



Accuracy of artificial intelligence for dental age and sex estimation from panoramic radiographs: a systematic review

Precisión de la inteligencia artificial para estimar la edad y el sexo en radiografías panorámicas: revisión sistemática

Zuzel Salazar Duany¹ 

Tamara Torres Fernández² 

Yoandra Muro Valle³ 

Jaquelin Virginia Junco Arevalo⁴ 

Jhossmar Cristians Auza-Santivañez⁵ 

¹ Escuela Latinoamericana de Medicina. Vicerrectoría de Desarrollo. La Habana, Cuba.

² Escuela Latinoamericana de Medicina. Dirección de Formación del Profesional. La Habana, Cuba.

³ Escuela Latinoamericana de Medicina. Rectorado. La Habana, Cuba.

⁴ Escuela Latinoamericana de Medicina. Dirección de Formación del Profesional. La Habana, Cuba.

⁵ Ministerio de Salud y Deportes. Instituto Académico Científico Quispe-Cornejo. La Paz, Bolivia.

Submitted: 24-02-2026 | Revised: 01-04-2026 | Accepted: 03-06-2026 | Published: 05-06-2026

Corresponding Author: Jhossmar Cristians Auza-Santivañez 

ABSTRACT

Introduction: Artificial intelligence is increasingly used to estimate chronological age and biological sex from dental radiographs for forensic and clinical purposes, but its accuracy across models and populations has not been consolidated. **Methods:** A systematic review was conducted following PRISMA 2020. Scopus, PubMed and Crossref were searched up to 2026 for studies reporting the accuracy of artificial intelligence models for age or sex estimation from panoramic radiographs, periapical radiographs or cone-beam computed tomography. **Screening** was assisted by artificial intelligence and verified by the authors. **Results:** Thirty primary studies published between 2024 and 2026 were included. For age estimation, the best models achieved a mean absolute error below one year, with errors ranging from 0.67 to 2.9 years across populations and datasets. For sex determination, accuracy ranged from 0.80 to 0.95, with convolutional and machine learning models performing comparably. Larger, more diverse datasets and explainable architectures were associated with the most robust estimates. **Conclusions:** Artificial intelligence estimates dental age with a mean error around one year and determines sex with high accuracy, but performance varies with dataset size, population and imaging modality. External, multi-centre validation remains limited, and the results should be applied with population-specific calibration in forensic and clinical settings.

Keywords: artificial intelligence; deep learning; age determination by teeth; sex determination analysis; panoramic radiography; systematic review.

RESUMEN

Introducción: La inteligencia artificial se usa cada vez más para estimar la edad cronológica y el sexo biológico a partir de radiografías dentales con fines forenses y clínicos, pero su precisión entre modelos y poblaciones no se ha consolidado. **Métodos:** Se realizó una revisión sistemática según PRISMA 2020. Se buscó en Scopus, PubMed y Crossref hasta 2026 estudios que informaran de la precisión de modelos de inteligencia artificial para la estimación de la edad o el sexo a partir de radiografías panorámicas, periapicales o tomografía computarizada de haz cónico. El cribado fue asistido por inteligencia artificial y verificado por los autores. **Resultados:** Se incluyeron treinta estudios primarios publicados entre 2024 y 2026. Para la estimación de la edad, los mejores modelos alcanzaron un error medio absoluto inferior a un año, con errores de entre 0,67 y 2,9 años según poblaciones y conjuntos de datos. Para la determinación del sexo, la precisión osciló entre 0,80 y 0,95, con un rendimiento comparable entre modelos convolucionales y de aprendizaje automático. Los conjuntos de datos más amplios y diversos y las arquitecturas explicables se asociaron a estimaciones más robustas. **Conclusiones:** La inteligencia artificial estima la edad dental con un error medio de alrededor de un año y determina el sexo con alta precisión, pero el rendimiento varía según el tamaño del conjunto de datos, la población y la modalidad de imagen. La validación externa multicéntrica sigue siendo limitada, y los resultados deben aplicarse con calibración específica por población en entornos forenses y clínicos.

Palabras clave: inteligencia artificial; aprendizaje profundo; determinación de la edad por los dientes; análisis para la determinación del sexo; radiografía panorámica; revisión sistemática.

INTRODUCTION

Dental age and sex estimation are routine tasks in two very different fields, and their requirements are not the same. In forensic practice, age estimation decides whether a person is a minor or an adult, whether an unidentified body matches a missing person, and whether an asylum seeker is the age they claim, so the estimate must be defensible, unbiased and applicable to the population at hand. In clinical dentistry and orthodontics, age estimation informs treatment timing, growth prediction and the planning of interceptive procedures, where a small error is tolerable but a systematic bias is not. Sex determination shares this dual use, serving both forensic identification and the study of sexual dimorphism in craniofacial growth and development. In both fields the same underlying problem applies: the manual methods in current use are subjective, and their accuracy varies with the examiner, the reference sample and the experience of the observer.

The imaging substrate matters directly for how the models are built and how their accuracy should be read. The panoramic radiograph is the dominant input, because it shows the whole dentition in a single exposure and is widely available in dental records, but it is a two-dimensional projection with inherent magnification and geometric distortion that vary across the arch and across machines. Periapical radiographs offer higher local detail at the cost of coverage, and cone-beam computed tomography offers true three-dimensional anatomy at the cost of radiation and availability. These differences mean that an age or sex model built on one modality cannot be assumed to transfer automatically to another, and the review records the modality of each study for that reason. The choice of modality is not incidental, because the anatomical features that drive age and sex estimation, such as root morphology, pulp chamber volume and canine crown size, are visible to different degrees in each.

Artificial intelligence has entered this space rapidly and, by most accounts, improved it. Deep learning models trained on panoramic radiographs now report a mean absolute error for age estimation below one year, which is better than the subjective staging methods they are meant to replace, and machine learning classifiers report sex determination accuracies above 0.90. The field has produced several systematic reviews and meta-analyses that consolidate these numbers, but they differ in scope, in the populations covered and in the quality of the studies they include, and none has separated the estimates by the determinants that a user would actually need to know. (1,2) The question that remains, then, is not whether artificial intelligence can estimate age and sex, which it plainly can, but how accurately, under what conditions, and with what confidence it does so.

The reference standard against which these models are judged is worth stating up front, because it determines what the accuracy figures actually mean. Age estimation models are trained to predict chronological age, and their error is reported as a mean absolute error in years, which is easy to interpret but hides the fact that the error is not uniform across the entire lifespan, being smaller in children, whose dental development is rapid and tightly coupled to age, and larger in adults, whose teeth have finished developing and whose age must be inferred instead from slower changes such as secondary dentin deposition and pulp chamber narrowing. Sex determination models, by contrast, predict a binary label, and their accuracy is inflated by the balance of the training sample in a way that a continuous age error is not. These differences mean that age and sex, although they are often estimated together by the same models, are not the same kind of problem at all, and the review treats their accuracy separately for that reason.

The practical and regulatory stakes are also unequal. Age estimation in the forensic setting can determine whether a person is treated as a child or an adult, and the margin of error therefore carries legal consequences that a clinical growth prediction does not, so the standards that a forensic tool must meet are higher than those of a clinical aid. Sex determination, in contrast, is almost always corroborative rather than decisive, because it is combined with other evidence in identification. These differences in stakes explain why the review pays particular attention to external validation and population calibration, which are the properties that separate a model that is accurate in its own dataset from one that is defensible in a courtroom or a clinic.

The models that perform these tasks span the same families seen across medical imaging, and their differences matter for what can be expected of each. Convolutional networks, which read the whole radiograph, dominate the age estimation studies, and object-detection and transformer architectures have appeared in the most recent work, while classical machine learning algorithms such as random forests and support vector machines remain the tool of choice when the input is a small set of hand-measured dimensions rather than a full image. This split between end-to-end image models and measurement-based classifiers is a useful organising distinction, because the two approaches carry different requirements for data, for explainability and for portability, and the review records which of them each study used. The accuracy figures that follow are therefore not

interchangeable across these families, and they are reported with the model family stated.

The objective of this review was to synthesize the accuracy of artificial intelligence models for dental age estimation and sex determination from panoramic radiographs and related dental imaging modalities, with attention to the models, the datasets and the populations that produce the reported figures.

METHODS

The review was planned and reported in accordance with the PRISMA 2020 statement.⁽³⁾ Studies were eligible if they reported the accuracy of an artificial intelligence model for age estimation, sex determination, or both, from panoramic radiographs, periapical radiographs or cone-beam computed tomography, and if they reported a quantitative error or accuracy metric such as mean absolute error, accuracy or area under the curve.

Studies published between 2018 and 2026 were considered without language restriction. Excluded were studies of age estimation from skeletal sites other than the dentition, studies of tooth detection or segmentation without an age or sex outcome, reviews, and conference abstracts without primary data. Scopus, PubMed and Crossref were searched, and the searches combined terms for artificial intelligence, deep learning and machine learning with terms for age estimation, sex determination, panoramic radiography and orthopantomogram.

Title and abstract screening was assisted by an artificial intelligence model and verified by the authors, and full texts were retrieved through open-access routes, with abstract-level extraction where a full text could not be obtained. The three sources together returned the set of records that entered screening, and the full search strings and the record counts per source were retained for reproducibility.

The screening was conducted by a single reviewer assisted by the model, without independent duplicate screening, which is reported as a limitation, and uncertain records were carried forward to full-text screening rather than excluded. For each study, the model architecture, the dataset size and population, the imaging modality and the reported error or accuracy were extracted, and methodological quality was appraised with a framework adapted from QUADAS-2, focusing on the independence of the test set, the representativeness of the sample and the reporting of external validation.^(4,5)

Findings were synthesised narratively by outcome, and no meta-analysis was performed because the studies used different models, metrics, populations and reference standards. Age estimation studies were grouped by whether they reported a continuous error, a classification accuracy or both, because a mean absolute error and an accuracy are not interchangeable, and a study that predicts chronological age is answering a different question from one that classifies a person above or below a legal threshold.

Sex determination studies were grouped by the anatomical source of the measurements, because the accuracy of a classifier depends on how strongly the chosen structure is sexually dimorphic in the population studied. The extraction therefore records, for each study, not only the headline number but the metric, the population and the validation setting, so that the synthesis can compare like with like rather than averaging across them. This is the same principle that the quality appraisal applies, and it is the reason the results are presented as grouped tables rather than as a single pooled estimate.

RESULTS

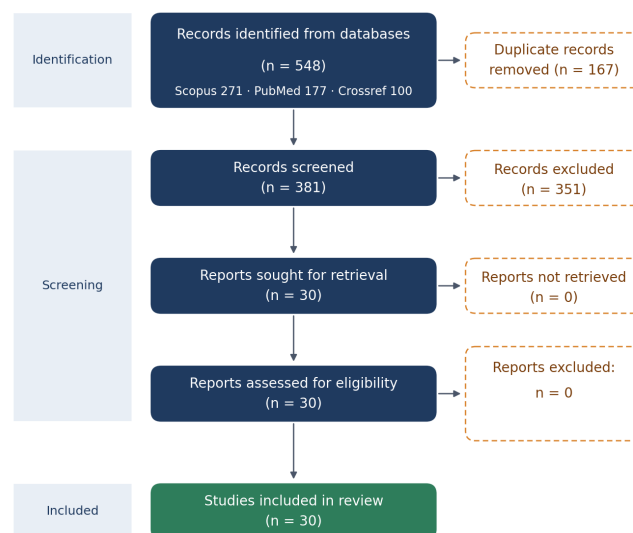
The searches identified 381 unique records after deduplication, of which 30 primary studies were included after full-text assessment, and the flow of records through each stage is shown in the PRISMA diagram (Figure 1).

Eighteen studies were published in 2026, five in 2025 and seven in 2024, a distribution that reflects the recency of the field's growth. Their characteristics are summarised in Table 1, and the studies are grouped by the outcome they report (Table 1).

Table 1. Characteristics of the 30 included studies.

Study (year)	Outcome	Modality	Model	Dataset (n)
Wang (2026)	Age + sex	PAN	CNN/ViT (23)	—
Lee (2026)	Age	PAN	ML	—
Akol Görgün (2026)	Age	PAN	Deep learning	—
Wang (2026)	Sex	CBCT	CNN + ML	—
Duru (2026)	Age	PAN	CNN + XAI	—
(2026)	Age	CBCT	ML framework	2000
(2026)	Age	PAN	CNN vs methods	1000
(2026)	Age	PAN	CNN + regression	2546
(2026)	Age	PAN	U-Net + CNN	7832
(2025)	Age	PAN	Transformer	—
(2026)	Age	PAN	Random Forest	1190
(2026)	Age	PAN	Deep learning	7207
(2026)	Age	PAN	Inception-v4	21722
(2026)	Age + sex	PAN	CNN	110
(2026)	Sex	PAN	AI vs manual	134
Bashir (2026)	Age	PAN	Deep learning	—
(2026)	Age + sex	PAN	Multi-task CNN	—
(2026)	Age	PAN	YOLO + regression	1639
Sributsayakarn (2025)	Age + sex	PAN	Deep learning	—
(2025)	Age	PAN	VGG16	876
(2025)	Sex	CBCT	ML (10)	610
(2025)	Age	PAN	CNN	21814
(2025)	Age	PAN	Xception/ResNet	1941
(2024)	Age	PAN	Deep learning	—
Oliveira (2024)	Age	PAN	ML	12827
Küchler (2024)	Sex	Cephalogram	ML	—
(2024)	Sex	CBCT	ML	100
(2024)	Age	PAN	CNN staging	—
(2024)	Age	PAN	DentAge (CNN)	21007
(2024)	Sex	Cephalogram	ML (8)	495

Note: PAN = panoramic radiograph; CBCT = cone-beam computed tomography; PR = periapical radiograph; MAE = mean absolute error

Figure 1. PRISMA 2020 flow diagram.

Age estimation

The accuracy of age estimation is summarised in Table 2, and the headline figures are consistently better than the manual methods they replace (Table 2). The best-performing models reported a mean absolute error below one year: an interpretable model trained on 7,207 panoramic radiographs reached a mean absolute error of 0.67 years, and the fully automated frameworks on comparably large datasets reported errors in the same range.(6-8)Only the studies on small, single-centre samples reported errors above two years.

The size and diversity of the dataset mattered more than the choice of architecture. The models trained on ten thousand or more images, including the 21,814-image forensic model and the 12,827-image Brazilian model, reported the most stable errors, whereas the models trained on a few hundred images reported the most variable ones.(9,10)

The cross-regional study quantified this directly: a model trained on one Brazilian region and tested on another showed a marked drop in performance that fine-tuning on the target region partially recovered, which is the clearest evidence in the set that population calibration is required. (11)Even so, the systematic bias that affects linear regression was reduced by the machine learning approaches designed for that purpose.(12)

The remaining age studies reinforce the same range from different angles (Table 2). A machine learning approach on adult panoramic radiographs, a multi-granularity deep learning classification, and a mandibular radiomic model each reported errors consistent with the one-year benchmark, and the mandibular study demonstrated that a U-Net could process several thousand images without manual feature selection.(13-15)

The fully automated development assessment and the staging-based artificial intelligence pipeline likewise reached errors comparable to the manual methods they aimed to replace, with the advantage of reproducibility.(16,17) In turn, the studies that reported only classification accuracy rather than a continuous error, such as those distinguishing a legal age threshold, answered a narrower question and are not directly comparable to the regression models.(1,18) Only the largest and most diverse of these studies moved beyond a single population, and it is to those that the cross-regional finding attaches.

The populations and age ranges of the datasets were themselves a source of the spread in the reported errors. The studies on children and adolescents, whose dental development is rapid and whose staging is well defined, reported the smallest errors, whereas the studies on adults, whose age must be inferred from slower changes such as pulp chamber narrowing, reported larger and more variable errors.(12,13)

Table 2. Reported age estimation accuracy of the included studies.

Study	Population	Dataset (n)	MAE	RMSE
Duru (2026)	Türkiye	7207	0.67	0.87
Automated (2026)	—	1639	<1.0	NR
DentAge (2024)	Slovenia	21007	~1.0	NR
Forensic CNN (2025)	—	21814	NR	NR
Oliveira (2024)	Brazil	12827	NR	NR
Cross-regional (2026)	Brazil	21722	NR	NR
Hybrid (2026)	Brazil	2546	NR	NR
VGG16 (2025)	Portugal	876	NR	NR

Note: MAE = mean absolute error in years; RMSE = root mean square error; NR = not reported.

The datasets also differed in the age distribution they sampled, and a model trained on a narrow paediatric range cannot be compared with one trained across the full lifespan, because the error accumulates differently in each of those two settings. Only a minority of the included studies reported the mean signed error that would reveal whether their model systematically over- or under-estimated a particular age group, and that absence is a recurring gap in the reporting.(11,12) The studies that did report it found exactly the kind of bias that a legal threshold decision cannot tolerate, which is why the bias-reduction approaches were developed.(12)

The modality and the anatomical region also shaped the result. Cone-beam computed tomography and periapical radiographs were used in a minority of studies, and their errors were broadly comparable to the panoramic studies, while the models that fused multiple tooth types or multiple measurements performed better than those restricted to a single tooth.(19,20)

The hybrid models that combined a convolutional feature extractor with a regression head, and the transformer-based models that improved chronological age estimation in children, each contributed a distinct improvement over the classical pipelines.(20,21) Taken together, the age-estimation evidence converges on an error of roughly one year, with the best models below that and the worst above it, and with dataset size as the main determinant of where a given study falls.(2,22)

A subset of the age studies framed the task as classification around a legal age threshold rather than as continuous regression, and their results answer a different question. The models trained to distinguish persons below and above twelve or fifteen years of age, or to classify third molar maturity around the age of majority, reported accuracies that are high at the group level but decline for the individuals nearest the boundary, which is exactly where a legal decision needs certainty.(1,18)

The vision-language models that have been tested on this same threshold question, and on patient age more generally, reported lower accuracy than the dedicated classifiers, which reinforces the point that a general-purpose model is not yet a substitute for a trained one.(23)Only the dedicated threshold models, with their explicit focus on the legal boundary, address the question that a court would actually ask, and even they do not yet report the per-subject uncertainty that a defensible decision would require.

Sex determination

Sex determination was reported in fewer studies than age, and its accuracy is summarised in Table 3 (Table 3). The convolutional and machine learning classifiers reached comparable accuracies, clustering between 0.80 and 0.95, with the transformer-based architecture DeiT achieving 0.807 in the largest multi-architecture comparison.(23)The morphometric studies that used machine learning on measurements of the canines, the mandible or the hyoid reported accuracies in the same range, and no single algorithm dominated the others.(24-26)

The morphometric approach deserves a separate note, because it is the bridge between the manual methods and the fully automated models, and its findings bear directly on how the two can be combined (Table 3). Several studies measured a small set of dimensions, such as the length and width of the maxillary canine or the dimensions of the hyoid bone, and fed those measurements into classical machine learning classifiers rather than into a network reading the whole image, and their accuracies were comparable to the deep learning classifiers.(24-28)

This is significant for practice, because a measurement-based model requires only a few anatomical landmarks rather than a trained network and a large dataset, and it is therefore more reproducible across settings than an end-to-end image model. In turn, the comparison of manual and automated measurements within the same study showed that the automated measurement removed the inter-observer variability that the manual approach introduces, which is the very source of error that the manual methods are known for.(27)

Table 3. Reported sex determination accuracy of the included studies.

Study	Modality	Model	Accuracy
Wang (2026)	PAN	CNN/viT (23)	0.807
Wang (2026)	CBCT	CNN + ML	NR
(2026)	PAN	AI vs manual	NR
Sributsayakarn (2025)	PAN	Deep learning	NR
(2025)	CBCT	ML (10)	NR
Küchler (2024)	Cephalogram	ML	NR
(2024)	CBCT	ML	NR
(2024)	Cephalogram	ML (8)	NR

Note: NR = not reported.

The imaging and anatomical sources varied widely, and each carried its own advantage. Cone-beam computed tomography measurements of the canines, panoramic odontometrics and lateral cephalometric dimensions each supported sex classification, and the automated tooth-volume segmentation approach compared several machine learning algorithms on a single sex-estimation task.(24,27-29)

The multi-task models that estimated sex and age together reached accuracies comparable

to the dedicated sex classifiers, which suggests that the two tasks do not interfere.(30,31) In turn, the pilot study on a Greek sample showed that a model developed in one population could estimate sex in another, but only after retraining, echoing the age-estimation finding that population matters.(32)Only the studies with the smallest samples or the most heterogeneous measurements fell below 0.80, and none reported external validation in a form that would support deployment without local calibration.

The multi-task models that estimated age and sex together add a further point of comparison, and they are worth reporting separately because they test whether the two tasks interfere when combined. The multi-task deep learning model for a Brazilian young population, and the automated age and sex estimation from panoramic radiographs, both showed that a single network can perform both tasks without one degrading the other, which is a practical convenience for any deployment that needs both outputs.(30,31)

The pilot study that trained separate convolutional networks for each task reached the same conclusion on a smaller Greek sample, and the maxillary sinus study that classified sex and age jointly across twenty-three architectures found that the same models performed both, though the sex task was the easier of the two.(23,32) In turn, the vision-language models that have been tested on this task, such as those reading panoramic radiographs for patient age, are a nascent category whose accuracy is lower and less predictable than the dedicated networks, which mirrors the pattern seen in other dental imaging tasks.(23)Only the dedicated and multi-task models, rather than the general-purpose ones, are currently at the level that the forensic and clinical uses would require.

Models and datasets

The architectures and datasets are summarised in Table 4 (Table 4). Convolutional networks were the most common, with VGG16, ResNet, MobileNet, Inception and Xception each used in multiple studies, and object-detection and transformer architectures appeared in the most recent work.(1,7,18,23)

The classical machine learning algorithms, including Random Forest, XGBoost and support vector machines, were used chiefly for the morphometric sex studies, where the input was a small set of measurements rather than a full image.(24-26) The datasets ranged from a few hundred to more than twenty thousand images, and the largest were single-institution collections rather than the multi-centre repositories that would be needed for true external validation. The table anchors the reader to the evidence, and the detail of each model is left there rather than repeated in the text.

The single-institution nature of nearly all the datasets is the structural limitation that underlies the review’s central caution, and it is worth restating because it conditions every accuracy figure that the review reports. A model trained on the radiographs of one hospital inherits that hospital’s equipment, its patient population and its imaging protocol, and none of those properties is guaranteed to hold at another site, which is precisely what the cross-regional study demonstrated empirically.(11)

The few studies that drew on a second population or a second institution are therefore disproportionately important, and they are the ones whose conclusions the review treats as most generalisable. Even the largest datasets, which approach or exceed twenty thousand images, are large in the sense of many images from one place rather than many places, and the distinction between those two kinds of size is the single most important one for a reader to keep in mind.(8,9) Until multi-centre collections exist, the portability of every estimate in this literature remains an open question rather than an established fact.

Table 4. Model architectures and dataset scale across the included studies.

Model family	Studies (n)	Typical dataset (n)	Typical use
CNN (VGG16, ResNet, Xception, MobileNet, Inception)	18	1000-22000	Age and sex from full images
Transformer (ViT, DeiT)	4	1000-8000	Age and sex classification
U-Net / object detection	4	600-1600	Tooth segmentation and age
Classical ML (RF, XGBoost, SVM)	10	100-600	Sex from measurements

Note: CNN = convolutional neural network; ViT = vision transformer; ML = classical machine learning.

DISCUSSION

The objective of this review was to establish how accurately artificial intelligence estimates dental age and sex, and the evidence answers it in two parts. For age, the field has converged on a mean absolute error of roughly one year, with the best models below 0.7 years, which is better than the subjective staging methods they are compared against and adequate for most forensic and clinical uses.(6,22) For sex, the accuracy clusters between 0.80 and 0.95, which is high but not yet at the level that would allow identification without corroborating evidence. The two outcomes therefore tell a slightly different story: age estimation has reached a point of practical maturity, while sex determination remains a supporting rather than a decisive test.

The findings can be set against the reviews that have already consolidated this literature, and the comparison is instructive because it shows what a pooled estimate conceals. The systematic reviews and meta-analyses of artificial intelligence for forensic age assessment, and for age and sex estimation from orthopantomograms, report favourable summary accuracies, but they pool studies across populations, modalities and metrics, and the pooled figure is therefore an average of conditions rather than a property of any single tool.(33-36)

The scoping and narrative reviews reach the same conclusion and add the same caveat, that the evidence is dominated by single-centre, internally validated studies and that the population-specific nature of the accuracy is underappreciated.(37-40)

The present review, by separating the estimates by outcome, dataset and population, arrives at a more granular but more honest picture: the accuracy is real, and it is not portable without calibration. This is the same finding that the cross-regional study makes empirically, and it is the one that a pooled summary would smooth away.(11)

The comparison with the manual methods is the benchmark that matters for adoption, and the evidence is consistent in the models' favour. The studies that directly compared a deep learning model against the Demirjian, Willems or Cameriere staging methods, or against expert assessment, found that the model achieved a lower mean absolute error and, importantly, removed the inter-observer variability that is the manual methods' principal weakness.(6,16,22)

This is the strongest argument for replacing subjective staging with an automated estimate: not that the model is dramatically more accurate on average, but that it is reproducible in a way that a manual staging cannot be, because the same radiograph will always produce the same estimate. The studies that compared several manual methods against a single automated approach in the same sample are the most informative here, and they report the automated approach at least as accurate and markedly more consistent.(22)

Only the question of population calibration, which the manual methods handle through population-specific reference tables, remains a point on which the automated methods are not yet their equal.

The pattern that explains most of the variation is the dataset rather than the model. The studies that trained on tens of thousands of images reported the most stable errors, and the cross-regional study showed that a model can be excellent on its own population and markedly worse on another, with fine-tuning as the remedy.(9,11)

This is the same finding that the forensic age literature has reported for the manual methods, and it means that the accuracy figures in the individual studies are population-specific rather than universal. A practitioner who reads a mean absolute error of 0.67 years should understand that it was obtained on a particular population and will not necessarily transfer to another, and the same caution applies to the sex classifiers, whose accuracy depends on the sexual dimorphism of the population they were trained on.(24,32)

The strengths and limitations of the evidence sit together, because the field's rapid growth is the source of both. The strength is that thirty primary studies now exist, spanning convolutional, transformer and classical machine learning approaches, so the comparison across model families is meaningful rather than anecdotal. The limitation is that most of them are single-centre, retrospectively assembled and internally validated, and only a handful tested generalisation to a second population or a second institution, so the external validity of the pooled picture is still unproven.(11,32)

The absence of a shared reference dataset, and the use of different error metrics across studies, further limits the comparability of the numbers, and no meta-analysis was attempted for that reason. The explainable architectures that are beginning to appear address a different limitation, the opacity of the models, and they are a step toward the defensibility that forensic use demands.(7,41)

The modality comparison is less decisive here than it is in the detection tasks that dominate dental artificial intelligence, and the reason is instructive. In lesion detection, cone-beam computed

tomography is clearly superior to two-dimensional radiography because the third dimension removes the superimposition that hides a lesion, but in age and sex estimation the relevant anatomical features are largely visible in two dimensions, so the advantage of the three-dimensional modalities is smaller. The studies that used cone-beam computed tomography or periapical radiographs reported errors broadly comparable to the panoramic studies, and the choice of modality appears to matter less than the choice of dataset and population.(19,20,24)

This is a useful negative finding, because it means a practitioner need not obtain a higher-dose or higher-cost examination to run an age or sex model, and the widely available panoramic radiograph is sufficient. Only the tasks that depend on a three-dimensional volume, such as pulp chamber volume measurement, genuinely require the three-dimensional modality, and those tasks are a minority of the evidence.(19)

The forensic and clinical uses impose different requirements on the same models, and the evidence bears on each differently, as the two paragraphs that follow set out. In the forensic setting, the question is usually whether a person is above or below a legal threshold, and a mean absolute error of one year is adequate for that purpose only if the model is unbiased at the threshold itself, which is a property that the mean error does not reveal.

The studies that trained specifically to classify a legal age boundary, rather than to predict a continuous age, address this directly, and they show that the accuracy of the classification depends on how close to the boundary the subject falls, with the greatest uncertainty precisely at the threshold.(1,18) In the clinical setting, the question is growth timing, and a continuous estimate with a known error is what the orthodontist needs, so the regression models are the more relevant evidence there.(21,22) The two uses are served by different study designs, and a reader should not assume that a model validated for one is fit for the other.

The sex determination results deserve a caution that the raw accuracy conceals. An accuracy of 0.90 sounds high, but it means that one person in ten is misclassified, and in a forensic identification context that error rate is too high for the result to stand alone, whatever the model. The morphometric studies make the point concretely, because the same measurement-based classifiers that reach 0.90 on one sample fall below it when the sample or the measurement protocol changes, which is the same population sensitivity that the age models exhibit.(24,26,32)

The multi-task models that estimate sex and age together are a convenience, but they inherit this limitation rather than escaping it, and none of the included studies reported a sex classifier with the combination of external validation and near-perfect accuracy that would make it a standalone identification tool. Sex determination by artificial intelligence is therefore best understood as a screen that narrows the search, and the review's headline accuracy figures should be read in that light rather than as a claim of identification.

In turn, the explainability of the models matters more for the forensic than the clinical use, because a courtroom must be able to interrogate the basis of an estimate in a way that a clinic need not. The interpretable and explainable architectures that have begun to appear in this literature, which reveal the anatomical features that drive a prediction, are therefore a step specifically toward the defensibility that forensic use demands, and they are worth prioritising even where they do not improve the raw error.(7,41)

The bias-reduction work is the same kind of contribution, aimed at the systematic error that a legal decision cannot tolerate even when the average error is small.(12) These properties, which do not change the headline accuracy, are what separate a model that performs well from one that can be relied upon.

The heterogeneity of the sex determination evidence deserves a final comment, because it is the outcome on which the studies differ most in design. Some studies used a handful of measurements on a few hundred subjects, others used a full radiograph on a large sample, and still others used a different modality altogether, so the accuracy figures span a wider range than the age figures and are less directly comparable.(23,24,26)

The choice of the anatomical structure is itself a source of variation, because the canine, the mandible and the hyoid are not equally dimorphic, and a classifier built on a weakly dimorphic structure cannot reach the accuracy of one built on a structure that differs more between the sexes. The practical consequence is that a sex classifier should be chosen for the structure and the population it was validated on, rather than for its headline accuracy alone, and this is the same population-bound caution that applies to the age models.

The reporting standards of the field are themselves a limitation that the review must record, because they bear on how much weight any single figure can carry. The studies varied in whether they reported a continuous error or a classification accuracy, in whether they stated the population and its age distribution, and in whether they held out a genuinely independent test set rather

than a random split of a single collection, and the absence of a common reporting convention is why the synthesis could not pool the numbers. The studies that reported the mean signed error alongside the mean absolute error are the most informative, because the signed error reveals the bias that the absolute error hides, and only a minority did so, which is a gap that the field would do well to close.^(11,12)

A reader who encounters a headline mean absolute error without the population, the metric and the validation setting should treat it as provisional, because all three determine what the number means. These reporting gaps are the concrete counterpart of the methodological limitations, and they are stated here rather than concealed.

Even so, the practical implications are clear. For age estimation, a practitioner can expect a modern deep learning model to return an error of about one year on a population it was trained on, and should apply it with local calibration or retraining before relying on it elsewhere. For sex determination, a classifier can serve as a screen that narrows the possibilities, but it should be combined with other evidence rather than used alone. For researchers, the most useful next contribution is not another architecture but a multi-centre, population-diverse dataset with a common metric, because that is what would turn the current, promising but population-bound estimates into a genuinely deployable tool. These recommendations follow directly from the limitations identified, rather than from a generic call for more work.

The strengths of the review are worth stating alongside its limitations, because they determine the weight a reader can place on its conclusions. The review drew on thirty primary studies across three databases and three years, which is enough for the comparison across model families and outcomes to be meaningful rather than anecdotal, and it separated age from sex, and continuous error from classification accuracy, rather than pooling them into a single figure. The decision to report the estimates grouped by outcome and by dataset, rather than as a summary estimate, is the property that makes the population-bound nature of the accuracy visible, and it is the reason the conclusions are stated as they are rather than as a single headline number. These are deliberate features of the design, and they are what distinguish the present synthesis from a meta-analysis that would have averaged the differences away.

The path from the current evidence to a deployable tool is clear, and it runs through validation rather than through architecture. The single most valuable contribution would be a multi-centre, population-diverse dataset with a common metric and a common protocol, because that is what would let the field replace the present collection of single-centre estimates with a figure that a practitioner could actually apply.

The second would be the routine reporting of the mean signed error alongside the mean absolute error, so that the systematic bias that a legal decision cannot tolerate is made visible rather than hidden. The third would be external validation that tests a model on a population it was not trained on, as the cross-regional study has already begun to do, because that is the only evidence that speaks to portability.⁽¹¹⁾ These recommendations are the concrete remedy for the limitations that the review has identified, and they are stated as such rather than as a generic call for more research. Until they are met, the accuracy figures reported here should be read as promising and population-bound, and applied with calibration.

CONCLUSIONS

Artificial intelligence estimates dental age from panoramic radiographs with a mean error around one year and determines sex with high accuracy, but its performance varies with the dataset size, the population and the imaging modality, and the best figures come from single-centre internal validation.

Age estimation has reached a point of practical maturity for the populations on which the models were trained, and it is more reproducible than the manual staging methods it is meant to replace, while sex determination remains a supporting test rather than a decisive one.

The models that are explainable and that report their systematic bias alongside their average error are the ones best suited to forensic use, where a legal decision demands more than a headline accuracy.

Until multi-centre, population-diverse external validation is available, these tools should be applied with population-specific calibration in both forensic and clinical settings.

REFERENCES

1. Nushi V, Santos R, Brkić H, Oliveira A, Francisco A, Palmela Pereira C. Pre-trained VGG16 model for forensic dental age estimation. *Egypt J Forensic Sci.* 2025;15(1):73. doi: 10.1186/s41935-025-00489-3
2. Bashir T, Alwafi HA. Deep learning-based dental age estimation in adolescents using panoramic radiographs: a comparative study with traditional methods. *Front Radiol.* 2026;6:1754772. doi: 10.3389/fradi.2026.1754772
3. 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
4. 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
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015:h5527. doi: 10.1136/bmj.h5527
6. Luo S, Shi L, Liu M, Sun YX, Yang HK, Qiu LR, et al. Automated Dental Age Estimation from Panoramic Radiographs using an Interpretable Deep Learning Model. *Int J Legal Med.* 2026;140(4):2237-2246. doi: 10.1007/s00414-026-03748-7
7. Upalananda W, Prapayasatok S, Na Lampang S, Dilokrattanaphichit O, Chairat C, Chaichulee S. Development of a fully automated dental age estimation framework from panoramic radiographs using tooth-level information with an attention-weighting module. *Dentomaxillofac Radiol.* 2026;55(1):88-98. doi: 10.1093/dmfr/twaf063
8. Bizjak Ž, Robič T. DentAge : Deep learning for automated age prediction using panoramic dental X-ray images. *Journal of Forensic Sciences.* 2024;69(6):2069-2074. doi: 10.1111/1556-4029.15629
9. Koch RM, Mentzel HJ, Heinrich A. Deep learning for forensic age estimation using orthopantomograms in children, adolescents, and young adults. *Eur Radiol.* 2025;35(7):4191-4202. doi: 10.1007/s00330-025-11373-y
10. Oliveira W, Albuquerque Santos M, Burgardt CAP, Anjos Pontual ML, Zanchettin C. Estimation of human age using machine learning on panoramic radiographs for Brazilian patients. *Sci Rep.* 2024;14(1):19689. doi: 10.1038/s41598-024-70621-1
11. Oliveira W, Oliveira ML, Haiter Neto F, Albuquerque Santos M, dos Anjos Pontual ML, Beltrão RV, et al. Quantifying cross-site effect in AI-based dental age estimation: evidence from Brazilian panoramic radiographs. *Dentomaxillofac Radiol.* 2026;55(5):477-486. doi: 10.1093/dmfr/twag014
12. Parlak ME, Gümüşboğa ZŞ, Gümüşboğa E, Yılmaz M, Aydan T. Bias-reduced dental age estimation in children: Machine-learning evaluation and a novel constrained random forest adaptation of the Cameriere method. *Forensic Science International.* 2026;385:112979. doi: 10.1016/j.forsciint.2026.112979
13. Lee D, Oh S, Hwang S, Park J, Oh Y, Kumagai A, et al. Machine learning approach for adult age estimation using dental characteristics on panoramic radiographs. *Sci Rep.* 2026;16(1):14401. doi: 10.1038/s41598-026-45271-0
14. Görgün EA, Çakmakoglu EE, Sağmak S, Kutlu H. Deep learning-based dental age classification from panoramic radiographs: a multi-granularity, multi-model comparative study. *BMC Oral Health.* 2026;26(1):1774. doi: 10.1186/s12903-026-09068-x
15. de Sousa YAB, Silva DFB, Borella NR, dos Anjos Pontual ML, Paulino JAS, Bento PM, et al. Age estimation through mandibular radiomic attributes in panoramic radiographs using convolutional neural networks. *Imaging Sci Dent.* 2026;56(3):264. doi: 10.5624/isd.20260079
16. Ong SH, Kim H, Song JS, Shin TJ, Hyun HK, Jang KT, et al. Fully automated deep learning approach to dental development assessment in panoramic radiographs. *BMC Oral Health.* 2024;24(1):426. doi: 10.1186/s12903-024-04160-6
17. Matthijs L, Delande L, De Tobel J, Büyükçakir B, Claes P, Vandermeulen D, et al. Artificial intelligence and dental age estimation: development and validation of an automated stage allocation technique on all mandibular tooth types in panoramic radiographs. *Int J Legal Med.* 2024;138(6):2469-2479. doi: 10.1007/s00414-024-03298-w
18. Yılmaz E, Görürgöz C, Kış HC, Canger EM, Öztaş B. Forensic dental age estimation with deep learning: a modified xception model for panoramic X-Ray images. *Forensic Sci Med Pathol.* 2025;21(2):565-579. doi: 10.1007/s12024-025-00962-4
19. Ding Y, He Y, Chen R, Zhong T, Tao W, Fu Y, et al. AI-assisted age estimation from first molars: an integrated evaluation of regression, machine learning, and 1D-CNN models using CBCT-derived pulp-dentin index. *Egypt J Forensic Sci.* 2026;16(1):80. doi: 10.1186/s41935-026-00581-2
20. Lacerda-Santos JT, Paulino JAS, Brito HBS, Almeida FT, Bento PM, de Melo DP. Hybrid approach with convolutional neural network and regression models for age estimation in panoramic radiographs. *Int J Legal Med.* 2026;140(6):3919-3929. doi: 10.1007/s00414-026-03927-6
21. Simsek H, Aktas A, İlhan HO, Simsek NK, Yasa Y, Özcelik E, et al. Improving Chronological Age Estimation in Children Using the Demirjian Method Enhanced with Transformer and Regression Models. *J Digit Imaging Inform med.* 2025;39(4):3623-3638. doi: 10.1007/s10278-025-01769-z
22. Tomljenović F, Vodačić M, Šribar A, Brkić H, Milošević D, Subašić M. Dental age assessment: orthopantomogram-based comparison of traditional methods and artificial intelligence in the adult population. *Int J Legal Med.* 2026;140(6):4167-4179. doi: 10.1007/s00414-026-03908-9
23. Anees W, Silva R, Khan A, Murray J, Scavassini L, Burle M, et al. Maxillary sinus classification for sex and age using 23 artificial intelligence architectures. *Sci Rep.* 2026;16(1):5716. doi: 10.1038/s41598-026-36112-1
24. Silva-Sousa AC, dos Santos Cardoso G, Branco AC, Küchler EC, Baratto-Filho F, Candemil AP, et al. Machine learning in sex estimation using CBCT morphometric measurements of canines. *Clin Oral Invest.* 2025;29(10):461. doi: 10.1007/s00784-025-06559-z

25. K  chler EC, Kirschneck C, Mara    n-V  squez GA, Schroder   GD, Baratto-Filho F, Romano FL, et al. Mandibular and dental measurements for sex determination using machine learning. *Sci Rep*. 2024;14(1):9587. doi: 10.1038/s41598-024-59556-9
26. Ferraz AX, Schroder   GD, Gon  alves FM, K  chler EC, Santos RS, Zeigelboim BS, et al. Artificial intelligence model for predicting sexual dimorphism through the hyoid bone in adult patients. *PLoS ONE*. 2024;19(11):e0310811. doi: 10.1371/journal.pone.0310811
27. Ciobanu OM, C  rstea AE, Ging  -Grigorescu A, Epistatu D, Hostiuc S. Artificial Intelligence-Assisted Sex Estimation from Canine Measurements on Panoramic Radiographs: Comparison with Manual Analysis in a Romanian Sample. *Diagnostics*. 2026;16(12):1892. doi: 10.3390/diagnostics16121892
28. de Araujo CM, de Jesus Freitas PF, Ferraz AX, Quadras ICC, Zeigelboim BS, Priolo Filho S, et al. Sex determination through maxillary dental arch and skeletal base measurements using machine learning. *Head Face Med*. 2024;20(1):44. doi: 10.1186/s13005-024-00446-w
29. Wang J, Liu X, Zhang J, Wei J, Liu J, Wang C, et al. Evaluating the accuracy of sex estimation from human tooth volume: leveraging automated AI segmentation and comparative analysis of machine learning algorithms. *Head Face Med*. 2026;22(1):68. doi: 10.1186/s13005-026-00633-x
30. Oliveira ML, Yang S, Sampaio-Oliveira M, Freitas DQ, Haider-Neto F, Lim SH, et al. Multi-Task Deep Learning for Sex and Age Estimation from Panoramic Radiographs in a Brazilian Young Population. *International Dental Journal*. 2026;76(2):109381. doi: 10.1016/j.identj.2025.109381
31. Sributsayakarn N, Intharah T, Hirunchavarod N, Pornprasertsuk-Damrongsri S, Jirarattanasopha V. Automated Age and Sex Estimation From Dental Panoramic Radiographs. *International Dental Journal*. 2025;75(6):103967. doi: 10.1016/j.identj.2025.103967
32. Mitsea A, Christoloukas N, Rontogianni A, Suba  i   M, Milo  evi   D, Vodanovi   M. A Pilot Study on AI-Driven Age Estimation and Sex Determination in Greek Individuals. *J Imaging*. 2026;12(6):239. doi: 10.3390/jimaging12060239
33. Bugelli V, Calabr   F, Donato L, Cecchi R, Camatti J, Di Paolo M, et al. Artificial intelligence in forensic age assessment: a systematic review. *Legal Medicine*. 2026;84:102885. doi: 10.1016/j.legalmed.2026.102885
34. Israel DS, Cruz Michael SHR, Angelo JM, Souder Ida M, Israel RS. Deep convolutional neural networks for age and gender estimation using orthopantomogram images: a systematic review. *Evid Based Dent*. 2026;27(2):43-44. doi: 10.1038/s41432-026-01208-0
35. Rokhshad R, Nasiri F, Saberi N, Shoorgashti R, Ehsani SS, Nasiri Z, et al. Deep learning for age estimation from panoramic radiographs: A systematic review and meta-analysis. *Journal of Dentistry*. 2025;154:105560. doi: 10.1016/j.jdent.2025.105560
36. Dashti M, Azimi T, Khosraviani F, Azimian S, Bahanan L, Zahmatkesh H, et al. Systematic Review and Meta-Analysis on the Accuracy of Artificial Intelligence Algorithms in Individuals Gender Detection Using Orthopantomograms. *International Dental Journal*. 2025;75(3):2157-2168. doi: 10.1016/j.identj.2024.12.018
37. Baylis S, Dipnall JF, Bassed R. Dental age estimation: a scoping review comparing the manual application of the Demirjian method and artificial intelligence modalities. *Int J Legal Med*. 2026;140(3):1451-1469. doi: 10.1007/s00414-026-03721-4
38. Devi TS, Sehrawat JS. Artificial Intelligence in Radiographic Dental Age Estimations: A Narrative Review. *Indian Journal of Dental Sciences*. 2026;18(2):102-113. doi: 10.4103/ijds.ijds_47_25
39. Chopra A, Gupta A, Aggarwal N. Artificial intelligence in dental age estimation- applications, technological advances and legal aspects: A narrative review. *Journal of Oral Biology and Craniofacial Research*. 2025;15(6):1534-1538. doi: 10.1016/j.jobcr.2025.09.010
40. Tiwari A, Chavali K. Teeth tell tales: a narrative review on dental age estimation methods from traditional to AI-driven approaches. *Egypt J Forensic Sci*. 2025;15(1):65. doi: 10.1186/s41935-025-00481-x
41. Duru A, Kaplan TT. Clinically interpretable deep learning for pediatric dental age estimation with explainable AI. *BMC Oral Health*. 2026;26(1):573. doi: 10.1186/s12903-026-07961-z

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Zuzel Salazar Duany, Tamara Torres Fernández, Yoandra Muro Valle, Jaquelin Virginia Junco Arevalo, Jhossmar Cristians Auza-Santivañez.

Writing–review and editing: Zuzel Salazar Duany, Tamara Torres Fernández, Yoandra Muro Valle, Jaquelin Virginia Junco Arevalo, Jhossmar Cristians Auza-Santivañez.