeya 2024	>0.87	NR
Alqahtani 2023	0.99	0.99
Chen 2023	87.12%	78.90%
Deng 2023	96%	NR
Fontenele 2022	0.95-0.99	0.91-0.99
Jang 2022	94.79%	NR
Nogueira-Reis 2022	99.3%	NR
Duan 2021	95.7% / 96.2%	NR
Lahoud 2021	NR	0.87 / 0.88
Shaheen 2021	NR	NR
Wang 2021	0.945 +/- 0.021	NR

Note: Dice = Dice similarity coefficient; IoU = intersection over union; HD = Hausdorff distance; ASSD = average symmetric surface distance.
Source: Compiled from data extracted from the included studies.

The overlap metrics and the surface distances are complementary. A high Dice indicates that the predicted and reference volumes overlap extensively, whereas a small average symmetric surface distance indicates that the predicted surface lies close to the reference surface everywhere, and a large Hausdorff distance indicates a localized outlier. A model can have a high Dice and a small average distance yet a large Hausdorff distance, which would point to a single mis-segmented region, and this pattern is visible in the studies that reported all three.(12,17) Reporting the full set is therefore more informative than reporting the Dice alone, and the field would benefit from a convention that includes at least one surface distance alongside the overlap coefficient.

The studies that reported a physical or an intraoral-scan reference deserve particular weight, because they evaluate the model against a criterion that is independent of any human operator. One such study, using extracted teeth scanned with an intraoral scanner as the ground truth, reported a root-mean-square surface deviation of 0.27 mm for the automatic segmentation.(25) This value, obtained against a physical rather than a manual reference, corroborates the high overlap metrics and provides the most convincing evidence that the models are not simply reproducing the biases of the manual annotation they were trained on. The scarcity of such independent references is itself a gap, because most studies evaluate against the same kind of manual segmentation they were trained to reproduce.

The architectures did not show a clear winner. The U-Net and nnU-Net models, the transformer hybrids, and the multi-stage hierarchical networks all converged on Dice coefficients in the same 0.9 to 0.99 range when applied to comparable datasets.(10,15,16,18,20) This convergence suggests that the architecture is not the limiting factor, and that the remaining error is driven by the data and by the inherent difficulty of the tooth boundaries, such as the thin root apex and the low-contrast root-to-bone interface. The nnU-Net self-configuring framework appeared in several of the studies and reported values near the top of the range, which reflects its advantage in automating the configuration choices that otherwise differ across studies and confound comparison.

The transformer-based and the nnU-Net models reported the highest values, but the differences are small enough that they may reflect dataset rather than architectural advantage, and a head-to-head comparison on a common benchmark would be needed to separate the two.

The temporal trend is likewise informative. The earliest included study, from 2021, reported a Dice around 0.94,(20) and the most recent studies report values in the same high range, with the gains since then coming from better datasets and from external validation rather than from a step change in the Dice itself.(10,11,15) The field has therefore reached a plateau in headline performance, and the frontier has moved from achieving a high Dice to demonstrating that the high Dice holds on unseen data and on the difficult tooth types, which is the more relevant question for clinical adoption.

DISCUSSION

The answer to the question posed by this review is that deep learning segments teeth in cone-beam computed tomography with a Dice similarity coefficient above 0.90 in most studies, and that the performance is therefore high enough to support the downstream digital workflows that depend on it. This is a clear and consistent finding across architectures and datasets, and it is the central result on which the recommendations of this review rest, and on which the clinical adoption of the technology can ultimately, and reasonably, be justified in routine clinical practice. The caveats are that performance varies with tooth type, with restored teeth and impacted canines being harder, and that external validation remains uncommon.

The high Dice coefficients carry a further implication that is easy to overlook. A Dice above 0.90 means the model has already surpassed the level of surface accuracy that most downstream applications require, because the planning software that consumes the segmentation usually smooths and simplifies the surface before use. The remaining error, concentrated at the root apex and at the restored and impacted teeth, is therefore unlikely to change the treatment plan in most cases, but it can change it in the cases that are already the most demanding. This asymmetry, high average performance with concentrated residual risk, is the essential character of the evidence reviewed here.

The comparison with the human operator is the standard against which the models should ultimately be judged, and the included studies that made this comparison found that the model error was comparable to or smaller than the inter-operator variability of the reference.(24,25) This is the more meaningful benchmark than the Dice alone, because the clinical question is not whether the model is perfect but whether it is no worse than the human process it replaces. By that standard, the evidence is reassuring for the routine case and unresolved for the difficult case, which is the same conclusion reached from the Dice alone.

The time savings, where reported, reinforce the case for adoption. The studies that measured the segmentation time found that the automatic method reduced it from the tens of minutes required for manual tracing to a few minutes of automated computation plus a short review, and this time reduction, more than the small gain in accuracy, is the primary economic argument for the technology. A tool that is no more accurate than a human but many times faster still changes the workflow, and it is this speed, combined with the acceptable accuracy, that makes automatic tooth segmentation an attractive proposition for practices that perform digital planning routinely.

The integration of these tools into clinical software is already underway, and the included studies that evaluated commercial and cloud platforms show that the technology has moved beyond the research laboratory.(11,22,23) This commercialization is itself evidence of maturity, but it also raises the bar for validation, because a tool sold for clinical use should be held to a higher standard of external, prospective evidence than a research prototype. The regulatory and integration questions that attend any clinical artificial intelligence system apply here in full, and they are the natural next subject for the literature to address.(32)

These findings align with the trajectory of the wider literature. A fully automatic system for tooth and alveolar bone segmentation reported strong performance and has become a widely cited reference point for the field,(33) and an earlier automated approach for implant planning reported the feasibility of the task.(34,35) Reviews of adjacent segmentation tasks in cone-beam computed tomography, such as alveolar bone segmentation, report similarly high Dice coefficients and the same recurring limitation of scarce external validation.(36,37) The present review confirms that tooth segmentation is among the most mature applications of deep learning in dental imaging.(32)

The contrast with the pre-deep-learning methods is instructive. Before convolutional networks, tooth segmentation relied on level sets, region growing, and statistical shape models,

which required careful initialization and struggled with the low contrast between the root and the surrounding bone.(6)A level-set method applied to the same task illustrates the earlier ceiling, and the deep learning methods reviewed here exceed it by a substantial margin on the standard overlap metrics.(38)The performance gain is therefore attributable to the learned representations rather than to any change in the imaging or in the clinical definition of the task, which explains the rapid adoption of deep learning for this purpose.(7)

This historical contrast matters for how the results should be read. The field has already crossed the threshold at which the model outperforms the earlier automated methods and approaches the reliability of a human operator, and the remaining question is therefore not whether deep learning works for tooth segmentation, but whether its performance transfers to the diverse clinical settings in which it will be used. The answer to that question, which requires external validation, is the main uncertainty that this review identifies and that future work should address. The answer will determine whether the technology remains a research tool that supports a human operator, or becomes a routine component of the digital dental workflow.

The heterogeneity among studies has identifiable sources. First, the tooth type and the clinical scenario differed, and a full-arch segmentation of a healthy dentition is easier than a single impacted tooth or a dentition with restorations. Second, the dataset and its annotation differed, and the studies that used large or virtually augmented training sets reported the highest values. (15,26) Third, the metric differed, because some studies reported the Dice coefficient, others the intersection over union, and still others surface distances, which are not interchangeable. Fourth, the reference standard differed, with most studies using manual segmentation and a few using a physical reference, and the inter-operator variability of the manual reference was not quantified in the included studies. Fifth, the scope of the segmentation differed, from a single tooth to the full dentition plus adjacent structures, and the scope changes both the difficulty and the reported metric.

These sources of heterogeneity are not merely statistical nuisances; they carry clinical meaning. A full-arch segmentation and a single impacted canine are different tasks with different failure modes, and a clinician needs to know which task a given accuracy figure refers to before relying on it. The studies that reported their results stratified by tooth type and by restoration status were the most informative, because they revealed where the residual error concentrates, and future studies should adopt this practice as a matter of course.

The clinical consequence is that a pooled Dice conceals the cases where segmentation is difficult. A clinician planning an implant or an orthodontic movement needs an accurate surface of the specific tooth in question, and the impacted canine, which is precisely the tooth that most needs three-dimensional localization for surgical planning, is also the tooth that the models segment least reliably.(12)This inverse relationship between clinical need and model performance is the most practically important finding of the review, and it argues for validating these tools on the specific clinical scenario in which they will be used.

The filling artifact is a second practical case worth isolating. Restorations produce radiopaque streaks that degrade the local contrast, and the two studies by the same group that examined this directly found a consistent, if modest, drop in the Dice for restored teeth.(31,39) This matters because the population that most needs digital planning, adults with a restored dentition, is also the population in which the segmentation is slightly less accurate. The effect is small enough that the tools remain usable, but large enough that a clinician should expect more manual correction on restored teeth. The finding also points to a general principle: the imaging artifacts that degrade human interpretation also degrade the model, because the model learns from the same pixel intensities, and any improvement in the imaging or in the artifact reduction pipeline is therefore likely to improve the segmentation as well.(40)

For research, the most informative next studies would validate models prospectively across multiple centers and devices, report the Dice by tooth type and by restoration status, and quantify the inter-operator variability of the reference so that the model error can be compared against the human variability it is meant to replace. A benchmark dataset with a public, expert-validated reference standard would also help, because it would allow studies to be compared directly rather than through the heterogeneous private datasets on which they currently rely. The adoption of deep learning in cone-beam computed tomography workflows depends on evidence of this kind, and on the resolution of the broader integration and regulatory questions that attend any clinical artificial intelligence system.(32,40)

The strengths of this review are the systematic search across three bibliographic databases, the explicit 2018 to 2025 window, the recording of values as reported without recalculation, and the separation of overlap metrics from surface distances. The limitations mirror those of the

primary studies. Full text was available for only a subset of reports, so some data were extracted from abstracts. Screening and extraction were performed by a single author with artificial intelligence assistance rather than two independent reviewers, which raises the risk of selection error and warrants a dual-reviewer design in future updates. A meta-analysis was not possible given the heterogeneous architectures, datasets, and tooth types, and publication bias may inflate the observed performance. Finally, the review did not assess the time efficiency of the tools, which is the other half of their clinical value, because the studies reported it inconsistently.

For clinical practice, the findings support using automatic tooth segmentation as a first pass that a clinician reviews and corrects, particularly for routine full-arch segmentations where the models are most reliable. The review then makes three recommendations. First, validate the tool on the specific clinical scenario, because a tool that segments a healthy full arch well may segment an impacted canine poorly. Second, inspect the restored teeth and the impacted teeth, where the residual error concentrates, and correct them manually. Third, treat the reported Dice as a lower bound on the clinical workload reduction, because the cases that need correction are precisely the ones where the error is largest.

For research and reporting, the field would benefit from a shared benchmark and a shared convention. A public, expert-validated reference dataset for tooth segmentation would allow direct comparison across studies, which is currently impossible across the heterogeneous private datasets, and a reporting convention that includes the Dice, the intersection over union, and at least one surface distance, stratified by tooth type and restoration status, would make the results interpretable. Progress on these two points would convert a large but heterogeneous literature into an evidence base on which clinical adoption can be justified, and would allow a future meta-analysis at the level of individual tooth types, which is where the clinical decisions reside.

Taken together, the evidence supports a clear conclusion. Deep learning tooth segmentation in cone-beam computed tomography is mature enough to serve as a reviewed first pass in routine digital workflows, and it is not yet validated well enough to be trusted without supervision on the specific difficult cases, the impacted and the restored teeth, where the clinical need is greatest. The distance between these two statements is a matter of validation rather than of capability, and it is the distance that the next generation of studies should close.

For the clinician, the actionable message is to adopt the technology as an assistant, to review and correct the impacted and the restored teeth in particular, and to validate the tool on the local scanner and population before relying on its output for surgical planning. The evidence reviewed here supports that adoption, and it sets a clear and specific expectation about where the residual risk lies. The field has reached the point where the remaining question is not whether the technology works, but whether it works for the specific tooth, the specific scanner, and the specific patient in front of the clinician.

For clinical practice, the findings support using automatic tooth segmentation as a first pass that a clinician reviews and corrects, particularly for routine full-arch segmentations where the models are most reliable. For research, the most informative next studies would validate models prospectively across multiple centers and devices, report the Dice by tooth type, and quantify the inter-operator variability of the reference so that model error can be compared against it. The adoption of deep learning in cone-beam computed tomography workflows depends on evidence of this kind, and on the resolution of the broader integration and regulatory questions that attend any clinical artificial intelligence system.(32,40)

CONCLUSIONS

Deep learning segments teeth in cone-beam computed tomography with a Dice similarity coefficient above 0.90 in most studies, which is sufficient for the downstream digital workflows that depend on segmentation. Performance varies with tooth type, with impacted canines and restored teeth being harder, and with the dataset, and external validation remains uncommon. The reference standard was manual segmentation in nearly all studies, and risk of bias was moderate. Automatic segmentation should therefore be used as a reviewed first pass, and prospective external validation with per-tooth reporting is required before unattended use in treatment planning can be recommended. The residual error concentrates where the clinical stakes are highest, at the impacted and the restored teeth, and this is the specific limitation that future validation must address. The architectures have largely converged on equivalent performance, so further progress will come from better data, independent references, and external validation rather than from any single architectural advance.

REFERENCES

1. Agrawal P. Applications of Cone Beam Computed Tomography in Orthodontics: An Overview. *ACT*. 2020;5(4). doi: 10.23880/act-16000198
2. Riyaz MA, Al-Ruthea MS. Cone Beam Computed Tomography in Prosthodontics. *IJCMR*. 2019;6(1). doi: 10.21276/ijcmr.2019.6.1.21
3. N M. Applications of Cone Beam Computed Tomography in Endodontics: A Review. *Dentistry*. 2014;04(07). doi: 10.4172/2161-1122.1000242
4. Qiu B, van der Wel H, Kraeima J, Glas HH, Guo J, Borra RJH, et al. Automatic Segmentation of Mandible from Conventional Methods to Deep Learning—A Review. *JPM*. 2021;11(7):629. doi: 10.3390/jpm11070629
5. Kallio-Pulkkinen S, Haapea M, Liukkonen E, Huuonen S, Tervonen O, Nieminen MT. Comparison between DICOM-calibrated and uncalibrated consumer grade and 6-MP displays under different lighting conditions in panoramic radiography. *Dentomaxillofacial Radiology*. 2015;44(5):20140365. doi: 10.1259/dmfr.20140365
6. Pei Y, Ai X, Zha H, Xu T, Ma G. 3D exemplar-based random walks for tooth segmentation from cone-beam computed tomography images. *Medical Physics*. 2016;43(9):5040-5050. doi: 10.1118/1.4960364
7. Dou W, Gao S, Mao D, Dai H, Zhang C, Zhou Y. Tooth instance segmentation based on capturing dependencies and receptive field adjustment in cone beam computed tomography. *Computer Animation & Virtual*. 2022;33(5):e2100. doi: 10.1002/cav.2100
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;n71. doi: 10.1136/bmj.n71
9. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009
10. Hu Y, Liu C, Liu W, Xiong Y, Zeng W, Chen J, et al. Fully automated method for three-dimensional segmentation and fine classification of mixed dentition in cone-beam computed tomography using deep learning. *Journal of Dentistry*. 2024;151:105398. doi: 10.1016/j.jdent.2024.105398
11. Dot G, Chaurasia A, Dubois C, Savoldelli C, Haghighat S, Azimian S, et al. DentalSegmentator: Robust open source deep learning-based CT and CBCT image segmentation. *Journal of Dentistry*. 2024;147:105130. doi: 10.1016/j.jdent.2024.105130
12. Ünal T, Kuran A, Gulsen IT, Kızılay FN, Gulsen E, Özüdoğru S, et al. Deep learning-based 3D automatic segmentation of impacted canines in CBCT scans. *BMC Oral Health*. 2025;25(1):1927. doi: 10.1186/s12903-025-07117-5
13. Shaheen E, Leite A, Alqahtani KA, Smolders A, Van Gerven A, Willems H, et al. A novel deep learning system for multi-class tooth segmentation and classification on cone beam computed tomography. A validation study. *Journal of Dentistry*. 2021;115:103865. doi: 10.1016/j.jdent.2021.103865
14. Duan W, Chen Y, Zhang Q, Lin X, Yang X. Refined tooth and pulp segmentation using U-Net in CBCT image. *Dentomaxillofacial Radiology*. 2021;50(6):20200251. doi: 10.1259/dmfr.20200251
15. Zhou W, Zhao D, Zhao B, Shen M, Liu X. A deep learning framework for automated dental segmentation and diagnostic report generation from cone-beam computed tomography. *Head Face Med*. 2025;21(1):80. doi: 10.1186/s13005-025-00555-0
16. Zhou Y, Xu Y, Khalil B, Nalley A, Tarce M. An open deep learning-based framework and model for tooth instance segmentation in dental CBCT. *Clin Oral Invest*. 2025;29(10):473. doi: 10.1007/s00784-025-06578-w
17. Chen Z, Chen S, Hu F. CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation. *Phys Med Biol*. 2023;68(17):175042. doi: 10.1088/1361-6560/acf026
18. Jang TJ, Kim KC, Cho HC, Seo JK. A Fully Automated Method for 3D Individual Tooth Identification and Segmentation in Dental CBCT. *IEEE Trans Pattern Anal Mach Intell*. 2022;44(10):6562-6568. doi: 10.1109/tpami.2021.3086072
19. Liu Y, Xie R, Wang L, Liu H, Liu C, Zhao Y, et al. Fully automatic AI segmentation of oral surgery-related tissues based on cone beam computed tomography images. *Int J Oral Sci*. 2024;16(1):34. doi: 10.1038/s41368-024-00294-z
20. Wang H, Minnema J, Batenburg KJ, Forouzanfar T, Hu FJ, Wu G. Multiclass CBCT Image Segmentation for Orthodontics with Deep Learning. *J Dent Res*. 2021;100(9):943-949. doi: 10.1177/00220345211005338
21. Nogueira-Reis F, Morgan N, Nomidis S, Van Gerven A, Oliveira-Santos N, Jacobs R, et al. Three-dimensional maxillary virtual patient creation by convolutional neural network-based segmentation on cone-beam computed tomography images. *Clin Oral Invest*. 2022;27(3):1133-1141. doi: 10.1007/s00784-022-04708-2
22. Sinard E, Gajny L, de La Dure-Molla M, Felizardo R, Dot G. Automated Cone Beam Computed Tomography Segmentation of Multiple Impacted Teeth With or Without Association to Rare Diseases: Evaluation of Four Deep Learning-Based Methods. *Orthod Craniofac Res*. 2025;28(3):433-440. doi: 10.1111/ocr.12890
23. Swait A, Elgarba BM, Morgan N, Ali S, Shujaat S, Borsci E, et al. Deep learning driven segmentation of maxillary impacted canine on cone beam computed tomography images. *Sci Rep*. 2024;14(1):369. doi: 10.1038/s41598-023-49613-0
24. Elsonbaty S, Elgarba BM, Fontenele RC, Swait A, Jacobs R. Novel AI-based tool for primary tooth segmentation on CBCT using convolutional neural networks: A validation study. *Int J Paed Dentistry*. 2025;35(1):97-107. doi: 10.1111/ipd.13204
25. Ntovas P, Sirirattanagool P, Asavanamuang P, Jain S, Tavelli L, Revilla-León M, et al. Accuracy and Time Efficiency of Artificial Intelligence-Driven Tooth Segmentation on CBCT Images: A Validation Study Using Two Implant Planning Software Programs. *Clinical Oral Implants Res*. 2025;36(10):1312-1323. doi: 10.1111/clr.70003

26. Takeya A, Watanabe K, Haga A. Fine structural human phantom in dentistry and instance tooth segmentation. *Sci Rep*. 2024;14(1):12630. doi: 10.1038/s41598-024-63319-x
27. Deng HH, Liu Q, Chen A, Kuang T, Yuan P, Gateno J, et al. Clinical feasibility of deep learning-based automatic head CBCT image segmentation and landmark detection in computer-aided surgical simulation for orthognathic surgery. *International Journal of Oral and Maxillofacial Surgery*. 2023;52(7):793-800. doi: 10.1016/j.ijom.2022.10.010
28. Ayidh Alqahtani K, Jacobs R, Smolders A, Van Gerven A, Willems H, Shujaat S, et al. Deep convolutional neural network-based automated segmentation and classification of teeth with orthodontic brackets on cone-beam computed-tomographic images: a validation study. *Eur J Orthod*. 2023;45(2):169-174. doi: 10.1093/ejo/cjac047
29. ElShebiny T, Abdelrauof D, Elattar M, Motro M, Retrouvey JM, El-Dawlatly M, et al. A deep learning model for multiclass tooth segmentation on cone-beam computed tomography scans. *American Journal of Orthodontics and Dentofacial Orthopedics*. 2025;168(2):191-198. doi: 10.1016/j.ajodo.2025.02.014
30. Lahoud P, EzEldeen M, Beznik T, Willems H, Leite A, Van Gerven A, et al. Artificial Intelligence for Fast and Accurate 3-Dimensional Tooth Segmentation on Cone-beam Computed Tomography. *Journal of Endodontics*. 2021;47(5):827-835. doi: 10.1016/j.joen.2020.12.020
31. Fontenele RC, Gerhardt MDN, Pinto JC, Van Gerven A, Willems H, Jacobs R, et al. Influence of dental fillings and tooth type on the performance of a novel artificial intelligence-driven tool for automatic tooth segmentation on CBCT images – A validation study. *Journal of Dentistry*. 2022;119:104069. doi: 10.1016/j.jdent.2022.104069
32. Liu TY, Lee KH, Mukundan A, Karmakar R, Dhiman H, Wang HC. AI in Dentistry: Innovations, Ethical Considerations, and Integration Barriers. *Bioengineering*. 2025;12(9):928. doi: 10.3390/bioengineering12090928
33. Cui Z, Fang Y, Mei L, Zhang B, Yu B, Liu J, et al. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nat Commun*. 2022;13(1):2096. doi: 10.1038/s41467-022-29637-2
34. Lee S, Woo S, Yu J, Seo J, Lee J, Lee C. Automated CNN-Based Tooth Segmentation in Cone-Beam CT for Dental Implant Planning. *IEEE Access*. 2020;8:50507-50518. doi: 10.1109/access.2020.2975826
35. Kurt Bayrakdar S, Orhan K, Bayrakdar IS, Bilgir E, Ezhov M, Gusarev M, et al. A deep learning approach for dental implant planning in cone-beam computed tomography images. *BMC Med Imaging*. 2021;21(1):86. doi: 10.1186/s12880-021-00618-z
36. Mohammed M, Fernandez-Medina T, Rajashekhar M, Baker S, Jennings E. Alveolar Bone Segmentation Methods in Assessing the Effectiveness of Periodontal Defect Regeneration Through Machine Learning of CBCT Data: A Systematic Review. *International Journal of Biomedical Imaging*. 2025;2025(1):9065572. doi: 10.1155/ijbi/9065572
37. Alaqla A, Khanagar SB, Albelaihi AI, Singh OG, Alfadley A. Application and performance of artificial intelligence-based models in the detection, segmentation and classification of periapical lesions: a systematic review. *Front Dent Med*. 2025;6:1717343. doi: 10.3389/fdmed.2025.1717343
38. Syuhada F, Rarasmaya Indraswari, Agus Zainal Arifin, Dini Adni Navastara. Multi-Projection Segmentation on Dental Cone Beam Computed Tomography Images Using Level Set Method. *J-Cosine*. 2021;5(2):130-139. doi: 10.29303/jcosine.v5i2.413
39. Fontenele RC, Gerhardt MDN, Willems H, Jacobs R. Can dental fillings affect the performance of an AI-driven innovative tool for automatic tooth segmentation in cone-beam computed tomography: A validation study?. *Journal of Dentistry*. 2022;121:103990. doi: 10.1016/j.jdent.2022.103990
40. Amirian M, Barco D, Herzig I, Schilling FP. Artifact Reduction in 3D and 4D Cone-Beam Computed Tomography Images With Deep Learning: A Review. *IEEE Access*. 2024;12:10281-10295. doi: 10.1109/access.2024.3353195

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Maria Fernanda Escalante Freitez.
Writing–review and editing: Maria Fernanda Escalante Freitez.