Open access • Journal Article • DOI:10.1007/S00414-020-02283-3

Towards fully automated third molar development staging in panoramic radiographs

Nikolay Banar, Jeroen Bertels, François Laurent, Rizky Merdietio Boedi ...+5 more authors

Institutions: University of Antwerp, Katholieke Universiteit Leuven, Diponegoro University, University of Pretoria

Published on: 01 Apr 2020 - International Journal of Legal Medicine (Springer Science and Business Media LLC)

Topics: Region of interest

This is a pre-print of an article published in International Journal of Legal Medicine. The final authenticated version is available online at: https://doi.org/10.1007/s00414-020-02283-3.

Fully automated third molar development staging in panoramic radiographs

Nikolay Banar · Jeroen Bertels · François Laurent · Rizky Merdietio Boedi · Jannick De Tobel · Patrick Thevissen · Dirk Vandermeulen

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, selected in consensus between three observers. The concepts of transfer learning, using pre-trained CNNs, and data augmentation were used to alleviate the burden of a limited dataset. In this work, a three-step procedure was proposed and the results were validated using five-fold 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.

N. Banar and J. Bertels have contributed equally to this work.

N. Banar
University of Antwerp - Computational Linguistics Group (CLiPS), Antwerp, Belgium
E-mail: nicolae.banari@uantwerpen.be

J. Bertels (corresponding author)
KU Leuven - Department of Electrical Engineering (ESAT/PSI), Leuven, Belgium
E-mail: jeroen.bertels@kuleuven.be

F. Laurent
KU Leuven - Department of Electrical Engineering (ESAT/PSI), Leuven, Belgium

R. Merdietio Boedi
Diponegoro University - Department of Dentistry, Semarang, Indonesia (10/2019 - Present)
KU Leuven - Department of Imaging and Pathology (Forensic Odontology), Leuven, Belgium (08/2017 - (07/2018)

J. De Tobel
KU Leuven - Department of Imaging and Pathology (Forensic Odontology), Leuven, Belgium

P. Thevissen
KU Leuven - Department of Imaging and Pathology (Forensic Odontology), Leuven, Belgium

D. Vandermeulen
KU Leuven - Department of Electrical Engineering (ESAT/PSI), Leuven, Belgium
University of Pretoria - Department of Anatomy - Pretoria, South-Africa

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 63.3 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 79 %. Taking into account the limited dataset size, this fully automated approach shows promising results compared to manual staging.

Keywords Dental age assessment · Third molar · Developmental stage · Localization · Segmentation · Classification.

1 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 to age. However, the manually performed staging's major drawback is a possible stage classification variability within and between observers. 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 [2]. However, related work in the field of automated dental age assessment 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.

1.1 Developmental stage assessment of teeth

De Tobel et al. [4] investigated different algorithms for the automated classification of the lower left third molar into its developmental stage. Their deep learning approach was superior compared to other algorithms. The OPGs were preprocessed using contrast-limited adaptive histogram equalization (CLAHE) [25]. The pretrained AlexNet [17] 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 82 %. Most misclassifications were found in neighbouring stages.

In a follow-up study, Merdietio et al. [23] investigated the added value of manual third molar segmentations for stage classification of the lower left third molar. In the same study population as De Tobel et al. [4], contours of the lower left third molar were manually delineated, removing the information around the tooth, which might confuse the staging. Using a DenseNet201 [11] CNN, segmented third molar information improved the stage classification accuracy from 54 % to 61 %, MAE

decreased from 0.61 to 0.53 stages, and LWK improved by 2 % compared to rectangular ROI information.

Yuma Miki et al. [24] also utilized the AlexNet [17] 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 to the non-deep learning method used by Hosntalab et al. [10]. Although the results of these studies were promising, the possibility to automatically classify the developmental stages directly from the presented OPGs remained unexplored.

1.2 Skeletal age assessment based on hand-wrist radiographs

Spampinato et al. [33] 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 [28] was used in a regression framework; (ii) a fine-tuning of a pre-trained CNN; (iii) an ad-hoc CNN, BoNet [33], trained from scratch. The assessment was conducted on the public Digital Hand Atlas Database System (DHADS) [7] containing 1391 radiographs of the left hands of children up to the age of 18 years. Compared to chronological age, 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. [18] tested a pre-trained deep residual CNN with 50 layers for age assessment from left hand radiographs. Their approach showed similar results compared to 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 [39].

Lee et al. [20] developed an automated system for bone age assessment from radiographs of left hand and wrist containing the following steps: (i) the LeNet-5 [19] 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 to the upper limits of the inter-observer variation obtained with the Greulich and Pyle (GP) method [8] in baseline Korean research [15].

Iglovikov et al. [12] 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 [27] CNN. They normalized image contrast and aligned hands by detecting key points with VGG-net [32]. 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 out-

performed the state-of-the-art BoneXpert software with 0.65 years and the work by Lee et al. [20], thereby obtaining an accuracy comparable to human observer performance.

1.3 Study rationale and aim

Despite well performing software for automated age assessment based on hand-wrist development, the implementation into forensic practice may be insufficient. After all, development ceases around the age of 18 [8], while in most countries, that age is the threshold from childhood to adulthood [30]. Thus, ideally, an age indicator is used 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 [29]. 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 [21]. With the upper end of the age range beyond the threshold of 18, this anatomical site holds the potential to differentiate better between minors and adults compared to the hand-wrist. The current study aimed to develop a fully automated system to classify a third molar into its developmental stage.

2 Materials and methods

2.1 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. [4] and later updated by Merdietio et al. [23]. The dataset consisted of 400 OPGs of varying sizes. These OPGs were selected in consensus between three observers, as to obtain 20 OPGs per sex and per developmental stage of the lower left third molar. Each stage was defined corresponding to a modified Demirjian et al. [5] staging technique proposed by De Tobel et al. [4], with a total number of 10 ordinal developmental stages (i.e. 0 to 9; Fig. 1 - top rows). The resulting study population ranged between 7 and 24 years old. The OPGs of different sizes and resolutions were cropped and resampled automatically to a common size of 1600x800 pixels. This explains the white and black spaces in the resulting ROIs (in all figures shown as all-white spaces for illustrative purposes).
In order to evaluate the performance of the current method, five-fold 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.

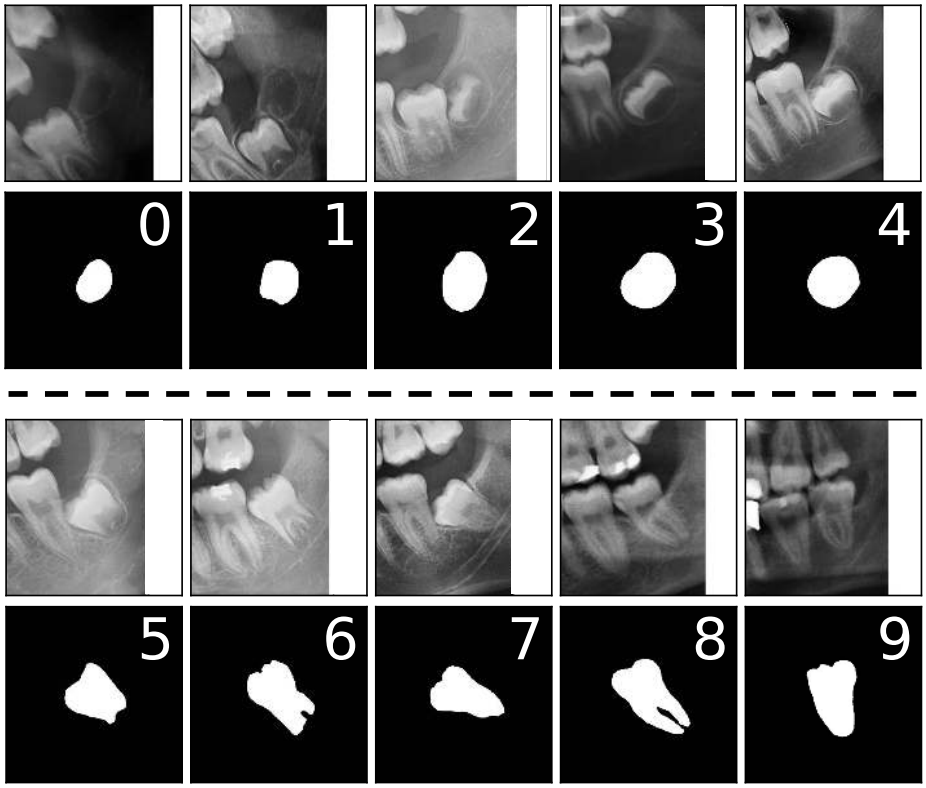

Fig. 1 Representative example of each of the 10 developmental stages of the lower left third molar (top rows) and their manual full segmentations according to Merdietio et al. [23] (bottom rows).

2.2 Three-step procedure

Based on recent work [4,12,18,20,33,23], the three-step procedure was proposed as presented in Fig. 2. First, a CNN located the considered third molar's center and established an ROI around it in the OPG. Second, another CNN segmented the third molar out of the established ROI. Third, a final CNN combined the third molar's ROI and its segmentation to classify the third molar's developmental stage.

2.2.1 Third molar localization

The first step (Fig. 2 - top) automatically extracts a bounding box of the lower left third molar. For each OPG, the full tooth manual segmentations as described in [23] were used as the training segmentations of the lower left third molars (Fig. 1 - bottom rows). The manual ROIs around the lower left third molar were obtained by automatically extracting a 448x448 pixels bounding box parallel to the image axes and centered around the geometrical center of the manual segmentations

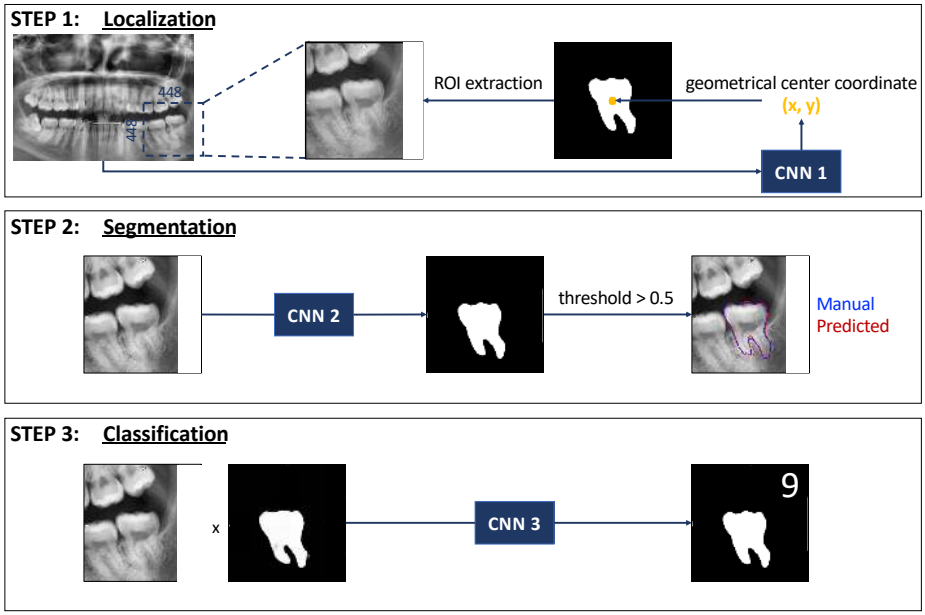

Fig. 2 A schematic overview of the proposed three-step procedure to automate third molar development staging. STEP 1: A first CNN detects a rectangular ROI around the third molar under assessment. STEP 2: Another CNN segments the third molar out of the established ROI. STEP 3: A final CNN combines the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

(annotated in yellow in Fig. 2). Unaligned ROIs of size 448x448 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 [26] CNN architecture was therefore utilized with minor modifications. Each image was divided into 25 cells and the cell containing the third molar and its geometrical center within this cell was predicted. For feature mapping, the DenseNet201 [11] CNN architecture, pre-trained on the ImageNet [28] dataset with dense layers suitable for the problem, was used. Hence, the sum of two mean squared error (MSE) objective functions - one for cell classification and one for geometrical center regression - was minimized. The combined loss function was optimized for the training set using the Adam optimizer [16] with default Keras [3] settings. The mean absolute error (MAE; Eq. 1) and Euclidean distance (Eq. 2) between the manual and predicted center coordinates in pixels were calculated:

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

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

where x_i and y_i, and $\hat{x}_i$ and $\hat{y}_i$ refer to the manual x and y, and predicted x and y coordinates of the geometrical center of the lower left third molar in the i^{th} OPG, respectively, 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 captures the manual segmentation of the third molar completely, partly or not, respectively.

2.2.2 Third molar segmentation

The second step (Fig. 2 - middle) automatically segments the lower left third molar, given a 448x448 bounding box around its geometrical center. For this purpose, a U-Net-like [27] CNN architecture was used. This model has been proven to work well across many segmentation tasks in medical imaging [27,13,1]. It processes the input image by successively applying linear convolutions and non-linear activations. The latter are necessary to avoid creating linear, and thus simple, features only. The convolutional kernels are of size 3x3 and leaky-ReLU [22] activations are used to create non-linear activations. 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. There are many loss functions, each having their own influence on how well the internal parameters converge to a solution. For segmentation, cross-entropy (CE), soft Dice (SD) or their linear combination (CS) are often used [1]. Here, a suitable loss function was identified on the manual ROIs and their segmentations by analyzing the segmentation performance in terms of pixel-wise accuracy (Eq. 3), precision (Eq. 4) and recall (Eq. 5), and the Dice score (Eq. 6) [1,34]:

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

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

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

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

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 number 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) in the medical field of image analysis [1].
The Adam optimizer [16] with default Keras [3] 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. In a second experiment, the most promising loss function was chosen to work directly forward on the localization output. That way, pixel-wise accuracy, precision and recall, and the Dice score were calculated for

each stage individually and with or without localization outliers.

2.2.3 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 [11] (as it was used in [23]). 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 prevent overfitting on the training set. Comparing the results of the simple CNN with the results of DenseNet201, sheds 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 on manual ROIs and segmentations was chosen and reported on the staging accuracy, MAE and LWK on the predicted ROIs and segmentations from the previous steps for each stage individually with and without localization outliers. These are frequently used metrics when evaluating staging performance and their definition can be found in [4].
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.

3 Results

The reported results were based on five-fold cross-validation of the complete procedure. For third molar segmentation and staging, the effect of different parameter setups was studied, working with manual 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.

3.1 Third molar localization

In Table 1 the localization results are given. The geometrical center of the third molar is localized with an average MAE of 75.42 pixels and an average Euclidean distance of 59.76 pixels. There is a trend for the detection algorithm to work better

Table 1 Quantitative results of the automated detection. "Wrong" cases are excluded from the results. Only the last column corresponds to the average performance with inclusion of wrong cases.

Metric ↓ \ Stage →	0	1	2	3	4	5	6	7	8	9	all
Euclidean distance (pixels)	77	89	56	52	48	47	57	52	58	62	63
MAE (pixels)	96	112	68	65	60	59	73	68	75	79	79

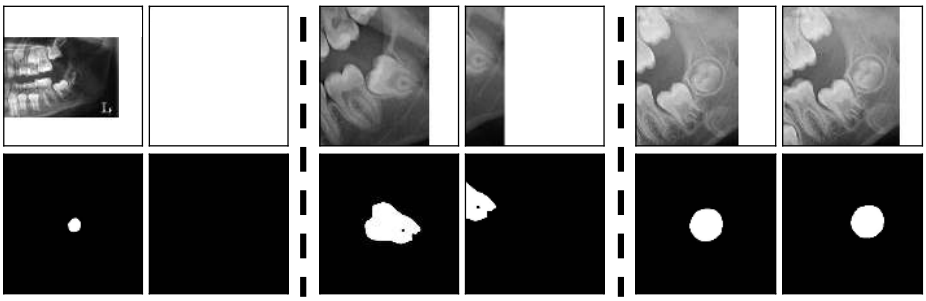

Fig. 3 Qualitative examples of detection (top row) and corresponding manual full segmentation (bottom row). The left column in each big box corresponds to the manual detection and the right column corresponds to the automated detection. The left two columns correspond to a "wrong" automated detection. 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.

for the middle stages. There are 393 good, 3 poor and 4 wrong localizations. In Fig. 3 one example of each is shown. Poor localizations are possible because of the following reasons: (i) the right 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 right cell.

3.2 Third molar segmentation

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 to the single losses (left side of Table 2). The use of this loss function for the predicted ROIs reveals a slight decrease compared to 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.

Table 2 Quantitative results of the automated segmentation. 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.

Experiment →	manual ROIs			predicted ROIs										
Loss →	CE	SD	CS	CS										
Stage →	all	all	all	0	1	2	3	4	5	6	7	8	9	all
Metric ↓														
Accuracy (%)	99	99	99	99	99	99	99	100	99	99	99	99	99	99
Precision (%)	95	95	95	88	95	92	97	97	95	96	95	96	96	95
Recall (%)	94	94	95	87	92	97	94	95	90	96	94	93	93	93
Dice (%)	94	94	94	85	93	94	95	96	91	96	95	93	93	93

3.3 Third molar staging

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 is a clear increase in performance when the segmentation is used (CO and MU compared to NO). MU seems to deliver the most promising results. Using this method for the staging on predicted ROIs with predicted segmentations, there is a drop in overall staging performance (right side of Table 3) compared to manual results. Zooming in on the results for each stage individually highlights the superior performance for lower stages. In Fig. 4 the staging output is illustrated for the segmentation examples.

Table 3 Quantitative results of the automated staging. 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.

Experiment →	manual segm's			predicted segmentations										
Method →	NO	CO	MU	MU										
Stage →	all	all	all	0	1	2	3	4	5	6	7	8	9	all
Metric ↓														
Accuracy (%)	55	57	60	85	85	62	45	51	42	64	20	37	45	54
MAE (Stages)	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
LWK (%)	81	83	84	/	/	/	/	/	/	/	/	/	/	79

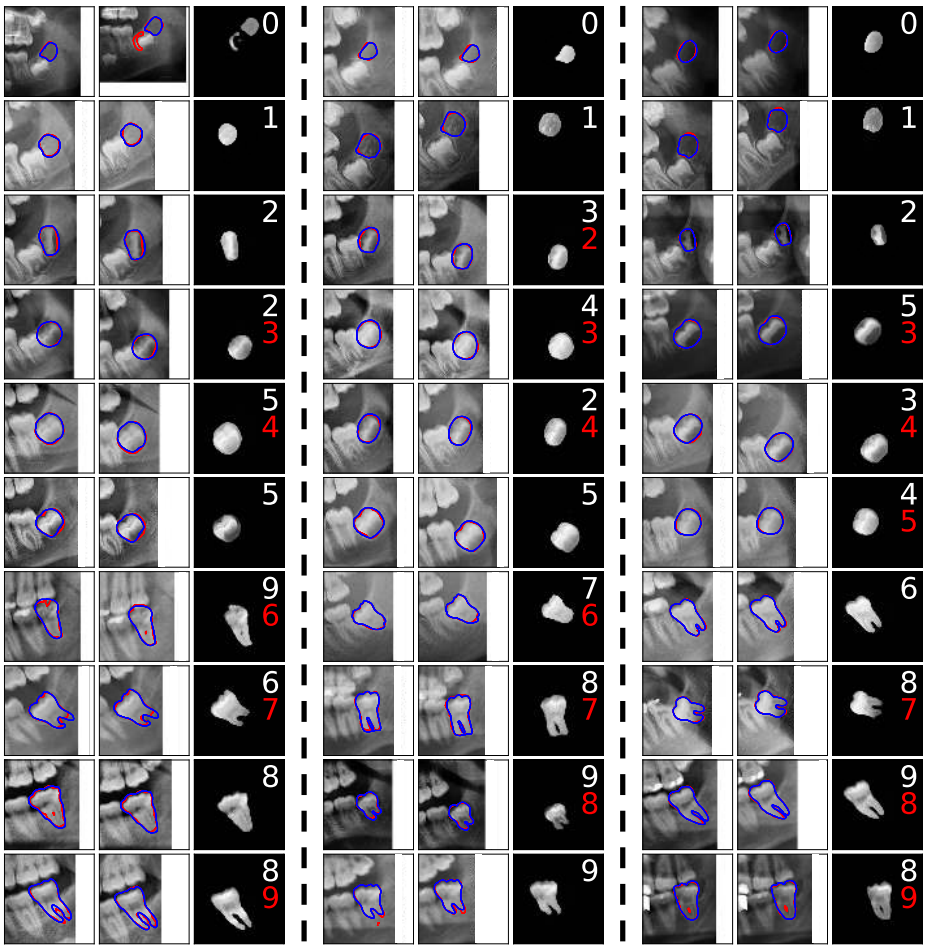

Fig. 4 Qualitative segmentation examples for each of the 10 developmental stages (rows; ordered top-bottom stage 0-stage 9). Left three, middle three and right three columns are the results for the first, second and third quartile of the Dice score using the predicted segmentations, respectively. The 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.

4 Discussion

4.1 Situation of findings in literature

Automated methods for skeletal age estimation have been used for over a decade [39]. Recently, the RSNA Pediatric Bone Age Challenge using hand radiographs demonstrated that different approaches to process the images can render similar results [9]. 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 79 % the current fully automated system for stage classification performed inferior compared to respectively 61 %, 0.53 stages and 84 % reported for the semi-automated system proposed by Merdietio et al. [23]. 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. Further optimization of all steps in the automated system is therefore recommended before its final application in forensic age estimation practice.

Moreover, to be 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) [9], based on the automated assessment of a hand-wrist radiograph. Assessing hand-wrist MRI, Tang et al. [38] 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. [37] 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 $\leq$ 18 years. Note that the reported MAEs in [9,38,37] 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. [35] reported an MAE of 0.85 years compared to chronological age, when assessing hand-wrist MRI. More recently, Stern et al. [36] 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 [35], N = 322 and age 13-25 years in [36]. 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.

4.2 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 the 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. [4] and Merdietio et al. [23]. 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. [40] 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. [41] 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. [6] 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. [23]. It may well be that more advanced strategies lead to superior performance compared to the early-fusion strategy explored in this work [14]. 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 to the results in Merdietio et al. [23] 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 one year need to be included [31]. 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 [39]. 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.

5 Conclusion

In this work, we proposed and validated a fully automated three-step procedure for third molar staging, directly starting from OPGs. The overall staging performance was close to previous research by De Tobel et al. [4] and Merdietio et al. [23], where only the final step was automated and manual ROI detection and third molar segmentation were necessary.

Acknowledgements.

J.B. is part of NEXIS, a project that has received funding from the European Union's Horizon 2020 Research and Innovations Programme (Grant Agreement #780026).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Bertels, J., Eelbode, T., Berman, m., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.: Optimizing the dice score and jaccard index for medical image segmentation: Theory & practice. Springer Verlag (2019)
2. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Xie, W., Rosen, G.L., et al.: Opportunities and obstacles for deep learning in biology and medicine. bioRxiv p. 142760 (2017)
3. Chollet, F., et al.: Keras. https://keras.io (2015)
4. De Tobel, J., Radesh, P., Vandermeulen, D., Thevissen, P.W.: An automated technique to stage lower third molar development on panoramic radiographs for age estimation : a pilot study. Journal of Forensic Odonto-Stomatology **35**(2), 49–60 (2017)
5. Demirjian, A., Goldstein, H., Tanner, J.: A new system of dental age assessment. Human biology pp. 211–227 (1973)
6. Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M.: Towards automatic bone age estimation from MRI: localization of 3D anatomical landmarks. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **17**(331239), 421–428 (2014)
7. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. Computerized medical imaging and graphics **31**(4-5), 322–331 (2007)
8. Greulich, W.W., Pyle, S.I.: Radiographic atlas of skeletal development of the hand and wrist. The American Journal Of The Medical Sciences **238**(3), 393 (1959)
9. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., Kitamura, F.C., Thodberg, H.H., Chen, L., Shih, G., Andriole, K., Kohli, M.D., Erickson, B.J., Flanders, A.E.: The rSNA pediatric bone age machine learning challenge. Radiology **290**(3), 498–503 (2019). DOI 10.1148/radiol.2018180736. URL http://pubs.rsna.org/doi/10.1148/radiol.2018180736
10. Hosntalab, M., Zoroofi, R.A., Tehrani-Fard, A.A., Shirani, G.: Classification and numbering of teeth in multi-slice ct images using wavelet-fourier descriptor. International journal of computer assisted radiology and surgery **5**(3), 237–249 (2010)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
12. Iglovikov, V., Rakhlin, A., Kalinin, A., Shvets, A.: Pediatric bone age assessment using deep convolutional neural networks. arXiv preprint arXiv:1712.05053 (2017)
13. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No New-Net. LNCS (2018). URL http://arxiv.org/abs/1809.10483
14. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-Scale Video Classification with Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition pp. 1725–1732 (2014). DOI 10.1109/CVPR.2014.223. URL https://www.computer.org/csdl/proceedings/cvpr/2014/5118/00/5118b725-abs.html
15. Kim, S.Y., Oh, Y.J., Shin, J.Y., Rhie, Y.J., Lee, K.H.: Comparison of the greulich-pyle and tanner whitehouse (tw3) methods in bone age assessment. Journal of Korean Society of Pediatric Endocrinology **13**(1), 50–55 (2008)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
18. Larson, D.B., Chen, M.C., Lungren, M.P., Halabi, S.S., Stence, N.V., Langlotz, C.P.: Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology **287**(1), 313–322 (2017)
19. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
20. Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B.A., Alkasab, T.K., Choy, G., Do, S.: Fully automated deep learning system for bone age assessment. Journal of digital imaging **30**(4), 427–441 (2017)
21. Liversidge, H.M.: Timing of human mandibular third molar formation. Annals of Human Biology **35**(3), 294–321 (2008). DOI 10.1080/03014460801971445

22. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. in ICML Workshop on Deep Learning for Audio, Speech and Language Processing **28** (2013)
23. Merdietio Boedi, R., Banar, N., De Tobel, J., Bertels, J., Vandermeulen, D., Thevissen, P.W.: Effect of Lower Third Molar Segmentations on Automated Tooth Development Staging using a Convolutional Neural Network. Journal of Forensic Sciences pp. 1556–4029.14182 (2019). DOI 10.1111/1556-4029.14182. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.14182
24. Miki, Y., Muramatsu, C., Hayashi, T., Zhou, X., Hara, T., Katsumata, A., Fujita, H.: Classification of teeth in cone-beam ct using deep convolutional neural network. Computers in biology and medicine **80**, 24–29 (2017)
25. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-Decem**, 779–788 (2016). DOI 10.1109/CVPR.2016.91
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
29. Schmeling, A., Dettmeyer, R., Rudolf, E., Vieth, V., Geserick, G.: Forensic Age Estimation. Deutsches Arzteblatt international **113**(4), 44–50 (2016). DOI 10.3238/arztebl.2016.0044. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4760148/pdf/Dtsch_Arztebl_Int-113-0044.pdf https://goo.gl/Qvw66p
30. Schmeling, A., Geserick, G., Reisinger, W., Olze, A.: Age estimation. Forensic Science International **165**(2-3), 178–181 (2007). DOI 10.1016/j.forsciint.2006.05.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0379073806003173
31. Schmeling, A., Grundmann, C., Fuhrmann, A., Kaatsch, H.J., Knell, B., Ramsthaler, F., Reisinger, W., Riepert, T., Ritz-Timme, S., Rösing, F.W., Rötzscher, K., Geserick, G.: Criteria for age estimation in living individuals. International Journal of Legal Medicine (2008). DOI 10.1007/s00414-008-0254-2
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
33. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in x-ray images. Medical image analysis **36**, 41–51 (2017)
34. Sϕrensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Det Kongelige Danske Videnskabernes Selskab **V**(4) (1948)
35. Stern, D., Ebner, T., Bischof, H., Grassegger, S., Ehammer, T., Urschler, M.: 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 (2014). DOI $10.1007/978 - 3 - 319 - 10470 - 6_28$
36. Štern, D., Payer, C., Giuliani, N., Urschler, M.: Automatic Age Estimation and Majority Age Classification from Multi-Factorial MRI Data. IEEE Journal of Biomedical and Health Informatics (2019). DOI 10.1109/JBHI.2018.2869606
37. Štern, D., Payer, C., Urschler, M.: Automated age estimation from MRI volumes of the hand. Medical Image Analysis (2019). DOI 10.1016/j.media.2019.101538
38. Tang, F.H., Chan, J.L., Chan, B.K.: 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 (2019). DOI 10.1007/s10278-018-0135-2
39. Thodberg, H.H., Kreiborg, S., Juul, A., Pedersen, K.D.: The bonexpert method for automated determination of skeletal maturity. IEEE transactions on medical imaging **28**(1), 52–66 (2009)
40. Unterpirker, W., Ebner, T., Stern, D., Urschler, M.: Automatic third molar localization from 3D MRI using random regression forests. In: T. Lambrou, X. Ye (eds.) Proceedings of the 19th Conference on Medical Image Understanding and Analysis, pp. 195–200. The British Machine Vision Association, United Kingdom (2015)

41. Vinayahalingam, S., Xi, T., Bergé, S., Maal, T., de Jong, G.: Automated detection of third molars and mandibular nerve by deep learning. Scientific Reports **9**(1), 9007 (2019). DOI 10.1038/s41598-019-45487-3. URL https://doi.org/10.1038/s41598-019-45487-3