



# Artificial intelligence and dental age estimation: development and validation of an automated stage allocation technique on all mandibular tooth types in panoramic radiographs

Lander Matthijs<sup>1</sup> · Lauren Delande<sup>1</sup> · Jannick De Tobel<sup>2</sup> · Barkin Büyükçakir<sup>3</sup> · Peter Claes<sup>3</sup> · Dirk Vandermeulen<sup>3</sup> · Patrick Thevissen<sup>4</sup>

Received: 5 November 2023 / Accepted: 16 July 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

## Abstract

Age estimation in forensic odontology is mainly based on the development of permanent teeth. To register the developmental status of an examined tooth, staging techniques were developed. However, due to inappropriate calibration, uncertainties during stage allocation, and lack of experience, non-uniformity in stage allocation exists between expert observers. As a consequence, related age estimation results are inconsistent. An automated staging technique applicable to all tooth types can overcome this drawback.

This study aimed to establish an integrated automated technique to stage the development of all mandibular tooth types and to compare their staging performances.

Calibrated observers staged FDI teeth 31, 33, 34, 37 and 38 according to a ten-stage modified Demirjian staging technique. According to a standardised bounding box around each examined tooth, the retrospectively collected panoramic radiographs were cropped using Photoshop CC 2021® software (Adobe®, version 23.0). A gold standard set of 1639 radiographs were selected ( $n_{31} = 259$ ,  $n_{33} = 282$ ,  $n_{34} = 308$ ,  $n_{37} = 390$ ,  $n_{38} = 400$ ) and input into a convolutional neural network (CNN) trained for optimal staging accuracy. The performance evaluation of the network was conducted in a five-fold cross-validation scheme. In each fold, the entire dataset was split into a training and a test set in a non-overlapping fashion between the folds (i.e., 80% and 20% of the dataset, respectively). Staging performances were calculated per tooth type and overall (accuracy, mean absolute difference, linearly weighted Cohen's Kappa and intra-class correlation coefficient). Overall, these metrics equalled 0.53, 0.71, 0.71, and 0.89, respectively. All staging performance indices were best for 37 and worst for 31. The highest number of misclassified stages were associated to adjacent stages. Most misclassifications were observed in all available stages of 31. Our findings suggest that the developmental status of mandibular molars can be taken into account in an automated approach for age estimation, while taking incisors into account may hinder age estimation.

**Keywords** Forensic age estimation · Dental development · Conventional radiography · Artificial intelligence

## Introduction

Estimating the age of young people lacking proof of birth and/or identity may be necessary to elucidate legal questions regarding the reliability of the stated age when applying for international protection or punishment in criminal offences. Based on the Children's Rights, unaccompanied minor refugees benefit from specific provisions when granted international protection [1–3]. In children and young adults, age estimates are established based on their developmental status. One reliable indicator of that developmental status is tooth development [4–7].

Expert human observers assess the extent of tooth development on medical images (e.g., panoramic radiographs) and

✉ Jannick De Tobel  
jannick.detobel@gmail.com

<sup>1</sup> Oral Health Sciences and Dentistry, KU Leuven, Leuven, Belgium

<sup>2</sup> Diagnostic Sciences – Radiology, Ghent University, Ghent, Belgium

<sup>3</sup> Electrical Engineering - Processing Speech and Images, KU Leuven, Leuven, Belgium

<sup>4</sup> Imaging and Pathology - Forensic Odontology, KU Leuven, Leuven, Belgium

correlate it with age. Thereto, several tooth development staging techniques have been established. They cover the entire tooth developmental track, from bud formation until closure of the root apices [8–13]. Although each consecutive developmental stage is well-defined in the various tooth development staging techniques, manual stage allocation remains prone to inter- and intra-observer variability [14]. The variability in stage allocation is due to (1) inappropriate calibration of investigators, causing difficulties in achieving compliance [15]; (2) uncertainties about stage allocation caused by ambiguous stage definitions, or an excessive or deficient number of stages [16, 17]; and (3) the observer's degree of experience [18]. Non-uniformity in stage allocation between expert observers and during repeated stage allocations leads to inconsistent age estimation outcomes. After all, dental age estimation methods were developed based on data obtained by applying a chosen tooth development staging technique on all subjects in a sample from a specific population. Those data represent a technique-specific reference for that given population and are used to compile population-specific age estimation atlases, tables, or models.

Due to the variability in stage allocation, the provided age estimations could be considered inadequate proof of the subject's age. In a legal context, this may lead to repeated court trials or appeals [19]. An automated stage allocation technique can be part of an automated age estimation method, providing uniformity in stage allocation and, consequently, undisputable age estimation outcomes.

Initially, such a technique was reported for third molar development [14]. A convolutional neural network (CNN) was used, and further improvements were made by altering the CNN architecture and the approach to abstracting relevant information from panoramic radiographs (i.e., partial and complete tooth segmentation) [14, 20]. Such a neural network is a learning system that can be trained to perform machine learning tasks, such as classification and regression. This ability to learn has allowed neural networks to be used in many fields, including medical data analysis [21, 22]. A subcategory of neural networks, the CNN, is particularly suitable for learning from image data, as popularly demonstrated by Krizhevsky, Sutskever, and Hinton in 2012 [23]. A CNN architecture, namely AlexNet, was able to classify images into one of 1000 categories with an accuracy of 0.85, showing the capability of the CNNs for image-based tasks [21]. New CNN architectures have been designed and deployed in many image-based tasks since [24, 25]. The CNN uses multiple layers of kernel convolutions applied to the input image to progressively extract higher-level features. This feature extraction phase is then followed by a conventional neural network to utilise these features for the task at hand, i.e. classification or regression. The parameters of these kernels and the neural network weights are tuned during training, allowing the network to learn and look for the relevant features in the images, thus maximising performance. The CNNs naturally lend themselves to use in dental staging

problems from panoramic radiographs, as in its origin, this is exactly an image-based classification task [20, 26, 27].

Later on, automated staging techniques were also reported for other tooth types [26, 28]. After all, combining tooth developmental information of permanent teeth and third molars improves the accuracy of the age estimates in children and young adults [29]. Despite reported overall accuracies ranging between 0.70 [26, 30] and 0.94 [31] for automated staging of all permanent teeth (including third molars), few previous studies reported tooth type-specific results, which may mask a tooth type-induced bias of the staging results. Moreover, none of the previous studies selected their reference population to obtain an evenly spread number of cases per stage per tooth type. Consequently, over- or underrepresentation of certain stages caused large accuracy discrepancies between stages [28].

Reflecting on possible differences in staging performance between tooth types allows determining whether automated staging can be an asset in automated age estimation, i.e. whether the staging results should be taken into account by the network to ameliorate the age estimation outcomes. Thus, the current study aims were (1) to select a reference population stratified by dental developmental stage per tooth type in the lower left quadrant; (2) to establish an integrated automated technique to stage the development of those teeth on panoramic radiographs using artificial intelligence (CNN); (3) to compare the staging performances of the different tooth types.

## Materials and methods

### Radiograph collection

Radiographs were selected from an available set of 4000 panoramic radiographs from UZ Leuven patient files registered between 2000 and 2015. All radiographs were processed anonymously, excluding all personal data except sex and age. Radiographs with the following criteria were excluded: (1) absence of the tooth type under study, (2) presence of orthodontic appliances, (3) inadequate image quality, (4) severe overlap with neighbouring tooth structures, and (5) abnormal position of the tooth under study.

The aim was to collect 400 radiographs per tooth type, 20 per stage per sex, thus establishing a gold standard data set. Therefore, two investigators sorted independently through the available set of panoramic radiographs. A third senior investigator decided about disagreements between both first investigators. To represent all tooth types (incisors, canines, premolars, molars, and third molars), teeth with Fédération Dentaire Internationale (FDI) [32] numbers 31, 33, 34, 37, and 38 were examined. A modified Demirjian staging technique described by De Tobel et al. (2017) [14] (Tables 1 and 2) was used to manually allocate a developmental stage.

**Table 1** Modified Demirjian descriptive allocation criteria for developmental stages in multi-rooted teeth

Stage	Allocation Criteria
0	The developmental crypt is visible in the jawbone. Calcification has yet to begin
1	Start of calcification is visible at the superior level of the crypt in the form of small cones. The calcified points are not fused together
2	The calcified points are fused and present a regularly outlined occlusal surface
3	a Occlusal enamel formation is complete. Extension towards the cervical region is seen b Dentinal deposition has started c The pulp chamber can be seen as a radiolucent curved outline at the occlusal border
4	a Crown formation is completed down to the cemento-enamel junction b Pulp chamber outline has a trapezoidal shape. The projections of the pulp horns create an umbrella like shape
5	a Beginning of root formation is seen as a spicule b The root length is less than the crown height c A calcified point or semi-lunar shape shows initial formation of the root bifurcation
6	a Roots are more defined with funnel shaped endings b Root length is equal to or greater than the crown height
7	Walls of the root canal are parallel; the apical end is still partially open
8	Walls of the root canal are converging at the apex; the apical end is still partially open
9	a The apical end of the root canal is completely closed b The periodontal ligament has a uniform width around the entire root

**Table 2** Modified Demirjian descriptive allocation criteria for developmental stages in single-rooted teeth

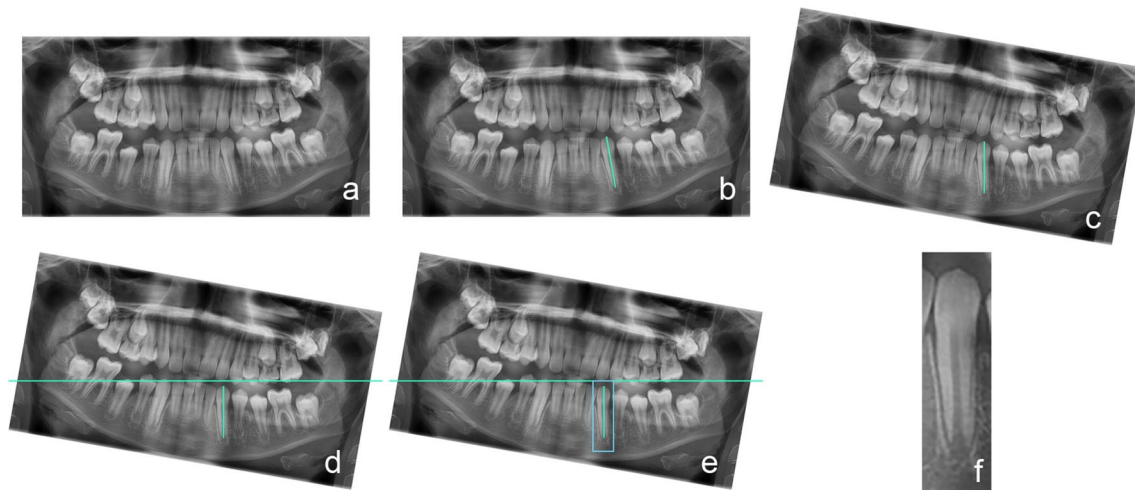
Stage	Allocation Criteria
0	The developmental crypt is visible in the jawbone. Calcification has yet to begin
1	Start of calcification is visible at the superior level of the crypt in the form of a cone
2	The calcified point forms one uniform cusp presenting a regularly outlined incisal edge
3	a Incisal enamel formation is complete. Extension towards the cervical region is seen b Dentinal deposition has started c The coronal outline of the pulp chamber can be seen as a radiolucent thick line in the center of the crown
4	a Crown formation is completed down to the cemento-enamel junction b Pulp chamber outline has a thick rectilinear shape
5	a Beginning of root formation is seen as an extension of dentine and cementum deposit downwards from the cemento-enamel junction b The root length is less than the crown height
6	Root length is equal to or greater than the crown height
7	Walls of the root canal are parallel; the apical end is still partially open
8	Walls of the root canal are converging at the apex; the apical end is still partially open
9	a The apical end of the root canal is completely closed b The periodontal ligament has a uniform width around the entire root

## Radiograph post-processing

The radiograph post-processing was conducted using Adobe Photoshop 2021® (version 23.0).

To obtain a uniform orientation of the examined teeth, the radiographs were rotated vertically by the long axis of the tooth (Fig. 1). For further processing, each tooth was extracted in a standardised bounding box, manually indicated on the rotated

radiograph. The standard dimensions of the bounding box were defined per tooth type by measuring the maximum width and length that occurred in the entire set. Each tooth was then centred within its box and positioned with its occlusal plane 8 pixels below the upper border of the bounding box. The images inside these standardized bounding boxes were then defined as input for the staging network. This procedure corresponded with the one described in detail by De Tobel et al. [14].



**Fig. 1** Illustration of radiograph post-processing to extract tooth 33. **a** Original total radiograph. **b** The vertical axis of the tooth is manually drawn. **c** The radiograph is rotated so that tooth 33's vertical axis corresponds with a true vertical. **d** A horizontal guideline is placed

8 pixels cranially from the tooth's incisal border. **e** The tooth is captured in a bounding box of a standard size. **f** The image is cropped to the bounding box. This image is used as input for the neural network

## CNN Network

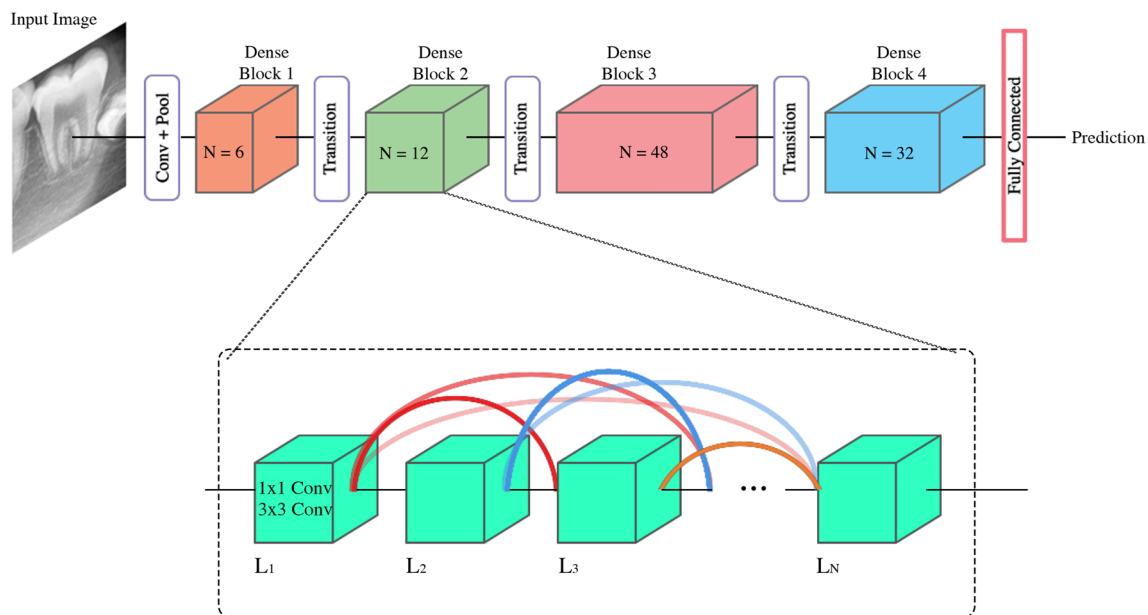
A deep, densely connected DenseNet201 CNN architecture consisting of four blocks of layers was used. The blocks consisted of 12, 24, 96 and 32 convolutional layers, respectively. This architecture uses the residual connection concept in which every layer is connected to each subsequent layer within the same block (Fig. 2). This allows the neural network to grow deeper without encountering the vanishing gradient problem [33, 34]. Within the network, the feature maps produced by each preceding layer are concatenated before being passed to the following layer. In an N-layer network, this concatenation produces  $N(N+1)/2$  connections, increasing the learning capability of the network without increasing the number of network parameters. The DenseNet201 architecture was initially trained on the ImageNet dataset, which has 1000 outputs on the output layer, and was modified to accommodate the ten developmental stages used in this study.

The input images for all teeth were resized such that the height of the image was 224 pixels and zero-padded to preserve the aspect ratio of the bounding boxes. Data augmentation was not necessary, due to the already large inherent variation of the dataset, and because performance evaluation of the network was conducted in a five-fold cross-validation scheme, which is robust against overfitting. In each fold, the entire dataset was split into training and test sets in a non-overlapping fashion between the folds (i.e., 80% and 20% of the dataset, respectively). During the training phase, the parameters of the network were

iteratively calculated to maximize classification accuracy on the training set. Since dental staging involves a multi-class classification problem, categorical cross-entropy was selected as the loss function during training. The training set for each fold was used for the stochastic training of the neural network, and the performance evaluation was carried out on the test data from the same fold. Throughout the cross-validation, a stochastic gradient descent optimisation algorithm with a learning rate of 0.001 and momentum of 0.9 was utilized. The models for each fold were trained for 150 epochs with a batch size of 8 samples. The performance metrics were then aggregated across folds. These NN learning parameters were kept constant over the different tooth types, while the parameters of the network itself were optimized for each tooth type independently.

## Evaluation of staging performance

Firstly, per tooth type, a confusion matrix of allocated stages between the human investigators (gold standard) and the automated software was constructed. Secondly, the classification performances were evaluated by measuring the Rank-N recognition rate (Rank-N RR), accuracy (expressed as Rank-1 RR), mean absolute difference (MAD, expressed in number of stages difference), linearly weighted Cohen's kappa (LWK), and intra-class correlation coefficient (ICC). Note that LWK and ICC produce similar results if the rating means and variations are close. Therefore, a discrepancy between LWK and ICC indicates more disparity between the rating means and variations.



**Fig. 2** An overview of the DenseNet-201 model architecture. This model consists of four dense blocks, each in turn made up  $N$  convolutional units of a  $1 \times 1$  convolution followed by a  $3 \times 3$  convolution. Within each block, all the units are densely inter-connected via residual connections, allowing the network to go deeper without facing vanishing gradients and to learn more complex features. The initial Conv+Pool is a  $7 \times 7$  convolution and a  $3 \times 3$  max-pooling,

which acts as a preliminary filter. The transition units all consist of  $1 \times 1$  convolution, and  $2 \times 2$  average pooling, which reduces the feature map size, retaining important patterns. The final fully connected layer projects high-dimensional feature maps onto 1000 dimension. All convolutions are followed by a ReLU activation, and the output layer uses a SoftMax activation function

## Results

In the youngest stages of teeth 31, 33, 34, and 37, the target of 20 radiographs per stage, sex and tooth type was not met (Table 3). Thousand six hundred and thirty-nine teeth ( $n_{31} = 259$ ,  $n_{33} = 282$ ,  $n_{34} = 308$ ,  $n_{37} = 390$ ,  $n_{38} = 400$ ) were selected from the initial dataset to act as gold standard input for the neural network. Figure 3 illustrates an example of each stage, present in the gold standard set for each tooth type.

The confusion matrices (Fig. 4) showed that, in general, the highest number of misclassified stages were allocated to the adjacent stages. Note that the spread in the confusion matrix (i.e. a higher occurrence of large stage discrepancies) was higher for teeth 31 and 38 than for the other teeth. Furthermore, the highest numbers of misclassifications were observed in all available stages of tooth 31, with proportions of perfect agreement ranging from 0.06 to 0.45. In contrast, tooth 37 showed the lowest numbers of misclassifications, with proportions of perfect agreement ranging from 0.43 to 0.86. Furthermore, tooth 33 demonstrated a range of 0 to 0.90, tooth 34 from 0 to 0.65, and tooth 38 from 0.40 to 0.70.

Similarly, all studied classification performance indices were worst for tooth 31 and best for tooth 37 (Tables 4 and 5). Note that the performance indices were very similar for teeth 33, 34, and 38, except for the MAD. Compared

to 33 and 34, the remarkably higher MAD for 38 is in line with the remarkably higher spread in its confusion matrix, which is also reflected in the discrepancy between the LWK and the ICC.

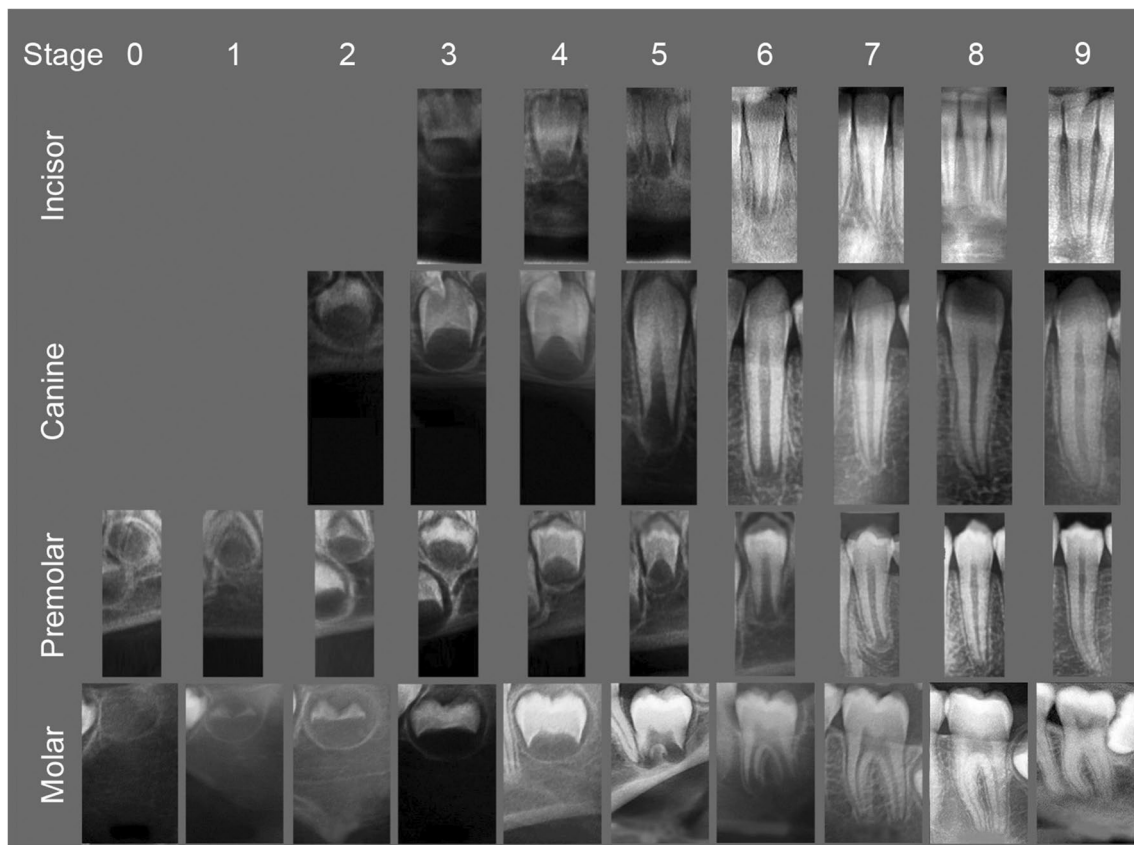
## Discussion

### Interpretation of results and comparison with literature

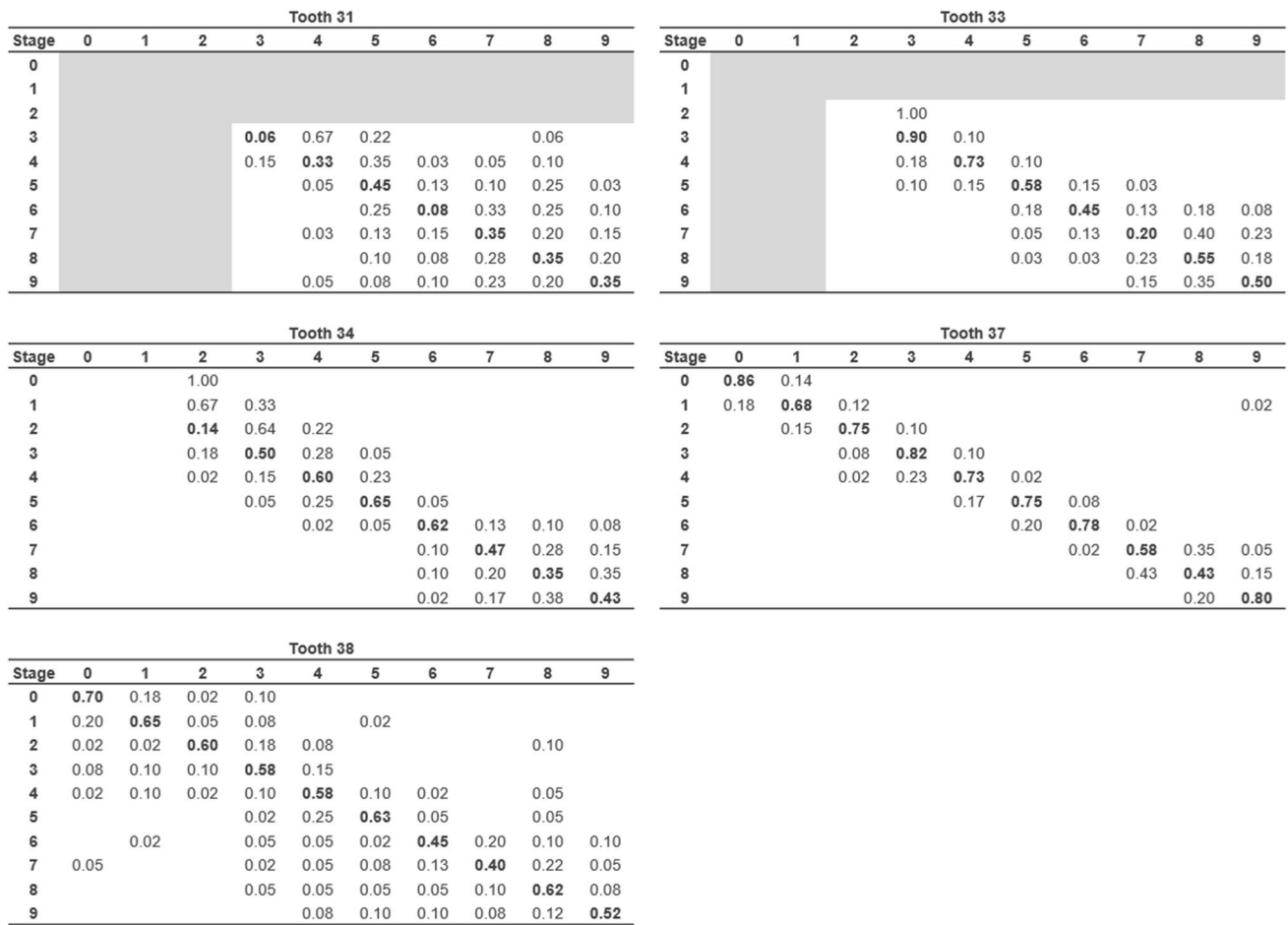
Three groups of automated staging performance were noted in the current study. Firstly, staging 31 proved to be the hardest for the network. Secondly, the network staged 33, 34, and 38 with moderate performance. Finally, staging 37 proved to be the easiest for the network. These results are in line with Ong et al. (2024) who reported staging performances on an upward trend from mandibular left incisors, over canines, premolars, and molars (third molars were not studied) [35]. However, Table 6 demonstrates contrasting results from other studies. In Aliyev et al. (2022) staging of 32 and 33 showed the worst performance, with an accuracy around 0.60 [26], while in Han et al. (2022) staging of 31, 33, and 37 showed the worst performance, with an accuracy around 0.55 [28].

**Table 3** Number of radiographs included per stage and per tooth type

Stage	Tooth type				
	31	33	34	37	38
0	0	0	0	30	40
1	0	0	3	40	40
2	0	2	25	40	40
3	19	40	40	40	40
4	40	40	40	40	40
5	40	40	40	40	40
6	40	40	40	40	40
7	40	40	40	40	40
8	40	40	40	40	40
9	40	40	40	40	40
<b>Total</b>	259	282	308	390	400

**Fig. 3** Visual representation of the various stages amongst the different tooth types, according to the modified Demirjian technique





**Fig. 4** Confusion matrices of allocated stages, showing the proportions per tooth type. Gold standard stages (rows) and stages allocated by the DenseNet201 CNN (columns) are shown. Bold values indicate

perfect agreement between automated and reference staging. Light grey cells show deficit stages per tooth type

**Table 4** Mean Rank-N recognition rate of automated stage allocation per tooth type

Tooth	The correct stage was ranked									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
31	0.30	0.23	0.17	0.14	0.09	0.04	0.02	0.00	0.00	0.00
33	0.55	0.26	0.13	0.04	0.02	0.00	0.00	0.00	0.00	0.00
34	0.51	0.27	0.15	0.03	0.02	0.01	0.00	0.00	0.00	0.00
37	0.71	0.25	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00
38	0.57	0.17	0.10	0.07	0.04	0.01	0.00	0.01	0.01	0.01

In contrast, accuracies around 0.55 corresponded with a moderate performance in the current study. Also in previous studies by our research group, the overall accuracy was lower than reported by other researchers (Table 7). This might be due to an over- or underrepresentation of teeth in certain stages and/or teeth of certain types (Table 6) [28, 31]. Still, tooth 37 was reported among the best suitable for automated staging by Aliyev et al. (2022) [26], which was confirmed by our current findings. From a radiological

and anatomical point of view, it makes sense that a molar would be easier to stage, for the network as well as for a human observer. Imaging artefacts mostly do not affect the molars, while they often hinder interpreting the central part of a panoramic radiograph. Moreover, the anatomy of molars is straightforward (as opposed to the anatomy of third molars) and they are at a distance from the mandibular canal (which might interfere with the image of the third molar) [36, 37]. Finally, when taking the step to estimate

**Table 5** Overall stage allocation performances in the different tooth types

Performance Index	Tooth Type				
	31	33	34	37	38
Accuracy	0.30	0.55	0.51	0.71	0.57
MAD	1.20	0.58	0.62	0.31	0.86
LWK	0.42	0.75	0.76	0.90	0.74
ICC	0.58	0.90	0.91	0.97	0.84

Accuracy is expressed as the first rank recognition rate (proportion of correctly allocated stages); Mean Absolute Difference (MAD) indicates the difference between automated and manual staging (expressed in number of stages); Linearly weighted Cohen's kappa coefficient (LWK) and intraclass correlation coefficient (ICC) quantify the overall agreement between the gold standard and the automated stages. LWK takes the ordinal character of staging into account. So does ICC, while it is generally higher than LWK, but it is more severely affected by large stage discrepancies

age based on dental staging, Aliyev et al. (2022) found that tooth 37 contributed the most of all permanent left mandibular teeth, even in their study population between 6 and 13 years old.

### Motivation for the network

Several factors led to the choice of DenseNet over other network architectures. A first significant advantage of DenseNet's dense connectivity lies in the feature map reuse, allowing for an increased number of connections with fewer parameters compared to conventional sequentially connected networks [25, 38]. This helps alleviate the vanishing gradient problem common in deep networks, allowing the usage of deeper models. Feeding previous feature maps to deeper layers reduces layer redundancy, ensuring that deeper layers still receive low-level feature map inputs. This is particularly applicable to staging (dental) development. For example, a fully formed crown indicates that part of the sample is above a certain stage. This feature, at the early layers of the model, allows the decision space to be limited, and also can be incorporated with other, more high level, features deeper into the network, helping increase decision performance. In conventional sequential networks, deeper layers may face

insignificant differences in feature maps, diminishing predictive capability.

A second advantage of dense connections is that they enhance training effectiveness by increasing parameter efficiency and improving gradient flow during training [25]. They act as a regularisation method, enhancing predictive performance and reducing overfitting. This is particularly beneficial in scenarios like dental staging on panoramic radiographs where data collection and manual operations are labour-intensive, and data may be scarce.

However, compared to conventional sequential approaches, the major limitation of DenseNet is the requirement of significantly more GPU memory. Due to the concatenation operations, larger feature map inputs are generated to each layer compared to, for example, the ResNet architecture, which employs residual connectivity but without dense connections. DenseNet also requires a longer training time, because it uses much smaller convolutions. A large convolution is a more compact operation than several small convolutions on a GPU, and including these small convolution layers significantly increases the training time. This introduces a trade-off between the predictive capability and the demand on hardware resources [39]. In the benefit of the outcome, we chose to prioritise the former over the latter.

### Limitations and future prospects

A first limitation of the current study lies in its focus only on the staging step, which represents one of three steps needed in a fully automated age estimation approach: (1) identifying the regions of interest; (2) assessing the developmental status of the anatomical structures of interest; (3) inferring an age estimate based on the developmental status of the individual (combining the information from different anatomical structures). Beware that the last step – i.e. the age estimate – may be biased by each of the former steps. Therefore, to exclude any bias induced by the first step, we opted for a standardised manual cropping of the teeth. Other studies have focused on automating this first step, be it by selecting bounding boxes around the teeth, or by segmenting the teeth (Table 7) [20, 40–42].

Next, the second step can induce bias because of staging performance differences of different tooth types. The

**Table 6** Literature overview of the accuracy of automated dental stage allocation per tooth type. The number of individuals per tooth type is displayed between brackets. However, Aliyev et al. (2022) did not report these numbers

Accuracy (Number of individuals)	Tooth Type							
	31	32	33	34	35	36	37	38
Aliyev et al. (2022)	0.70	0.58	0.63	0.74	0.72	0.79	0.77	-
Han et al. (2022)	0.53 (49)	0.67 (46)	0.57 (135)	0.71 (406)	0.80 (826)	0.67 (649)	0.55 (645)	0.97 (5126)
Current work	0.30 (259)	-	0.55 (282)	0.51 (308)	-	-	0.71 (390)	0.57 (400)



**Table 7** Literature overview of automated dental stage allocation

Study	Tooth type	Number of individuals per tooth type	Number of stages (technique)	Localisation method	Staging method	Validation method	Performance metrics				
							Acc	AUC	MAD	LWK	ICC
<b>De Tobel et al. (2017)</b>	38	400	10 (modified Demirjian)	Manual bounding box	AlexNet	Train-test split	0.51	0.60	0.60	0.82	-
<b>Merdietio et al. (2019)</b> <sup>a</sup>	38	400	10 (modified Demirjian)	Manual bounding box	DenseNet201	Train-test split	0.54	0.61	0.61	0.82	-
<b>Merdietio et al. (2019)</b> <sup>b</sup>	38	400	10 (modified Demirjian)	Manual segmentation	DenseNet201	Train-test split	0.61	0.53	0.53	0.84	-
<b>Banar et al. (2020)</b>	38	400	10 (modified Demirjian)	YOLO-like CNN for localisation	DenseNet201	Fivefold cross-validation	0.54	0.69	0.69	0.79	-
<b>Aliyev et al. (2022)</b>	31, 32, 33, 34, 35, 36, 37	475	8 (Demirjian)	Manual segmentation	Custom CNN	Tenfold cross-validation	0.70	-	-	-	-
<b>Han et al. (2022)</b>	31, 32, 33, 34, 35, 36, 37, 38	10,257	8 (Demirjian)	Faster R-CNN bounding box	ResNet-101	Train-validate-test split	0.87	0.17	0.17	0.80	-
<b>Mohammad et al. (2022)</b>	34, 35	240	8 (Demirjian, with only 6 stages present in their age group)	Manual bounding box for localisation	DCNN using the Python TensorFlow and Keras libraries	Train-test split	0.78	-	-	0.58	-
<b>Dong et al. (2023)</b>	entire dentition (32 teeth)	673*	9 (modified Demirjian)	YOLOv3 bounding box	Custom CNN SOS-Net	Train-validate-test split	0.94	0.98§	-	-	0.97§
<b>Upalananda et al. (2023)</b>	38, 48°	4000	8 (Demirjian, with only 5 stages present in their age group)	Manual bounding box	GoogLeNet	Train-test split	0.82	-	-	-	-
<b>Kokomoto et al. (2024)</b>	entire dentition (32 teeth)	8023	8 (modified Gleiser and Hunt)	Scaled-YOLO v4 P6 bounding box	EfficientNetV2-M	Fivefold stratified cross-validation	0.70	-	-	-	-
<b>Milani et al. (2024)</b>	38, 48°	3422	9 (modified Demirjian)	EfficientNet-B0 bounding box	EfficientNet	Train-test split	0.84	-	-	-	-
<b>Ong et al. (2024)</b>	31, 32, 33, 34, 35, 36, 37	5133	8 (Demirjian)	YOLOv5 for localisation	EfficientNet	Train-test split	-	-	-	-	-
<b>Current work</b>	31, 33, 34, 37, 38	259, 282, 308, 390, 400, resp.	10 (modified Demirjian)	Manual bounding box	DenseNet201	Fivefold cross-validation	0.53	0.71	0.71	0.71	0.89

Acc = accuracy; AUC = area under the curve; DP-AC = dynamic programming—active contour; ICC = intraclass correlation coefficient; LWK = linearly weighted kappa; MAD = mean absolute difference

\* Detailed number per tooth type included in a graph; not lower than 500 for all teeth except third molars; not higher than 300 for third molars

§ For comparison with other studies, only the results for mandibular teeth were depicted here

° Together considered as same tooth type

current study suggests that including the developmental information of all mandibular tooth types might hinder age estimation. This was in line with previous studies (Table 6). Although some previous studies integrated the staging step and the age estimation step, none of them unfortunately studied the effect that only including the developmental information of certain tooth types would have on age estimation. Such a study would contribute to the explainability of the automated approach using neural networks, which is essential for an automated approach to be put into practice [27]. Furthermore, the age estimation performance of any automated approach needs to be compared with the current gold standard, verifying its superiority (e.g. regarding speed, reliability, performance) over manual approaches.

A second limitation lies in the relatively small study sample, with the number of individuals per tooth type ranging from 259 to 400. Table 7 demonstrates that a higher accuracy was obtained in studies with a larger study sample. Still, our sample was stratified by stage per tooth type, which allows drawing firmer conclusions about staging than studies with samples stratified by age or with non-stratified samples.

## Conclusion

Our findings suggest that the developmental status of mandibular molars can be taken into account in an automated approach for age estimation, which would increase explainability. Canines, premolars, and third molars can be taken into account with caution. Finally, incisors are the hardest to assess for an automated approach, and taking them into account may hinder age estimation.

**Data Availability** The dataset consisting of four thousand panoramic radiographs generated and analyzed during the current study is available through the corresponding author upon reasonable request.

## References

- Office of the Commissioner General for Refugees and Stateless Persons (2019) Guide for unaccompanied minors who apply for asylum in Belgium. [https://www.cgrs.be/sites/default/files/brochures/asiel\\_asile\\_-\\_nbmv\\_mena\\_-\\_unaccompaniedforeign-minor\\_-\\_eng\\_2.pdf](https://www.cgrs.be/sites/default/files/brochures/asiel_asile_-_nbmv_mena_-_unaccompaniedforeign-minor_-_eng_2.pdf). Accessed 1 Feb 2022
- UNHCR (2019) Access to education for refugee and migrant children in Europe. <https://www.unhcr.org/neu/wp-content/uploads/sites/15/2019/09/Access-to-education-europe-19.pdf>. Accessed 3 Feb 2022
- Office of the Commissioner General for Refugees and Stateless Persons (2021) Children in the asylum procedure. <https://www.cgrs.be/en/asylum/children-asylum-procedure>. Accessed 7 Dec 2021
- Yan J, Lou X, Xie L, Yu D, Shen G, Wang Y (2013) Assessment of dental age of children aged 3.5 to 16.9 years using Demirjian's method: a meta-analysis based on 26 studies. *PLoS One* 8:e84672. <https://doi.org/10.1371/journal.pone.0084672>
- Lewis JM, Senn DR (2015) Forensic Dental Age Estimation: An Overview. *J Calif Dent Assoc* 43:315–319
- Sukhia RH, Fida M (2010) Correlation among chronologic age, skeletal maturity, and dental age. *World J Orthod* 11:e78–84
- Panainte I, Pop SI, Mártha K (2016) Correlation Among Chronological Age, Dental Age and Cervical Vertebrae Maturity in Romanian Subjects. *Rev Med Chir Soc Med Nat Iasi* 120:700–710
- Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. *Hum Biol* 45:211–227
- Gleiser I, Hunt EE Jr (1955) The permanent mandibular first molar: its calcification, eruption and decay. *Am J Phys Anthropol* 13:253–283
- Köhler S, Schmelzle R, Loitz C, Puschel K (1994) Development of wisdom teeth as a criterion of age determination. *Ann Anat* 176:339–345
- Moorrees CF, Fanning EA, Hunt EE Jr (1963) Age variation of formation stages for ten permanent teeth. *J Dent Res* 42:1490–1502
- Nanda RS, Chawla TN (1966) Growth and development of dentitions in Indian children. I. Development of permanent teeth. *Am J Orthod* 52:837–853. [https://doi.org/10.1016/0002-9416\(66\)90253-3](https://doi.org/10.1016/0002-9416(66)90253-3)
- Nolla CM (1952) The development of permanent teeth. University of Michigan Ann Arbor
- De Tobel J, Radesh P, Vandermeulen D, Thevissen PW (2017) An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *J Forensic Odontostomatol* 35:42–54
- Kullman L, Tronje G, Teivens A, Lundholm A (1996) Methods of reducing observer variation in age estimation from panoramic radiographs. *Dentomaxillofac Radiol* 25:173–178
- Lynnerup N, Belard E, Buch-Olsen K, Sejrsen B, Damgaard-Pedersen K (2008) Intra- and interobserver error of the Greulich-Pyle method as used on a Danish forensic sample. *Forensic Sci Int* 179(242):e1–6. <https://doi.org/10.1016/j.forsciint.2008.05.005>
- Dhanjal KS, Bhardwaj MK, Liversidge HM (2006) Reproducibility of radiographic stage assessment of third molars. *Forensic Sci Int* 159(Suppl 1):S74–S77. <https://doi.org/10.1016/j.forsciint.2006.02.020>
- Wittschieber D, Schulz R, Vieth V et al (2014) Influence of the examiner's qualification and sources of error during stage determination of the medial clavicular epiphysis by means of computed tomography. *Int J Legal Med* 128:183–191. <https://doi.org/10.1007/s00414-013-0932-6>
- Youth Justice Legal Centre (2015) Age assessment. <http://www.yjlc.uk/wpcontent/uploads/2015/01/Age-Assessment-Legal-summary.pdf>. Accessed 20 Oct 2020
- Merdietio Boedi R, Banar N, De Tobel J, Bertels J, Vandermeulen D, Thevissen PW (2020) Effect of Lower Third Molar Segmentations on Automated Tooth Development Staging using a Convolutional Neural Network. *J Forensic Sci* 65:481–486. <https://doi.org/10.1111/1556-4029.14182>
- Gore JC (2020) Artificial intelligence in medical imaging. *Magn Reson Imaging* 68:A1–a4. <https://doi.org/10.1016/j.mri.2019.12.006>
- Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK (2018) Medical Image Analysis using Convolutional Neural Networks: A Review. *J Med Syst* 42:226. <https://doi.org/10.1007/s10916-018-1088-1>
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L, Weinberger K (eds) *Advances in Neural Information Processing System*. The MIT Press Cambridge, Massachusetts USA

24. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556.
25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. (2017) Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–8.
26. Aliyev R, Arslanoglu E, Yasa Y, Oktay AB. (2022) Age estimation from pediatric panoramic dental images with CNNs and LightGBM. 2022 Medical Technologies Congress (TIPTEKNO). IEEE. pp. 1–4.
27. Büyükcakır B, Bertels J, Claes P, Vandermeulen D, de Tobel J, Thevissen PW (2024) OPG-based dental age estimation using a data-technical exploration of deep learning techniques. *J Forensic Sci* 69:919–931. <https://doi.org/10.1111/1556-4029.15473>
28. Han M, Du S, Ge Y et al (2022) With or without human interference for precise age estimation based on machine learning? *Int J Legal Med* 136:821–831. <https://doi.org/10.1007/s00414-022-02796-z>
29. Metsaniitty M, Waltimo-Siren J, Ranta H, Fieuws S, Thevissen P (2019) Dental age estimation in Somali children and sub-adults combining permanent teeth and third molar development. *Int J Legal Med* 133:1207–1215. <https://doi.org/10.1007/s00414-019-02053-w>
30. Kokomoto K, Kariya R, Muranaka A, Okawa R, Nakano K, Nozaki K (2024) Automatic dental age calculation from panoramic radiographs using deep learning: a two-stage approach with object detection and image classification. *BMC Oral Health* 24:143. <https://doi.org/10.1186/s12903-024-03928-0>
31. Dong W, You M, He T et al (2023) An automatic methodology for full dentition maturity staging from OPG images using deep learning. *Appl Intell* 53:29514–29536
32. Leatherman G (1971) Two-digit system of designating teeth—FDI submission. *Aust Dent J* 16:394. <https://doi.org/10.1111/j.1834-7819.1971.tb03438.x>
33. Balduzzi D, Frean M, Leary L, Lewis J, Ma KW-D, McWilliams B. (2017) The shattered gradients problem: If resnets are the answer, then what is the question? International Conference on Machine Learning. PMLR. pp. 342–50.
34. Jastrzębski S, Arpit D, Ballas N, Verma V, Che T, Bengio Y (2017) Residual connections encourage iterative inference. arXiv preprint arXiv:171004773.
35. Ong SH, Kim H, Song JS et al (2024) Fully automated deep learning approach to dental development assessment in panoramic radiographs. *BMC Oral Health* 24:426. <https://doi.org/10.1186/s12903-024-04160-6>
36. Johan NA, Khamis MF, Abdul Jamal NS, Ahmad B, Mahanani ES (2012) The variability of lower third molar development in Northeast Malaysian population with application to age estimation. *J Forensic Odontostomatol* 30:45–54
37. Rickne CS, Weiss G (2017) *Woelfel's Dental Anatomy*. Wolters Kluwer Philadelphia, Pennsylvania, USA
38. Wan K, Yang S, Feng B, Ding Y, Xie L. (2019) Reconciling feature-reuse and overfitting in densenet with specialized dropout. 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI). IEEE. pp. 760–7.
39. Zhang C, Benz P, Argaw DM et al. (2021) Resnet or densenet? Introducing dense shortcuts to resnet. Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3550–9.
40. Banar N, Bertels J, Laurent F et al (2020) Towards fully automated third molar development staging in panoramic radiographs. *Int J Legal Med* 134:1831–1841. <https://doi.org/10.1007/s00414-020-02283-3>
41. Niu L, Zhong S, Yang Z et al (2024) Mask refinement network for tooth segmentation on panoramic radiographs. *Dentomaxillofac Radiol* 53:127–136. <https://doi.org/10.1093/dmfr/twad012>
42. Leite AF, Gerven AV, Willems H et al (2021) Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clin Oral Investig* 25:2257–2267. <https://doi.org/10.1007/s00784-020-03544-6>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.