Medical Image Analysis 58 (2019) 101538

Contents lists available at ScienceDirect

Medical Image Analysis

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

Automated age estimation from MRI volumes of the hand

Darko Štern[a,d], Christian Payer[a,b], Martin Urschler[a,c,*]

[a] Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria
[b] Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria
[c] School of Computer Science, The University of Auckland, Auckland, New Zealand
[d] BioTechMed-Graz, Medical University Graz, Graz, Austria

ARTICLE INFO

Article history:
Received 13 July 2018
Revised 21 February 2019
Available online 31 July 2019

Keywords:
Biological age (BA) estimation
Age regression
Convolutional neural network
Random forest

ABSTRACT

Highly relevant for both clinical and legal medicine applications, the established radiological methods for estimating unknown age in children and adolescents are based on visual examination of bone ossification in X-ray images of the hand. Our group has initiated the development of fully automatic age estimation methods from 3D MRI scans of the hand, in order to simultaneously overcome the problems of the radiological methods including (1) exposure to ionizing radiation, (2) necessity to define new, MRI specific staging systems, and (3) subjective influence of the examiner. The present work provides a theoretical background for understanding the nonlinear regression problem of biological age estimation and chronological age approximation. Based on this theoretical background, we comprehensively evaluate machine learning methods (random forests, deep convolutional neural networks) with different simplifications of the image information used as an input for learning. Trained on a large dataset of 328 MR images, we compare the performance of the different input strategies and demonstrate unprecedented results. For estimating biological age, we obtain a mean absolute error of 0.37 ± 0.51 years for the age range of the $subjects \leq 18$ years, i.e. where bone ossification has not yet saturated. Finally, we validate our findings by adapting our best performing method to 2D images and applying it to a publicly available dataset of X-ray images, showing that we are in line with the state-of-the-art automatic methods for this task.

1. Introduction

The progress of physical maturation of young individuals can be used as a biological marker connected to aging. Estimating *biological age* (BA) from physical development is therefore a highly important topic in both clinical and legal (forensic) medicine applications. In clinical medicine, BA estimation is motivated by diagnosis of endocrinological diseases like accelerated or delayed development in adolescents (Martin et al., 2011), or for optimally planning the time-point of pediatric orthopedic surgery interventions while bones are still growing. Examples of such interventions include leg-length discrepancy correction (Lee et al., 2013) or spinal deformity correction surgery (Wang et al., 2009). In legal medicine, when identification documents of children or adolescents are missing, as may be the case in asylum seeking procedures (Schmeling et al., 2011) or in criminal investigations (Schmeling et al., 2004), estimation of physical maturation is used as an approximation to assess unknown *chronological age* (CA).

Established radiological methods for estimation of physical maturation are based on the visual examination of bone ossification in X-ray images (Greulich and Pyle, 1959; Tanner et al., 1983). Ossification is best followed in the long bones and radiologists mainly examine hand bones due to the large number of assessable bones that are visible in X-ray images of this anatomical region, together with the fact that aging progress is not simultaneous for all hand bones. More specifically, carpal and distal phalanges are the first bones to finish ossification, while in radius and ulna, maturation can be followed up to an age of around 18 years. From the level of ossification assessed by the radiologist, the estimation of physical maturation of an individual is then quantified by associating its maturity to the age of the subjects in the reference atlas who showed the same level of ossification. In the remainder of this manuscript, we will refer to this quantification as *biological age as estimated by radiologists* (BAR).

[*] Corresponding author at: Ludwig Boltzmann Institute for Clinical Forensic Imaging, Graz, Austria.
E-mail address: martin.urschler@cfi.lbg.ac.at (M. Urschler).

https://doi.org/10.1016/j.media.2019.101538

A major drawback of the widely used BAR approach is exposure to ionizing radiation, which cannot be justified in legal medicine applications for examining healthy children and adolescents without any diagnostic reasons. Furthermore, the dependency on subjective visual comparison to reference X-ray images makes these methods prone to high inter- and intra-rater variability (Bull et al., 1999; Kaplowitz et al., 2011). Finally, the population used in building the reference atlas dates from the middle of the last century (Greulich and Pyle, 1959), which, due to changes in nutritional habits as well as faster development of modern population, leads to this method being considered inaccurate and outdated (Cole et al., 1988; Hackman and Black, 2013; Zabet et al., 2014).

Over the recent years, research in age estimation has exhibited tremendous interest in magnetic resonance imaging (MRI) as an ionizing radiation free alternative to develop new schemes for following ossification progress corresponding to the development of bones (Dvorak et al., 2007; Terada et al., 2013; Tomei et al., 2014). However, all methods currently developed for estimation of physical maturation from MRI are still based on visual examination by a radiologist (Serinelli et al., 2015; Urschler et al., 2016; De Tobel et al., 2019), thus suffering from the same drawback of inter- and intra-rater variability as the established methods based on X-ray images. The present work conducts the first comprehensive evaluation of different strategies that our group has developed for *automatically* estimating the progress of physical maturation from 3D MRI.

1.1. Related work

In the age estimation method proposed by Greulich-Pyle (GP), all hand bones are simultaneously compared to the best matching reference image from an atlas of radiographs (Greulich and Pyle, 1959). The GP method is preferred by radiologists as it is easy to use and fast to apply. However, radiologists have to visually extract and mentally fuse information from different bones, which makes the GP method highly prone to intra- and inter-rater variability. To reduce subjectivity, Tanner-Whitehouse (TW2) proposed the visual examination of 13 selected hand bones that is simplified by separating their ossification process into stages according to textual and visual descriptions (Tanner et al., 1983). Individual stage estimates are then transferred to an ossification score and fused according to a pre-defined nonlinear function. Both, the ossification score and the fusion function, were derived from characteristics of a sample population. Although the TW2 method is more objective and slightly more accurate compared to GP (Cole et al., 1988), it is used less frequently in clinical and forensic practice since it is more time consuming.

Aiming for reliable and accurate estimation without subjective influence of the examiner, automated image analysis methods for age estimation were developed (Tanner and Gibbons, 1994; Pietka et al., 2001). Most prominently, the BoneXpert method (Thodberg et al., 2009) uses Active Appearance Models (Cootes et al., 2001) to automatically segment hand bones and employs principal component analysis to reduce dimensionality of age relevant shape and appearance feature information before regressing age. To reproduce the BAR, the fusion of individual estimations per bone is calibrated using the same pre-defined nonlinear function developed for TW2. Very recently, deep convolutional neural networks (DCNN) have shown to be immensely successful in solving diverse machine learning and computer vision problems, mainly due to their ability to automatically learn task relevant features from large training datasets (LeCun et al., 2015). Spampinato et al. (2017) investigated several deep learning architectures to assess BAR automatically on a publicly available dataset of subjects with different genders and ethnicities in the CA range from 0 to 18 years. Recently, Radiological Society of North America (RSNA) organized a Pediatric Bone Age Challenge intended to show the potential of machine learning and artificial intelligence in learning BAR from 14,036 clinical hand radiographs and corresponding reports obtained from two children's hospitals (Larson et al., 2017). Evaluated on 200 images, whose reference BAR annotation was done by three radiologists, the winner of the competition used the deep Inception V3 CNN (Szegedy et al., 2016) with additional gender information.

A severe drawback of radiographic age estimation techniques is exposure to ionizing radiation. This drawback is further amplified in legal medicine applications, where several anatomical structures are examined to extend the age estimation range beyond 18 years, e.g. for adolescent asylum seekers lacking valid identification documents. For such multi-factorial age estimation, dental radiography of wisdom teeth (Demirjian et al., 1973) and computed tomography of clavicle bones (Kellinghaus et al., 2010) are employed in addition to hand radiographs, thereby significantly increasing the required amount of ionizing radiation. Thus, MRI specific radiological staging schemes are currently seeing a lot of research interest (Serinelli et al., 2015; Baumann et al., 2015). To the best of our knowledge, our group is the only one developing *automatic* MRI-based methods for age estimation in order to simultaneously overcome the problems of (1) exposure to ionizing radiation, (2) necessity to define new, MRI specific staging systems and (3) subjective influence of the examiner. Working with data from an adolescent age range, in Štern et al. (2014) we used a random forest (RF) (Breiman, 2001) to separately regress CA from image intensity based features of 11 selected hand bones. There, a decision tree excluding metacarpal and phalanx information from older subjects served as a heuristic fusion strategy for age estimation, making this method *ad hoc* and dependent on parameter tuning. In Štern and Urschler (2016), we explored the capability of RFs for information fusion by allowing it to internally decide from which bones to learn a subject's CA. Thus, we treated aging as a global developmental process without the need for heuristic fusion schemes as in Štern et al. (2014), or pre-defined nonlinear functions as in Tanner et al. (1983) or Thodberg et al. (2009). However, by introducing more bones into the RF framework, the space from which handcrafted features are generated was increased, making the regression task harder. To simplify this regression problem, unlike Kashif et al. (2016) where generic feature generators like SIFT (Lowe, 2004) were evaluated for bone age assessment, we instead utilized a task specific image preprocessing step emphasizing epiphyseal plates before generating Haar-like features. Following current research trends of replacing handcrafted features in RF with automatically learned ones, in Štern et al. (2016) we proposed a DCNN architecture to combine age information from individual bones in an automatic fashion by letting the DCNN learn directly the features relevant for age estimation.

1.2. Contribution

In recent years, our group has initiated the development of fully automatic age estimation methods from 3D MRI scans of the hand, in order to simultaneously overcome problems of exposure to ionizing radiation, necessity to define new, MRI specific staging systems and to eliminate the subjective influence of the examiner. In this work, we improve our MRI-based age estimation strategies and comprehensively evaluate their performance on a large dataset. Thus, we extend our previous work in the following aspects:

- We provide a theoretical background for understanding the high dimensional, nonlinear regression problem of biological age (BA) estimation from 3D hand MRI.

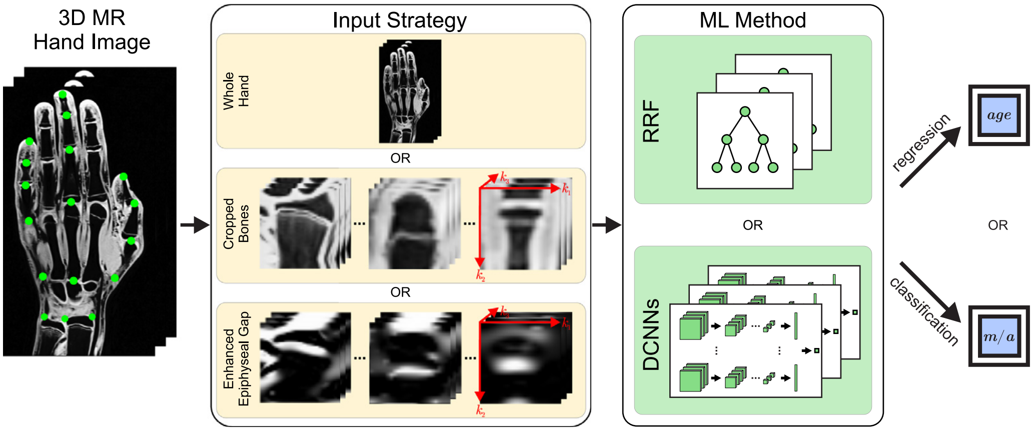


Fig. 1. Overview of our investigated approaches for automatic age estimation from 3D hand MRI. For both regressing age (*age*) and distinguishing minors from adults (*m*/*a*), we evaluate the performance of random regression forests (RRF) and deep convolutional neural networks (DCNNs), when three different input strategies are used for training these state-of-the-art machine learning (ML) methods.

- We thoroughly investigate different strategies for simplifying this regression problem by comparing the performance of RFs using handcrafted feature extraction with DCNNs promising automatic learning of task specific filters, see Fig. 1.
- With the DCNN approach, we inspect the predictor by visualizing the changes in anatomical regions that provide the age relevant information during development.
- We extend our previously proposed RF and DCNN approaches by improving data augmentation and increasing robustness of predictions.
- We evaluate both approaches on a much larger dataset of 328 subjects.
- We significantly improve on our previous age estimation results with both approaches and present the new state-of-the-art method for 3D hand MRI with our DCNN approach.

Additionally, since to the best of our knowledge, our methods are the only ones developed for automatic age estimation from 3D MR images, we validate our findings by adapting our best performing method and applying it to a publicly available dataset comprising of 2D X-ray images. This allows us to compare our results with the state-of-the-art automatic age estimation methods for 2D X-ray images.

2. Method

2.1. Biological age (BA) estimation

Biological age (BA) z of a subject with hand bone maturity $\boldsymbol{x}$ is the expected CA Y of the healthy population that has the same degree of hand bone maturity $\boldsymbol{X} = \boldsymbol{x}$:

$$z = E(Y \mid \boldsymbol{X} = \boldsymbol{x}). \tag{1}$$

For now, let us assume that hand bone maturity $\boldsymbol{X}$ is a known biological marker. In the following Section 2.2, we present different strategies for extracting age relevant features $\boldsymbol{x}$ from hand images.

In statistical decision theory, the conditional expectation $E(Y \mid \boldsymbol{X} = \boldsymbol{x})$ in (1) is referred to as the *regression* function, since the best prediction of Y at any point $\boldsymbol{X} = \boldsymbol{x}$ is the conditional mean, when best is measured by average squared error (Hastie et al., 2009). Thus, given the dataset of pairs $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \in (\boldsymbol{X}, Y)$ from N healthy subjects, the conditional expectation in (1) can be modeled with a regression function $f(\cdot; \boldsymbol{\theta})$, whose parameters $\boldsymbol{\theta}$ are estimated as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|Y - f(\boldsymbol{X}; \boldsymbol{\theta})\|^2. \tag{2}$$

For test subject $j \in \{1, \ldots, M\}$, estimated BA $\hat{z}_j$ can then be obtained from the maturity of its hand bones $\boldsymbol{x}_j$ as:

$$\hat{z}_j = f(\boldsymbol{x}_j; \hat{\boldsymbol{\theta}}), \tag{3}$$

Among different functions $f(\cdot; \boldsymbol{\theta})$ that can be used to model the conditional expectation in (1), for the best performing one in terms of generalization, the mean absolute error (MAE) between estimated BA $\hat{z}$ and true BA z is minimal. Unfortunately, the BA z given with (1) is a latent variable, and therefore neither available for training the prediction model nor for its evaluation. Thus, to evaluate different models, the estimated BA $\hat{z}$ can solely be compared to the available CA y, which leads to an MAE $\xi_{CA\to CA}$ defined as:

$$\begin{aligned} \xi_{CA\to CA} = \frac{1}{M}\sum_{j=1}^{M}\left|\hat{z}_j - y_j\right| &= \frac{1}{M}\sum_{j=1}^{M}\left|\hat{z}_j - z_j + z_j - y_j\right| \\ &\leq \frac{1}{M}\sum_{j=1}^{M}\left|\hat{z}_j - z_j\right| + \frac{1}{M}\sum_{j=1}^{M}\left|z_j - y_j\right| \\ &= \xi^{FIT} + \xi^{BV}. \end{aligned} \tag{4}$$

From (4) we see that the estimation error is composed of two components. The MAE between model prediction and BA (ξ^{FIT}) depends on how good a prediction function $f(\cdot; \boldsymbol{\theta})$ models the conditional expectation in (1). While ξ^{FIT} reflects the goodness of fit of the prediction function, the second component (ξ^{BV}), which is derived from the absolute differences between CA and BA of subjects, solely depends on the M subjects studied in the dataset. Known as biological variation (BV), this difference varies in the population up to ±2 years (Ritz-Timme et al., 2000), making comparison of different prediction functions unreliable, when distinct datasets with a limited number of subjects are involved in experiments.

To eliminate the influence of biological variation on the evaluation error, the performance of prediction models can be compared to the *biological age as estimated by radiologists* (BAR) by utilizing the widely accepted Greulich and Pyle (1959) or Tanner et al. (1983) methods. When using BAR as defined by these methods, an error on the testing dataset (ξ^{OVtest}) due to subjective interpretation of the respective method is introduced. However, this error, which is reported around 0.5 years (Lynnerup et al., 2008; Martin et al., 2011), is much smaller than the error due to biological variation, $\xi^{OVtest} << \xi^{BV}$. Nevertheless, an additional estimation error (ξ^{B}) due to the bias towards the datasets used for building the atlas in Greulich and Pyle (1959) or the nonlinear function in Tanner et al. (1983) is also introduced:

$$\xi_{CA \rightarrow BAR} = \xi^{FIT} + \xi^{OVtest} + \xi^{B}. \tag{5}$$

Up until now, we have presumed that CA Y is statistically dependent on the degree of hand bone maturity $\boldsymbol{X}$ over the whole age range. However, this dependence is only valid until the CA of around 18 years, since information on physical development is saturated in hand images after all hand bones have finished ossification. Since all the subjects with saturated ossification information have the same hand bone maturity $\boldsymbol{x}$, a model $f(\cdot\,; \boldsymbol{\theta})$ trained with average squared error in (2) learns to predict the average CA of the subjects older than 18 years in the training dataset. This average value solely depends on the age range of the training data and therefore it is not related to the biological development of the subjects with saturated ossification information. Although the inclusion of multi-factorial age relevant information extracted from wisdom teeth and clavicle bones can extend the range of prediction to CA beyond 18 years as done in legal medicine applications, in this work we focus on age estimation from a single site, i.e. hand bones.

To avoid the problem of saturation of age relevant information, a common approach in the automatic age estimation methods from 2D X-ray images is to use BAR instead of CA in (2) when training the model. Such an approach also provides means for more reliable comparison of different models, since the bias of the datasets used for building the Greulich and Pyle (1959) atlas or the nonlinear function in Tanner et al. (1983) presented in (5) is now integrated into the model and therefore no longer contributes to the evaluation error:

$$\xi_{BAR \rightarrow BAR} = \xi^{FIT} + \xi^{OVtest} + \delta \cdot \xi^{OVtrain}. \tag{6}$$

However, in addition to the observer error on the testing dataset ξ^{OVtest}, if the BAR of a different observer is used during training, an error $\xi^{OVtrain}$ due to his subjective interpretation of the respective reference method (Greulich and Pyle, 1959; Tanner et al., 1983) has to be taken into account. Since the predictions of the two observers are expected to be correlated, we introduce a multiplicative coefficient $\delta \in [0, 1]$ in Eq. (6), which is inversely related to their inter-observer variability.

Finally, it is important to understand that due to the bias being integrated in the model, the estimated age of a model trained on BAR using Greulich and Pyle (1959) or Tanner et al. (1983) is not the BA, but an estimation of the BAR. Furthermore, in legal medicine, the estimation of BAR is often used to approximate CA, even though it is associated with the largest error:

$$\xi_{BAR \rightarrow CA} = \xi^{FIT} + \xi^{OVtrain} + \xi^{B} + \xi^{BV}. \tag{7}$$

2.2. Bone maturity feature extraction

Mapping the maturity of bones $\boldsymbol{x}$ to BA z can be seen as a regression task, where a machine learning method is employed to learn a regression function $f(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$ from annotated training data as defined in (2). However, the maturity of hand bones $\boldsymbol{x}$ is not a single hand bone feature that can be physically measured, but a combination of appearance features that defines the current stadium of a continuous physical process of hand bone maturation. Thus, to learn the regression function $f(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$, the maturity of the hand bones $\boldsymbol{x}$ has to be extracted from each hand image I in a dataset.

In traditional machine learning methods like support vector machines (Cortes and Vapnik, 1995) or random forests (Breiman, 2001), handcrafted feature extraction is employed as a preprocessing step to the regression task. Thus, to generate handcrafted feature extractors that are essential for performing successful age estimation, prior knowledge of the potential image information that define bone maturity is highly important. By automatically learning the task specific filters, state-of-the-art DCNN approaches promise that no separate preprocessing step is needed. However, to achieve that, DCNN models have a significantly higher number of parameters $\boldsymbol{\theta}$ that have to be learned in order to simultaneously extract the relevant image features and perform regression. Consequently, compared to RF, training of a DCNN requires a larger training dataset to prevent over-fitting.

In order to compare the performance of different machine learning methods for 3D MRI age estimation when a limited dataset is available, we conducted an extensive evaluation of three different strategies depending on the simplification level of this high dimensional, nonlinear regression problem, see Fig. 1. Thus, we evaluated the performance of RF and DCNN based methods (Section 2.3) when trained on (1) the *whole raw input image*, (2) the *cropped age relevant structures* to reduce variations in pose (Section 2.2.2), and (3) the response of hand-crafted *filter based enhancement of the age relevant structures* to reduce variation in pose and appearance (Section 2.2.3).

2.2.1. Whole image

Following the spirit of deep CNNs to learn complex tasks solely given raw input images, as also successfully explored by participants at the RSNA Pediatric Bone Age Challenge for age estimation from X-ray images, this strategy takes the whole MRI volume as input for nonlinear regression, without any simplification of the regression problem. This approach, which is highly dependent on the availability of large datasets, is the most demanding to train a prediction model, since age relevant information has to be extracted from the high dimensional space corresponding to all voxels in the whole MRI volume.

2.2.2. Cropping age relevant structures

In this preprocessing step, the input image domain is simplified for the regression task by eliminating hand pose variations in the images. As shown in Fig. 1, we extract the same 13 bones $b \in (b_1, \ldots, b_{13})$ that radiologists use in the TW2 method for BAR estimation. To localize bones in the 3D MRI volumes, we developed a fully automatic method able to predict the location of anatomical landmarks defined between the hand and wrist bones, see Fig. 1. Details of this method are explained in Payer et al. (2016, 2019). The obtained anatomical landmarks are used to crop a cube for each individual bone $I^b = I^b(k_1, k_2, k_3)$ used in the experiments. The size of each cube is chosen in a way that guarantees encapsulation of the epiphysial and metaphysial parts of the bone inside of the cube. By cropping the bones, they are all brought into the same canonic position with the k_2 axis of the cube corresponding to the main bone axis, see Fig. 1. Thus, all bones are aligned and used either directly to learn the regression function in (2) or as an input to the filter based enhancement of the age relevant structures.

2.2.3. Filter based enhancement of age relevant structures

For our investigated age range between 13 and 25 years, the main aging information is contained within the developmental level of the epiphyseal gaps of long bones, since ossification in carpal bones has already been completed. Therefore, this preprocessing step enhances the appearance of the epiphyseal plate compared to surrounding anatomical structures in the 3D image of the cropped bone. Due to the planar shape of the epiphyseal plate, which is orthogonal to the k_2 axis of the cropped bone image (see Fig. 1), the epiphyseal plate can be enhanced by filtering the bone image $I^b = I^b(k_1, k_2, k_3)$ with a 1D Laplacian of Gaussian (LoG) filter along the k_2 axis:

$$\nabla^2 G(k_2; \sigma) = \left(\frac{k_2^4}{\sigma^4} - \frac{1}{\sigma^2} \right) e^{-\frac{k_2^2}{2\sigma^2}}, \tag{8}$$

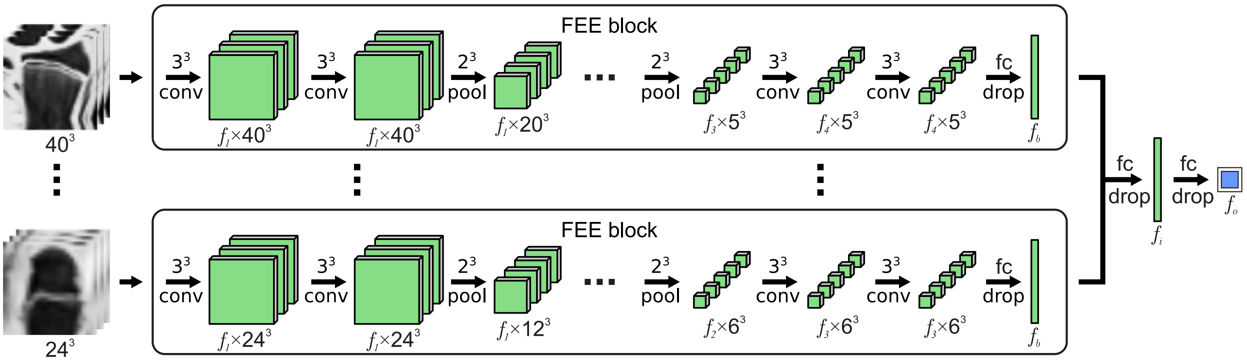


Fig. 2. Inspired by how radiologists perform visual age estimation independently from a set of selected bones, our DCNN architecture consists of one feature extraction and encoding (FEE) block per bone followed by two fully convolutional layers (fc) combining individual bone specific information into the final prediction (f_o) for the hand.

where the parameter σ corresponds to the width of the epiphyseal plate. Since the width of the epiphyseal plate decreases with aging, we sum the filtering responses for different σ to ensure the enhancement of plates with different widths:

$$I_F^b = \sum_{\sigma \in \{\sigma_{\min}, \sigma_{\max}\}} \nabla^2 G(k_2; \sigma) * I^b(k_1, k_2, k_3). \tag{9}$$

2.3. Machine learning methods

We experimented with the two currently dominant machine learning approaches in medical image analysis applications by investigating the RF approach that uses handcrafted feature generation and the DCNN method that promises automatic feature extraction when trained for age estimation. As an input to our RF and DCNN methods, we used either the 13 cropped bone images (see Section 2.2.2), or epiphyseal gap enhanced images of the 13 bones (see Section 2.2.3). Additionally, we experimented with DCNN trained on whole hand images (see Section 2.2.1). Both methods were evaluated for the regression task, where either CA or BAR is used as a target. When estimating whether a subject is a minor or an adult, we used our DCNN architecture trained for a classification task.

2.3.1. RF

In terms of performance, the great success of the RF approach to modeling complex nonlinear functions lies in the ensemble of weak learner models. Each of the trees in an RF is trained in a hierchical fashion to give a prediction with a set of simple decisions from weak learners, with each of them performing better than by random guess. The key aspect of the RF to generalize well even in the presence of a limited dataset is that each tree provides uncorrelated predictions when combined into an ensemble prediction. Nevertheless, compared to DCNNs, which automatically learn features relevant for the task, features used by the RF have to be constructed manually. Since after cropping, the epiphyseal plate is orthogonal to the k_2 axis of the bone image, we generate features for the RF as an average image value along a randomly generated line of length d parallel to the k_2 axis $((k_1, k_2 = K_2, k_3), (k_1, k_2 = K_2 + d, k_3))$. Thus, in each tree node of the RF, a feature is generated based on either the cropped bone images I^b (see Section 2.2.2) for method *rf-reg* or on the epiphyseal plate enhanced bone images I_F^b (see Section 2.2.3) for method *rf-reg-enh*. Specifically, feature generation occurs by randomly selecting a hand bone b out of the set of 13 possible bones, followed by randomly generating a bone image coordinate $(k_1 = K_1, k_2 = K_2, k_3 = K_3)$ as well as a length d along the k_2 axis.

Regression RF: Similar to Criminisi et al. (2013) and Štern and Urschler (2016), a most informative feature/threshold pair is found at each node of the T trees of a forest, such that the training samples reaching the node are best discriminated over the ages. This feature is determined by maximinizing information gain IG:

$$IG = Var(S_n) - \sum_{i \in \{L,R\}} \frac{|S_{n,i}|}{|S_n|} Var(S_{n,i}). \tag{10}$$

Here $Var(\cdot)$ is the variation of age in the set of subjects S_n whose hand bone images reached the node n. The bone images of the subjects in the set S_n are pushed to the left (L) or right (R) child node according to the splitting decision made by thresholding a feature response. The selected feature and threshold from a pool of randomly generated N_F features and N_T thresholds are stored in each node to be used during testing. When the maximum tree depth N_D is reached, or there is no further improvement in IG, the recursive splitting procedure is finished. In the leaf nodes of each tree, we store a list of ages of all subjects that reached the node.

During testing, the feature response is computed for a test subject based on the stored feature parameters, and the subject is passed to the left or right child node according to the result of thresholding. Differently to our previous RF method (Štern and Urschler, 2016), where the estimated age was calculated by averaging the mean ages of subjects stored in the reached leaf nodes, in this work we robustly estimate the age from the combined set of all ages accumulated from the reached leaf nodes of all trees in the RF. Estimated age is calculated as a truncated mean of this set after discarding 5% of ages with highest and 5% with lowest values, respectively.

2.3.2. DCNN

Due to their ability to automatically learn task relevant features, DCNNs have recently attracted great interest in both the computer vision and medical image analysis communities. To prevent overfitting due to the large amount of training parameters, a common approach when working with limited datasets is to finetune a DCNN architecture pretrained on a large training dataset (Spampinato et al., 2017). However, this is only possible when working with 2D images, as no pretrained networks like ImageNet (Russakovsky et al., 2015) exist for 3D images. Therefore, in this work, we focus on exploiting the flexibility of building a task specific DCNN architecture $\phi(\boldsymbol{\theta})$ suitable for working with a limited dataset of 3D images.

Inspired by the idea of the best performing radiological TW2 method that estimates ossification stages of hand bones separately before combining them with a nonlinear function, our proposed DCNN architecture shown in Fig. 2 consists of per-bone feature extraction and encoding (FEE) blocks that are combined with two fully connected layers to model a nonlinear fusion function. Due to their excellent capabilities for extracting task relevant features, our FEE blocks have a similar structure as proposed in the widely used LeNet (LeCun et al., 1998) or VGG (Simonyan and Zisserman, 2015) architectures. Thus, the repeating structure of our FEE block consists of two consecutive convolution layers and one max-pooling layer, where Rectified Linear Units (*ReLUs*) are used as nonlinear activation functions (Nair and Hinton, 2010). Depending on the bone it is related to, our FEE block is either four levels deep for radius and ulna, or three levels deep for the remaining hand long bones, see Fig. 2. To extract features from 3D MR images, we use $3 \times 3 \times 3$ filters in the convolutional layers. The FEE block finishes with a fully connected layer, leading to an encoded output feature representation for each cropped bone f_b. To fuse feature representations into a single feature vector f_i, we use a fully connected layer followed by a *ReLU* activation unit. Finally, the age estimate is obtained after a fully connected layer with a single continuous age prediction output. In our experiments with whole hand images as input, we used the same architecture but with only one FEE block that is four levels deep.

We studied different numbers of filter outputs in our FEE block and outputs of the fully connected layers, but we could not find a single combination of hyperparameters that showed superior performance. Moreover, due to a limited dataset combined with strong data augmentation, the prediction performance remained dependent on when training was stopped long after training has reached a plateau, even when different levels of dropout (Srivastava et al., 2014) were included in the fully connected layers. Differently from our previous work (Štern et al., 2016), where a single combination of hyperparameters was used to generate a DCNN, in this work we followed the traditional and widely used idea of neural network ensembles (Hansen and Salamon, 1990) to improve on generalization capabilities. In common DCNN ensemble approaches, uncorrelated predictions are achieved by each DCNN being trained on different bootstrap samples of the original training dataset, i.e. a bagging strategy (Breiman, 1996), and with different random initializations of the DCNN parameters. Following recent findings in Lee et al. (2015) and Lakshminarayanan et al. (2017) that different random initializations are more beneficial for decorrelating ensemble models, in our approach we additionally use a different set of hyperparameters for each DCNN $\phi_e(\boldsymbol{\theta}_e)$:

$$\phi_E(\boldsymbol{\theta}_E) = \frac{1}{N_e}\sum_{e=1}^{N_e}\phi_e(\boldsymbol{\theta}_e), \tag{11}$$

with $\boldsymbol{\theta}_E = \left[\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_e}\right]$ being the set of parameters of the N_e DCNNs in the ensemble $\phi_E(\boldsymbol{\theta}_E)$.

Regression DCNNs: Associated with the regression target y_n, a training sample $s_n, n \in \{1, \dots, N\}$ can be the 13 cropped bone images I^b (Section 2.2.2) in the *cnn-reg* experiments, the 13 epiphyseal plate enhanced bone images I_F^b (Section 2.2.3) in the *cnn-reg-enh* experiments or a whole hand image I in the *cnn-reg-hand* experiments. Optimizing a regression ensemble of DCNN architectures ϕ_E with ensemble parameters $\boldsymbol{\theta}_E$ is performed by stochastic gradient descent by minimizing an L_2 loss:

$$\hat{\boldsymbol{\theta}}_E^{age} = \text{argmin}_{\boldsymbol{\theta}_E}\frac{1}{2}\sum_{n=1}^{N}\left\|\phi_E(s_n;\boldsymbol{\theta}_E) - y_n\right\|^2 + \lambda\left\|\boldsymbol{\theta}_E\right\|^2, \tag{12}$$

where λ is a parameter determining the strength of the weight decay regularization term.

Classification DCNNs: When estimating whether a subject is a minor or an adult, we compare the classification results derived by thresholding the result of the regression DCNN (*cnn-reg*) with the classification results obtained by training the same architecture with a multinomial logistic classification loss computed as softmax:

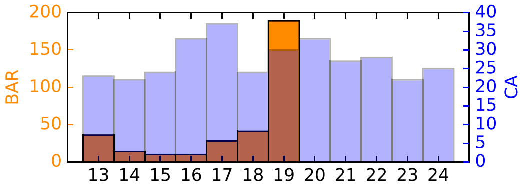


Fig. 3. Biological age as estimated by radiologists (BAR) and chronological age (CA) distributions of our $N = 328$ dataset.

$$\hat{\boldsymbol{\theta}}_E^{m/a} = \text{argmin}_{\boldsymbol{\theta}_E}\sum_{n=1}^{N}\sum_{j\in\{m,a\}} -y_n^j \log\frac{e^{\phi_E^j(s_n;\boldsymbol{\theta}_E)}}{\sum_{k\in\{m,a\}} e^{\phi_E^k(s_n;\boldsymbol{\theta}_E)}} + \lambda\left\|\boldsymbol{\theta}_E\right\|^2. \tag{13}$$

In this *cnn-class* experiment, the binary variable y_n is defined as 1 for a minor (m) and 0 for an adult (a).

2.4. Materials

3D MRI dataset: Our 3D MRI dataset was collected at the Ludwig Boltzmann Institute for Clinical Forensic Imaging in Graz as part of a study investigating the role of MRI in forensic age estimation, which was performed in accordance with the Declaration of Helsinki and approved by the ethical committee of the Medical University of Graz (EK 21,399 ex 09/10). All eligible participants provided written informed consent. For underage participants written consent of the legal guardian was additionally obtained. Exclusion criteria were history of endocrinal, metabolic, genetic or developmental disease. The dataset consists of $N = 328$ 3D hand MRI volumes acquired from male Caucasian volunteers with known CA between 13 and 25 years, with $N = 141$ subjects aged less than or equal to 18 years. All MRI examinations were performed on a 3.0 T scanner (Tim Trio, Siemens Healthineers, Germany), in prone position with outstretched fixed left arm using a head and a neck coil together to cover the hand and the wrist simultaneously. The MRI protocol used for acquiring the 3D volumetric images was an isotropic T1 weighted 3D VIBE sequence with a water selective prepulse, at a resolution of $0.45 \times 0.45 \times 0.9 mm^3$ and image size $294 \times 512 \times 72$ voxels. Image acquisition time was 4 min, however, this time can be reduced by parallel imaging and undersampling (Neumayer et al., 2018). BAR was determined by consensus of two radiologists experienced in age estimation from X-ray and MR images. The distribution of CA and BAR in our MRI dataset is given as a histogram with bins of one year in Fig. 3.

2D X-ray dataset: In order to reproduce our findings from 3D MR age estimation on a publicly available dataset of 2D X-ray images, we extracted all 835 images of subjects older than 10 years from the publicly available *Digital Hand Atlas Database*.[1] This allowed us to compare our method to other state-of-the-art automatic methods developed specifically for 2D X-ray data on the same dataset. The maximal CA of the subjects in this dataset is 19 years and all subjects show an approximately uniform age distribution. The average size of these images is 1563×2169 pixels and we used a histogram matching preprocessing step to normalize intensities regarding contrast and brightness.

[1] http://ipilab.usc.edu/BAAweb/, as of Jun. 2017.

3. Experimental setup

3.1. Implementation details

For all experiments, we scaled and shifted the intensity values of raw hand images such that intensity values are in the range from -1 to 1. In the experiments with cropped bone images, we ensured that the epiphysis and metaphysis part of the bone were contained within the cube. In experiments with 3D MR data, we used trilinear interpolation to resample the cropped cubes to a size of $40 \times 40 \times 40$ voxels for radius and ulna, and due to training time and memory restrictions, $24 \times 24 \times 24$ voxels for the remaining hand long bones. When conducting experiments with 2D X-ray imaging data, the size of the resampled region was 80×80 pixels for all bones. In the whole hand experiments, 3D MR images were resampled to the size of $96 \times 128 \times 32$ voxels and when working with 2D X-ray images only the region that contained the hand was resampled to 256×256 pixels.

In DCNN experiments with both MRI and X-ray datasets, we used an ensemble of five architectures with different numbers of intermediate outputs in the FEE block. In the experiments with 13 cropped hand bones, we used the following $N_e = 5$ combinations of intermediate outputs in the FEE block: (16,16,32,32), (8,16,32,64), (32,32,32,32), (16,32,32,64) and (16,32,64,64). The fully connected layer f_b at the end of the FEE block has 256 outputs generating the feature vector of the individual bone and the fully connected layer f_i generates 1024 outputs. In both the MRI and X-ray image based DCNN experiments with the whole hand, we used only one FEE block with increased number of intermediate output filters: (16,32,64,128), (32,32,64,64), (16,64,64,64), (16,64,64,128) and (16,32,32,128). We also increased the number of outputs of the fully connected layer f_b at the end of the FEE block to 1024. The fully connected layer f_i generates 1024 outputs. In all experiments with DCNN, the fully connected layer f_o generates either one output for regression, or two outputs for classification. Optimization of DCNN was done with the TensorFlow framework (Abadi et al., 2016) using the optimizer ADAM (Kingma and Ba, 2015) with a maximum of 20,000 iterations when trained using cropped bone images and 40,000 iterations when whole hand images were used for training. We used a mini-batch size of 8, a learning rate of 10^{-5}, and an L_2 weight decay parameter $\lambda = 0.0005$ for regularization. In all our experiments with RF, we used $T = 100$ trees with maximal tree depth of $N_D = 25$. At each split node, $N_F = 200$ randomly generated features and $N_T = 10$ candidate thresholds were generated.

In the 3D MR experiments with enhanced epiphyseal plates, the LoG filters from (8) were applied to the images with σ between $\sigma_{\min} = 0.5\,\text{mm}$ and $\sigma_{\max} = 3\,\text{mm}$.

Data augmentation: Differently to our previous work (Štern et al., 2016), where the size of the training dataset was increased by a fixed number of images that were augmented, in this work images were additionally transformed on-the-fly in a data augmentation step using values randomly sampled from a uniform distribution within the following intervals. The intensity values were shifted by $[-0.1, 0.1]$ and scaled by [0.8,1.2]. Additionally, the cropped 3D MR or 2D X-ray bone images were geometrically transformed using translation by $[-2\,\text{mm}, 2\,\text{mm}]$, scaling by [0.85, 1.15], and rotation by $[-5^\circ, 5^\circ]$ in each dimension.

3.2. Evaluation setup

In all our evaluation experiments, we used a four fold cross-validation such that each sample from our dataset was tested exactly once. In each cross-validation fold, the tested samples resembled the same age distribution as in our whole dataset. During training, subjects were randomly sampled from both 3D MR and 2D X-ray datasets such that the age distribution was uniform over the whole training age range. When training DCNN for majority age classification, we randomly sampled subjects such that the two classes were equally represented.

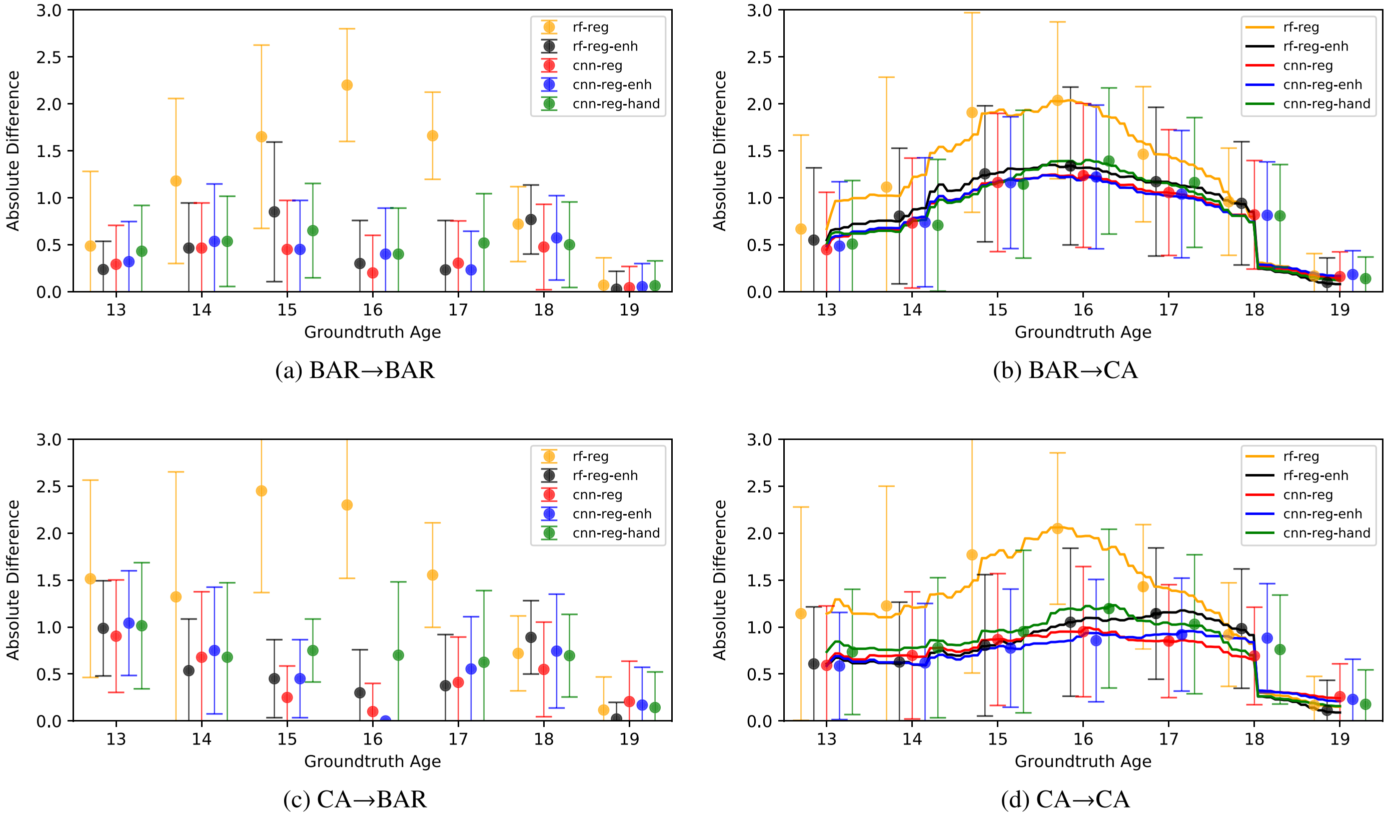
3D MRI age regression experiment: In our MRI age regression experiment, we differentiated four cases depending on which age the method was trained and evaluated on: the method trained on BAR and evaluated on BAR (BAR → BAR), trained on BAR and evaluated on CA (BAR → CA), trained on CA and evaluated on BAR (CA → BAR), and finally trained on CA and evaluated on CA (CA → CA). Due to the limitation in the range of prediction caused by the saturation of the hand bones (see Section 2.1), in the regression experiment CA → BAR and CA → CA, the CA of subjects older than 19 years was clamped to 19 years. We compute mean and standard deviation of absolute differences between predicted and ground truth age as our error measure (MAE). Additionally, we generate a discretized distribution of differences between predicted and ground truth age in a range from -3 years to $+3$ years with a bin size of 1 year.

2D X-ray age regression experiment: To compare performance with other recently published methods on the 2D X-ray hand images, we adapted our DCNN architectures for whole hand images (*cnn-2d-hand*) and cropped 13 hand bones (*cnn-2d-bones*) to work with 2D images. All our 2D DCNN architectures were trained separately for male and female subjects older than 10 years in the X-ray *Digital Hand Atlas Database*. Since the carpal bones are an important source of age relevant information for the subjects in this age range, we also investigated a DCNN architecture (*cnn-2d-bones-carpal*) that in addition to the 13 hand bone images also takes into account an image of the cropped carpal bones. Due to the flexibility of our DCNN architecture (see Fig. 2), this was achieved by adding an additional FEE block to our *cnn-2d-bones* architecture. We evaluated our results in terms of the MAE and discretized distribution between predicted and different ground truth ages (BAR1, BAR2, CA) provided in the dataset.

3D MRI classification experiment: With the majority age threshold of 18 years, we defined minors as positive samples and adults as negative samples. We evaluate classification experiments by inspecting true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). From the true positive rate (TPR, sensitivity) and the true negative rate (TNR, specificity) for different thresholds of the classifiers, we derive receiver operating characteristic (ROC) curves by plotting sensitivity over 1 - specificity. Since we defined correctly classified minors as true positives, sensitivity indicates the percentage of minors that are correctly classified as minors. To compare classifiers, we use the area under the ROC curve (AUC).

4. Results

In our experiments for age estimation from 328 hand MRI volumes, we investigated the performance of RF and DCNN based methods when trained on *whole raw input images, cropped age relevant structures* to reduce variation in pose, or the response of the handcrafted *filter based enhancement of age relevant structures* to reduce variation in pose and appearance. Comprehensive cross-validation results in terms of MAE are shown in Table 1 for all subjects and separately for subjects where the age relevant information is not yet saturated, i.e. with a BAR between 13 and 18 years. Additionally to the MAE, in Table 1 we also show the discretized distribution of the difference between predicted and ground truth age. We compare the results with our previous work (Štern et al., 2016), where a single DCNN was trained either on the same 13 cropped bone images (*cnn-reg-2016*) or 13 epiphyseal plate enhanced bone images (*cnn-reg-enh-2016*). Additionally, we give more detailed insight into the error of our compared methods in Fig. 4.



(a) BAR→BAR

(b) BAR→CA

(c) CA→BAR

(d) CA→CA

Fig. 4. Mean absolute difference of predicted and groundtruth age shown separately for one year age groups for (a) BAR → BAR, (c) CA → BAR, and continuously for (b) BAR → CA and (d) CA → CA (see also Table 1).

There, we show the results separately for each one year age group when predicted age is compared to BAR (Fig. 4a and 4c), and continuously when predicted age is compared to CA (Fig. 4b and 4d).

In Fig. 5, we show the feature response of convolutional layer f_3 (see Fig. 2), when *cnn-reg-hand* is trained on BAR. The heatmap value in the images corresponds to the accumulated feature response of all subjects in our 3D MR dataset on the position relative to a reference hand pose of a subject whose hand contour is superimposed to the image.

In Table 2, we show the results of our DCNN methods for 2D X-ray images evaluated on subjects in the *Digital Hand Atlas Database* older than 10 years and compare results with the state-of-the-art DCNN-based BoNet method of Spampinato et al. (2017).

For majority age classification, the ROC curves and their corresponding AUC are shown in Fig. 6 for our best performing *cnn-reg* method by varying the threshold of the regression predictions. Furthermore, for the same architecture trained with a classification loss (*cnn-class*), ROC curves and AUC are shown by varying the threshold of the prediction output. Additionally, we show the performance when BAR, i.e. the radiological estimation of BA utilizing the widely accepted Greulich and Pyle (1959) method, was thresholded at 18 years ($\text{BAR} \geq 18$) or 19 years ($\text{BAR} = 19$) to distinguish minors from adults. Finally, in Table 3 we compare *cnn-reg*, *cnn-class* and radiological assessment based majority age classifiers in terms of their specificity for several fixed sensitivities, i.e. 99%, 95%, 89% ($\text{BAR} \geq 19$) and 69% ($\text{BAR} \geq 18$), which are also visualized as horizontal lines in the ROC plot in Fig. 6.

Training DCNNs for one fold of the cross-validation was around 6 h for 3D MRI data and 2 h for 2D X-ray images, respectively, while testing a single volume/image takes less than a second on our system with Intel Core i7 CPU and NVIDIA Geforce GTX 1080 GPU with 8 GB of RAM.

5. Discussion

In this work, we presented our fully automatic age estimation method for 3D MRI scans of the hand, which is an objective, software-based solution for age estimation that avoids the need for ionizing radiation. Extending our previous works (Urschler et al., 2015; Štern and Urschler, 2016; Štern et al., 2016), we thoroughly investigated the two most powerful machine learning methods used in medical image analysis, i.e. RF and DCNN, regarding their age estimation accuracy on our unique dataset of 328 MR images.

5.1. 3D MRI age regression

With an absolute deviation of 0.20 ± 0.42 years for our *cnn-reg* method when training on BAR and comparing predictions with BAR, our method based on mimicking radiologists performing TW2 outperforms by a large margin all our previous results for age estimation from MR images (see Table 1). Moreover, on our full dataset (*all subjects*) similar results were also obtained for the *rf-reg-enh* method (0.23 ± 0.45 years). Performance gains compared to our previous work (*cnn-reg-enh-2016* with MAE of 0.36 ± 0.30 years) were achieved by improving data augmentation and increasing robustness of predictions in both methods. Regarding data augmentation, improvements were due to on-the-fly processing and increasing the range of possible intensity and geometrical transformations. Regarding robustness, for our RF-based method we introduced a novel scheme for robustly averaging all the ages stored in the forest's leaf nodes during training as explained in Section 2.3.1. For the DCNN based method, we achieved robustness by using an ensemble of neural networks, following recent developments in medical image analysis (Suk et al., 2017; Benou et al., 2017).

Compared to our previous work, we also increased the evaluation dataset to 328 MR images, which allows us to draw stronger

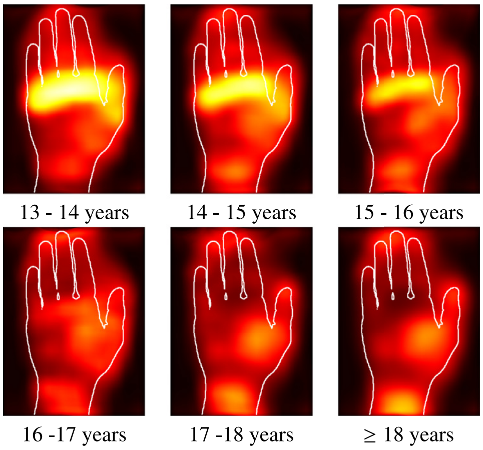


Fig. 5. Visualization of the activation of convolutional layer f_3 for different ranges of BAR predictions, when training our DCNN on the whole hand 3D MRI data. Accumulated feature responses from all datasets are registered to the same reference image, whose contours are superimposed on the image.

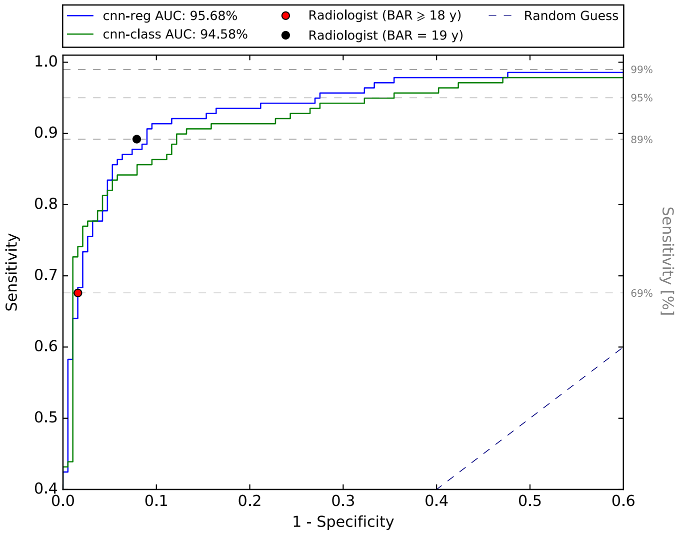


Fig. 6. Results for distinguishing minors from adults when training our best performing DCNN architecture directly for classification (*cnn-class*), when thresholding estimations from our age regression approach (*cnn-reg*), or when using two distinct age thresholds on the radiologist's prediction. Defining correctly classified minors as true positives, we visualize ROC curves and area under the curve (AUC).

statistical conclusions from our findings. This full dataset is roughly uniformly distributed regarding CA, however, when working with BAR the data are highly biased towards subjects, where age relevant information saturates, i.e. subjects with a BAR of 19 years (see Fig. 3). Nevertheless, when evaluating solely on the subset of subjects where the age relevant information is not yet saturated, i.e. with a BA between 13 and 18 years as determined by radiologists, the result of our *cnn-reg* method shows an unprecedented accuracy (0.37 ± 0.51 years). In this age range, with an absolute deviation of 0.48 ± 0.56 years the RF-based method (*rf-reg-enh*) is no longer competitive compared with the DCNN result, which indicates the good performance of RF on subjects where age information is saturated, but underlines its lower capability to discriminate different ages during development, see Fig. 4a for each age group separately. Nevertheless, the results of both investigated methods are below the best reported inter-rater variabilities of radiologists, which are around 0.5 years (Lynnerup et al., 2008; Martin et al., 2011).

Facing a lack of methods for age estimation from 3D MRI outside of our group, we performed an extensive evaluation of our investigated machine learning methods with three different strategies depending on the simplification level of the high dimensional, nonlinear age regression problem: (1) providing the *whole raw input image* to the method, (2) *cropping age relevant structures* to reduce variations in pose, and (3) further simplification by handcrafted *filter based enhancement of the age relevant structures*. The detailed results of our experiments on the subjects where age information is not yet saturated show that for the RF-based methods, the filter based enhancement of the age relevant structures is a crucial preprocessing step (*rf-reg-enh* vs. *rf-reg*, see *subjects* $\leq$ *18* in Table 1 and Fig. 4a for each age group separately). However, given the same preprocessing, it is outperformed by *cnn-reg-enh* with an absolute deviation of 0.41 ± 0.54 years. Interestingly, the best performance of all our experiments (0.37 ± 0.51 years) can be seen for the DCNN based method without any filter based enhancement but solely with cropping age relevant structures (*cnn-reg*). This finding can be interpreted with the handcrafted preprocessing filter being tailored to the enhancement of plate-like structures, which is beneficial during early stages of epiphyseal gap development, but in later stages when the epiphyseal gap loses this characteristic, this filter may eliminate age relevant information. On the other hand, CNNs trained on the whole raw input image information are capable of adapting their convolution filters to extract age relevant features for the whole age range. We did not see this behaviour in our previous work (Štern et al., 2016). This can be explained with the stronger data augmentation used in the present work that drastically increased the training samples, thus allowing the DCNN to learn age relevant features directly from raw image information. Both preprocessing strategies, *cropping age relevant structures* and *filter based enhancement* depend on the location of anatomical structures, for which we have already developed landmark localization methods showing state-of-the-art accuracy and robustness when tested on the same 3D MRI dataset (Payer et al., 2019; Urschler et al., 2018). When skipping the cropping of age relevant structures, but instead applying the DCNN on the whole raw input image (*cnn-reg-hand*), we achieve a slightly worse performance of 0.48 ± 0.53 years. This indicates that despite extensive data augmentation, our training dataset of 328 MR images is not sufficient to directly learn the possible range of anatomical variations simultaneously with pose variations from the whole input images. Nevertheless, our visualization experiment (see Fig. 5) shows that *cnn-reg-hand* is able to automatically learn the convolutional features for extracting age information from different regions depending on the age range. While in younger subjects, the DCNN focuses on the region of the epiphyseal gap in metacarpals and proximal phalanges, for older subjects the remaining age relevant information is located in the epiphyseal gaps of radius and ulna. This visualization also confirms that our cropping of age relevant structures simplifies the age regression problem, thus enabling a DCNN to achieve state-of-the-art performance even in the presence of a limited dataset.

5.2. 2D X-ray age regression

Due to the lack of publicly available MRI datasets or different methods for automatic age estimation from MRI, we additionally adapted our DCNN to 2D images and compared its performance with other recently published methods on the widely used X-ray *Digital Hand Atlas Database*. From Table 2, we see that our proposed method outperforms the BoNet approach of Spampinato et al. (2017) in all experiments regarding BAR estimation, i.e. for male, female and male-female combined evaluation setups. Again, cropping the bones was beneficial (*cnn-2d-hand* vs. *cnn-2d-bones*). Moreover, due to the flexibility of our architecture, we were able to incorporate the carpal bones (*cnn-2d-bones-carpal*), which further improved our results since these bones are relevant for the

Table 1
Results of our automatic age estimation methods for 3D MRI hand images, when trained and evaluated on biological age as estimated by radiologists (BAR) as well as chronological age (CA). We distinguish the evaluation between all subjects in our database, and the subset of subjects less than 18 years, and compare results to our previous work (Štern et al., 2016) as well as using the radiologist's BAR to approximate CA.

	subjects ≤ 18				all subjects	
	BAR → BAR (mean ± std [years]) error histogram [years] (|−3|−2|−1|0|+1|+2|+3|)	**CA → BAR** (mean ± std [years]) error histogram [years] (|−3|−2|−1|0|+1|+2|+3|)	**BAR → CA** (mean ± std [years]) error histogram [years] (|-3|-2|−1|0|+1|+2|+3|)	**CA → CA** (mean ± std [years]) error histogram [years] (|−3|−2|−1|0|+1|+2|+3|)	**BAR → BAR** (mean ± std [years]) error histogram [years] (|−3|−2|−1|0|+1|+2|+3|)	**CA → BAR** (mean ± std [years]) error histogram [years] (|−3|−2|−1|0|+1|+2|+3|)
radiologist	– –	1.02 ± 0.79 |6|13|37|44|29|9|3|	– –	– –	– –	– –
cnn−reg−2016 (Štern et al., 2016)	– –	– –	– –	– –	0.42 ± 0.36 |0|2|24|156|53|5|0|	0.60 ± 0.47 |0|3|11|134|81|11|0|
cnn−reg−enh−2016 (Štern et al., 2016)	– –	– –	– –	– –	0.36 ± 0.30 |0|1|28|176|35|0|0|	0.56 ± 0.44 |1|2|13|128|89|7|0|
rf-reg	1.12 ± 0.84 |8|32|42|32|18|0|0|	1.38 ± 1.02 |14|31|57|24|1|0|0|	1.47 ± 0.98 |16|31|39|21|23|5|2|	1.48 ± 1.00 |12|35|59|21|6|1|0|	0.51 ± 0.78 |8|32|47|212|27|2|0|	0.63 ± 0.93 |14|31|60|197|19|2|0|
rf-reg-enh	0.48 ± 0.56 |0|1|31|73|24|3|0|	0.67 ± 0.60 |0|7|50|52|21|2|0|	1.09 ± 0.79 |4|20|35|42|25|12|2|	0.96 ± 0.74 |5|20|46|42|23|5|0|	0.23 ± 0.45 |0|1|38|258|27|4|0|	0.30 ± 0.52 |0|7|56|239|23|3|0|
cnn-reg	**0.37** ± **0.51** |0|1|26|85|19|1|0|	0.53±0.61 |0|7|35|70|19|1|0|	0.99±0.72 |4|20|36|44|25|11|1|	0.82±0.65 |4|9|48|55|21|4|0|	**0.20** ± **0.42** |0|1|33|267|25|2|0|	0.34±0.53 |0|7|38|226|54|3|0|
cnn-reg-enh	0.41 ± 0.54 |0|2|25|81|23|1|0|	0.68 ± 0.64 |1|5|41|54|26|5|0|	0.98 ± 0.72 |3|16|38|44|24|15|0|	0.83 ± 0.62 |3|11|45|51|26|5|0|	0.21 ± 0.44 |0|2|31|262|31|2|0|	0.38 ± 0.57 |1|5|44|216|55|7|0|
cnn-reg-hand	0.48 ± 0.53 |0|1|41|70|19|1|0|	0.73 ± 0.70 |1|16|47|54|14|0|0|	1.04 ± 0.75 |7|16|42|44|22|9|1|	0.96 ± 0.77 |5|23|44|49|15|4|0|	0.25 ± 0.45 |0|1|47|249|29|2|0|	0.39 ± 0.60 |1|16|50|221|38|2|0|

investigated age range that starts at 10 years. On the same dataset, but on the age range from 2–17 years, the commercial BoneXpert (Thodberg et al., 2009) method, which was trained on an in-house dataset of more than 1500 images, has a root-mean-square error (RMSE) of 0.64 years for the male population (Thodberg and Saevendahl, 2010). Stated as an RMSE, we achieve an error of 0.77 years in our age range between 10 and 19 years, however results are not directly comparable due to the restricted age range used for our method, and due to BoneXpert not being able to reliably predict the challenging last two years when different bones saturate in ossification. Nevertheless, the BoneXpert method is among the top five at the RSNA challenge[2], with no significant difference compared with the top performing, heavily tuned deep CNN methods. The method of (Larson et al., 2017), which was among the top ten at the RSNA challenge, also was evaluated on the *Digital Hand Atlas Database* (Larson et al., 2017). It showed an RMSE of 0.73 years but solely when trained on all X-ray images from the RSNA challenge, while their result decreased to an RMSE of 1.08 years after training on a subset of 1558 images. Thus, our approach, which was not originally developed for 2D hand X-rays, may be considered in line compared with the state-of-the-art in this application.

5.3. BAR vs. CA in age estimation

As explained in Section 2.1, the methods trained with biological age as estimated by radiologists (BAR) are not predicting the "true" biological age, but reproduce the estimation of radiologists when using an atlas based Greulich and Pyle (1959) or a staging based Tanner et al. (1983) method. Instead, as defined by (1), to estimate biological age, methods have to be trained with CA on large datasets representing the population. However, due to the problem of saturation (see Section 2.1), the prediction of a model trained on CA corresponds to BA only for the subjects whose hand bone ossification is not finished. Therefore, from here on we discuss the results only for the subjects younger than 18 years, see *subjects ≤ 18* in Table 1 and Fig. 4. When evaluated on BAR, our methods trained with CA as regression target show an increased error compared with the methods trained with BAR (0.53 ± 0.61 vs. 0.37 ± 0.51 years for *cnn-reg*). This increased error was to be expected, since our dataset of 328 MR images is not large enough to represent the whole population. However, from the distribution of errors on subjects below 18 years shown in Table 1, it can be seen that there is an increase in subjects being underestimated, which supports the observation that children nowadays mature earlier (Zabet et al., 2014). Thus, our results indicate that established radiological BA estimation methods are no longer estimating the average CA of the children that have the same level of bone maturity but introduce a systematic error ξ^B (see Eq. (5)).

When using the model trained on CA for predicting CA, the overall error of 0.82 ± 0.65 years for *cnn-reg* is significantly higher than when compared to BAR. This is due to the biological variation in maturity of subjects having the same chronological age. In legal medicine, this difference due to biological variation is neglected and BAR is used to approximate CA (1.02 ± 0.79 years), which we were able to objectively and automatically reproduce when training *cnn-reg* with BAR (0.99 ± 0.72 years). Moreover, our *cnn-reg* trained on CA is more accurate compared with using BAR to approximate CA. Thus, our results show that DCNNs are not only capable of reproducing age estimation as performed by a radiologist, but also to provide more accurate and objective estimation of CA.

[2] http://rsnachallenges.cloudapp.net/competitions/4, as of Jun. 2018.

Table 2
Results of our automatic age estimation methods adapted to work with 2D images, when applied to a publicly available 2D X-ray image dataset and compared to the recent work of Spampinato et al. (2017) (BoNet). Methods were trained and evaluated for males (M), females (F) and both combined (ALL) using two radiological readings (BAR1,BAR2) as well as chronological age (CA) of 835 subjects older than 10 years.

<table>
<tr><th></th><th></th><th>BAR1 → BAR1
(mean ± std [years])
error histogram [years]
(|−3|−2|−1|0|+1|+2|+3|)</th><th>BAR2 → BAR2
(mean ± std [years])
error histogram [years]
(|−3|-2|−1|0|+1|+2|+3|)</th><th>CA → CA
(mean ± std [years])
error histogram [years]
(|−3|−2|-1|0|+1|+2|+3|)</th></tr>
<tr><td>M</td><td>BoNet</td><td>0.68</td><td>0.74</td><td>–</td></tr>
<tr><td></td><td>(Spampinato et al., 2017)</td><td>–</td><td>–</td><td>–</td></tr>
<tr><td></td><td>cnn-2d-bones</td><td>0.54 ± 0.60</td><td>**0.57 ± 0.60**</td><td>0.86 ± 0.63</td></tr>
<tr><td></td><td></td><td>|2|9|102|214|82|9|0|</td><td>|0|11|89|197|68|4|0|</td><td>|6|28|95|140|93|21|0|</td></tr>
<tr><td></td><td>cnn-2d-bones-carpal</td><td>**0.49 ± 0.60**</td><td>**0.57 ± 0.60**</td><td>0.77 ± 0.60</td></tr>
<tr><td></td><td></td><td>|1|13|66|238|92|8|0|</td><td>|1|3|57|198|100|11|0|</td><td>|0|23|89|157|93|20|0|</td></tr>
<tr><td></td><td>cnn-2d-hand</td><td>0.58 ± 0.64</td><td>0.66 ± 0.65</td><td>0.89 ± 0.67</td></tr>
<tr><td></td><td></td><td>|2|14|97|207|86|10|2|</td><td>|1|12|99|175|69|13|0|</td><td>|8|40|92|137|90|15|1|</td></tr>
<tr><td>F</td><td>BoNet</td><td>0.79</td><td>0.75</td><td>–</td></tr>
<tr><td></td><td>(Spampinato et al., 2017)</td><td>–</td><td>–</td><td>–</td></tr>
<tr><td></td><td>cnn-2d-bones</td><td>0.70 ± 0.61</td><td>0.68 ± 0.65</td><td>1.00 ± 0.73</td></tr>
<tr><td></td><td></td><td>|1|13|105|157|124|17|0|</td><td>|3|5|73|173|137|25|1|</td><td>|5|20|80|124|107|33|7|</td></tr>
<tr><td></td><td>cnn-2d-bones-carpal</td><td>**0.66 ± 0.61**</td><td>**0.60 ± 0.62**</td><td>0.90 ± 0.70</td></tr>
<tr><td></td><td></td><td>|0|7|80|174|132|24|0|</td><td>|3|8|102|194|97|13|0|</td><td>|2|29|67|153|92|29|5|</td></tr>
<tr><td></td><td>cnn-2d-hand</td><td>0.89 ± 0.75</td><td>0.77 ± 0.70</td><td>1.20 ± 0.96</td></tr>
<tr><td></td><td></td><td>|2|41|108|135|95|29|7|</td><td>|2|31|107|156|96|23|2|</td><td>|16|37|63|89|47|21|3|</td></tr>
<tr><td>ALL</td><td>BoNet</td><td>0.73</td><td>0.74</td><td>–</td></tr>
<tr><td></td><td>(Spampinato et al., 2017)</td><td>–</td><td>–</td><td>–</td></tr>
<tr><td></td><td>cnn-2d-bones</td><td>0.62 ± 0.61</td><td>0.62 ± 0.63</td><td>0.93 ± 0.69</td></tr>
<tr><td></td><td></td><td>|3|22|207|371|206|26|0|</td><td>|3|16|162|370|205|29|1|</td><td>|11|48|175|264|200|54|7|</td></tr>
<tr><td></td><td>cnn-2d-bones-carpal</td><td>**0.57 ± 0.61**</td><td>**0.58 ± 0.61**</td><td>0.83 ± 0.66</td></tr>
<tr><td></td><td></td><td>|1|20|146|412|224|32|0|</td><td>|4|11|159|392|197|24|0|</td><td>|2|52|156|310|185|49|5|</td></tr>
<tr><td></td><td>cnn-2d-hand</td><td>0.73 ± 0.72</td><td>0.72 ± 0.68</td><td>1.03 ± 0.82</td></tr>
<tr><td></td><td></td><td>|4|55|205|342|181|39|9|</td><td>|3|43|206|331|165|36|2|</td><td>|24|77|155|226|137|36|4|</td></tr>
</table>

Table 3
Majority age classification performance of regression (*cnn-reg*), classification DCNNs (*cnn-class*), and radiologist evaluated as false positive rates (FPR) and the corresponding number of false positives (FP) for fixed false negative rates.

	FPR (FP)			
	FNR = 31% (BAR ≥ 18)	FNR = 11% (BAR = 19)	FNR = 5%	FNR = 1%
cnn-reg	2.1% (4)	9.0% (17)	**27.5%** (**52**)	**67.2%** (**127**)
cnn-class	**1.1%** (**2**)	12.2% (23)	35.4% (67)	81.0% (153)
radiologist	5.3% (10)	**7.9%** (**15**)	100.0% (189)	100.0% (189)

5.4. 3D MRI majority age classification

To determine legal majority age in forensic applications, a highly relevant question is to classify whether a subject is a minor below 18 years. Therefore, we evaluated our best performing DCNN approach regarding its classification performance either by thresholding the estimation of *cnn-reg*, or by retraining the same architecture with a classification loss (*cnn-class*). Regarding classification performance as measured by AUC, we see from our results in Fig. 6 that thresholding the estimate of *cnn-reg* is superior compared to *cnn-class*. This indicates that the regression loss (12) is a better discriminator, since, unlike the pure 0–1 distance classification loss (13), it also takes into account the distance of the predicted age from the majority age threshold during training.

In Table 3 and with the dots in Fig. 6, we also show the results of BA performed by radiologists being used to distinguish minors from adults. When the majority age threshold is set to 18 years of BAR or more, we can see that more than 30% of minors are wrongly classified as adults, which is the ethically more relevant misclassification that needs to be avoided. This error can be reduced by increasing the threshold to 19 years of BAR, but still the FNR is more than 10%. For further reduction of minors wrongly classified as adults, the GP method performed by radiologists cannot be used. On the other hand, with our method we can more accurately follow closure of epiphyseal gaps until the moment when all bones have finished ossification and minors can no longer be distinguished from adults based solely on hand images. Thus, until that moment our method is able to further reduce misclassifying minors as adults in a tradeoff between sensitivity and FPR. However, due to biological variation, this tradeoff will always result in substantial numbers of misclassifications, and has to be taken into account when using biological development of the hand for estimating chronological age.

6. Conclusion

In this work, an objective software-based solution for automatic age estimation from 3D MRI scans of the hand was presented. Compared with other groups working on automatic age estimation, our approach does not depend on the need for ionizing radiation, which is critical in legal medicine applications involving healthy individuals. Our thorough evaluation of different machine learning methods revealed that our DCNN based regression approach achieves the new state-of-the-art accuracy compared with previous MRI-based methods. Moreover, when adapted for 2D images, the same method is in line with state-of-the-art methods developed specifically for X-ray data.

On our MRI dataset of 328 images, we have shown that the introduction of prior knowledge on where age relevant anatomical information is located is beneficial for the DCNN to learn this high dimensional nonlinear regression problem. Nevertheless, it may be assumed that in the presence of a much larger training dataset, the need for such a preprocessing step could be overcome, which is also indicated by the results of the recent RSNA challenge on age estimation from 14,036 2D hand X-ray images, where the best performing deep learning methods used the whole hand images as input.

While our age regression method has shown to be accurate for the age range up to 18 years, the legally important classification results for determining whether a subject is a minor or an adult are biologically limited due to the saturation of hand bones around 18 years. Recent works indicate that combining estimations from complementary anatomical sites could extend the age estimation to the legally relevant age range up to 25 years. In future work, we will investigate the capabilities for automated multi-factorial MRI-based age estimation to improve the classification accuracy.

Acknowledgments

We thank Thomas Widek for aiding in the preparation of the MRI age estimation dataset, radiologists Sabine Grassegger and Thomas Ehammer for providing the biological age estimates, as well as Bridgette Webb for proofreading this manuscript. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used in this research. This work was supported by the Austrian Science Fund (FWF): P28078-N33.

Declaration of Competing Interest

None.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A System for Large-scale Machine Learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. USENIX Association, Berkeley, CA, USA, pp. 265–283.

Baumann, P., Widek, T., Merkens, H., Boldt, J., Petrovic, A., Urschler, M., Kirnbauer, B., Jakse, N., Scheurer, E., 2015. Dental age estimation of living persons: comparison of MRI with OPG. Forensic Sci. Int. 253, 76–80. doi:10.1016/j.forsciint.2015.06.001.

Benou, A., Veksler, R., Friedman, A., Riklin Raviv, T., 2017. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. Med. Image Anal. 42, 145–159. doi:10.1016/j.media.2017.07.006.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (421), 123–140. doi:10.1007/BF00058655.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324.

Bull, R.K., Edwards, P.D., Kemp, P.M., Fry, S., Hughes, I.A., 1999. Bone age assessment: a large scale comparison of the Greulich and Pyle, and tanner and whitehouse (TW2) methods. Arch. Disease Childh. 81 (2), 172–173. doi:10.1136/adc.81.2.172.

Cole, A.J.L., Webb, L., Cole, T.J., 1988. Bone age estimation: a comparison of methods. Br. J. Radiol. 61 (728), 683–686. doi:10.1259/0007-1285-61-728-683.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685. doi:10.1109/34.927467.

Cortes, C., Vapnik, V., 1995. Support-Vector networks. Mach. Learn. 20 (3), 273–297. doi:10.1023/A:1022627411411.

Criminisi, A., Robertson, D., Konukoglu, E., Shotton, J., Pathak, S., White, S., Siddiqui, K., 2013. Regression forests for efficient anatomy detection and localization in computed tomography scans. Med. Image Anal. 17 (8), 1293–1303. doi:10.1016/j.media.2013.01.001.

De Tobel, J., Hillewig, E., de Haas, M.B., Van Eeckhout, B., Fieuws, S., Thevissen, P.W., Verstraete, K.L., 2019. Forensic age estimation based on T1 SE and VIBE wrist MRI: do a one-fits-all staging technique and age estimation model apply? Eur. Radiol. 29 (6), 2924–2935. doi:10.1007/s00330-018-5944-7.

Demirjian, A., Goldstein, H., Tanner, J.M., 1973. A new system of dental age assessment. Human Biol. 45 (2), 211–227.

Dvorak, J., George, J., Junge, A., Hodler, J., 2007. Age determination by magnetic resonance imaging of the wrist in adolescent male football players. Br. J. Sports Med. 41 (1), 45–52. doi:10.1136/bjsm.2006.031021.

Greulich, W.W., Pyle, S.I., 1959. Radiographic Atlas of Skeletal Development of the Hand and Wrist, 2nd Stanford University Press, Stanford, CA.

Hackman, L., Black, S., 2013. The reliability of the Greulich and Pyle atlas when applied to a modern scottish population. J. Forensic Sci. 58 (1), 114–119. doi:10.1111/j.1556-4029.2012.02294.x.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12 (10), 993–1001. doi:10.1109/34.58871.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer doi:10.1007/b94608.

Kaplowitz, P., Srinivasan, S., He, J., McCarter, R., Hayeri, M.R., Sze, R., 2011. Comparison of bone age readings by pediatric endocrinologists and pediatric radiologists using two bone age atlases. Pediatric Radiol. 41 (6), 690–693. doi:10.1007/s00247-010-1915-0.

Kashif, M., Deserno, T.M., Haak, D., Jonas, S., 2016. Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? a general question answered for bone age assessment. Comput. Biol. Med. 68, 67–75. doi:10.1016/j.compbiomed.2015.11.006.

Kellinghaus, M., Schulz, R., Vieth, V., Schmidt, S., Pfeiffer, H., Schmeling, A., 2010. Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int. J. Legal Med. 124, 321–325. doi:10.1007/s00414-010-0448-2.

Kingma, D. P., Ba, J. L., 2015. Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations, pp. 1–15.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Proceedings of the 31st Conference on Neural Image Processing Systems (NIPS).

Larson, D.B., Chen, M.C., Lungren, M.P., Halabi, S.S., Stence, N.V., Langlotz, C.P., 2017. Performance of a deep-Learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 287 (1), 313–322. doi:10.1148/radiol.2017170236.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. doi:10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., Batra, D., 2015. Why m heads are better than one: training a diverse ensemble of deep networks. arXiv preprint arXiv: 1511.06314

Lee, S.C., Shim, J.S., Seo, S.W., Lim, K.S., Ko, K.R., 2013. The accuracy of current methods in determining the timing of epiphysiodesis. Bone Joint J. 95-B (7), 993–1000. doi:10.1302/0301-620X.95B7.30803.

Lowe, D.G., 2004. Distinctive image features from scale-Invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Lynnerup, N., Belard, E., Buch-Olsen, K., Sejrsen, B., Damgaard-Pedersen, K., 2008. Intra- and interobserver error of the Greulich-Pyle method as used on a danish forensic sample. Forensic Sci. Int. 179, 242.e1–242.e6. doi:10.1016/j.forsciint.2008.05.005.

Martin, D.D., Wit, J.M., Hochberg, Z., Saevendahl, L., van Rijn, R.R., Fricke, O., Cameron, N., Caliebe, J., Hertel, T., Kiepe, D., Albertsson-Wikland, K., Thodberg, H.H., Binder, G., Ranke, M.B., 2011. The use of bone age in clinical practice - Part 1. Hormone Res. Paediatr. 76 (1), 1–9. doi:10.1159/000329372.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 807–814. 10.1.1.165.6419.

Neumayer, B., Schloegl, M., Payer, C., Widek, T., Tschauner, S., Ehammer, T., Stollberger, R., Urschler, M., 2018. Reducing acquisition time for MRI-based forensic age estimation. Sci. Rep. 8, 2063. doi:10.1038/s41598-018-20475-1.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2016. Regressing Heatmaps for Multiple Landmark Localization Using CNNs. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI 2016. Springer, Cham, Athens, pp. 230–238. doi:10.1007/978-3-319-46723-8_27.

Payer, C., Štern, D., Bischof, H., Urschler, M., 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. 54, 207–219. doi:10.1016/j.media.2019.03.007.

Pietka, E., Gertych, A., Pospiech, S., Cao, F., Huang, H.K., Gilsanz, V., 2001. Computer-Assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction. IEEE Trans. Med. Imaging 20 (8), 715–729. doi:10.1109/42.938240.

Ritz-Timme, S., Cattaneo, C., Collins, M.J., Waite, E.R., Schuetz, H.W., Kaatsch, H.J., Borrman, H.I.M., 2000. Age estimation: the state of the art in relation to the specific demands of forensic practise. Int. J. Legal Med. 113 (3), 129–136. doi:10.1007/s004140050283.

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., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. doi:10.1007/s11263-015-0816-y.

Schmeling, A., Garamendi, P.M., Prieto, J.L., Landa, M.I., 2011. Forensic Age Estimation in Unaccompanied Minors and Young Living Adults. In: Vieira, D.N. (Ed.), Forensic Medicine - From Old Problems to New Challenges. InTech, pp. 77–120. doi:10.5772/19261.

Schmeling, A., Olze, A., Reisinger, W., Geserick, G., 2004. Forensic age diagnostics of living people undergoing criminal proceedings. Forensic Sci. Int. 144 (2–3), 243–245. doi:10.1016/j.forsciint.2004.04.059.

Serinelli, S., Panebianco, V., Martino, M., Battisti, S., Rodacki, K., Marinelli, E., Zaccagna, F., Semelka, R.C., Tomei, E., 2015. Accuracy of MRI skeletal age estimation for subjects 12–19. potential use for subjects of unknown age. Int. J. Legal Med. 129 (3), 609–617. doi:10.1007/s00414-015-1161-y.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR).

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R., 2017. Deep learning for automated skeletal bone age assessment in X-ray images. Med. Image Anal. 36, 41–51. doi:10.1016/j.media.2016.10.010.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Štern, D., Ebner, T., Bischof, H., Grassegger, S., Ehammer, T., Urschler, M., 2014. Fully automatic bone age estimation from left hand MR Images. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI. Springer, Cham, Boston, pp. 220–227. doi:10.1007/978-3-319-10470-6_28.

Štern, D., Payer, C., Lepetit, V., Urschler, M., 2016. Automated age estimation from hand MRI Volumes using deep learning. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention MICCAI. Springer, Cham, Athens, pp. 194–202. doi:10.1007/978-3-319-46723-8_23.

Štern, D., Urschler, M., 2016. From individual hand bone age estimates to fully automated age estimation via learning-based information fusion. In: Proceedings of the IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, Prague, pp. 150–154. doi:10.1109/ISBI.2016.7493232.

Suk, H.I., Lee, S.W., Shen, D., 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med. Image Anal. 37, 101–113. doi:10.1016/j.media.2017.01.008.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. doi:10.1109/CVPR.2016.308.

Tanner, J.M., Gibbons, R.D., 1994. A computerized image analysis system for estimating tanner-Whitehouse 2 bone age. Hormone Res. 42, 282–287. doi:10.1159/000184210.

Tanner, J.M., Whitehouse, R.H., N, C., Marshall, W.A., Healy, M.J.R., Goldstein, H., 1983. Assessment of Skeletal Maturity and Predicion of Adult Height (TW2 method), 2nd ed. Academic Press.

Terada, Y., Kono, S., Tamada, D., Uchiumi, T., Kose, K., Miyagi, R., Yamabe, E., Yoshioka, H., 2013. Skeletal age assessment in children using an open compact MRI system. Mag. Res. Med. 69 (6), 1697–1702. doi:10.1002/mrm.24439.

Thodberg, H.H., Kreiborg, S., Juul, A., Pedersen, K.D., 2009. The bonexpert method for automated determination of skeletal maturity. IEEE Trans. Med. Imaging 28 (1), 52–66. doi:10.1109/TMI.2008.926067.

Thodberg, H.H., Saevendahl, L., 2010. Validation and reference values of automated bone age determination for four ethnicities. Acad. Radiol. 17 (11), 1425–1432. doi:10.1016/j.acra.2010.06.007.

Tomei, E., Semelka, R.C., Nissman, D., 2014. Text-Atlas of Skeletal Age Determination: MRI of the Hand and Wrist in Children. Wiley-Blackwell, Hoboken doi:10.1002/9781118692202.

Urschler, M., Ebner, T., Štern, D., 2018. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med. Image Anal. 43 (1), 23–36. doi:10.1016/j.media.2017.09.003.

Urschler, M., Grassegger, S., Štern, D., 2015. What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. Ann. Human Biol. 42 (4), 358–367. doi:10.3109/03014460.2015.1043945.

Urschler, M., Krauskopf, A., Widek, T., Sorantin, E., Ehammer, T., Borkenstein, M.H., Yen, K., Scheurer, E., 2016. Applicability of Greulich-Pyle and tanner-Whitehouse grading methods to MRI when assessing hand bone age in forensic age estimation: A Pilot study. Forensic Sci. Int. 266, 281–288. doi:10.1016/j.forsciint.2016.06.016.

Wang, W.W.J., Xia, C.W., Zhu, F., Zhu, Z.Z., Wang, B., Wang, S.F., Yeung, B.H.Y., Lee, S.K.M., Cheng, J.C.Y., Qiu, Y., 2009. Correlation of risser sign, radiographs of hand and wrist with the histological grade of iliac crest apophysis in girls with adolescent idiopathic scoliosis. Spine 34 (17), 1849–1854. doi:10.1097/BRS.0b013e3181ab358c.

Zabet, D., Rérolle, C., Pucheux, J., Telmon, N., Saint-Martin, P., 2014. Can the Greulich and Pyle method be used on french contemporary individuals? Int. J. Legal Med. 129 (1), 171–177. doi:10.1007/s00414-014-1028-7.