Towards fully automated third molar development staging in panoramic radiographs

Nikolay Banar[1], Jeroen Bertels[2,*], François Laurent[2]; Rizky Merdietio Boedi[3,4],

Jannick De Tobel[4], Patrick Thevissen[4] and Dirk Vandermeulen[2]

1. Computational Linguistics and Psycholinguistics Research Center (CLiPS), University of Antwerp, Antwerp, Belgium
2. Department of Electrical Engineering (ESAT/PSI), KU Leuven, Leuven, Belgium
3. Department of Dentistry, Diponegoro University, Semarang, Indonesia
4. Department of Imaging and Pathology (Forensic Odontology), KU Leuven, Leuven, Belgium
5. Department of Anatomy, University of Pretoria, Pretoria, South Africa

*Correspondence to Jeroen Bertels. Email: nicolae.banari@uantwerpen.be

Abstract

Staging third molar development is commonly used for age assessment in sub-adults. Current staging techniques are, at most, semi-automated and rely on manual interactions prone to operator variability. The aim of this study was to fully automate the staging process by employing the full potential of deep learning, using convolutional neural networks (CNNs) in every step of the procedure. The dataset used to train the CNNs consisted of 400 panoramic radiographs (OPGs), with 20 OPGs per developmental stage per sex, staged in consensus between three observers. The concepts of transfer learning, using pre-trained CNNs, and data augmentation were used to mitigate the issues when dealing with a limited dataset. In this work, a three-step procedure was proposed and the results were validated using fivefold cross-validation. First, a CNN localized the geometrical center of the lower left third molar, around which a square region of interest (ROI) was extracted. Second, another CNN segmented the third molar within the ROI. Third, a final CNN used both the ROI and the segmentation to classify the third molar into its developmental stage. The geometrical center of the third molar was found with an average Euclidean distance of 63 pixels. Third molars were segmented with an average Dice score of 93%. Finally, the developmental stages were classified with an accuracy of 54%, a mean absolute error of 0.69 stages, and a linear weighted Cohen's kappa coefficient of 0.79. The entire automated workflow on average took 2.72 s to compute, which is substantially faster than manual staging starting from the OPG. Taking into account the limited dataset size, this pilot study shows that the proposed fully automated approach shows promising results compared with manual staging.

Keywords: Dental age assessment; Third molar; Developmental stage; Localization; Segmentation; Classification

Introduction

In forensic practice, dental age assessment is commonly conducted by well-trained forensic odontologists using panoramic radiographs (OPGs). The registered degree of development is classified using specific tooth development staging techniques and correlated with age.

However, the manually performed staging's major drawback is a possible stage classification variability within and between observers. This has been comprehensively reviewed and reported in [5], with kappa values ranging between 0.52 and 1.00 for different staging techniques. To counter this drawback, automated age assessment methods have been proposed, especially since recent applications of deep learning in the context of medical imaging have shown to give promising results [3]. Recent research by Stern et al. [37] used MRI data of the hand, clavicle, and teeth to fuse age-relevant information from three anatomical sites to achieve a mean absolute prediction error in regressing chronological age of 1.01 ± 0.74 years. However, related work in the field of automated dental age assessment, using X-ray imaging, is limited. By contrast, in the field of bone age assessment, an automated method has been established and validated based on hand-wrist radiographs. Hence, both fields were explored and conclusions were drawn for the current study design.

Developmental stage assessment of teeth

De Tobel et al. [5] investigated different algorithms for the automated classification of the lower left third molar into its developmental stages. Their deep learning approach was superior compared with other algorithms. The OPGs were preprocessed using contrast-limited adaptive histogram equalization (CLAHE) [26]. The pre-trained AlexNet [18] CNN architecture was retrained on a small dataset of 400 rectangular ROIs, carefully extracted by experts from their corresponding OPGs. The authors did not report the age range of the study population, but the entire developmental span of the third molar was covered. They reported a mean accuracy of 51%, a mean absolute error (MAE) of 0.60 stages, and a mean linearly weighted kappa (LWK) of 0.82. Most misclassifications were found in neighboring stages.

In a follow-up study, Merdietio et al. [24] investigated the added value of manual third molar segmentations for stage classification of the lower left third molar. On the same data as used by De Tobel et al. [5], contours of the lower left third molar were manually delineated using two approaches: Rough Segmentation (RS) and Full Segmentation (FS). Both methods removed the information around the tooth, which might confuse the staging. The process of evaluating and segmenting OPGs requires a dental expert to determine whether the surrounding anatomical parts provide any assistance in stage allocation depending on the alveolar eruption. The FS approach provided a high-quality segmentation; however, it was tedious and the time spent for each segmentation was 8 min on average. The average time for the RS approach was 5 min and a bounding box around the third molar was obtained in 2 min. Using a DenseNet201 [12] CNN, inclusion of the FS third molar segmentations improved the stage classification accuracy from 54 to 61%, MAE decreased from 0.61 to 0.53 stages, and LWK improved by 0.02 compared with only rectangular ROI information.

Yuma Miki et al. [25] also utilized the AlexNet [18] CNN to classify ROIs, extracted from 52 dental cone-beam computed tomography images into seven tooth types. First, the smallest possible bounding box enclosing a tooth was placed manually on the CT volume. Then, the middle 60% axial ROIs were used as input for the CNN. The average classification accuracy was 89% and was comparable with a conventional non-deep learning method used by Hosntalab et al. [11]. Although the results of these studies were promising, the possibility to

automatically classify the developmental stages directly from the presented OPGs remained unexplored.

Skeletal age assessment based on hand-wrist radiographs

Spampinato et al. [34] were the first to conduct research on automated skeletal bone age assessment using deep learning. They tested several approaches on a public dataset: (i) a CNN pre-trained on ImageNet [29] was used in a regression framework; (ii) a fine-tuning of a pre-trained CNN; (iii) an ad hoc CNN, BoNet [34], trained from scratch. The assessment was conducted on the public Digital Hand Atlas Database System (DHADS) [8] containing 1391 radiographs of the left hands of children up to the age of 18 years. Compared with the bone age obtained by human observers, they reported an MAE of 1.15, 0.80, and 0.79 years for the three approaches, respectively. The latter two outperformed state-of-the-art methods from previous years.

Larson et al. [19] tested a pre-trained deep residual CNN with 50 layers for age assessment from left hand radiographs. Their approach showed similar results compared with human observers. The root mean squared error (RMSE) on the DHADS was 0.73 years, slightly worse than the RMSE of 0.61 obtained by BoneXpert [40].

Lee et al. [21] developed an automated system for bone age assessment from radiographs of left hand and wrist containing the following steps: (i) the LeNet-5 [20] CNN was utilized for image segmentation to remove redundant information around the hand; (ii) a classification CNN pre-trained on ImageNet was applied. A mean accuracy of 57% and 60%, and an RMSE of 0.93 and 0.82 years, was obtained for males and females respectively. These numbers are somewhat comparable with the upper limits of the inter-observer variation obtained with the Greulich and Pyle (GP) method [9] in baseline Korean research [16].

Iglovikov et al. [13] also presented an automated framework for bone age assessment. They applied deep learning to a dataset of left hand radiographs, labelled by pediatric radiologists from a pediatric bone age challenge. First, radiographs were segmented using a U-Net-like [28] CNN. They normalized image contrast and aligned hands by detecting key points with VGG-net [33]. Both regression and classification CNNs from the VGG-net family of CNNs were applied, with classification CNNs slightly outperforming regression CNNs. An ensemble of regional CNNs showed superior performance with an MAE of 0.51 years. This result outperformed the state-of-the-art BoneXpert software with 0.65 years and the work by Lee et al. [21], thereby obtaining an accuracy comparable with human observer performance.

Study rationale and aim

Although well performing software for automated age assessment based on hand-wrist development exists, the implementation into forensic practice may be insufficient. Indeed, hand-wrist development ceases around the age of 18 [9], while in most countries, that age is the threshold from childhood to adulthood [31]. Thus, ideally, an age indicator is required that helps to discern minors from adults. Therefore, international guidelines state that besides development of the hand-wrist, also third molars and the clavicles should be taken

into account [30]. In the current study, our focus was on third molars, whose developmental span has been described to start around the age of 7 and end around the age of 21 [22]. With the upper end of the age range beyond the threshold of 18, this anatomical site holds the potential to better differentiate between minors and adults compared with the hand-wrist. The current study aimed to develop a fully automated system to classify a third molar into its developmental stage.

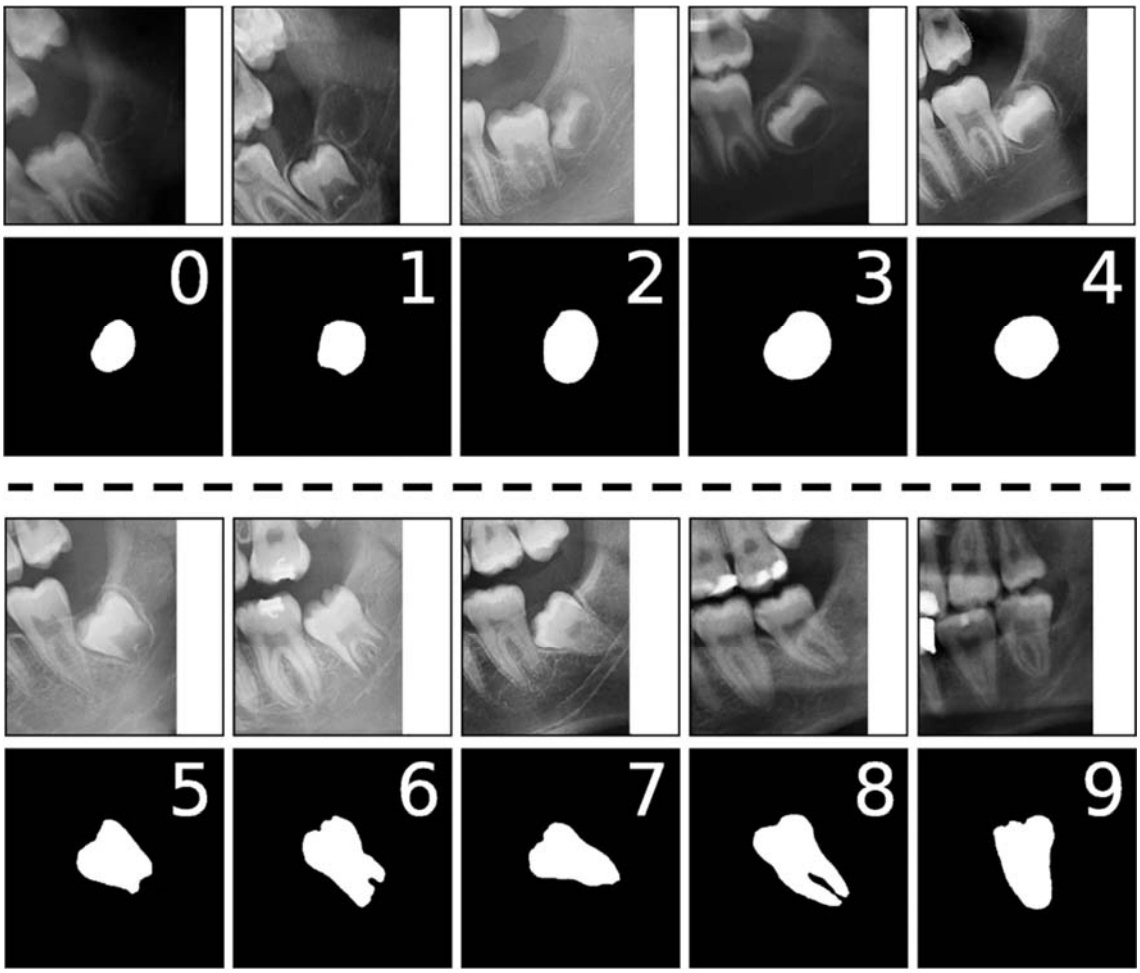

Fig. 1. Representative example of each of the 10 developmental stages of the lower left third molar (top rows) and their manual full segmentations according to Merdietio et al. [24] (bottom rows)

Materials and methods

Dataset

To develop and train the fully automated deep learning based system, a dataset of annotated OPGs was required. The dataset of OPGs was collected at the University Hospitals Leuven, Belgium, and was first used by De Tobel et al. [5] and later updated by Merdietio et al. [24]. The dataset consisted of 400 OPGs of varying sizes, with 20 OPGs per sex and per developmental stage of the lower left third molar, resulting in a chronological age range between 7 and 24 years. These OPGs were selected in consensus between three observers

as follows. Developmental stages were allocated to each lower left third molar by two observers, R.M.B. and J.D.T., with 2 and 7 years of experience evaluating dental radiographs, respectively. A third observer, P.W.T., had over 10 years of experience in dental age estimation, and acted as a referee. The developmental stages were allocated corresponding to a modified Demirjian et al. [6] staging technique proposed by De Tobel et al. [5], obtaining a total number of 10 ordinal developmental stages (i.e., 0 to 9; Fig. 1—top rows). Reaching the consensus stage was reported to take up to 15 min per case, depending on the difficulty and thus expertise required. The OPGs of different sizes and resolutions were cropped and resampled automatically to a common size of 1600 × 800 pixels. This cropping and resampling account for the appearance of white and black spaces in some of the resulting ROIs (shown as all-white spaces for illustrative purposes).

In order to evaluate the performance of the current method, fivefold cross-validation was used. Therefore, the dataset was randomly split into five equally large validation sets of 80 OPGs with four OPGs per sex and per developmental stage of the lower left third molar. In each fold, the remaining 320 OPGs were used to train the CNNs from the proposed procedure.

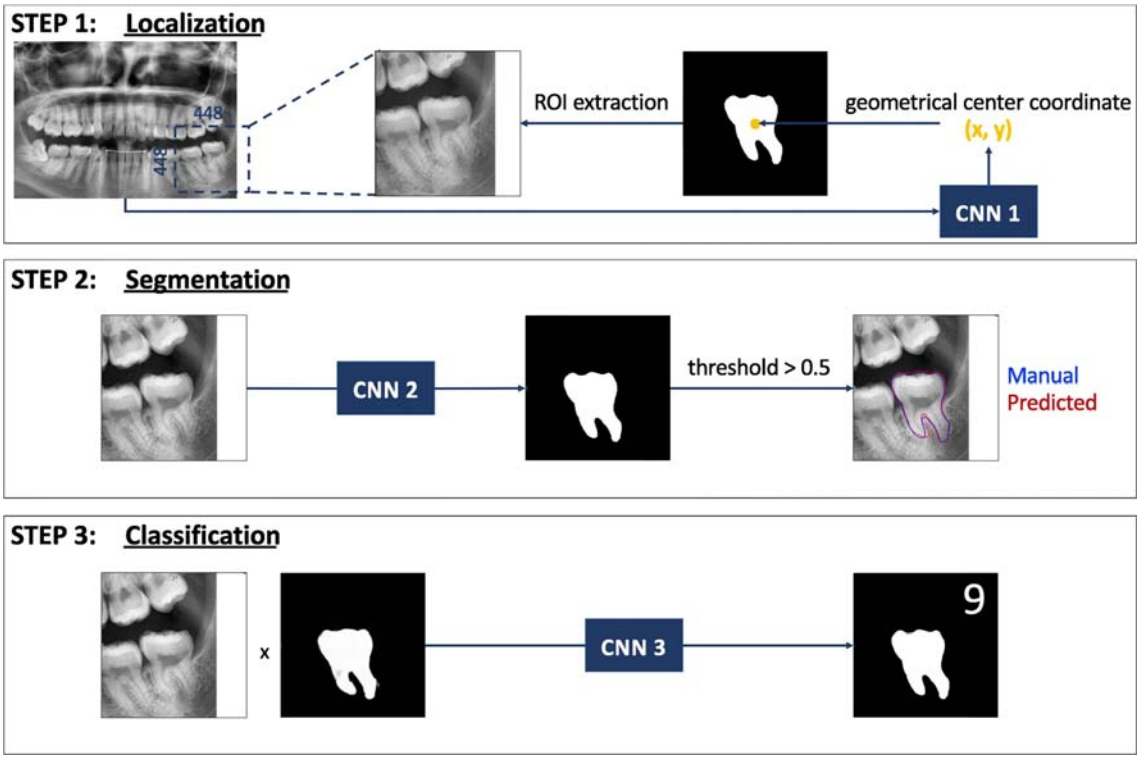

Fig. 2. A schematic overview of the proposed three-step procedure to automate third molar development staging. STEP 1: A first CNN detects a rectangular ROI around the third molar under assessment. STEP 2: Another CNN segments the third molar out of the established ROI. STEP 3: A final CNN combines the third molar's ROI and its segmentation to classify the third molar's developmental stage

Three-step procedure

Based on recent work [5, 13, 19, 21, 24, 34], the following three-step procedure was proposed as presented in Fig. 2. First, a CNN estimated the center of the lower left third molar in the OPG and defined a fixed region of interest (ROI) around it. Second, another CNN segmented the third molar within the proposed ROI. Third, a final CNN combined the image content within this ROI and the associated automated segmentation to classify the third molar's developmental stage.

Third molar localization

The first step (Fig. 2—top) automatically extracts a bounding box or ROI around the lower left third molar. The training ROIs were defined as a 448 × 448 pixels bounding box parallel to the image axes and centered around the geometrical center of the manual full segmentations (FS) as described in [24] (Fig. 1—bottom row, annotated in yellow in Fig. 2). In contrast to [5, 24], where training ROIs were carefully aligned with the third molar, here, larger unaligned ROIs (in combination with rotational augmentation) were used to trade off spatial noise (e.g., surrounding teeth can trick the final staging) and localization performance (e.g., larger ROIs have higher chances to capture a minimal surface of the third molar). A YOLO-like [27] CNN architecture was therefore utilized with minor modifications. Each image was divided into 25 adjacent non-overlapping cells and the cell containing the third molar and its geometrical center within this cell was predicted. For feature mapping, the DenseNet201 [12] CNN architecture, pre-trained on the ImageNet [29] dataset with dense layers suitable for the problem, was used. The sum of two mean squared error (MSE) losses—one for cell classification and one for geometrical center regression— was defined as objective function to be minimized. This combined loss function was optimized for 10 epochs using the Adam optimizer [17] with default Keras [4] settings. The mean absolute error (MAE; Eq. 1) and mean Euclidean distance (MED; Eq. 2) between the manual and predicted center coordinates in pixels were calculated:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}(|y_i-\hat{y}_i|+|x_i-\hat{x}_i|)}{N}, \quad (1)$$

$$\mathrm{MED} = \frac{\sum_{i=1}^{N}\sqrt{(y_i-\hat{y}_i)^2+(x_i-\hat{x}_i)^2}}{N}, \quad (2)$$

where (x_i, y_i) and ($\hat{x}_i$, $\hat{y}_i$) refer to the manually annotated and predicted coordinates, respectively, of the geometrical center of the lower left third molar in the i th OPG, and N = 400 refers to the total number of OPGs. A qualitative localization measure of "good", "poor," or "wrong" was further defined when the predicted ROI overlapped with the manual segmentation of the third molar completely, partly or not, respectively.

Third molar segmentation

The second step (Fig. 2—middle) automatically segmented the lower left third molar, given a 448 × 448 bounding box around its geometrical center. For this purpose, a U-Net-like [28] CNN architecture was used. This model has been proven to work well across many

segmentation tasks in medical imaging [2, 14, 28]. It processes the input image by successively applying linear convolutions using kernels of size 3 × 3 and non-linear leaky-ReLU [23] activations. The latter are necessary to avoid creating linear, and thus simple, features only. Before final classification (segmentation can be seen as classifying each pixel as being foreground or background), this successive pattern should result in local and global features that are informative for the state of a certain pixel.

In order to train the CNN, its internal parameters need to be optimized with respect to a certain loss function (i.e., the optimization objective), which directly compares the automatic and manual segmentations. For segmentation, cross-entropy (CE), soft Dice (SD), or their linear combination (CS) are often used [2]. Here, a suitable loss function was identified on the manual ROIs and their associated segmentations by analyzing the segmentation performance in terms of pixel-wise accuracy (3), precision (4) and recall (5), and the Dice score (6) [2, 35]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{5}$$

$$\text{Dice} = \frac{2TP}{2TP+FP+FN}, \tag{6}$$

where TP, TN, FP, and FN refer to the pixels labelled correctly as third molar or background, or incorrectly as third molar or background, respectively. This way, accuracy represents the proportion of pixels classified correctly, precision represents the fraction of pixels being classified as tooth correctly, and recall represents the fraction of tooth pixels that are correctly identified as being tooth. The Dice score is a commonly used intersection-over-union measure used to compare two segmentations (here manual and predicted) [2].

The Adam optimizer [17] with default Keras [4] settings for 150 epochs was found to work well for convergence. The initial learning rate was set at 10^{-3} and reduced by a factor of 10 every 50 epochs. These settings were found by visually inspecting the Dice score on the validation set for the first fold only. Convergence was defined successful when the Dice score had plateaued (i.e., no increase anymore due to a sufficient number of epochs and an appropriate decay scheme, and no decrease yet, which could point to over-fitting). In a second experiment, the most promising loss function was chosen to work directly forward on the localization output. Finally, pixel-wise accuracy, precision and recall, and the Dice score were calculated for each stage individually and with or without localization outliers.

Third molar classification

The third step (Fig. 2—bottom) automatically classifies the lower left third molar into its developmental stage. Hence, given the bounding box, the task is to classify the ROI into one of 10 developmental stages. First, experiments were conducted with the manual ROIs and segmentations with two CNNs: a simple ad hoc CNN with 10 layers and the more complex

DenseNet201 [12] (as it was used in [24]). Both CNNs process the information in a similar way and with the same principles as explained in the previous section for U-Net. Before final classification—in this case multi-class staging—informative image features should have been derived. The more complex the CNN, the more complex patterns it could detect in the input images but the more data is generally needed in order to detect generalized features [1]. Comparing the results of the simple CNN with the results of DenseNet201 shedded light on the interplay between these two aspects for this particular dataset.

Apart from the CNN used for classification, experiments were conducted with three types of input, as a way to incorporate the available information: the ROI only (NO), and the ROI and segmentations concatenated (CO) or multiplied (MU). Finally, the most promising of those methods trained on manual ROIs and segmentations was chosen and staging accuracy, MAE and LWK on the predicted ROIs and segmentations from the previous steps for each stage individually with and without localization outliers were reported. These are frequently used metrics when evaluating staging performance and their definition can be found in [5].

For all experiments, the parameters of the CNN were optimized with respect to the CE loss function for the training set using stochastic gradient descent (SGD) for 150 epochs. The initial learning rate was set at 10^{-3} and reduced by a factor of 10 every 50 epochs.

Results

The reported results were based on fivefold cross-validation of the complete procedure. For third molar segmentation and staging, the effect of different parameter setups was studied, using the manual input (ROI and segmentation) as training information. Subsequently, for the optimal parameter setup, results of working with manual information were compared with results obtained when using the output(s) of the automated three-step procedure.

Third molar localization

The automated localization, including extracting and storing of the rectangular ROI, took 2.08 s on average per image. In Table 1, the localization results are given. The geometrical center of the third molar was localized with a MAE of 79 pixels and a MED of 63 pixels. There was a trend for the detection algorithm to work better for the middle stages. There were 393 good, 3 poor, and 4 wrong localizations. In Fig. 3, one example of each is shown. Poor localizations were possible because of the following reasons: (i) the correct cell (i.e., the left side of the patient) is misclassified and the coordinates are predicted relative to the wrong cell; (ii) the regression prediction is not bounded to the cell and, hence, it may lead to the coordinates located far from the correct cell.

Table 1 Quantitative results of the automated detection

Metric ↓	Stage →	0	1	2	3	4	5	6	7	8	9	All	All*
MED (pix.)	mean	77	89	56	52	48	47	57	52	58	62	60	63
	std	30	35	31	24	25	22	36	26	33	34	32	47
	min	16	20	12	14	6	3	14	20	8	14	3	3
	max	137	142	123	125	108	95	199	108	188	178	188	492
MAE (pix.)	mean	96	112	68	65	60	59	73	68	75	79	75	79
	std	38	45	37	32	30	30	48	35	46	42	41	59
	min	18	28	16	15	8	3	18	21	9	16	3	3
	max	177	197	143	147	121	133	254	149	266	230	266	620

"Wrong" cases are excluded from the results. Only the last column corresponds to the average performance with inclusion of wrong cases

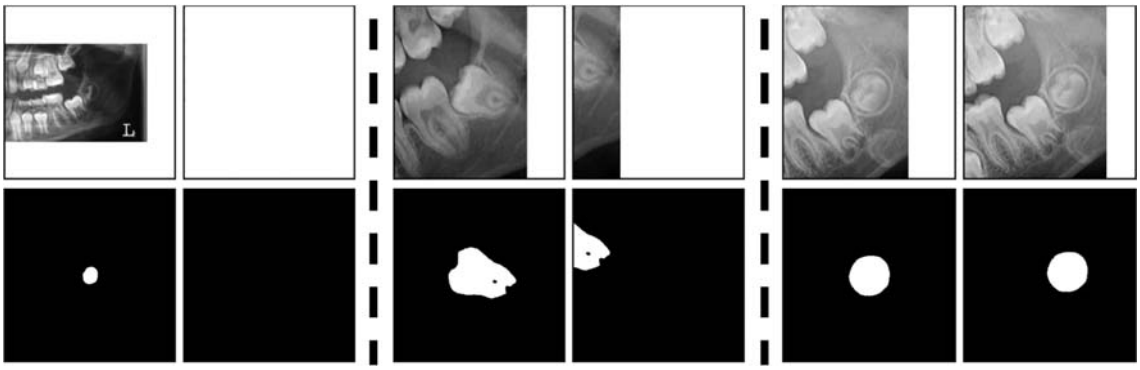

Fig. 3. Qualitative examples of detection (top row) and corresponding manual full segmentation (bottom row). The left column in each big box corresponds to the manual detection and the right column corresponds to the automated detection. The left two columns correspond to a "wrong" automated detection. In this case, the corresponding manual segmentation results in an empty ROI. The middle two columns correspond to a "poor" detection, where only part of the segmentation is retained within the automatically detected ROI. The right two columns correspond to a "good" detection since the entire segmentation is captured

Third molar segmentation

The automated segmentation, including storing the segmentation masks, took 0.13 s per image on average. In Table 2, the segmentation results are given for both the manual ROIs and the automatically predicted ROIs from the previous step. All cases were included in the calculation. Considering the manual ROIs, the linear combination (CS) of cross-entropy (CE) with soft Dice (SD) loss performed slightly superior compared with the single losses (left side of Table 2). The use of this loss function for the predicted ROIs revealed a slight decrease compared with the performance on manual ROIs (right side of Table 2). Zooming in on the results for each stage individually highlights an inferior performance for stage 0. In Fig. 4, segmentation results are illustrated for each stage, per quartile of the Dice score on the predicted ROIs.

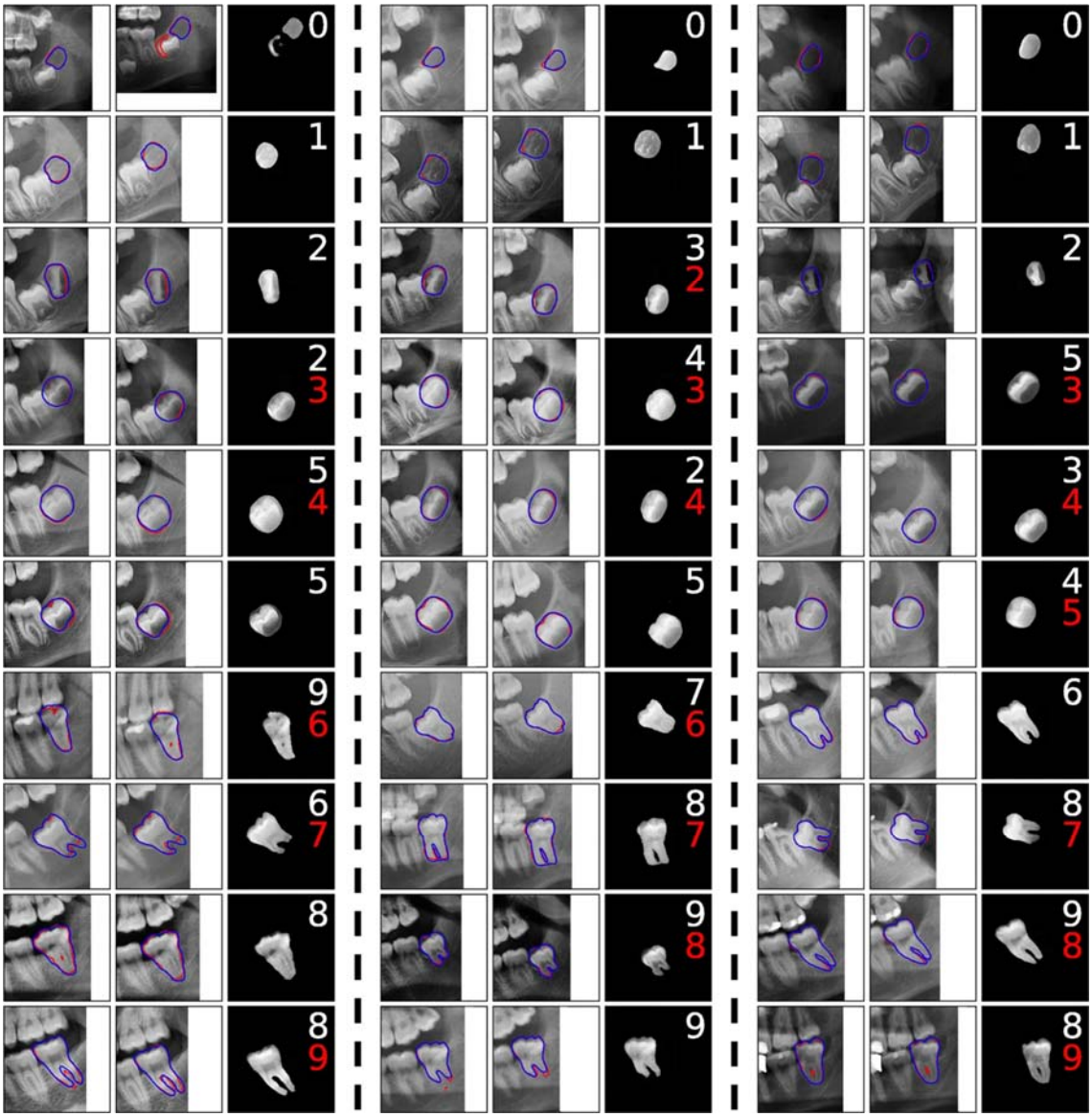

Fig. 4. Qualitative segmentation examples for each of the 10 developmental stages (rows; ordered top-bottom stage 0–stage 9). Left three, middle three, and right three columns are the results for the first, second, and third quartile of the Dice score using the predicted segmentations, respectively. The manual ROI is given first, followed by the automatically predicted ROI, and finally the predicted ROI multiplied by the predicted segmentation. The blue contour of the predicted segmentation is overlaid on the red contour of the manual segmentation. When only blue is visible, the delineation is (almost) perfect. The resulting fully automated stage prediction is annotated in white, while red indicates the manual stage when the predicted stage was incorrect. Although the ROI was localized and the third molar segmented successfully in most cases, this was not true for the ROI in the first row, second column. It can be assumed that the automated segmentation is more difficult for the initial third molar stages because they lack (or present minimal) calcified tooth parts

Table 2 Quantitative results of the automated segmentation

	Exp. →	Manual ROIs			Predicted ROIs										
	Loss →	CE	SD	CS	CS										
Metric ↓	Stage →	All	All	All	0	1	2	3	4	5	6	7	8	9	All
Accuracy (%)	mean	99	99	99	99	99	99	99	100	99	99	99	99	99	99
Precision (%)	mean	95	95	95	88	95	92	97	97	95	96	95	96	96	95
Recall (%)	mean	94	94	95	87	92	97	94	95	90	96	94	93	93	93
Dice (%)	mean	94	94	94	85	93	94	95	96	91	96	95	93	93	93
	std	7	7	7	19	4	4	3	2	16	3	3	15	15	11
	min	0	0	8	0	80	82	88	92	17	84	83	2	1	0
	max	99	99	99	100	98	100	99	98	98	98	100	98	100	100

On the manual ROIs and corresponding segmentations, three different loss functions (i.e., cross-entropy (CE), soft Dice (SD) are tested and their sum (CS)) and the overall result are reported. For CS, the results are analyzed for the predicted ROIs both overall and for each stage separately

Third molar staging

The automated staging, including combining the ROI and segmentation mask, took 0.51 s per image on average. In Table 3, the staging results are given for manual ROIs with manual segmentations and the automatically predicted ROIs with predicted segmentations from the previous steps. All cases were included in the calculation and the segmentation masks were combined with the detected ROIs in the following ways: (i) concatenation (CO); (ii) multiplication (MU); (iii) no combination (NO); hence, only the detected ROI is used. Looking at the results on the manual ROIs with manual segmentations (left side of Table 3), there was a clear increase in performance when the segmentation is used (CO and MU compared with NO). MU seemed to deliver the most promising results. Using this method for the staging on predicted ROIs with predicted segmentations showed a drop in overall staging performance (right side of Table 3) compared with manual results. Zooming in on the results for each stage individually highlights the superior performance for lower stages. In Table 4, the results for MU for fully automated predictions are presented as a confusion matrix. In Fig. 4, the staging output is illustrated for the segmentation examples.

Table 3 Quantitative results of the automated staging

	Exp. →	Manual segm's			Predicted segmentations										
	Method →	NO	CO	MU	MU										
Metric ↓	Stage →	All	All	All	0	1	2	3	4	5	6	7	8	9	All
Accuracy (%)	mean	55	57	60	85	85	63	45	50	43	65	20	35	45	54
MAE (stages)	mean	0.62	0.58	0.51	0.55	0.25	0.62	0.60	0.54	0.75	0.67	1.20	0.78	0.98	0.69
	std	0.90	0.82	0.79	1.75	0.73	1.23	0.58	0.59	0.80	1.02	1.15	0.68	1.33	1.08
	max	6	7	8	9	4	7	2	2	3	3	7	2	7	9
LWK (%)	mean	0.81	0.83	0.84	/	/	/	/	/	/	/	/	/	/	0.79

On the manual ROIs and corresponding manual segmentations, three different combination types are tested (i.e., no combination and thus only use of ROI (NO), concatenation (CO), and multiplication (MU; this type of combination is performed in Fig. 4)) and the overall result are reported. For MU, the results are analyzed for the predicted ROIs with corresponding predicted segmentations both overall and for each stage separately

Table 4 Cross tabulation of allocated stages by the fully automated system (rows) and by the human observers as a consensus stage (columns)

		Manual									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Automated	0	*0.85*	0.10	0	0	0	0	0	0.025	0	0
	1	0.075	*0.85*	0.05	0	0	0	0	0	0	0
	2	0	0	*0.625*	0.2	0.05	0.025	0	0	0	0.025
	3	0	0.025	0.25	*0.45*	0.25	0.05	0	0	0	0
	4	0	0	0.025	0.3	*0.5*	0.4	0	0	0	0
	5	0.05	0.025	0.025	0.05	0.2	*0.425*	0	0.025	0	0.025
	6	0	0	0	0	0	0.05	*0.65*	0.225	0.15	0
	7	0	0	0	0	0	0.025	0.15	*0.2*	0.325	0.2
	8	0	0	0	0	0	0.025	0.1	0.3	*0.35*	0.3
	9	0.025	0	0.025	0	0	0	0.1	0.225	0.175	*0.45*

Normalized by the total number of samples N = 400 as to represent fractions

Discussion

Situation of findings in literature

Automated methods for skeletal age estimation have been used for over a decade [40]. Recently, the RSNA Pediatric Bone Age Challenge using hand radiographs demonstrated that different approaches to process the images can render similar results [10]. Although localization and segmentation seemed to be commonly used in the automated approaches, stage classification—as is done by human observers—was not described by most automated systems. Nonetheless, one might hypothesize that adding the stage classification step might further ameliorate age estimation performance. Starting from the stage classification, the automated method may only need to interpret and further classify the sequence of developmental changes within the considered stage, which might reduce the computational burden of the automated system.

With a stage classification accuracy of 54 %, an MAE of 0.69 stages, and a LWK of 0.79, the current fully automated system for stage classification performed inferior compared with respectively 61%, 0.53 stages and 0.84 reported for the semi-automated system proposed by Merdietio et al. [24]. They only automated the final step (i.e., stage classification) while tooth localization and segmentation were done manually, and where the latter is generally considered tedious and prone to observer variability. More specifically, the fully automated system took 2.72 s to compute on average. This is substantially faster than manual staging with or without qualitative segmentations, respectively taking more than 2 min or 10 min. However, further optimization of all steps in the automated system is recommended before its final application in forensic age estimation practice.

Moreover, before being applied in practice, the next step that needs to be added to the proposed automated system is the age estimation step itself. Regarding skeletal age assessment, the lowest MAE reached in the RSNA challenge was 4.26 months (= 0.36 years) [10], based on the automated assessment of a hand-wrist radiograph. Assessing hand-wrist MRI, Tang et al. [39] reported MAEs of 0.13 years for males and 0.08 for females. However, their study population was very small, with only 79 individuals. Moreover, they only

included participants between 12 and 17 years old, while in forensic age estimation studies, a sufficient portion of the study population should be well over 18. Unfortunately, also the population of the RSNA challenge only included a very small portion of adults. By contrast, Stern et al. [38] studied hand-wrist MRIs of males between 13 and 25 years old. They reported an MAE of 0.34 years in their total population, and 0.53 years in participants ≤ 18 years. Note that the reported MAEs in [10, 38, 39] were errors between the automatically estimated age and the bone age determined by radiologists. Conversely, in forensic age estimation, the errors between estimated age and chronological age are relevant. In their pilot paper, Stern et al. [36] reported an MAE of 0.85 years compared with chronological age, when assessing hand-wrist MRI. More recently, Stern et al. [37] combined hand-wrist MRI with clavicles and third molars MRI, obtaining an MAE of 1.01 years. The larger error in the more recent paper might seem unexpected but can be explained by differences in study population: N = 56 and age 13–19 years in [36], N = 322 and age 13–25 years in [37]. Thus, the latter study is more relevant to forensic age estimation. Moreover, it is the only one presenting a fully automated system for dental age estimation in adolescents and young adults, albeit embedded in the multi-factorial system. Unfortunately, the studies by Stern et al. only included men, which poses their major shortcoming.

Limitations and future prospects

The proposed three-step procedure for fully automated staging of the lower left third molar has some shortcomings, which lend themselves for improvement and should be addressed in future studies. First, the OPGs were of different sizes and resolutions. Therefore, the OPGs have white and black spaces due to resampling and cropping (in all figures shown as all-white spaces for illustrative purposes). This strategy might be considered suboptimal and may have led to incorrect predictions further downstream (e.g., the “wrong” localization in Fig. 3). Second, the ROIs used in this work were quite large and not aligned, as opposed to the ones used by De Tobel et al. [5] and Merdietio et al. [24]. This was necessary to alleviate poor localization performance and retain sufficient segmentation area within the ROI (as to reduce the number of “wrong” and “poor” localizations in Fig. 3). Although a similar performance was obtained (note the results for manual segmentations using the MU method in Table 3), which justified our choice, a better localization step is necessary and may lead to an improved performance due to expected superior segmentations (e.g., partly false segmentation in Fig. 4—top left).

In work by Unterpirker at al. [41], a localization error of 3.55 ± 2.62 mm was reported when detecting third molars as landmarks on MRI and using random regression forests (RRFs). A further optimization of the currently used localization step might be to predict the third molar’s location based on anatomical landmarks of other structures. To achieve this, skeletal landmarks seem more suitable than dental landmarks, since the former are broadly constant between individuals (e.g., the presence of the inferior alveolar nerve and the foramen mentale), while the latter are highly variable (e.g., extractions, restorations, tooth movement). In recent work, Vinayalingam et al. [42] use the location of the inferior alveolar nerve relative to the roots of lower third molars to study risk assessment of third molar removal. In another study by Ebner et al. [7], a two-step procedure was proposed with a landmark localization algorithm also using RRFs in hand MRI. Their two-step procedure included a coarse RRF estimation followed by a refined estimation of the landmark. Large

anatomical variations were found on radius and ulna, creating the highest mean error of the evaluated hand MRI. The landmark was chosen based on a constraint on the surrounding structures. This process could however have limitations when applied to third molars, due to the large anatomical variation. Hence, choosing a consistent anatomical landmark will affect the localization process and its quest is left for future research.

In light of stage classification, only simple combinations of concatenation and multiplication were tested regarding the combination of the ROI with segmentation information, following the research by Merdietio et al. [24]. It may well be that more advanced strategies lead to superior performance compared with the early-fusion strategy explored in this work [15]. Discerning adjacent stages—especially near the end of development (stages 7 to 9)—remains a challenging task, even for an automated deep learning approach. Nonetheless, those final stages occur around the age of 18, making them especially relevant in forensic age estimation, when minors need to be discerned from adults. Thus, further optimization of the classification step is desirable, which can only be achieved by adding more training data. This will affect the learning process for stage classification directly, as well as indirect improvement due to related ameliorations in the automated segmentation.

It is clear that multiple factors may have led to an inferior performance compared with the results in Merdietio et al. [24] with all manual information. However, given that the entire workflow of detecting, segmenting, and staging a third molar has been automated, we believe these results are promising and ready to be used before integrating the obtained third molar stage into an age assessment model. An interesting part of future research will be to transfer this procedure to all third molars (i.e., 18, 28, and 48), and possibly to other developing permanent teeth in younger individuals. Thus, increasing the number of age indicators, which in the end might increase age estimation performance. Similarly, in older individuals, several degenerative changes (e.g., secondary dentin, periodontosis, root resorption) might be detected automatically, and their information might be combined automatically to derive an age estimate.

Thus, future research should focus on complementing the proposed three-step procedure with an age estimation, rendering a comprehensive four-step procedure. The reference population for such a study needs to represent all relevant age categories uniformly. Recommendations state that at least ten individuals per sex per age category of 1 year need to be included [32]. However, to train a deep CNN for age estimation, the reference population should be as large as possible. For instance, BoneXpert was based on 1559 hand/wrist radiographs [40]. Although the numbers of cases per age category were not specified, the graphs in their original paper reflect a more or less uniform age distribution.

Conclusion

In this work, we proposed and validated a fully automated three-step procedure for third molar staging, directly starting from OPGs. The automated system outputs a third molar's stage in under 3 s, which is substantially faster than manual stage allocation by experts. Taking into account the limited dataset size, this fully automated approach shows promising results compared with manual staging.

Acknowledgments

The computational resources were partly provided by the Flemish Supercomputer Center (VSC).

Funding

This work is funded in part by Internal Funds KU Leuven (grant no. C24/18/047).

References

1. Berrada L, Zisserman A, Kumar MP (2018) Smooth loss functions for deep top-k classification. In: 6Th international conference on learning representations, ICLR 2018 - conference track proceedings
2. Bertels J, Eelbode T, Berman m, Vandermeulen D, Maes F, Bisschops R, Blaschko M (2019) Optimizing the dice score and jaccard index for medical image segmentation: theory & practice. In: MICCAI 2019. Springer, Verlag
3. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin A, Do BT, Way GP, Ferrero E, Agapow PM, Xie W, Rosen GL et al (2017) Opportunities and obstacles for deep learning in biology and medicine. bioRxiv. pp 142760
4. Chollet F et al (2015) Keras. https://keras.io
5. 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 Odonto-Stomatol 35(2):49–60
6. Demirjian A, Goldstein H, Tanner J (1973) A new system of dental age assessment. Human Biol 45:2:211–227
7. Ebner T, Stern D, Donner R, Bischof H, Urschler M (2014) Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. Medical image computing and computer-assisted intervention: MICCAI. Int Conf Med Image Comput Comput-Assist Intervent 17(331239):421–428
8. Gertych A, Zhang A, Sayre J, Pospiech-Kurkowska S, Huang H (2007) Bone age assessment of children using a digital hand atlas. Comput Med Imaging Graph 31(4-5):322–331
9. Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of the hand and wrist. Amer J Med Sci 238(3):393
10. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT, Abdala N, Kitamura FC, Thodberg HH, Chen L, Shih G, Andriole K, Kohli MD, Erickson BJ, Flanders AE (2019) The rSNA pediatric bone age machine learning challenge. Radiology 290(3):498–503. https://doi.org/10.1148/radiol.2018180736
11. Hosntalab M, Zoroofi RA, Tehrani-Fard AA, Shirani G (2010) Classification and numbering of teeth in multi-slice ct images using wavelet-fourier descriptor. Int J Compute Assist Radiol Surgery 5(3):237–249
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: CVPR, vol 1, pp 3
13. Iglovikov V, Rakhlin A, Kalinin A, Shvets A (2017) Pediatric bone age assessment using deep convolutional neural networks. arXiv:1712.05053

14. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH (2018) No New-Net. LNCS. arXiv:1809.10483
15. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 1725–1732. https://doi.org/10.1109/CVPR.2014.223. https://www.computer.org/csdl/proceedings/cvpr/2014/5118/00/5118b725-abs.html
16. Kim SY, Oh YJ, Shin JY, Rhie YJ, Lee KH (2008) Comparison of the greulich-pyle and tanner whitehouse (tw3) methods in bone age assessment. J Korean Soc Pediat Endocrinol 13(1):50–55
17. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
18. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
19. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP (2017) Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 287 (1):313–322
20. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
21. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S (2017) Fully automated deep learning system for bone age assessment. J Digit Imaging 30(4):427–441
22. Liversidge HM (2008) Timing of human mandibular third molar formation. Ann Hum Biol 35(3):294–321. https://doi.org/10.1080/03014460801971445
23. Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In ICML Workshop on Deep Learning for Audio, Speech and Language Processing, pp 28
24. Merdietio Boedi R, Banar N, De Tobel J, Bertels J, Vandermeulen D, Thevissen PW (2019) Effect of lower third molar segmentations on automated tooth development staging using a convolutional neural network. J Forensic Sci 14182:1556–4029. https://doi.org/10.1111/1556-4029.14182
25. Miki Y, Muramatsu C, Hayashi T, Zhou X, Hara T, Katsumata A, Fujita H (2017) Classification of teeth in cone-beam ct using deep convolutional neural network. Comput Biol Med 80:24–29
26. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, ter Haar Romeny B, Zimmerman JB, Zuiderveld K (1987) Adaptive histogram equalization and its variations. Comput Vis Graph Image Process 39(3):355–368
27. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 779–788. https://doi.org/10.1109/CVPR.2016.91
28. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241

29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
30. Schmeling A, Dettmeyer R, Rudolf E, Vieth V, Geserick G (2016)) Forensic Age Estimation. Deutsch Arzteblatt Int 113(4): 44–50. https://doi.org/10.3238/arztebl.2016.0044. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4760148/pdf/Dtsch_Arztebl_Int-113-0044.pdfhttps://goo.gl/Qvw66p
31. Schmeling A, Geserick G, Reisinger W, Olze A (2007) Age estimation. Forensic Sci Int 165(2-3):178–181. https://doi.org/10.1016/j.forsciint.2006.05.016. https://linkinghub.elsevier.com/retrieve/pii/S0379073806003173
32. Schmeling A, Grundmann C, Fuhrmann A, Kaatsch HJ, Knell B, Ramsthaler F, Reisinger W, Riepert T, Ritz-Timme S, Rösing FW, Rötzscher K, Geserick G (2008) Criteria for age estimation in living individuals International Journal of Legal Medicine. https://doi.org/10.1007/s00414-008-0254-2
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
34. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R (2017) Deep learning for automated skeletal bone age assessment in x-ray images. Med Image Anal 36:41–51
35. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Det Kongelige Danske Videnskabernes Selskab V(4):1–34
36. Stern D, Ebner T, Bischof H, Grassegger S, Ehammer T, Urschler M (2014) Fully automatic bone age estimation from left hand MR images. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention. https://doi.org/10.1007/978-3-319-10470-6_28
37. Štern D, Payer C, Giuliani N, Urschler M (2019) Automatic age estimation and majority age classification from multi-factorial MRI data. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2018.2869606
38. Štern D, Payer C, Urschler M (2019) Automated age estimation from MRI volumes of the hand. Medical Image Analysis. https://doi.org/10.1016/j.media.2019.101538
39. Tang FH, Chan JL, Chan BK (2019) Accurate age determination for adolescents using magnetic resonance imaging of the hand and wrist with an artificial neural Network-Based approach. Journal of Digital Imaging. https://doi.org/10.1007/s10278-018-0135-2
40. Thodberg HH, Kreiborg S, Juul A, Pedersen KD (2009) The bonexpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 28(1):52–66
41. Unterpirker W, Ebner T, Stern D, Urschler M (2015) Automatic third molar localization from 3D MRI using random regression forests. In: Lambrou T, Ye X (eds) Proceedings of the 19th Conference on Medical Image Understanding and Analysis. The British Machine Vision Association, UK, pp 195–200
42. Vinayahalingam S, Xi T, Bergé S, Maal T, de Jong G (2019) Automated detection of third molars and mandibular nerve by deep learning. Sci Rep 9(1):9007. https://doi.org/10.1038/s41598-019-45487-3