Expert Systems With Applications 189 (2022) 116038

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automated estimation of chronological age from panoramic dental X-ray images using deep learning

Denis Milošević [a,*], Marin Vodanović [b], Ivan Galić [c], Marko Subašić [a]

[a] Faculty of Electrical Engineering and Computing, University of Zagreb, Unksa ul. 3, Zagreb, Croatia
[b] School of Dental Medicine, University of Zagreb, Gundulićeva ul. 5, Zagreb, Croatia
[c] University Hospital Centre Split, Croatia

ARTICLE INFO

Keywords:
Age estimation
Convolutional neural network
Deep learning
Forensic odontology
Image processing
Medical image analysis

ABSTRACT

Age estimation is a key component in forensic analysis, be it in legal proceedings or archeological research. Current methods in forensic odontology are based on manual measurements of a wide array of morphometric parameters, typically from dental x-ray images, and occasionally from material remains. While those parameters follow a set progression during human development, thereby allowing current methods to precisely estimate the age of juveniles, estimation for adults and seniors proves to be more difficult. In this study, we explore the applicability of deep learning to the problem of chronological age estimation. We determine the best convolutional neural network model derived from state-of-the-art architectures, we determine the best performing model parameters using pretrained general-purpose vision model parameters as the starting point, and we perform ablation experiments to highlight which anatomical regions of the dental system contribute the most to the estimation. The proposed approach attains the lowest estimation error in literature for adult and senior subjects, which we verify on one of the largest datasets of panoramic dental x-ray images in literature. The dataset consists of 4035 panoramic dental x-ray images of male and female subjects with ages between 19 and 90 years. This study also evaluates the feasibility of the proposed model for age estimations of individual teeth, achieving an estimation error competitive with current methods while being fully automated. The estimation error is verified on our dataset of 76416 individual tooth images, which is the largest dataset to date in forensic odontology literature. Unlike current methods, dental alterations, decay, illnesses, or missing teeth do not pose a problem to the proposed model. With a median estimation error of 2.95 years for panoramic dental x-ray images and 4.68 years for individual teeth, and by deriving the model from state-of-the-art architectures, verifying those results on the largest dataset in forensic odontology literature and demonstrating the importance of different anatomical regions of the dental system for estimation, this study sets the baseline for future research of automated chronological age estimation in forensic odontology.

1. Introduction

Age estimation is one of the fundamental steps of the forensic process. Knowing the age is an important factor considered in many facets of a person's life and death. Age estimation is used in legal proceedings to protect the rights of people without proper documentation, be it for seeking asylum or when taking care of a found child. Although there are many ways to estimate the age, teeth and jaw analyses perform exceptionally well with their ratio of indicativeness, non-invasiveness, and durability. Teeth outlast all other tissue when it comes to decomposition, making it a prime forensic target for identification, age, and sex assessment (Dudar, Pfeiffer, & Saunders, 1993). This feature makes them useful in age estimation for archeological research, where demographic data and the changes therein can be gathered from important historical sites. X-ray imaging also allows for a non-invasive approach to estimation, avoiding permanent injury or harm to a person.

Current estimation methods rely on manual measurements and human estimations. This allows for human error to influence the results, especially in disaster situations where the workload for forensic experts far exceeds reasonable amounts. In addition, whereas estimating the age of minors is relatively straight-forward due to a wide selection of developmental markers, estimation in adults and seniors still poses a big problem in forensic odontology. The lack of developmental indicators is compounded with the simple fact of life that things get

* Corresponding author.
E-mail addresses: denis.milosevic@fer.hr (D. Milošević), vodanovic@sfzg.hr (M. Vodanović), ivan.galic.st@gmail.com (I. Galić), marko.subasic@fer.hr (M. Subašić).

https://doi.org/10.1016/j.eswa.2021.116038
Received 19 February 2021; Received in revised form 23 August 2021; Accepted 3 October 2021
Available online 16 October 2021
0957-4174/© 2021 Elsevier Ltd. All rights reserved.

damaged over time, which includes the human body. Dental corrections, illnesses, and the loss of teeth and the skeletal support structure of the jaw contribute to the problems posed to manual measurement of dental parameters. With the advance of computing and automation in medicine, this problem seems like a good candidate for automation through modern deep learning techniques. Deep learning models can determine the most useful indicators for age estimation in the form of image features, whether developmental, morphometric, morphological or of any other type.

This study explores the applicability of deep learning for the forensic odontology task of chronological age estimation for adults and seniors. As mentioned, current methods are based on manual measurements, which can introduce variability to the estimation results and a lack of transparency. An automated system based on deep learning reduces the processing time, while at the same time offering consistent, reproducible results. Deep learning systems have the distinct advantage to discover a number of features and feature relations that would be infeasible for manual calculation. In contrast, features of current approaches are manually extracted, and their relations to age are linear. While such an approach allows for easier manual calculations, estimations tend to be rougher approximations. A large set of non-linear features allows for a finer estimation and lower overall estimation errors, which we demonstrate in this study.

To ensure that the trained neural network is reliable, stable, and well-trained, and to ensure that the results are trustworthy, a large dataset is required. Smaller, custom architecture networks can be used to tackle any image analysis problem, including this one, but without exhaustive analysis, biases and other systematic issues can enter the system, producing good results on one dataset (usually smaller in size), but failing in practice. This study tackles the problem of chronological age estimation by using established and proven neural network architectures as a starting point. Those networks are made as general-purpose vision systems, trained to work in the domain of photographs and not medical images. Despite this difference, general-purpose vision models discover general features for image processing in their lower convolutional layers. Therefore such models can be adapted and used as a starting point for the training of specialized networks, even for medical images. This study exhaustively evaluates state-of-the-art convolutional neural network architectures as feature extractors and uses their general image features as a starting point for creating a specialized model for age estimation that outperforms current methods in forensic odontology.

Another issue is the natural decay, illnesses, and dental alterations that can occur. Forensic odontology studies disqualify any samples that are imperfect, reducing the practicality of those methods. Neural networks can process huge amounts of data, allowing for the discovery of features that work even when teeth are not in perfect condition. This allows for broader application of the method in general, and especially in the domain of adults and seniors who are more likely to have such problems. In this study, we show that those samples can be included and that they do not negatively affect the performance of the model.

We also show that this model can be applied to X-ray images of individual teeth. In archeological contexts, it is more common to find single teeth as an artifact than entire jaws or skeletons. In legal proceedings, given that teeth are the most resilient part to decay and destruction through force and fire (Dudar et al., 1993) state that often only a handful of teeth can be found as evidence. This study tackles this problem too, demonstrating competitive performance with current methods while being reproducible and faster.

Additionally, we perform a series of ablation experiments to determine which anatomical structures contribute the most to the correctness of an estimation. Those experiments focus on the contribution of the skeletal structures and the contribution of teeth by themselves. We demonstrate that both structures can be used independently to estimate the age, and show that their interaction produces better estimations than they produce on their own.

This study is divided into five main parts. In Section 2 of this study, we first give an overview of the current state of age estimation in forensic odontology. We give a full analysis of the dataset used and its properties in Section 3, followed by a detailed explanation of the exhaustive model training, the hyperparameters and augmentations used, the ablation experiments, and the performance evaluation approach in Section 4. Section 5 describes the results of all experiments and performance of all trained models in detail, followed by a detailed analysis and discussion of those results in Section 6.

2. Related works

Current age estimation methods are based on developmental indicators, which are used in assessing how far a person is in dental development. Early work shows the usefulness of dental indicators for age estimation (Saunders, 1837). One of the first methods based on this principle (Nolla et al., 1952) developed a 10 stage chart for comparison to estimate the age of a person. Other methods soon followed, establishing the foundation of the field (Demirjian, Goldstein, & Tanner, 1973; Haavikko, 1970). However, whereas those methods work well on estimating the age of a child, they cannot provide detailed insight into the age of adults. Estimation of age in children is a fundamentally easier problem, as tooth development follows a strict, genetically defined, schedule (Cameriere, Ferrante, & Cingolani, 2006; Gleiser & Hunt, 1955; Moorrees, Fanning, & Hunt, 1963). This is supported by the average error reported by studies doing age estimation in children, where the error is measured in months, not years. The most known studies for age estimation in children (Cameriere, Ferrante, De Angelis, Scarpino, & Galli, 2008; Demirjian et al., 1973; Haavikko, 1970; Nolla et al., 1952) all report errors of less than 1 year, with many studies having a mean error of less than 0.5 years.

Three studies set the foundation for the estimation of age in adults. Kvaal, Kolltveit, Thomsen, and Solheim (1995) developed a method based on the discovery that the dental pulp cavity is reduced with advancing age due to secondary dentine deposit. They use a set of ratios between dimensions of the pulp, root, and tooth to construct a model for age estimation, successfully estimating the age of adults. Drusini, Toso, and Ranzato (1997) bases their method on the correlation of reduction of the coronal pulp cavity and chronological age but uses it in conjunction with the tooth-coronal index (Ikeda, Umetsu, Kashimura, Suzuki, & Oumi, 1985). Cameriere, Ferrante, and Cingolani (2004) estimate the age of adults from single-rooted teeth, specifically the single-rooted maxillary right canine. They use a set of ratios based on pulp and tooth length, width, and area to estimate the age. Newer studies have tried using CT instead of radiographic imaging (Yang, Jacobs, & Willems, 2006), estimating the age by volume matching, with varying results. A common factor in all those studies is the exclusion of teeth with illnesses or dental alterations, as they measure the indicators of the progress of natural processes which in the case of illness or alterations get irrecoverably damaged. Newer methods are establishing dental development atlases in children and adolescents with systemic diseases (Pereira, Russell, de Pádua Fernandes, Alves da Silva, & de Sousa Santos, 2019).

Other branches of forensics estimate age too. Age estimation by skeletal means is another approach (Greulich & Pyle, 1959; Roche, Thissen, & Chumlea, 1988). However, Ciapparelli (1992) shows that age estimation by dental indicators might be more applicable than through skeletal structures. Additionally, Dudar et al. (1993) shows that post-death changes are the slowest progressing on dental tissue, making it a prime candidate for identification and age estimation. Destructive methods can also be useful for age estimation, like methods that require tooth sectioning (Solheim, 1984). Those indicators show better accuracy, but their application requires either permanent injury to a person or the destruction of evidence which in most cases is not justifiable. Radiographic imaging (Röntgen, 1895) allows for measurements that would previously be impossible to perform

non-destructively. Early studies showed the viability for radiographic imaging in dental applications (Eckert & Garland, 1984; Matsikidis & Schultz, 1982), opening the gates for dental research without causing injury to living people (Carvalho, Silva, Lopes-Júnior, & Peres, 2009; Limdiwala & Shah, 2013; Marroquin, Karkhanis, Kvaal, Vasudavan, Kruger, & Tennant, 2017; Panchbhai, 2011; Singaraju & Sharada, 2009). There are also approaches that focus only on the categorization between juvenile or adult, giving only rough information about the age of a person (Cameriere et al., 2008).

The question of whether ethnicity or geographic location influence the applicability of methods was not answered by any of the mentioned studies. In addition to the overarching studies, localized studies have been done worldwide confirming the viability of the mentioned methods across geographical and environmental differences (Ardakani, Bashardoust, & Sheikhha, 2007; Babar, Iqbal, & Jan, 2008; Farah, Booth, & Knott, 1999; Galić, Nakaš, Prohić, Selimović, Obradović, & Petrovečki, 2010; Galić, Vodanović, Cameriere, Nakaˇ s, Galić, Selimović, & Brkić, 2011; Gulsahi, Yuzugullu, Imirzalı oglu, & Genç, 2008; Selmanagić, Ajanović, Kamber-ÄĘesir, Redžepagić-Vražalica, Jelešković, & Nakaš, 2020). Those studies use well-established methods but derive specific model parameters for their population.

With the advances in the field of computer vision, machine learning and deep learning more and more repetitive medical tasks are getting automated or semi-automated tools. Deep learning and computer vision methods are being applied in a wide selection of medical fields (Litjens, Kooi, Bejnordi, Setio, Ciompi, Ghafoorian, Van Der Laak, Van Ginneken, & Sánchez, 2017). Those applications can be of clinical use, like classification of lung nodules in CT scans (Shen, Zhou, Yang, Yang, & Tian, 2015) or hemorrhage detection in fundus images (Grinsven, Ginneken, Hoyng, Theelen, & Sánchez, 2016). They can also be used for research purposes, like discovering differences in brains (Xin, Zhang, Tang, & Yang, 2019) or the important factors for determining the myofiber orientation in cardiac high-resolution phase-contrast CT images (Baličević, Lončarić, Cárdenes, Gonzalez-Tendero, Paun, Crispi, Butakoff, & Bijnens, 2015). More recently, machine and deep are used for rapid diagnosis and containment of the COVID-19 epidemic, like recognizing the disease from X-ray chest radiographs (Altan & Karasu, 2020) and diagnosis using only cough recordings (Laguarta, Hueto, & Subirana, 2020).

As for forensics, some progress is made with deep learning. Applications in age estimation are present, like the estimation of age with automated skeletal assessment (Spampinato, Palazzo, Giordano, Aldinucci, & Leonardi, 2017). For forensic odontology, some advances have been made. Sex estimation from panoramic dental X-ray images achieves significant accuracy (Milošević, Vodanović, Galić, & Subašić, 2019). Progress is also being made in the field of panoramic dental X-ray image segmentation of teeth (Jader, Fontineli, Ruiz, Abdalla, Pithon, & Oliveira, 2018; Silva, Oliveira, & Pithon, 2018). Foundations for automated tooth detection (Betul Oktay, 2017) and numbering (Tuzoff, Tuzova, Bornstein, Krasnov, Kharchenko, Nikolenko, Sveshnikov, & Bednenko, 2019) have been set. There is also work done for estimation of age based on panoramic dental X-ray images with a dataset that focuses primarily on the estimation of age in children and young adults with dental development still ongoing (Vila-Blanco, Carreira, Varas-Quintana, Balsa-Castro, & Tomas, 2020). In Banjšak, Milošević, and Subašić (2020) age estimation with deep learning is being explored for archeological samples, where the model estimations are compared to the age categories estimated by experts. Some research has been done into classifying samples into age groups, either from panoramic dental X-ray images (Guo, Han, Chi, Long, Zhang, Yang, Yang, Chen, & Du, 2021) or from first molar images (Kim, Lee, Noh, Park, & Auh, 2021). Classification of the final five stages of Demirjian's classification of tooth development has also been tackled by deep learning in Upalananda, Wantanajittikul, Na Lampang, and Janhom (2021). Deep learning and machine learning progress in dentistry and implantology is on the rise, as can be seen in the recent review by Kang, Duong, and Park (2020).

Table 1
Detailed overview of data samples per age group.

Age group	Orthopantomographs		Individual teeth	
	Female	Male	Female	Male
[18, 20)	21	8	367	250
[20, 25)	337	221	6240	4601
[25, 30)	419	264	7691	4411
[30, 35)	399	234	7652	4755
[35, 40)	325	247	6668	4843
[40, 45)	246	188	4997	3577
[45, 50)	167	137	3116	2673
[50, 55)	185	118	2872	2162
[55, 60)	132	98	1925	1542
[60, 65)	75	79	1465	1408
[65, 70)	30	35	648	840
[70, 75)	24	24	565	561
[75, 80)	6	11	152	299
[80, 85)	2	3	43	69
[85, 90)	0	1	0	0
Subtotal	**2368**	**1667**	**44401**	**32015**
Total	**4035**		**76416**	

3. Data

Our dataset consists of 4035 anonymized panoramic dental X-ray images. The age of a person is calculated from the date of birth and the date of imaging. The age resolution is one day, but in all our experiments we express age as a floating-point number of years. The age range is from 19 to 90 years, with a female to male ratio of 58.7%:41.3%. The samples are collected from multiple locations in Croatia and belong to the collection of the Department of Dental Anthropology School of Dental Medicine University of Zagreb. The use of this collection for research purposes has been approved by the ethics committee School of Dental Medicine University of Zagreb.

The data is a sample of "real world" data, i.e. no filtering was done to include only high-quality samples. A uniformly sampled subset of the data, 983 panoramic dental X-ray images, have been labeled with teeth status. In this subset, 962 images had some kind of alteration, either dental fillings, caries, bridges, implants, retainers, orthodontic appliances, missing teeth, or crowns. Our samples contain various pathologies and loss of mandibular molars or anomalous molar and teeth, unlike studies usually conducted in the field of forensic odontology (Badran, Othman, Thnaibat, & Amin, 2015). While this is just a fraction of the entire dataset, it has the same age and sex distribution, and can therefore be used as a rough estimate of those metrics for the entire dataset.

The raw X-ray readings are processed into panoramic dental X-ray images by converting them to 8-bit images in JPEG format. As the images are taken from a wide array of orthopantomographs, the resulting images have a width in the range of 1127 px to 3260 px and height in the range of 553 px to 1536 px. Samples by sex and age brackets can be seen in Fig. 1.

The dataset of full panoramic dental X-ray images consists of 2346 female and 1647 male samples belonging to 3994 individuals. However, the distribution of samples by age groups and gender is not uniform. The distribution of samples by age and sex is shown in Table 1. The data is biased towards younger samples and slightly biased towards female samples. However, the average age is still 38.17 years, placing our overall samples firmly into the "adult" category from the dental development perspective. The age groups of 75+ years have a significantly smaller sample size, which is taken into account during result analysis.

Another important factor to consider is the distribution of teeth. With age, teeth accumulate damage which can lead to major alterations and even tooth loss, whether from negligence or external factors. When considered as a dataset of individual teeth, this dataset contains 76416 individual teeth, with 44401 belonging to female and 32015 belonging

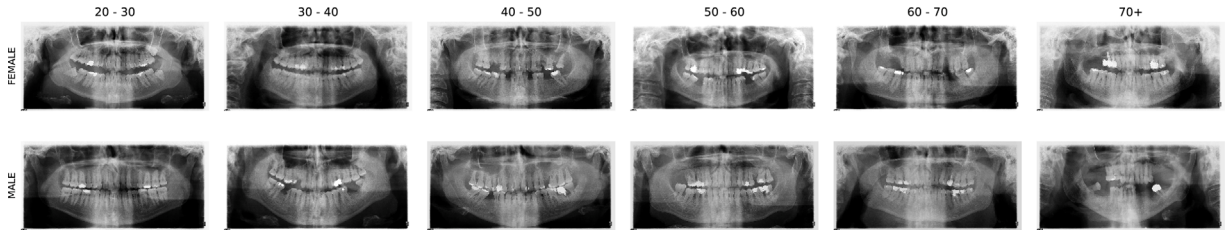

Fig. 1. X-ray sample images. Example of panoramic dental X-ray images per sex and age group. Images of older individuals tend to have more damage, missing teeth, and dental corrections.

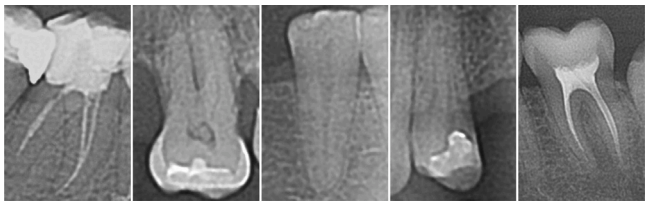

Fig. 2. Samples of individual teeth. Individual teeth are clipped from the full panoramic dental X-ray image by their bounding box.

male samples. Individual teeth images were extracted from a subset of 2683 manually annotated panoramic dental X-ray images. A detailed overview of tooth count per age group can be seen in Table 1. A general trend towards a decrease in teeth count with age can be noticed.

A part of the research was done on individual teeth images clipped from the panoramic dental X-ray images. Those teeth are divided into 16 classes, based on their location and type. As per standard notation, each tooth is identified with two numbers — the quadrant in which it is located and its position within the quadrant. For this study, teeth are considered of the same category if they are in the same jaw (upper or lower) and in the same position within a quadrant. Our notation consists of two parts, the jaw side and tooth index. Therefore teeth 11 and 21 belong to the same category under the name "up-1", and teeth 31 and 41 belong to the same category under the name "down-1". This way of tooth grouping is used in two founding works of modern forensic odontology (Cameriere et al., 2004; Kvaal et al., 1995).

The data is split into 3 subsets — train, validation, and test sets. The ratio of train/validation/test size is 80% : 10% : 10%. Those subsets are sampled uniformly from the dataset, thereby preserving the age and sex distribution in all subsets. For hyperparameter and method selection only results on the validation set were taken into account, and all results are reported on the test set that had no influence on research decisions. Special care was given to prevent data leakage. Multiple images in the dataset can come from the same person, which could skew the results if such a set of images got distributed over both the train and test set. For that purpose, each image is associated with an identity hash that can be used to determine if an image comes from the same patient, but that cannot be used to determine the identity of the person.

Each image is resized to the size of 512px x 512px and pre-processed as required by the convolutional feature extractor used. Preservation of aspect ratio has been shown to not improve model performance (Milošević et al., 2019), therefore anisotropic scaling was used.

4. Method

This study approaches the problem of age estimation from panoramic dental X-ray images for forensic odontology as an image analysis problem. Instead of manual measurements and modeling estimators by established dental parameters, this approach leverages the advancements made in the field of automated image analysis, allowing it to determine features specific to this problem. This is realized by designing a convolutional neural network as a chronological age estimation model. This model takes an image as the input and produces one floating-point number in the range of 0 to 100 as the output, representing the estimation of age in years.

An exhaustive hyperparameter search was used to determine the best performing deep learning model architecture for this task, based on performance measured on the validation set. Models for age estimation were trained on the entire panoramic dental X-ray images and specific regions of interest, as well as on individual teeth.

4.1. Model and training

The high-level structure of the model can be divided into four parts. The first part of the network is the convolutional neural network with pretrained weights provided by the base network. The following pre-trained convolutional networks were tested in our experiments for transfer learning approach: DenseNet201 (Huang, Liu, van der Maaten, & Weinberger, 2017), InceptionResNetV2 (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017), ResNet50 (He, Zhang, Ren, & Sun, 2016), VGG16, VGG19 (Simonyan & Zisserman, 2015) and Xception (Chollet, 2017). The second part is a 1×1 convolutional layer used to adjust the number of channels in the final feature map. The third part is an optional attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017). The fourth and last part consists of two fully connected layers. The second to last layer is variable in size as defined by its hyperparameter, and the last is a fully connected layer with a single unit that produces the estimate. Batch normalization (Ioffe & Szegedy, 2015) is used after the first fully-connected layer. Pretrained feature extractor are not altered, and batch normalization is used within those architectures as described in their studies. All activation functions, including the output activation, are ReLU (Schmidhuber, 2015). A diagram of the model can be seen in Fig. 3.

A detailed overview of the hyperparameter search space is given in Table 2. In this study, experiments have been performed for 6 different pretrained feature extractors, 28 different values for the number of channels in the final feature map, 16 different values for the size of the intermediate fully-connected layer, and with and without an attention mechanism.

The difference in grid size for the number of channels in the final feature map and the size of the intermediate fully-connected layer is a conscious decision based on preliminary experiments. Preliminary experiments have shown that the model performance is much more sensitive to the number of channels in the final feature map in comparison to the size of the penultimate fully-connected layer. We have therefore dedicated more compute time to the discovery of the number of channels in the final convolutional layer.

All models were trained with Adam (Kingma & Ba, 2015) as the optimizer, with a learning rate of $3.24 \cdot 10^{-4}$. The age estimation task is formulated as a regression problem. The sum of squared differences is used as the loss function.

Hyperparameter search and proper model training share the same training hyperparameters, with the only difference being the number

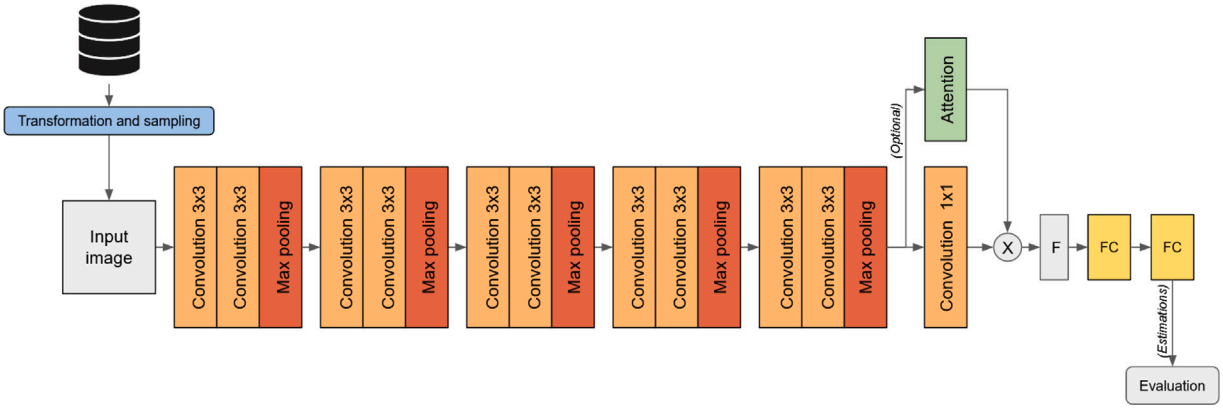

Fig. 3. X-ray image analysis and model training approach. For training, images are sampled from their respective subset and transformed as described in Section 4.3. The prepared images are then processed by the trained convolutional and fully connected layers (VGG16 is shown as an example). The final estimation is then collected and evaluated to show model performance.

Table 2
Overview of model hyperparameters used for the grid search.

Hyperparameter	Search space	Best value
Pretrained feature extractor	DenseNet201 (Huang et al., 2017), InceptionResNetV2 (Szegedy et al., 2017), ResNet50 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2015), VGG19 (Simonyan & Zisserman, 2015), Xception (Chollet, 2017)	VGG16
Number of channels in the final feature map	Between 5 to 1000	40
Presence of attention mechanism	Present or not present	Not present
Size of intermediate fully connected layer	Between 1 and 2048	128

of epochs the model is allowed to train. The training during hyperparameter search is limited to 100 epochs, which proves to be enough for hyperparameter selection. Fine-tuning is trained for 1000 epochs, which was enough time for the model to converge.

Our preliminary experiments have shown that training a state-of-the-art network architecture with just 4000 panoramic dental X-ray images leads to significant overfitting. Also, training a model with smaller capacity results in worse overall performance, therefore transfer learning (Tan, Sun, Kong, Zhang, Yang, & Liu, 2018) was used. Transfer learning allows for the use of bigger architectures with good general-purpose image analysis features that can then be fine-tuned for panoramic dental X-ray images. The current state-of-the-art convolutional neural network architectures are used as a feature extractor for this study. The parameters of those architectures are pretrained on ImageNet as provided by the Tensorflow library.

Those models are pretrained on the ImageNet dataset, which consists of 3 channel color images, but panoramic dental X-ray images are grayscale images. To be able to use those pretrained models, our grayscale images are stored as 3 channel color images with identical color values. Images are resized to 512px x 512px, as models with this image size performed best in our preliminary experiments, and as has been shown in Milošević et al. (2019).

4.2. Exploration of age indicators in dental X-ray images

This study focuses on age estimation from panoramic dental X-ray images and X-ray images of individual teeth. The foundational studies (Cameriere et al., 2004; Kvaal et al., 1995) designed regression models with measurements from a single tooth and regression models with measurements from multiple teeth. Our approach follows a slightly different path, as we are not explicitly extracting any information prior to the age estimation. In an effort to interpret our results we perform some experiments on specific anatomical regions only, but those experiments, though informative, did not provide our best age estimation results.

For panoramic dental X-ray images, four different variants are tested. One model was trained on panoramic dental X-ray images with no obstructions, where the entire skeletal and dental structure is visible, with all pathologies and dental alterations, if present, visible. This differs from region-based approaches that are becoming more popular. Region-based approaches like (Wang, Li, Shu, & Li, 2019) divide the input image into multiple regions that get processed independently and then aggregated for further processing. Our approach is purely motivated by interpretability. By training models on specific anatomical regions, we can gain insight into what information our final model uses for its prediction.

To determine which regions contain useful age indicators, three variants of the panoramic dental X-ray image model were trained. One model is trained on images only containing teeth, having the surrounding skeletal structure blocked out. Another model is trained on images where the central region containing teeth is blocked out, retaining the surrounding skeletal structure. The third model is an improvement on the tooth-blocking variant. Instead of just blocking the region where teeth are, every single tooth is blocked out individually, ensuring that no roots or oddly placed teeth can appear in the image. Those tooth regions come from the subset of data manually annotated with tooth locations, as mentioned in Section 3. The regions are blocked out with a black bounding rectangle. The idea was to exclude the influence of teeth, while simultaneously including as much of the skeletal structure around teeth as possible. Samples X-ray images of the mentioned variants can be seen in Fig. 4.

For single tooth models, images of each tooth present in the full X-ray image have been cropped out into individual images. For each of the 16 types of teeth mentioned in Section 3, a model was trained. This enables the model to focus solely on the indicators of a single tooth, like (Cameriere et al., 2004; Kvaal et al., 1995). A sample of single tooth images is shown in Fig. 2.

Fig. 4. Samples of the panoramic dental X-ray image variants. Models are trained on either images with the teeth area covered (left), images where only teeth area is visible (middle), or images where each tooth is individually covered (right).

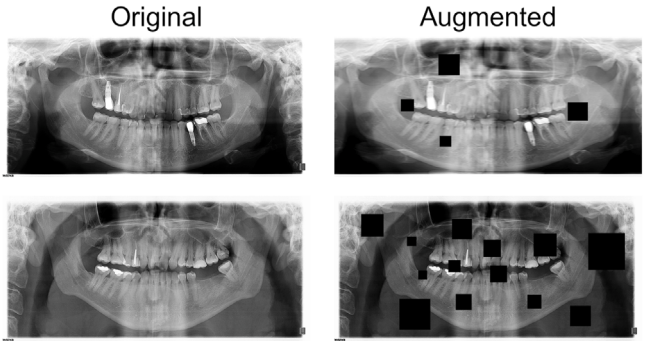

Fig. 5. Examples of image augmentation. The left column shows unmodified images, the right column shows those images augmented randomly. The first row shows a significant change in illumination achieved with gamma correction, a high degree of blurring, and a subtle affine scaling in the *y*-axis. The second row shows the opposite effect of gamma correction where the image got darkened, the blurring is more subtle and there is no affine scaling on the *y*-axis. However, there is a high occurrence rate of coarse dropout on the second row, whereas the first row has a low occurrence rate of coarse dropout.

Table 3
List of augmentations and hyperparameters used during training.

Augmentation	Hyperparameters
Left–right flip	p: 50%
Coarse dropout	Image rescale factor: 2% to 5% Amount of dropped pixels: 0% to 1%
Average blur	Square kernel size in px.: 0 to 2
Gamma contrast	gamma: 0.85 to 1.15
Rescaling	x: 1 to 1.4 y: 1 to 1.2

4.3. Model stability with stratified sampling and augmentation

To prevent bias and to improve performance and increase model stability, stratified sampling and augmentation are used on the training set. The validation and test set remain untouched in regard to these techniques.

Augmentation is known to improve model performance and stability by introducing noise to the data, thereby allowing the deep learning model to reach a better local optimum (Shorten & Khoshgoftaar, 2019). Images in the training set are randomly augmented during training, ensuring that the network never sees two exact same samples. The magnitude range of those augmentations is held constant during training, but the specific augmentation magnitude is randomly selected for every image batch from a predefined interval. A detailed overview of augmentations used and their hyperparameters can be seen in Table 3, and an example of augmentation can be seen in Fig. 5.

The dataset is biased towards younger samples, which leaves older age groups underrepresented in the learning process. Research shows that a non-uniform distribution of samples can lead to degraded performance of estimators (Shahrokh Esfahani & Dougherty, 2014). Therefore, stratified sampling was used to bring the non-uniformity of sample count per age group in line. Age groups are defined as 5 year long subgroups starting with the age of 20, with an exception for the age group below 20 spanning only two years. The age group with the biggest sample size was selected as the baseline size for age groups for stratified sampling. For each age group, a multiplication factor was determined so that the final number of samples in each age group is equal or close to equal. It is calculated for each age group as the ratio of baseline size and samples we have per age group. For example, if the multiplication factor is 1.7, then that data point will be selected once, and have an additional 70% chance to be selected again.

4.4. Evaluation approach

All models are evaluated from a general performance standpoint, as well as on a per-age group basis. For models estimating based on panoramic dental X-ray images, the test set remains unchanged throughout all experiments, including the ablation experiments. The test sets for individual teeth models are different between every tooth type, as those datasets differ from type to type. The performance is measured with three different metrics: the average absolute error in years, the median absolute error in years, and the coefficient of determination (R^2).

The mean absolute error is the difference (in years) between the estimated and actual age of a sample. It gives insight into the statistically expected value of the error. The median absolute error is the middle value of the sorted array of absolute errors. It provides a look at the absolute error value without the influence of extreme outliers. R^2 is used in regression models to give insight into the proportion of variance between the estimated values and independent variables from which the estimation was made (Cameriere et al., 2004; Drusini et al., 1997; Kvaal et al., 1995; Vila-Blanco et al., 2020).

5. Results

As mentioned, the hyperparameter search has given a general performance overview of different models. The best performing model was then extensively trained on full panoramic dental X-ray images, individual teeth, and specific regions of the panoramic dental X-ray image.

To find the best set of network parameters, hyperparameter searching was performed. A reasonable hyperparameter space was established, which was divided into a uniformly spaced grid, and the performance of each hyperparameter value at grid point was evaluated. This search method is called grid search (LaValle, Branicky, & Lindemann, 2004), which is one of the usual approaches for hyperparameter searching. The search space consisted of well-established backbone networks, usual intervals for the depth of the final convolutional layer, and the number of units in the second to last fully connected layer. The resulting mean absolute errors range from 5.67 years up to 20 years, depending on the hyperparameters selected. Models with the attention mechanism have consistently underperformed. The best performing model, as shown in Table 2, has the following hyperparameters: VGG16 backbone network, with 40 channels in the final convolutional layers, with 128 units in the second to last fully-connected layer, and with no attention mechanism and with batch normalization.

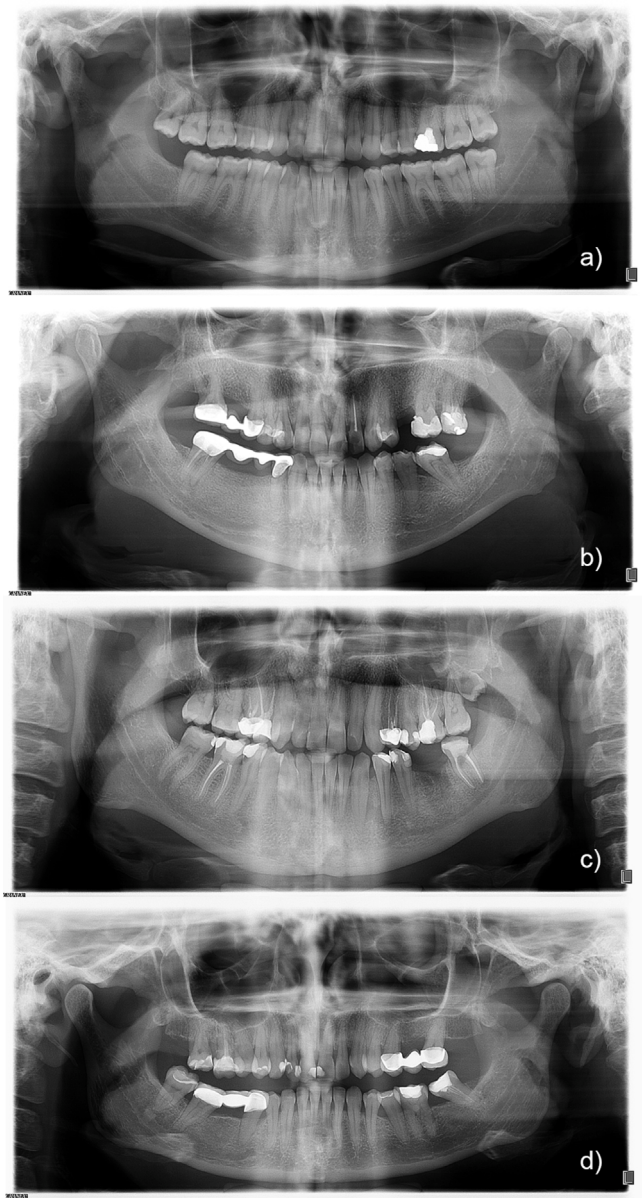

Fig. 6. Samples of successful and unsuccessful age estimation. These four examples are randomly selected from the test dataset. Two images were selected as success cases (error less than or equal to 2 years), and two were selected as failure cases (error bigger than or equal to 7 years). Samples a) and b) are success cases, while samples c) and d) failure cases. Data: a) has a true age of 27.1 years while the model estimated 27.6 years; b) has a true age of 51.4 years while the model estimated 49.4 years; c) has a true age of 41.5 years while the model estimated 49.9 years; and d) has a true age of 24.3 years while the model estimated 34.2 years. As can be seen, alterations of teeth do not necessitate a worse prediction result. However, heavy modifications, multiple missing teeth, and other deformation and illnesses can cause worse performance on some samples.

Models trained on the full panoramic dental X-ray images reach a mean absolute error of 4.06 years, a median absolute error of 3.11 years, and an R^2 of 0.8405, with the age group of 20 to 25 years of age achieving a mean and median absolute error of 2.82 years. When augmentation is added to the training data the mean absolute error falls to 3.96 years, the median absolute error to 2.95 years, and R^2 rises to 0.8439, with reduced errors of the higher age groups. The performance in the best performing age group changes, with the mean absolute error slightly increasing to 2.97 years, but the median absolute error falling to 2.77 years. With stratified sampling, the model reaches an overall mean absolute error of 4.02 years, a median absolute error of 3.17 years, and an R^2 of 0.8363, but with performance gains in the higher age groups. An overview of full panoramic dental X-ray image models is shown in Fig. 8, and an overview of the relationship between true age and estimations of the best performing model is shown in Fig. 7. A detailed view per age group is shown in the first three columns of Table 4. Samples of success and failure cases can be seen in Fig. 6.

In addition to models for full panoramic dental X-ray images, models for specific regions of interest have been made to determine which part of the jaw/teeth expresses the age information better. For models estimating age from the central region where only teeth are visible, the mean absolute error reaches 4.65 years, the median absolute error reaches 3.62 years and R^2 reaches 0.7820, with the age group of 25 to 30 years performing best with a mean absolute error of 3.04 years and a median absolute error of 2.19 years. For models estimating age from the outer regions of the X-ray, where no teeth are visible, the mean absolute error is 5.06 years, the median absolute error is 4.12 years and R^2 is 0.7481. If that model is modified to estimate age on the entire image except for the individual tooth locations, the mean absolute error is 5.42 years, the median absolute error is 4.40 years and the R^2 is 0.7308. An overview of the performance of the region of interest models is shown in Fig. 9. A detailed view per age group is shown in the last three columns of Table 4.

A family of models was trained on X-ray images for each specific individual tooth. Age estimates of those models are on average worse than estimations made on the full panoramic dental X-ray image. Depending on the tooth, the mean absolute errors range from 6.30 years to 8.68 years, and the median absolute error range from 4.68 years to 7.33 years, with R^2 ranging from 0.3277 (down-3) to 0.5541 (down-7). An overview of estimation performance is shown in Fig. 10, and a detailed view per age group is shown in Tables 5 and 6.

As for computational performance, all models processing an entire panoramic dental X-ray image have the same inference time, regardless of which data variant the model was trained on, and it holds true across all age groups. Training time itself for full panoramic dental X-ray images does not differ significantly, as the most processing-intensive image transformations are done as preprocessing. The same holds true for the per-tooth image family, where there is no significant performance difference between teeth models. Full panoramic dental X-ray image models can process 100 images in 1.65 s, while single tooth models can process 100 images in 1.61 s. Conventional, manual age estimation methods can take up to 20 to 30 min per image.

6. Analysis and discussion

When taking age groups into account, a clear trend of performance decrease with higher age is noticeable. Two factors contribute to this trend. One stems from the data itself. As shown in Section 3, the collected data is biased towards younger and female samples. The optimization process will therefore be biased towards finding more discriminative features for younger age groups as small improvements in estimations compound and have a bigger total effect on the value of the loss function. In addition, more data allows for better discriminative features to be found. The other factor is based on the differences in changes teeth undergo over the lifetime of a person. As age advances, tooth development stops and calcification and decay (Sayegh & Reed, 1968) become the main driving force of change. To complicate matters further, externally caused damage accumulates over time, and dental intervention causes changes to teeth and bone, weakening or completely destroying age indicators. Children lack this wear-and-tear over time, and paired with the presence of developmental indicators makes age estimation for children an easier problem, thereby reducing the overall average error of studies that include child subjects. This effect can be observed for all described models, including in variants with stratified sampling in regard to age.

An important question in age estimation from panoramic dental X-ray images is whether only teeth contribute to the final assessment,

Table 4
Mean absolute error of full-image models per age group. μ is the mean absolute error, $\hat{y}$ is the median absolute error.

Age	Baseline		Augmentation		Strat. sampling		Teeth only		Masked [rough]		Masked [fine]	
	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$
[15, 20)	10.37	11.01	11.80	12.19	6.45	6.74	12.04	13.06	11.80	11.96	11.31	10.91
[20, 25)	2.82	2.82	2.97	2.77	3.58	2.32	3.67	3.45	5.17	4.59	3.56	3.80
[25, 30)	2.97	2.37	3.02	2.12	3.50	2.38	3.04	2.19	4.61	3.61	5.22	4.15
[30, 35)	3.72	2.99	3.27	2.42	3.29	2.61	4.20	3.41	4.44	3.35	4.19	3.49
[35, 40)	3.67	2.53	3.63	3.03	3.80	3.64	4.43	3.20	3.26	2.73	5.14	4.55
[40, 45)	4.65	3.94	4.47	3.20	4.45	4.33	4.84	3.76	5.39	5.67	5.30	3.52
[45, 50)	5.62	5.52	5.54	4.45	3.76	3.54	5.72	3.96	6.26	4.73	4.54	4.01
[50, 55)	4.07	2.72	4.25	4.24	4.24	2.70	4.68	4.40	4.58	3.52	6.30	6.55
[55, 60)	6.03	4.97	5.76	4.66	4.16	3.30	7.17	6.81	5.68	3.83	5.92	5.73
[60, 65)	3.84	3.61	3.29	2.70	4.20	3.59	4.52	4.69	5.51	3.73	7.56	6.13
[65, 70)	11.78	13.75	11.25	12.10	9.08	7.86	13.57	15.48	10.02	10.03	10.61	10.78
[70, 75)	5.01	5.67	3.90	4.98	12.68	12.90	6.80	7.07	14.12	12.33	15.38	16.76
[75, 80)	11.85	12.02	12.97	11.68	11.03	11.01	20.18	18.45	16.90	17.35	18.05	19.86
All ages	4.06	3.11	**3.96**	**2.95**	4.02	3.17	**4.65**	**3.62**	5.06	4.12	5.42	4.40

Table 5
Mean absolute error per age group of per-tooth models of the maxillary teeth. μ is the mean absolute error, $\hat{y}$ is the median absolute error.

Age	Up-1		Up-2		Up-3		Up-4		Up-5		Up-6		Up-7		Up-8	
	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$
[15, 20)	8.75	8.35	9.99	10.87	8.05	7.17	12.17	11.95	7.35	8.83	7.55	4.94	6.62	6.73	9.87	9.76
[20, 25)	11.16	11.03	11.11	10.36	10.32	9.47	10.65	10.27	8.36	7.91	5.83	5.36	8.51	8.18	7.31	7.24
[25, 30)	7.33	6.25	7.24	6.79	6.04	4.88	7.34	6.46	5.40	4.20	3.90	3.58	5.45	4.23	5.15	3.90
[30, 35)	4.39	3.60	4.80	3.62	4.35	3.21	5.48	4.54	4.69	4.12	4.70	3.60	4.47	3.98	5.29	4.35
[35, 40)	3.61	3.26	4.00	3.31	4.91	4.09	4.74	3.13	4.42	3.75	5.49	5.36	4.57	3.38	4.74	4.14
[40, 45)	4.28	3.72	4.56	3.75	5.71	5.54	5.16	5.09	5.99	5.50	6.78	6.19	5.31	5.05	6.37	5.94
[45, 50)	7.22	7.34	6.60	5.88	5.67	5.05	6.55	5.74	7.11	6.07	7.98	7.05	4.99	4.32	7.35	7.42
[50, 55)	10.77	10.87	9.84	8.74	10.97	10.11	7.92	6.93	11.68	11.52	12.48	11.75	8.66	8.74	13.89	15.23
[55, 60)	12.83	13.51	12.41	12.39	12.45	11.78	11.67	10.40	13.60	13.80	12.88	12.99	10.78	11.02	13.89	14.96
[60, 65)	15.95	15.91	14.98	14.98	14.17	14.02	12.61	12.69	14.65	13.11	14.56	13.62	11.66	11.51	15.16	17.24
[65, 70)	21.89	23.83	20.06	19.32	23.00	23.24	21.02	20.84	22.44	20.77	24.99	23.53	18.05	16.51	21.18	22.50
[70, 75)	25.42	25.11	20.78	22.53	24.45	23.89	18.41	16.99	19.01	22.42	22.80	23.31	22.76	23.67	35.20	36.02
[75, 80)	30.59	29.81	27.56	25.78	26.88	27.44	27.57	29.29	23.35	24.51	32.25	34.31	24.40	25.35	29.31	29.31
Total	8.30	6.52	8.11	6.79	8.14	6.43	8.01	6.77	7.65	6.11	7.52	5.69	**6.87**	**5.68**	7.32	5.89

Table 6
Mean absolute error per age group of per-tooth models of the mandibular teeth. μ is the mean absolute error, $\hat{y}$ is the median absolute error.

Age	Down-1		Down-2		Down-3		Down-4		Down-5		Down-6		Down-7		Down-8	
	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$	μ	$\hat{y}$
[15, 20)	8.67	8.26	12.60	11.10	12.38	11.48	13.75	15.49	13.37	10.68	9.19	7.87	8.30	5.54	4.91	5.56
[20, 25)	8.98	9.00	9.44	8.95	10.84	10.37	11.01	10.06	9.30	8.76	7.13	6.60	6.56	6.30	4.03	3.67
[25, 30)	6.21	5.64	6.22	5.75	7.77	7.58	7.57	6.90	6.89	6.08	4.74	3.57	5.37	4.10	3.94	3.19
[30, 35)	5.01	4.13	4.38	3.55	4.68	3.87	4.75	4.44	4.45	3.72	3.68	2.83	4.37	3.36	4.09	3.07
[35, 40)	4.00	3.78	4.09	3.29	4.45	3.54	4.62	3.84	4.50	3.62	4.21	3.19	4.45	3.87	4.57	3.61
[40, 45)	5.83	5.54	5.49	5.09	5.44	4.82	5.43	5.26	4.68	4.31	5.55	5.46	4.90	4.29	6.31	4.83
[45, 50)	6.29	5.57	7.28	7.51	6.27	6.41	7.18	6.85	6.40	6.39	7.41	7.37	6.04	4.90	7.76	7.44
[50, 55)	12.89	12.90	12.23	11.72	11.42	11.59	10.71	10.38	9.60	8.26	11.59	11.08	8.14	7.74	12.00	13.27
[55, 60)	12.42	12.55	13.50	13.24	12.44	11.71	11.36	11.21	11.47	13.12	12.82	13.96	11.00	12.26	11.71	11.55
[60, 65)	14.48	14.39	17.05	18.12	15.61	15.85	13.00	11.87	15.97	15.11	13.31	12.83	11.81	12.47	16.40	14.76
[65, 70)	20.45	20.55	23.18	22.72	23.50	23.79	20.47	21.93	20.83	21.62	22.49	22.20	18.92	20.52	22.64	24.72
[70, 75)	25.23	26.03	24.53	24.14	24.54	26.87	23.53	23.83	22.27	19.47	23.48	24.80	21.28	21.67	31.95	31.95
[75, 80)	38.81	37.22	31.30	30.93	32.50	35.44	34.93	35.60	30.17	28.92	24.71	25.38	22.75	21.58	26.48	26.48
Total	8.23	6.64	8.43	6.71	8.65	7.33	8.43	6.96	7.87	6.40	6.87	5.39	6.53	5.13	**6.39**	**4.68**

given that classical age estimation methods mostly take tooth parameters into account. As can be seen in Fig. 9 and Table 4, among the selected specific regions, the best performance is achieved with the model of images containing only teeth, followed by the model of images that had the average tooth region-blocked out leaving only the surround skeletal structure, followed by the model of images with stringently covered teeth. This suggests that teeth in and of themselves contain the most informative age indicators, but that the surrounding skeletal structure contains useful age indicators too.

To explain the discrepancy between the models for roughly and stringently covered teeth, one must understand the difference in masking. As shown in Fig. 4, the roughly covered images have the middle region blacked out — the region that *mostly* contains all teeth. This leaves room for edges or parts of the tooth root to still be visible in some images, providing the model with some tooth-related indicators. When each tooth gets covered individually, no pixel in the resulting image belongs to any tooth, effectively leaving only the indicators from the skeletal structure. We believe that this could be an explanation for the performance drop with stringently covered teeth. However, significantly better estimations are made when using unobstructed images. This suggests that while both tooth and skeletal indicators are useful, their interaction result in indicators that are superior to both individually.

To lessen the impact of data bias in regard to model performance, stratified sampling is used. As oversampling can lead to overfitting and unstable models, augmentation is added to the training process to

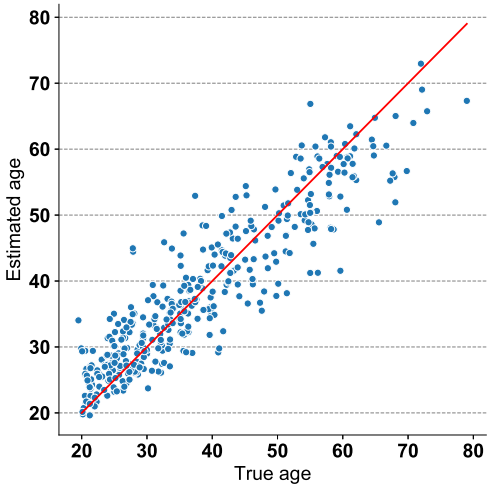

Fig. 7. Model estimations compared to real age. Each dot represents one sample from the test set, with its position defined as (real age, estimated age). The red line represents the ideal location for estimations.

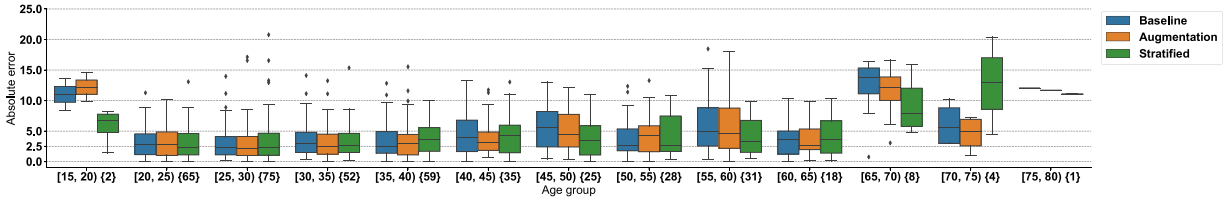

Fig. 8. Errors of estimates for panoramic dental X-ray image models. An overview of absolute errors of model predictions, by data transformation used. The *x*-axis shows to which age group the results belong, and the number in the curly brackets represents the number of samples in that age group.

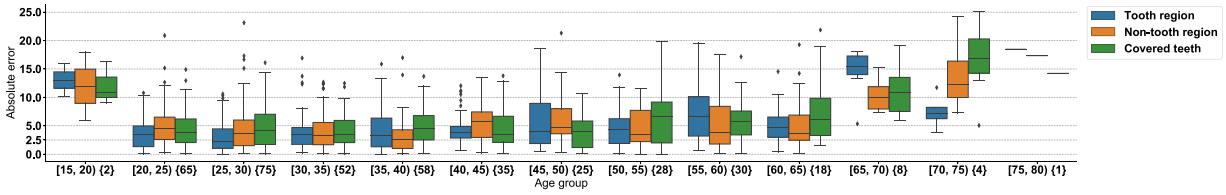

Fig. 9. Errors of estimates for ablation experiments of panoramic dental X-ray image models. An overview of absolute errors of model predictions for all performed ablation experiments. As for Fig. 8, the *x*-axis shows to which age group the results belong, and the number in the curly brackets represents the number of samples in that age group.

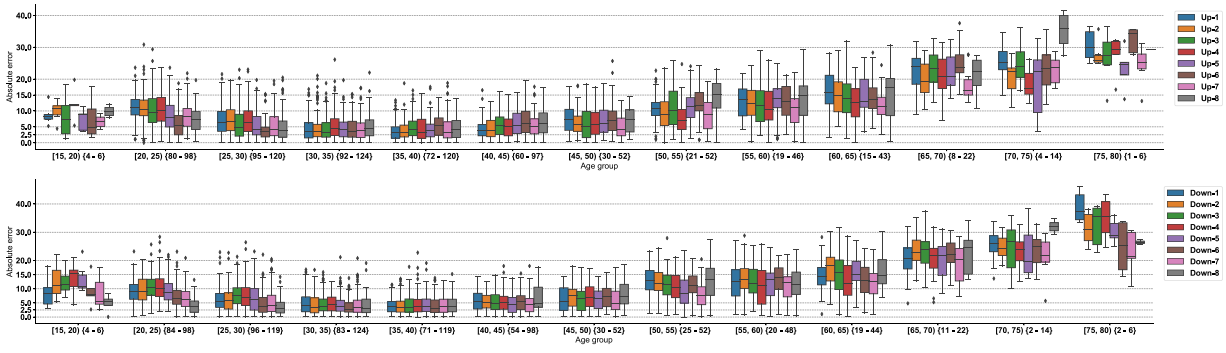

Fig. 10. Errors of estimates for models of X-ray images of individual teeth. An overview of absolute errors per tooth type. The upper figure shows the performance of the maxillary teeth, while the lower figure shows the performance of the mandibular teeth. As in Figs. 8 and 9, the *x*-axis shows to which age group the results belong, and the interval in the curly brackets represents the minimum and the maximum number of samples in that age group, depending on the tooth.

counter those side effects. Augmentation is a commonly used technique in deep learning, and applying it by itself gave improvements across the board, lowering the error in total and in most age groups. With stratified sampling added, results seem to indicate an increase in estimation performance compared to the naive approach but decreased slightly on average in comparison to the augmentation-only variant. However, as can be seen in Table 4, it is a trade-off where a bit of general performance was sacrificed to improve performance for higher aged samples. An overview of performance improvements with augmentation and stratified sampling can be seen in Fig. 8.

In general, these results show that a deep learning approach is not only viable but that the performance of such an approach is

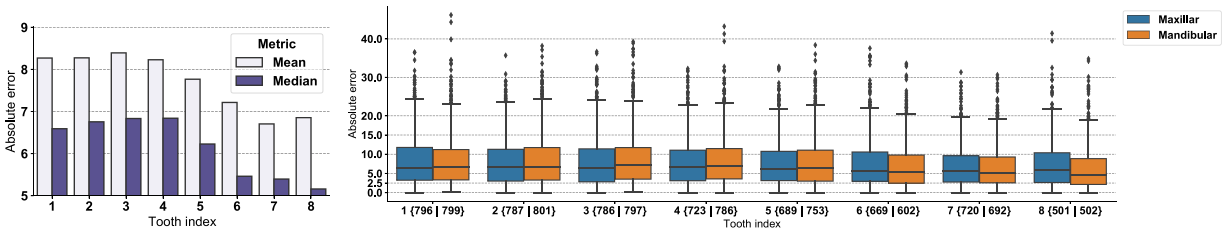

Fig. 11. Age estimation error per tooth type and side. An overview of absolute prediction errors, by tooth type (left) by tooth type and side (right). (Left) indicates a trend whereby molars show lower estimation errors. (Right) show that there is no significant difference between maxillary and mandibular teeth in regards to age estimation.

Table 7
Comparison with findings of (Kvaal et al., 1995).

Tooth type	(Kvaal et al., 1995)		Proposed method	
	R^2	S.S.E.	R^2	S.S.E.
Up-1	0.70	9.5y	0.37	10.68y
Up-2	0.67	10.0y	0.42	10.30y
Up-5	0.60	11.0y	0.43	9.83y
Down-2	0.57	11.5y	0.35	10.88y
Down-3	0.56	11.5y	0.33	11.00y
Down-4	0.64	10.5y	0.36	10.77y

better then conventional methods. While the performance in the 65+ age group does not produce results as usable as for the younger age groups, it is still an improvement. Therefore, this study shows that there are non-developmental indicators in later age groups that can be used to determine a more precise estimate. This performance gain, together with the full automation that this approach offers, highlights the improvements that can be gained in the field of forensic odontology.

The most used current methods in forensic odontology are based on Cameriere et al. (2004), Kvaal et al. (1995). Kvaal et al. (1995) achieve a standard error of estimate of 8.6 years with an R^2 of 0.76 on a sample size of 100 people. The proposed full panoramic dental X-ray image model achieves a standard error of estimate of 5.28 years with an R^2 of 0.8439. Cameriere et al. (2004) achieve a *median* absolute error of 3.7 years and an R^2 of 0.851 with a sample size of 100 people, whereas the proposed method performs with the *median* absolute error of 2.95 years with an R^2 of 0.844. Both classical methods require manual measurements, whereas the proposed model is automated.

One problem when comparing (Cameriere et al., 2004; Drusini et al., 1997; Kvaal et al., 1995) to the proposed model is that the classical methods use individual or group measurements of teeth to estimate the age, whereas the proposed models take the entire dental area into account. Another problem is their usage of healthy, unaltered teeth. As seen in Section 3, we estimate that a significant portion of our dataset contains teeth with alterations, which might lead to decreased performance, as there might not be enough dental markers left intact to determine the age. The family of individual-tooth models is trained to give better insight into the age information present in each tooth alone.

Kvaal et al. (1995) report results for tooth types up-1, up-2, up-5, down-2, down-3 and down-4. The results are compared to the proposed method in Table 7 using the same metrics as in Kvaal et al. (1995). The proposed method performs better just on some of the teeth, but Kvaal et al. have a consistently higher R^2 value. Two reasons might be the cause of this. First is the difference in sample size. Kvaal et al. (1995) have radiographs of 100 dental patients. The second reason is the difference of selection by tooth status. Kvaal et al. (1995) use only healthy, non-modified teeth, while the proposed study does not differentiate between "perfect" teeth, teeth with illnesses, filling, crowns, etc.

Drusini et al. (1997) uses the tooth-coronal index (Ikeda et al., 1985) for estimating age specifically on adults. The study achieves a standard error in the range of 5.88 years to 6.66 years, and an R^2 between 0.77 to 0.89 on their dataset of 433 samples. As with Kvaal et al. (1995), the sample size and explicit selection of full intact teeth should be considered when interpreting the results.

Cameriere et al. (2004) use the maxillary canine (up-3 in our notation) in their method. As mentioned, their method produces a *median* absolute error of 3.7 years. The proposed model for up-3 achieves a *median* absolute error of 6.43 years with an R^2 of 0.3978. Interestingly, our experiments with per tooth age estimation obtained the worst average results on up-3 (Table 5), and down-3 also (Table 6). Again, the discrepancy in the results might be caused by the difference in sample size and tooth selection process, similarly as in Drusini et al. (1997), Kvaal et al. (1995).

Vila-Blanco et al. (2020) have a similar approach by using deep neural networks for age estimation. Their approach is based on a custom-designed convolutional neural network trained from scratch on a dataset of 2289 panoramic dental X-ray images. They report a mean absolute error of 2.84 years on their dataset. Their dataset covers a wider range of ages than ours, going as young as 4.5 years, whereas our starts with 19 years. Like our dataset, most of their samples are concentrated in the younger age groups. Specifically, given Table 1 in Vila-Blanco et al. (2020), the average age of their sample is between 17.1 years and 25.5 years. Our dataset is biased towards younger samples too, but our samples have an average age of 38.17 years, firmly placing them into the "adult" and "post-development" categories. Direct comparison of performance per age group is not possible, as performance per age group from their study is not available due to the results per age group being reported with an open lower bound. Given the vastly different nature of the datasets (estimating age in mostly children and young adults with still developing dentition vs age estimation for adults), such a comparison would ultimately be of little use. A comparison of regions of interest cannot be directly done, as our study uses regression models on which GradCAM (Selvaraju, Cogswell, Das, Vedantam, Parikh, & Batra, 2017) cannot be applied. Class activation map methods such as GradCAM, show the response of the network in relation to classification units in the final layer. While the methods themselves can be applied to a regression network, the results cannot be validly interpreted.

The studies Guo et al. (2021), Kim et al. (2021) approach the problem of age estimation as a classification problem. Instead of estimating the age of a person, they estimate the age group. As the estimation data of those models differ highly from ours (categorical vs continuous data), as well as the objective being fundamentally different, direct comparison is not possible. The training objective of those studies puts a lot of emphasis on the difference between, for example, a 49 and a 50 year old sample (the edges of an age group), valuing that error much higher than the difference between a 30 and a 31 year old (samples within the same age group). The training objective of our study considers an error of 1 year equally across all age groups. Banjšak et al. (2020) approach the problem similarly. Additionally, Banjšak et al. (2020) work on the domain of panoramic dental X-ray images of jaws from the 11th century. Those images contain samples that have been deceased for a long time, that in addition to the wear-and-tear of life contain damage from post-mortem changes, be they biological (decay) or mechanical (broken skulls, lost teeth, or jaw segments). This again disallows for direct compassion of results. Upalananda et al. (2021) on the other hand does not estimate age directly, but one of the final five stages of Demirjian's classification of tooth development.

As our study focuses on adults and seniors, Demirjian's classification of tooth development is not applicable.

It is important to note how well the proposed method works in regards to the tooth type. As shown in Tables 5 and 6, and in Fig. 11 (left), a clear trend favoring molars is observable, which is in contrast with Cameriere et al. (2004) where the maxillary canine is used. As seen in Section 3, the trend does not follow the number of teeth available for each type. When compared to the trend observed with full image models, a higher count of samples does not correlate with a better performing estimator in this case. Molars tend to be missing more often, and they tend to have more dental interventions. Nonetheless, their estimations significantly outperform other tooth types, indicating that their age indicators are more useful for age estimation. It is also discernible that there is no significant difference in performance between maxillary teeth and mandibular teeth, as shown in Fig. 11 (right). In regards to performance per age group, the same trend of decreasing performance with age is observable with all our models.

7. Conclusion

Age estimation is a fundamental task in the forensic process, with teeth and mandible being the perfect indicator due to their physical durability and resistance to post-death decay. Most of the current methods for age estimation include a demanding process of manual measurement of a wide array of morphometric parameters which are then compared to reference values based on the assumed sex and ethnicity. Developmental markers are the strongest indicators for age estimation, but their effectiveness diminishes with age. That is the main reason why age estimation in juveniles is always more successful regardless of the estimation method used. For the same reason, age estimation in adults is a more difficult problem with higher expected errors.

We propose a fully automated approach for age estimation by using deep learning with panoramic dental X-ray images. Experiments have shown that an automated approach with convolutional neural networks is not only viable but works well on adults. In addition, experiments have shown that both teeth and the surrounding skeletal structure contain useful age indicators that can be used independently, but produce better results when used in conjunction. Per-tooth experiments have shown that age indicators can be extracted from images of individual teeth, but with a worse overall performance than the full image model. The fully automated deep learning model for full panoramic dental X-ray images achieves a mean absolute error of 3.96 years, a median absolute error of 2.95 years, and an R^2 of 0.8439.

CRediT authorship contribution statement

Denis Milošević: Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Marin Vodanović:** Methodology, Data curation, Supervision, Validation, Writing – review & editing. **Ivan Galić:** Data curation, Validation. **Marko Subašić:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS). This research has been partially supported by the Croatian Science Foundation, Republic of Croatia (project IP-2020-02-9423). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Altan, A., & Karasu, S. (2020). Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique. *Chaos, Solitons & Fractals*, *140*, Article 110071.

Ardakani, F., Bashardoust, N., & Sheikhha, M. (2007). The accuracy of dental panoramic radiography as an indicator of chronological age in Iranian individuals. *Journal of Forensic Odontostomatology*, *25*(2), 25.

Babar, M., Iqbal, S., & Jan, A. (2008). Essential guidelines for forensic dentistry. *Pakistan Oral and Dental Journal*, *27*, 79–84.

Badran, D. H., Othman, D. A., Thnaibat, H. W., & Amin, W. M. (2015). Predictive accuracy of mandibular ramus flexure as a morphologic indicator of sex dimorphism in Jordanians. *International Journal of Morphology*, *33*(4), 1248–1254. http://dx.doi.org/10.4067/S0717-95022015000400009, URL http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-95022015000400009&lng=en&nrm=iso&tlng=en.

Baličević, V., Lončarić, S., Cárdenes, R., Gonzalez-Tendero, A., Paun, B., Crispi, F., et al. (2015). Assessment of myofiber orientation in high resolution phase-contrast CT images. In *Lecture notes in computer science*, *Functional imaging and modeling of the heart* (pp. 111–119). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-20309-6_13.

Banjšak, L., Milošević, D., & Subašić, M. (2020). Implementation of artificial intelligence in chronological age estimation from orthopantomographic X-ray images of archaeological skull remains. *Bulletin of the International Association for Paleodontology*, *14*(2), 122–129.

Betul Oktay, A. (2017). Tooth detection with convolutional neural networks. In *2017 medical technologies national congress (TIPTEKNO)* (pp. 1–4). http://dx.doi.org/10.1109/TIPTEKNO.2017.8238075.

Cameriere, R., Ferrante, L., & Cingolani, M. (2004). Variations in pulp/tooth area ratio as an indicator of age: a preliminary study. *Journal of Forensic Science*, *49*(2), 1–3. http://dx.doi.org/10.1520/JFS2003259, URL http://www.astm.org/DIGITAL_LIBRARY/JOURNALS/FORENSIC/PAGES/JFS2003259.htm.

Cameriere, R., Ferrante, L., & Cingolani, M. (2006). Age estimation in children by measurement of open apices in teeth. *Journal of Legal Medicine*, *120*(1), 49–52. http://dx.doi.org/10.1007/s00414-005-0047-9, URL https://doi.org/10.1007/s00414-005-0047-9.

Cameriere, R., Ferrante, L., De Angelis, D., Scarpino, F., & Galli, F. (2008). The comparison between measurement of open apices of third molars and Demirjian stages to test chronological age of over 18 year olds in living subjects. *Journal of Legal Medicine*, *122*(6), 493–497. http://dx.doi.org/10.1007/s00414-008-0279-6, URL https://doi.org/10.1007/s00414-008-0279-6.

Carvalho, S. P. M., Silva, R. H. A. d., Lopes-Júnior, C., & Peres, A. S. (2009). Use of images for human identification in forensic dentistry. *Radiologia Brasileira*, *42*(2), 125–130. http://dx.doi.org/10.1590/S0100-39842009000200012, URL http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100-39842009000200012&lng=en&nrm=iso&tlng=pt.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 1800–1807). Honolulu, HI: IEEE, http://dx.doi.org/10.1109/CVPR.2017.195, URL http://ieeexplore.ieee.org/document/8099678/.

Ciapparelli, L. (1992). The chronology of dental development and age assessment. *Practical Forensic Odontology*, 22–42.

Demirjian, A., Goldstein, H., & Tanner, J. M. (1973). A new system of dental age assessment. *Human Biology*, *45*(2), 211–227, URL https://www.jstor.org/stable/41459864.

Drusini, A. G., Toso, O., & Ranzato, C. (1997). The coronal pulp cavity index: A biomarker for age determination in human adults. *American Journal of Physical Anthropology*, *103*(3), 353–363. http://dx.doi.org/10.1002/(SICI)1096-8644(199707)103:3<353::AID-AJPA5>3.0.CO;2-R, https://onlinelibrary.wiley.com/doi/abs/10.1002/.

Dudar, J. C., Pfeiffer, S., & Saunders, S. (1993). Evaluation of morphological and histological adult skeletal age-at-death estimation techniques using ribs. *Journal of Forensic Science*, *38*(3), 677–685.

Eckert, W. G., & Garland, N. (1984). The history of the forensic application in radiology. *The American Journal of Forensic Medicine and Pathology*, *5*(1), 53–56, URL https://journals.lww.com/amjforensicmedicine/Abstract/1984/03000/The_history_of_the_forensic_application_in.10.aspx.

Farah, C., Booth, D., & Knott, S. (1999). Dental maturity of children in Perth, Western Australia, and its application in forensic age estimation. *Journal of Clinical Forensic Medicine*, *6*(1), 14–18.

Galić, I., Nakaš, E., Prohić, S., Selimović, E., Obradović, B., & Petrovečki, M. (2010). Dental age estimation among children aged 5–14 years using the Demirjian method in Bosnia-Herzegovina. *Acta Stomatologica Croatica*, *44*(1), 17–25.

Galić, I., Vodanović, M., Cameriere, R., Nakaˇ s, E., Galić, E., Selimović, E., et al. (2011). Accuracy of Cameriere, Haavikko, and Willems radiographic methods on age estimation on Bosnian–Herzegovian children age groups 6–13. *Journal of Legal Medicine*, *125*(2), 315–321.

Gleiser, I., & Hunt, E. E. (1955). The permanent mandibular first molar: its calcification, eruption and decay. *American Journal of Physical Anthropology*, *13*(2), 253–283.

Greulich, W. W., & Pyle, S. I. (1959). *Radiographic atlas of skeletal development of the hand and wrist*. Stanford University Press.

Grinsven, M. J. J. P. v., Ginneken, B. v., Hoyng, C. B., Theelen, T., & Sánchez, C. I. (2016). Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, *35*(5), 1273–1284. http://dx.doi.org/10.1109/TMI.2016.2526689.

Gulsahi, A., Yuzugullu, B., Imirzalı oglu, P., & Genç, Y. (2008). Assessment of panoramic radiomorphometric indices in Turkish patients of different age groups, gender and dental status. *Dentomaxillofacial Radiology*, *37*(5), 288–292.

Guo, Y.-c., Han, M., Chi, Y., Long, H., Zhang, D., Yang, J., et al. (2021). Accurate age classification using manual method and deep convolutional neural network based on orthopantomogram images. *Journal of Legal Medicine*, http://dx.doi.org/10.1007/s00414-021-02542-x, URL https://doi.org/10.1007/s00414-021-02542-x.

Haavikko, K. (1970). The formation and the alveolar and clinical eruption of the permanent teeth. An orthopantomographic study. *Suomen Hammaslaakariseuran Toimituksia=Finska Tandlakarsallskapets Forhandlingar*, *66*(3), 103–170.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ikeda, N., Umetsu, K., Kashimura, S., Suzuki, T., & Oumi, M. (1985). Estimation of age from teeth with their soft X-ray findings. *Japanese Journal of Legal Medicine*, *39*(3), 244–250.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd international conference on international conference on machine learning: Vol. 37*, (pp. 448–456). Lille, France: JMLR.org.

Jader, G., Fontineli, J., Ruiz, M., Abdalla, K., Pithon, M., & Oliveira, L. (2018). Deep instance segmentation of teeth in panoramic X-ray images. In *2018 31st SIBGRAPI conference on graphics, patterns and images* (pp. 400–407). http://dx.doi.org/10.1109/SIBGRAPI.2018.00058, ISSN: 2377-5416.

Kang, D.-Y., Duong, H. P., & Park, J.-C. (2020). Application of deep learning in dentistry and implantology. *Implantology*, 148–181. http://dx.doi.org/10.32542/implantology.202015, URL www.implantology.or.kr/articles/article/RvNO/.

Kim, S., Lee, Y.-H., Noh, Y.-K., Park, F. C., & Auh, Q.-S. (2021). Age-group determination of living individuals using first molar images based on artificial intelligence. *Scientific Reports*, *11*(1), 1073. http://dx.doi.org/10.1038/s41598-020-80182-8, URL https://www.nature.com/articles/s41598-020-80182-8.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, 2015, San Diego, CA, USA, May 7–9, 2015, Conference track proceedings*. URL http://arxiv.org/abs/1412.6980.

Kvaal, S. I., Kolltveit, K. M., Thomsen, I. O., & Solheim, T. (1995). Age estimation of adults from dental radiographs. *Forensic Science International*, *74*(3), 175–185. http://dx.doi.org/10.1016/0379-0738(95)01760-G, URL http://www.sciencedirect.com/science/article/pii/037907389501760G.

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 275–281. http://dx.doi.org/10.1109/OJEMB.2020.3026928, Conference Name: IEEE Open Journal of Engineering in Medicine and Biology.

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *International Journal of Robotics Research*, *23*(7–8), 673–692.

Limdiwala, P. G., & Shah, J. S. (2013). Age estimation by using dental radiographs. *Journal of Forensic Dental Sciences*, *5*(2), 118–122. http://dx.doi.org/10.4103/0975-1475.119778, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3826039/.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Marroquin, T., Karkhanis, S., Kvaal, S., Vasudavan, S., Kruger, E., & Tennant, M. (2017). Age estimation in adults by dental imaging assessment systematic review. *Forensic Science International*, *275*, 203–211. http://dx.doi.org/10.1016/j.forsciint.2017.03.007, URL https://linkinghub.elsevier.com/retrieve/pii/S0379073817301081.

Matsikidis, G., & Schultz, P. (1982). Altersbestimmung nach dem Gebiss mit Hilfe des Zahnfilms. *Zahnarztl Mitt*, *72*(2524), 2527–2528.

Milošević, D., Vodanović, M., Galić, I., & Subašić, M. (2019). Estimating biological gender from panoramic dental X-ray images. In *2019 11th international symposium on image and signal processing and analysis* (pp. 105–110). http://dx.doi.org/10.1109/ISPA.2019.8868804, ISSN: 1849-2266.

Moorrees, C. F., Fanning, E. A., & Hunt, E. E. (1963). Age variation of formation stages for ten permanent teeth. *Journal of Dental Research*, *42*(6), 1490–1502.

Nolla, C. M., et al. (1952). *The development of permanent teeth*. (Ph.D. thesis), University of Michigan.

Panchbhai, A. (2011). Dental radiographic indicators, a key to age estimation. *Dentomaxillofacial Radiology*, *40*(4), 199–212. http://dx.doi.org/10.1259/dmfr/19478385, URL http://www.birpublications.org/doi/10.1259/dmfr/19478385.

Pereira, C. P., Russell, L. M., de Pádua Fernandes, M., Alves da Silva, R. H., & de Sousa Santos, R. F. V. (2019). Dental age estimation based on development dental atlas assessment in a child/adolescent population with systemic diseases. *Acta Stomatologica Croatica*, *53*(4), 307–317. http://dx.doi.org/10.15644/asc53/4/1, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6993474/.

Roche, A. F., Thissen, D., & Chumlea, W. (1988). *Assessing the skeletal maturity of the hand-wrist: Fels method*. Thomas.

Röntgen, W. C. (1895). Uber eine neue Art von Strahlen. *Sitzungsber Phys Med Ges Wurtzburg*, *9*, 132–141.

Saunders, E. (1837). *The teeth a test of age, considered with reference to the factory children: addressed to the members of both houses of parliament*. Renshaw.

Sayegh, F., & Reed, A. (1968). Calcification in the dental pulp. *Oral Surgery, Oral Medicine, Oral Pathology*, *25*(6), 873–882.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Selmanagić, A., Ajanović, M., Kamber-Ćesir, A., Redžepagić-Vražalica, L., Jelešković, A., & Nakaš, E. (2020). Radiological evaluation of dental age assessment based on the development of third molars in population of bosnia and herzegovina. *Acta Stomatologia Croatica*, *54*(2), 161–167.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*.

Shahrokh Esfahani, M., & Dougherty, E. R. (2014). Effect of separate sampling on classification accuracy. *Bioinformatics*, *30*(2), 242–250. http://dx.doi.org/10.1093/bioinformatics/btt662, URL https://doi.org/10.1093/bioinformatics/btt662.

Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *Lecture notes in computer science, Information processing in medical imaging* (pp. 588–599). Springer International Publishing.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. http://dx.doi.org/10.1186/s40537-019-0197-0, URL https://doi.org/10.1186/s40537-019-0197-0.

Silva, G., Oliveira, L., & Pithon, M. (2018). Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives. *Expert Systems with Applications*, *107*, 15–31.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Singaraju, S., & Sharada, P. (2009). Age estimation using pulp/tooth area ratio: A digital image analysis. *Journal of Forensic Dental Sciences*, *1*(1), 37.

Solheim, T. (1984). Dental age estimation. an alternative technique for tooth sectioning. *The American Journal of Forensic Medicine and Pathology*, *5*(2), 181–184.

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., & Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. *Medical Image Analysis*, *36*, 41–51.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 4278–4284). San Francisco, California, USA: AAAI Press.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In *Lecture notes in computer science, Artificial neural networks and machine learning – ICANN 2018* (pp. 270–279). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01424-7_27.

Tuzoff, D. V., Tuzova, L. N., Bornstein, M. M., Krasnov, A. S., Kharchenko, M. A., Nikolenko, S. I., et al. (2019). Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, *48*(4), Article 20180051.

Upalananda, W., Wantanajittikul, K., Na Lampang, S., & Janhom, A. (2021). Semi-automated technique to assess the developmental stage of mandibular third molars for age estimation. *Australian Journal of Forensic Sciences*, 1–11.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc., URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vila-Blanco, N., Carreira, M. J., Varas-Quintana, P., Balsa-Castro, C., & Tomas, I. (2020). Deep neural networks for chronological age estimation from OPG images. *IEEE Transactions on Medical Imaging*, *39*(7), 2374–2384. http://dx.doi.org/10.1109/TMI.2020.2968765.

Wang, J., Li, X., Shu, X., & Li, W. (2019). Region-manipulated fusion networks for pancreatitis recognition. CoRR URL: arXiv:1907.01744.

Xin, J., Zhang, Y., Tang, Y., & Yang, Y. (2019). Brain differences between men and women: Evidence from deep learning. *Frontiers in Neuroscience*, *13*, http://dx.doi.org/10.3389/fnins.2019.00185, URL https://www.frontiersin.org/articles/10.3389/fnins.2019.00185/full.

Yang, F., Jacobs, R., & Willems, G. (2006). Dental age estimation through volume matching of teeth imaged by cone-beam CT. *Forensic Science International*, *159*, S78–S83. http://dx.doi.org/10.1016/j.forsciint.2006.02.031, URL http://www.sciencedirect.com/science/article/pii/S0379073806000788.