



# Automated age estimation of young individuals based on 3D knee MRI using deep learning

Markus Auf der Mauer<sup>1</sup> · Eilin Jopp-van Well<sup>2</sup> · Jochen Herrmann<sup>3</sup> · Michael Groth<sup>3</sup> · Michael M. Morlock<sup>4</sup> · Rainer Maas<sup>5</sup> · Dennis Säring<sup>1</sup>

Received: 23 March 2020 / Accepted: 9 November 2020 / Published online: 17 December 2020  
© The Author(s) 2020

## Abstract

Age estimation is a crucial element of forensic medicine to assess the chronological age of living individuals without or lacking valid legal documentation. Methods used in practice are labor-intensive, subjective, and frequently comprise radiation exposure. Recently, also non-invasive methods using magnetic resonance imaging (MRI) have evaluated and confirmed a correlation between growth plate ossification in long bones and the chronological age of young subjects. However, automated and user-independent approaches are required to perform reliable assessments on large datasets. The aim of this study was to develop a fully automated and computer-based method for age estimation based on 3D knee MRIs using machine learning. The proposed solution is based on three parts: image-preprocessing, bone segmentation, and age estimation. A total of 185 coronal and 404 sagittal MR volumes from Caucasian male subjects in the age range of 13 and 21 years were available. The best result of the fivefold cross-validation was a mean absolute error of  $0.67 \pm 0.49$  years in age regression and an accuracy of 90.9%, a sensitivity of 88.6%, and a specificity of 94.2% in classification (18-year age limit) using a combination of convolutional neural networks and tree-based machine learning algorithms. The potential of deep learning for age estimation is reflected in the results and can be further improved if it is trained on even larger and more diverse datasets.

**Keywords** Age estimation · Magnetic resonance imaging · Knee · Machine learning · Convolutional neural networks

## Introduction

The determination of certain age limits plays a crucial role in asylum applications, criminal proceedings, and professional youth sport. It can have important consequences for the persons in question. For example, special benefits apply for underage refugees [1, 2], specific laws are enforced to accused subjects [3], or exclusion from tournaments can occur for young athletes [4–6]. The retrieval of the chronological age is required whenever there is a lack of documentation or doubt about the alleged age [7–9].

The European Asylum Support Office (EASO) has presented guidelines on age estimation methods [10].

---

Markus Auf der Mauer and Eilin Jopp-van Well shared first authorship.

---

✉ Markus Auf der Mauer  
markusalexander.adm@gmail.com

Extended author information available on the last page of the article.

The recommendation is to perform the assessment of the chronological age using first non-medical and then medical methods. However, the results from non-medical methods, such as personal interviews or psychological assessments, are often inconclusive [11]. Hence, medical methods are necessary. These are based on the visual inspection of growth plate ossification of multiple long bones in medical images. The results from the inspection are then compared with reference data, e.g., Greulich and Pyle [12] or Tanner-Whitehouse 2 [13] hand/wrist atlases, to estimate the age of the subject. Other anatomical sites used for age estimation are the teeth [14], the clavicle [15], and the knee [16]. X-rays are frequently used for image acquisition as they offer a fast and inexpensive examination. But radiation exposure is considered “harm to the body” and should be used as a last option [10]. Moreover, the visual inspection is labor-intensive and subjective to the radiologist or expert conducting the analysis.

To overcome the disadvantages of X-ray-based age estimation, non-invasive, automated, and unbiased methods are required. In recent years, artificial intelligence (AI)

has developed rapidly and gained popularity by enabling automated workflows and setting state of the art results in many medical applications [17, 18]. In age estimation, artificial neural networks (ANNs) seem to be particularly useful because they have the ability to independently extract and learn relevant information from data for a specific task [19]. Recent studies have delivered encouraging outcomes [20–30].

## Goal of this work

The purpose of the current work was to develop a method based on machine learning (ML) for the automated, computer-based, and non-invasive age estimation of male adolescents and young adults using three-dimensional (3D) magnetic resonance images (MRIs) of the knee.

In addition, the current work is dedicated to validating the results of a previous work [26] from 2018, but on a larger database, and then to developing a *new* robust and automated framework for age assessment based on knee MRIs. Furthermore, it can serve as a good comparison to a similar study by Dallora et al. [30] published in 2019, which also used deep learning and knee MRIs. Finally, the motivation to show that the promising results of Stern et al. [27]—who used deep learning for age estimation on the hand, collarbone, and teeth—are also suitable for another anatomical site, namely the knee.

## Materials

The proposed automated and computer-based method for age estimation was applied and evaluated on two datasets:  $n_{cor} = 185$  coronal and  $n_{sag} = 404$  sagittal 3D MRIs of the knee. The coronal dataset DS<sub>cor</sub> was acquired from two prospective studies [16] and [31] and included MRIs of 79 male Caucasian subjects in the age range of 14.41 and 21.66 years (mean  $\pm$  sd: 17.60  $\pm$  1.53 years; minors:

61.62%). The sagittal dataset DS<sub>sag</sub> was acquired from the above-mentioned prospective studies and expanded with retrospective data. It included MRIs of 297 male Caucasian subjects in the age range of 13.00 and 21.83 years (mean  $\pm$  sd: 17.42  $\pm$  2.28 years; minors: 57.92%). The chronological age of the subjects was calculated as the difference between the birth and MRI acquisition dates. The age distributions for both datasets are shown in Fig. 1.

To attain a homogeneous study population, the following criteria were established for the subject recruitment: Caucasian, male, middle to high socio-economic status, raised in Hamburg (Germany) or surroundings, between 13 and 21 years of age, and no chronic diseases or severe bone injuries.

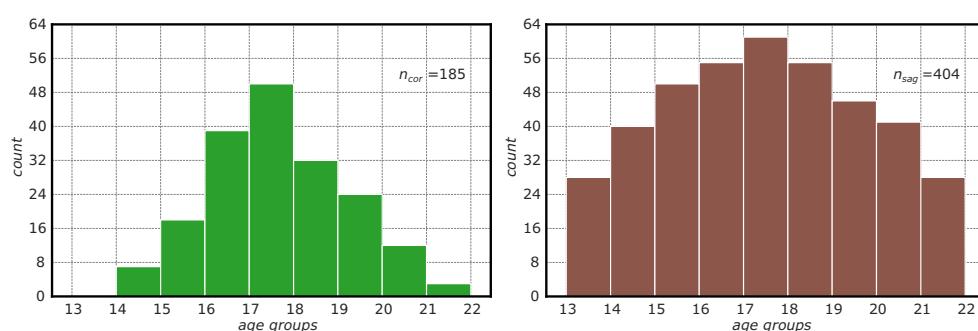
## MRI protocol

MR images were retrieved with six different MRI scanners, four with 3.0T and two with 1.5T field strength, and included sequences in both coronal and sagittal orientation. The protocols used were T1-weighted TSE (Turbo Spin Echo) and T1-weighted SENSE (SENSitivity Encoding). The coronal sequences had the following specifications: TR 600–850 ms, TE 10–20 ms, flip angle 90°. The sagittal sequences were executed with the parameters: TR 590–5000 ms, TE 10–70 ms, flip angle 90–150°.

The acquired coronal MRIs had a resolution in voxels which ranged between  $512 \times 512 \times 24$  and  $800 \times 800 \times 41$  and a voxel size in mm<sup>3</sup> extending from  $0.19 \times 0.19 \times 2.2$  until  $0.49 \times 0.49 \times 4.9$ . The sagittal images had resolutions of  $512 \times 512 \times 24$  voxels– $1050 \times 1050 \times 50$  voxels and the voxel sizes were in the range of  $0.17 \times 0.17 \times 2.2$  mm<sup>3</sup> to  $0.49 \times 0.49 \times 5.2$  mm<sup>3</sup>.

## Additional subject-related data

In addition to the MRI data, additional *subject-related data* were collected for subjects in DS<sub>cor</sub>. This data was



**Fig. 1** Age distribution for datasets DS<sub>cor</sub> (left) and DS<sub>sag</sub> (right). The imbalance in number of individuals per age group was mitigated by data augmentation (Section “Age estimation”)

composed of anthropometric measurements (AM) and the ossification stages (OS) of the growths plates of the knee. AM included the weight, standing and sitting height, and lower leg length (LLL). The weight was measured with a standard body scale device, the standing and sitting height were acquired with an anthropometer corresponding to the standardized measuring length [32, 33], and the LLL was measured using an anthropometric device [34, 35].

For the OS, the three-stage system by Jopp et al. [16] was selected and applied to the epiphyses of the distal femur (DF), proximal tibia (PT), and proximal fibula (PF). The stages of the individual bones were also summed to generate the *score of the knee joint* (SKJ), similar to Galić et al. [36]. A statistical analysis of the classification system on a sub-sample of DS<sub>cor</sub> can be found in a previous study [31].

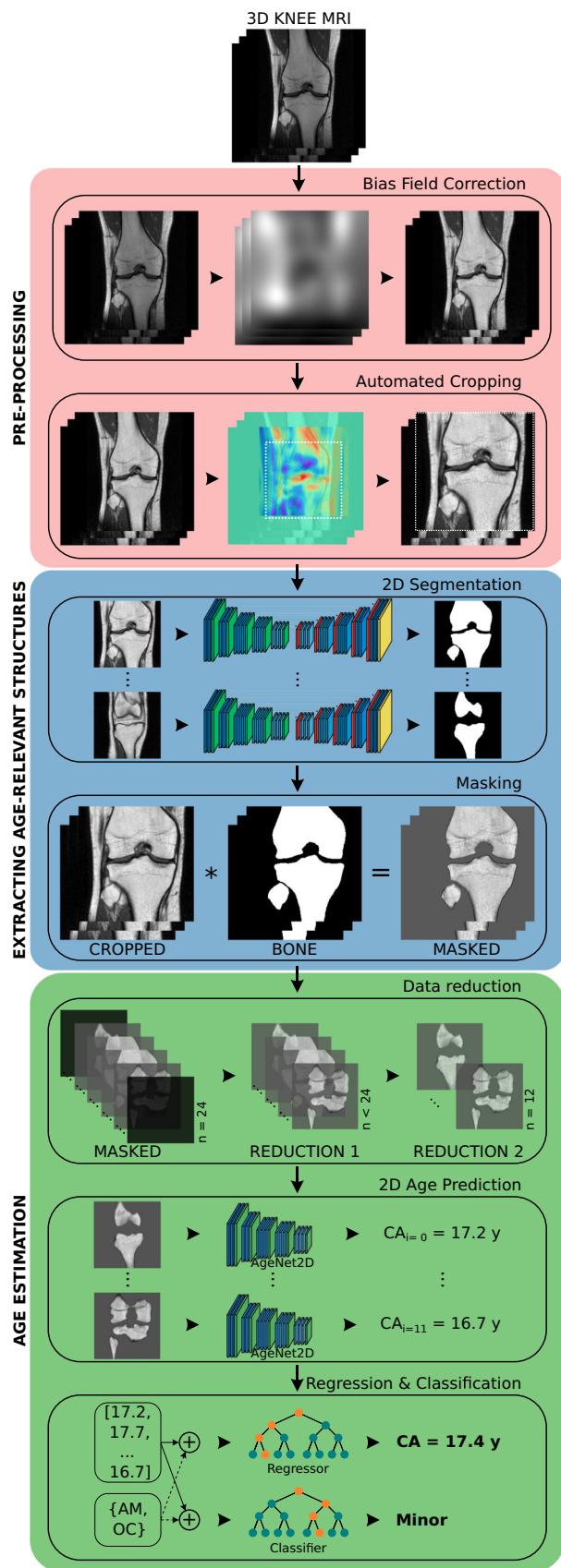
## Methods

The proposed approach of the current work for automated age estimation is presented in Fig. 2. The workflow is composed of three major parts: *Pre-processing* (Section “*Pre-processing*”) of the 3D knee MRIs, *Extracting age-relevant structures* (Section “*Extracting age-relevant structures*”) in the images, and finally, *Age estimation* (Section “*Age estimation*”) based on the pre-processed images and subject-related data acquired from the study participants.

The approach resembles the method presented in a previous work [26] but introduces the following important novelties regarding data, pre-processing, and age estimation:

- Larger DS<sub>cor</sub> and a newly acquired DS<sub>sag</sub>
- A more robust automated cropping approach as part of the pre-processing, capable of cropping both coronal and sagittal knee MRIs
- Improved selection of the most informative images from the MRI volumes
- Age-stratified image augmentation to account for imbalance in the age distribution
- New and robust deep learning model architecture for age estimation on knee MRI
- Classification of minors and adults
- A *repeated* stratified fivefold cross-validation to account for the stochastic nature of deep learning

More details can be found in the following subsections of this manuscript. For a complete and thorough presentation of all developed and tested methods, especially the design and testing of various machine learning algorithms and deep learning network architectures for age estimation, see [37].



**Fig. 2** Proposed solution for automated age estimation

## Pre-processing

*Bias field correction* (BFC) was a necessary step to correct low-frequency intensity non-uniformities in the MRIs caused by inhomogeneities in the magnetic field (Fig. 2—salmon-colored box). N4ITK [38], the de facto standard in the medical image processing field, was used for this purpose. Before the application of N4ITK, all images were downsampled to a common in-plane resolution of  $448 \times 448$  pixels. This step was essential to address the issue of different image resolutions—caused by the use of different MRI scanners and sequences—and to accelerate the processing time of the N4ITK algorithm.

*Automated cropping* was used to generate standardized volume of interests (VOI) from the bias-field-corrected images due to differences in leg position and selection of the field of view (FOV) during MR examination (Fig. 2—salmon-colored box). For this purpose, an approach based on *template matching* (e.g., [39, 40]) was developed. It computed the best correlation between a predefined characteristic region and equally sized regions in the image in a sliding-window fashion. The characteristic region defined for coronal MRIs was a small *patch* showing the tubercles of intercondylar eminence and for sagittal MRIs, a region around the posterior cruciate ligament (PCL) (Fig. 3). Differences in in-plane resolution, resulting from data acquisition with different MRI scanners, were taken into account by resampling the pre-defined patches accordingly. Finally, the standardized VOI was extracted using the point of the highest correlation as the center and then extending 130 mm in *x* and *y* directions and along the entire depth of the MRI volume.

## Extracting age-relevant structures

After pre-processing the acquired images, there were two options to train a deep learning network for age estimation: (a) to feed the pre-processed and almost “raw” MRIs or (b) to feed only age-relevant structures such as bones

and growth plates to simplify the age estimation problem from appearance information. Both options were tested using a convolutional neural network (CNN) described in Section “Age estimation.” In option (a), the training for age estimation led to an underperforming learning progression (Fig. 5). The error of the network for unknown data (orange line) did not improve significantly after hundreds of training cycles, while the error on the data on which the network learned (blue line), i.e., the training data, decreased steadily. The differences between the error curves is known as a *generalization gap* and indicates that the training data are not representative and therefore the model cannot generalize to new data.

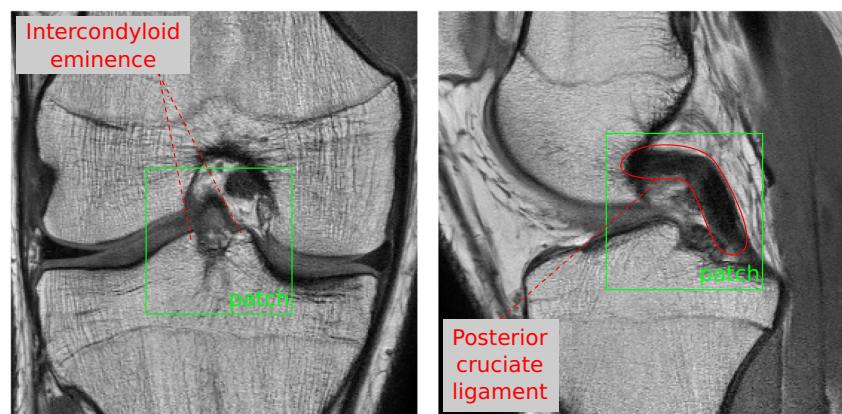
To be able to test option (b), the bones and growth plates had first to be extracted from the images. As in the previous work [26], training a CNN to detect these age-relevant structures proved to be an excellent choice (Fig. 2—blue box).

For this purpose, ground truth structures had to be labelled in the original MRIs in a first step. A sub-set of 100 coronal and 25 sagittal ground truth segmentation maps of the bones were generated using semi-automatic region growing and manual correction by a researcher (MAdM) well experienced in the analysis of 3D knee MRIs and subsequently approved by a radiologist (JH).

To solve the bone segmentation task, the standardized VOIs from the pre-processing had to be further processed to be suitable for the CNN. Each slice was first *downsampled* to a size of  $224 \times 224$  pixels and then *normalized* to transform all pixel intensities to a similar range of values. The *standard score* was used to normalize the input to zero mean and a standard deviation of one.

The final input to the CNN was standardized, downsampled, and normalized two-dimensional image slices of an MRI volume. The architecture of the CNN resembled U-Net [41] and was adapted for the bone segmentation in knee MRIs. This network type is known as *autoencoder*. It first compresses the input to learn features at different scales and then restores the input to predict a segmentation map in

**Fig. 3** Characteristic regions defined in coronal (left) and sagittal MRIs (right) for the template matching algorithm of the automated cropping



the original size. The final architecture for the segmentation task was identical to a previous work [26] (please refer to the paper for further details).

For the CNN-based segmentation a total of 100 three-dimensional coronal MRIs and their corresponding ground truths were available. The data was randomly split into the training set (70%), the validation set (15%), and the test set (15%). To increase the robustness of the network, the size of the training set was virtually increased via *augmentation*. In contrast to the previous work [26], training images were not only translated but additionally rotated, flipped, and cropped using three different VOI sizes.

The training setup, i.e., optimizer, learning rate, loss function, metrics, batch size, epochs, and early stopping, matched the one in [26]. Differently, CNN-based segmentation was not only applied to coronal but to sagittal MRIs as well. To reduce the number of ground truths for training, the CNN for sagittal data was initialized with the weights learned from the coronal model and then retrained for 5 epochs.

Finally, the segmentation maps predicted by the CNNs were used to extract age-relevant structures from the standardized VOIs, i.e., bones and growth joints, by *masking* (Fig. 2—blue box). This step generated the desired input for the age estimation part described next.

The extraction of age-relevant structures as in option (b) led to better and more stable learning curves (Fig. 5) than option (a), since the learned knowledge about bone segmentation could be incorporated into the age estimation network using *transfer learning*.

## Age estimation

The proposed age estimation part of the current work was composed of three parts (Fig. 2—green box): part one was the *data reduction* of the masked MRIs, part two was the *2D age prediction* based on the reduced images using a CNN, and part three was the ultimate *age regression and majority classification* based on the CNN age predictions and the subject-related data using machine learning algorithms. For simplicity, the pipeline is explained for coronal images in this sub-section but was performed for sagittal MRIs as well.

### Data reduction

Due to empty image slices in the masked MRIs and to further simplify the age estimation problem, the image volumes per subject were reduced in two phases. *Reduction 1* removed image slices with no or sparse bone content, only keeping slices with more than 2% bone information related to the total size of the volume. In *Reduction 2*, all MRIs were further reduced to a predefined minimum of 12 slices. This

final reduction ensured a balance of data per subject, similar bone information per slice, and the possibility to train a neural network based on the 3D context. Reduction 2 was implemented by first computing a bone-amount-per-slice distribution and then evenly selecting image slices along the volume with a reference at the center of gravity (CoG) of the distribution.

### 2D age prediction (AgeNet2D)

Prior to designing and training a neural network for age estimation, the 185 coronal, masked, and reduced MRI volumes ( $n_{2D} = 185 \times 12 = 2220$ ) were split into three new sub-sets: training ( $\approx 63\%$ ), validation ( $\approx 18\%$ ), and test ( $\approx 19\%$ ) sets. This *age-stratified random split* ensured that all sets had a similar age distribution. Moreover, it was assured that the training set included the entire age range (14–21 years) of the coronal data, such that the model could effectively predict any age in that range.

Due to the imbalance of the age groups in DS<sub>cor</sub>, 2D image augmentation was applied only to the images of the less represented age groups of the training set. The types of augmentation included the generation of two additional FOVs per image, horizontal translation, and rotation. This increased the number of training samples from 1392 to 2412 two-dimensional images.

*AgeNet2D* is the CNN that was developed as part of the age estimation task in the current work. It was created as a variant of the contracting path, i.e., the encoder, of the segmentation network designed in the previous work [26]. The encoder is the part of a CNN that compresses the input images.

*AgeNet2D* included several different types of layers such as Convolution (*conv*), Exponential Linear Unit (*elu*), Batch-Normalization (*bn*), Dropout (*do*), Max-Pooling (*mp*), Global Average Pooling (*gap*), Fully Connected (*fc*), and Linear Activation (*lin*).

Convolution is the primary operation in CNNs, which convolves a set of small learnable filters, i.e., kernels, across the entire width and height of the image. The outputs of a convolutional layer are known as *feature maps*. ELU [42] is an activation function appended to a convolution operation to introduce a non-linearity into the network. Batch Normalization [43] is a popular technique that standardizes the inputs of a layer which results in more stable and faster training and reduced generalization error due to its regularization effect [44]. Dropout [45] randomly drops convolutional kernels and their connection to other layers during training—controlled by the dropout rate  $p$ —and enables the network to learn independent features [45, 46]. Thus, Dropout acts as an effective regularization measure against overfitting. Max-Pooling reduces the size of feature maps, allowing the network to effectively learn

features at different scales. Furthermore, it has the benefit of decreasing the amount of calculations and lowering the risk of overfitting [47]. GAP [48, 49] and FC layers are often used at the end of CNNs to reduce the dimensionality of the feature maps before the output of the network is generated via a final activation function, such as a Linear Activation.

The final architecture of AgeNet2D (Fig. 4) resulted from a comprehensive *grid search*, which tested different number, order, and combinations of layers as well as different values of hyperparameters (e.g., dropout rate or kernels).

It was composed of five “Down Blocks” with the following sequence of layers: two sets of “ $3 \times 3$  conv → elu → bn → do ( $p = 0.1 \times d$ )” followed by one “ $2 \times 2$  mp”. The *mp*-layer halved the feature maps’ dimension at each network depth  $d$ , thus compressing the input image from  $224^2$  to  $7^2$  pixels. The last part of the network contained the layers sequence: “gap → do ( $p = 0.5$ ) → fc → lin.” The *gap*-layer was used to reduce the dimensionality of the features maps from  $7 \times 7 \times 256$  to  $1 \times 1 \times 256$  by averaging each feature map. Finally, a sequence of *do* ( $p = 0.5$ ), *fc*, and *lin* layers was used to regress the chronological age from the last 256 features. The entire network contained a total of 1.18 million trainable parameters.

Due to the complexity of the age regression problem, the number of kernels per convolutional layer had to be doubled with network depth, starting from 8 up to 256 kernels, in order to find correlating features in the images. In addition, the initial dropout rate of  $p = 0.1$  in the Down Blocks was multiplied with the network depth  $d$  to account for the increasing number of parameters.

After defining the final architecture, the weights and biases of the network layers, i.e., the trainable parameters, had to be initialized. The options were either to assign random values (the standard approach) or to initialize the parameters with “pre-trained values” learned from a different but similar problem. The latter was possible because a neural network had already been trained on knee MRIs, but for the segmentation problem. Both cases were

tested, but the use of transfer learning led to more stable training processes and better results.

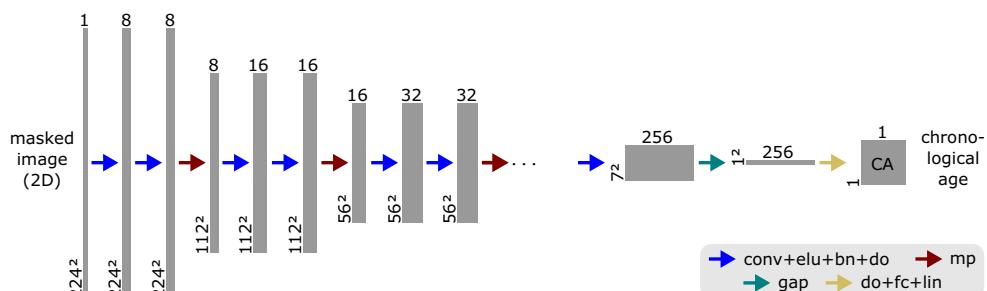
To enable this transfer of knowledge, the segmentation network of the previous work [26] had to be slightly modified to correspond to the layer sequences as shown in Fig. 4, and retrained. Then the learned parameters from the encoder were used to initialize the weights and biases of AgeNet2D.

After initializing the network, the training process was started to learn the best possible 1.18 million parameters for age estimation from knee MRIs. To achieve this, an optimization algorithm must iteratively change parameter values to minimize the error or loss of the task. Using the Adam optimizer [50] and the loss function of the mean squared error (MSE) proved to be a good choice for solving the given problem. MSE is defined as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1)$$

where  $n$  are the number of samples,  $y_i$  the true age of the  $i$ th subject, and  $\hat{y}_i$  the predicted age by the network. To control how much the optimizer modifies the parameters depending on the loss after a certain number of training samples, i.e., batch, the *learning rate* had to be set. The default value of Adam is 0.001, but due to initialization with pre-trained parameters, the learning rate was reduced to 0.0001 to “fine tune” the CNN for the age estimation task. The optimal *batch size*, after which a weight update of the network had to be performed, was 16, and resulted from several test runs. The last hyperparameter that was set was the number of training cycles or *epochs*. Due to the complexity of the problem and the low learning rate, AgeNet2D was trained over 1000 epochs.

Finally, after CNN training, all pre-processed, masked, and reduced 2D MRI could be fed into the network to predict chronological age.



**Fig. 4** AgeNet2D—the CNN developed for age estimation—uses 2D masked knee MR slices (only bone and growth plates) and predicts the chronological age of a subject. The network compresses the input from  $224^2$  to  $7^2$  pixels using Convolution (*conv*) and Max-Pooling (*mp*) operations to compute features at different scales. Exponential

Linear Unit (*elu*) is an activation function, Batch-Normalization (*bn*) standardizes layer inputs, Dropout (*do*) randomly drops convolutional kernels, Global Average Pooling (*gap*), and Fully Connected (*fc*) are further layers that condense the features and finally, the Linear activation (*lin*) layer regresses the chronological age (adapted from [37])

## Regression and classification

To regress the final age of a subject based on knee MRIs, the 12 age predictions of AgeNet2D had to be combined. Instead of taking the average or median or even using the minimum-age concept [51], machine learning algorithms (MLAs) were trained on the 12 age predictions of AgeNet2D to finalize the age estimate per subject. This approach not only led to more accurate results but also enabled the integration of the subject-related data in the age estimation process of DS<sub>cor</sub> (Fig. 2—last part of the green box). For DS<sub>sag</sub>, this additional data was not available.

Classification between minors and adults is an essential task in forensic age estimation. Consequently, this problem was also investigated. The straightforward approach would be to use the regression estimate for classification. However, classification based on initial tests showed fewer errors when an MLA based on 12-age predictions (+ subject-related data) was trained to separate minors from adults. It was particularly useful for more accurate classification of subjects near the 18-year age limit.

The MLAs trained for this work included Support-Vector Machines (SVM), Linear Regression (LR), K-Nearest Neighbors Classifier (KNC), and tree-based algorithms such as Random Forests (RF) and Gradient Boosting (GB).

## Model evaluation

The evaluation of the trained models was performed on the *test set*, i.e., on the fraction of the data from which the model did not learn. To obtain a robust and unbiased estimate of model performance, a repeated stratified  $k$ -fold cross-validation (CV) with  $k = 5$  folds was performed.

During the generation of the fivefold, it was ensured that each test set was unique. The stratification also guaranteed that the folds and sets had a similar age distribution. Finally, due to the stochastic nature of depth learning and most machine learning algorithms, the evaluation was repeated ten times for each fold. This provided a more reliable assessment of the performance of each model.

The *mean absolute error* (MAE) was the principal metric used to evaluate the regression models:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |(y_i - \hat{y}_i)|, \quad (2)$$

where  $y_i$  is the true chronological age of the  $i$ th subject,  $\hat{y}_i$  the prediction by the model, and  $n$  the total number of subjects. Further metrics tracked during the evaluation were the standard deviation of the absolute error, and the 95% confidence interval of the mean (95% CI).

For the classification of the 18-year age limit, the accuracy, the sensitivity, and the specificity were the main evaluation metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5)$$

with TP as true positives, TN true negatives, FP false positives, and FN false negatives. All minors, i.e., below 18 years of age, were defined as TPs. Thus, the accuracy represents the number of correct predictions (whether minor or adult) over the total samples of the test set. The sensitivity corresponds to the number of minors that were correctly classified as such, while specificity is the number of adults that were correctly classified. The metrics range between 0 and 1, where values closer to 1 represent better performance.

Another metric used to compare the performance of different models was the *Area Under the Curve* (AUC) score. It represents the area under the *Receiver Operating Characteristic* (ROC) curve and indicates the capacity of the model to discriminate between minors and adults. The score can take any value between 0 and 1. A perfect classifier has a score of 1 while a “naive” classifier has 0.5. The latter one would simply classify by chance.

## Statistical reference

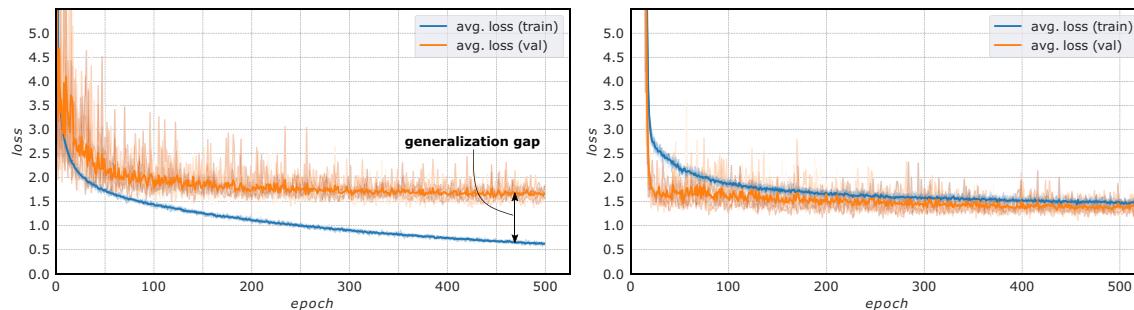
To confirm the added value of the rather complex age estimation framework presented in the current work, the trained models were compared with the *best available guess* [52].

For regression, the best available guess was defined as the mean age of the training set distribution.

In contrast, the best available guess for classification was to predict all subjects as minors, i.e., a “naive” classifier. This resulted in a sensitivity of 1.0, a specificity of 0.0, and an AUC score of 0.5.

## Results

Simplifying the age estimation problem by extracting age-relevant structures in knee MRIs was successful (Fig. 5). The training progression improved and the model was able to generalize on unseen data using masked knee MRIs in comparison to using “raw” images. This was verified quantitatively by evaluating the two scenarios—“raw” or *unsegmented* vs. *masked* MRIs—on a single fold of DS<sub>cor</sub>. Age predictions with a CNN improved markedly from an average MAE of  $0.97 \pm 0.84$  years (“raw”) to  $0.81 \pm 0.65$



**Fig. 5** Training progression of AgeNet2D using “raw” or “unsegmented” knee MRIs (left) vs. using *masked* images (right). The mean squared error *loss* is plotted for the known data (blue), i.e., the training set, and the unknown data (orange), i.e., the validation set, over multiple epochs. Using masked knee MRIs improved

years (masked) on coronal knee MRIs when trained several times on a single fold.

## Model performance on regression

The results for age regression based on AgeNet2D—with and without MLAs—using the underlying DS<sub>cor</sub> and DS<sub>sag</sub> datasets are enlisted in Table 1. The results for each dataset are divided into three scenarios: best available guess, repeated CV results (*rep.*), and the best CV results across all five folds (*best*).

All regression models trained on DS<sub>cor</sub> outperformed the best available guess ( $1.63 \pm 0.99$  years) of the training sets. In addition, the introduction of an MLA to finalize the age prediction improved all metrics by more than 10%

training progression and avoided the *generalization gap* created by CNN training on “raw” images. This supports the assumption that the age estimation problem could be simplified by extracting age-relevant structures from knee MRIs (adapted from [37])

compared to the use of AgeNet2D as the sole predictor. The inclusion of the subject-related data did not boost the performance of the models but rather hurt the MAE. Overall, the best-performing regression model across all folds for DS<sub>cor</sub> combined AgeNet2D with an *Extremely Randomized Trees Regressor* (ETR) as MLA and neglected the subject-related data. It achieved a MAE of  $0.67 \pm 0.49$  years and had a 95% CI of [0.62; 0.72]. The predicted ages by the model over all folds are plotted over the true chronological ages in Fig. 6.

The best available guess on DS<sub>sag</sub> yielded inferior metrics compared to that on DS<sub>cor</sub> due to a larger dataset, a wider age range, and a more uniformly distributed training set. Similar to the coronal results, the regressors on sagittal MRIs outperformed the results using the best

**Table 1** Age regression performance on the test sets in a (repeated) fivefold cross-validation (CV) using masked knee MRIs and subject-related data (srd). Metrics are presented in years and include the mean absolute error (MAE), the standard deviation (SD), and the 95% confidence interval of the mean (95% CI)

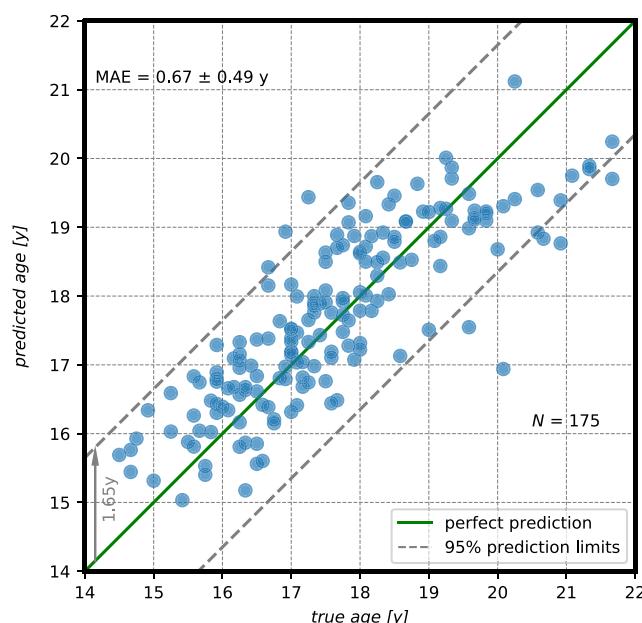
CV	Model	MAE $\pm$ SD	95% CI
DS <sub>cor</sub>			
-	Best available guess <sup>a</sup>	$1.63 \pm 0.99$	[1.59; 1.67]
rep. <sup>b</sup>	AgeNet2D	$0.81 \pm 0.65$	[0.80; 0.83]
	AgeNet2D+MLA <sup>c</sup>	$0.71 \pm 0.55$	[0.71; 0.72]
	AgeNet2D+MLA <sub>srd</sub>	$0.73 \pm 0.55$	[0.71; 0.74]
best <sup>d</sup>	AgeNet2D	$0.79 \pm 0.62$	[0.74; 0.83]
	AgeNet2D+MLA	$0.67 \pm 0.49$	[0.62; 0.72]
	AgeNet2D+MLA <sub>srd</sub>	$0.69 \pm 0.47$	[0.63; 0.74]
DS <sub>sag</sub>			
-	Best available guess	$1.93 \pm 1.20$	[1.90; 1.96]
rep.	AgeNet2D	$0.92 \pm 0.73$	[0.91; 0.94]
	AgeNet2D+MLA	$0.81 \pm 0.62$	[0.79; 0.83]
best	AgeNet2D	$0.89 \pm 0.70$	[0.82; 0.96]
	AgeNet2D+MLA	$0.79 \pm 0.57$	[0.67; 0.90]

<sup>a</sup>Mean age of the training set

<sup>b</sup>Repeated cross-validation for all five folds

<sup>c</sup>Machine learning algorithm

<sup>d</sup>Best cross-validation for all five folds



**Fig. 6** Predicted vs. true chronological age of test subjects from all five folds ( $n = 35 * 5 = 175$ ) using AgeNet2D followed by an Extremely Randomized Trees Regressor (ETR) trained on masked coronal knee MRIs. The green central line highlights a perfect prediction, while the two parallel grey lines encompass 95% of the model predictions

available guess ( $1.93 \pm 1.20$  years) of the training set significantly ( $p < 0.001$ ). The introduction of MLAs improved regression metrics by a large extent. The best-performing model across all folds was AgeNet2D followed by an ETR as MLA with a MAE of  $0.79 \pm 0.57$  years and a 95% CI of  $[0.67; 0.90]$ . Although the metrics were inferior compared to the ones on  $DS_{cor}$ , the sagittal models were based on a broader and larger population.

### Model performance on classification

All listed models achieved above 80% in accuracy, sensitivity, specificity, and AUC on majority classification (Table 2). The best-performing model on  $DS_{cor}$  surpassed 89% in accuracy and combined AgeNet2D and a *Random Forests Classifier* (RFC) as MLA. The model using MLA without subject-related data had a slightly higher average sensitivity and AUC in comparison to the MLA using it. This is the desired outcome to correctly classify more minors.

Surprisingly, the models trained on  $DS_{sag}$  had better classification metrics although the regression results were inferior in comparison to  $DS_{cor}$ . Once more, an RFC delivered the best average metrics when trained on AgeNet2D's age predictions from sagittal MRIs. It attained an accuracy of 90.9%, a sensitivity of 88.6%, a specificity of 94.2%, and an AUC of 94.4% over all folds. The ROC curve

of the model indicates that it has the potential to increase its sensitivity, at the cost of specificity, to increase the amount of correctly classified minors (Fig. 7).

## Discussion

The proposed method enabled a non-invasive, computer-based, and fully automated solution for age estimation based on an actual and homogeneous population. In contrast, current methods in practice rely on labor-intensive and subjective visual assessments of the growth plate ossification stages. This assumes that the stages provide sufficient criteria to distinguish minors from adults. Contradictory results can be observed in studies claiming that the staging system used allows for separation between minors and adults [53–55] while other studies claim the opposite [16, 31, 56, 57]. Further information and differences of studies using MRI for age estimation can be found in the systematic review by De Tobel et al. [58].

### Principal findings

The idea to use bone segmentation to simplify age estimation based on appearance information was successful. AgeNet2D was able to find correlating features with age in the images in a more robust and generalizable manner when using masked instead of “raw” knee MRIs. In [28], the authors had similar findings when cropping age-relevant structures from hand MRIs instead of using the whole image volume as input of a CNN. The benefit of executing bone segmentation *prior* to age estimation in the current work was the possibility to transfer knowledge from one problem to the next. Stern et al. [21] and Dallora et al. [30] were also successful in improving age prediction by using *transfer learning*.

Age predictions by AgeNet2D—based on 2D masked knee images—were effectively combined to a final age per subject using ML-based algorithms to enhance age regression further. An alternative to avoid this two-step age estimation would be to use 3D CNNs on the knee MRI volumes. First tests on both datasets  $DS_{cor}$  and  $DS_{sag}$  showed inferior results in comparison to the combination of CNNs and ML-based regressors. Nevertheless, a 3D-CNN has the capacity to collect more contextual information from the MRI volume and this potential should be evaluated in future work.

When comparing the age regression methods of the current work, the coronal MRI-based models were superior to the models on sagittal MRI. The difference could be related to a larger sagittal dataset with a broader age range and a more uniform age distribution of  $DS_{sag}$ , which could have introduced a higher possibility of variance and outliers.

**Table 2** Performance on majority classification ( $\geq 18$  years) on the test sets in a (repeated) fivefold cross-validation (CV) using masked knee MRIs and subject-related data. Metrics are presented in percentage and include accuracy (Acc.), sensitivity (Sens.), specificity (Spec.), and area under the receiver operating characteristic curve (AUC)

CV	Model	Acc.	Sens.	Spec.	AUC
$DS_{cor}$	- Best available guess <sup>a</sup>	49.3	100.0	0.0	50.0
	rep. <sup>b</sup> AgeNet2D+MLA <sup>c</sup>	85.7	86.4	84.6	90.8
	AgeNet2D+MLA <sub>srd</sub>	83.5	81.4	87.1	89.6
	best <sup>d</sup> AgeNet2D+MLA	89.1	89.1	89.2	92.5
	AgeNet2D+MLA <sub>srd</sub>	89.7	88.2	92.3	92.0
$DS_{sag}$	- Best available guess	52.3	100.0	0.0	50.0
	rep. AgeNet2D+MLA	87.5	88.4	86.1	94.3
	best AgeNet2D+MLA	90.9	88.6	94.2	94.4

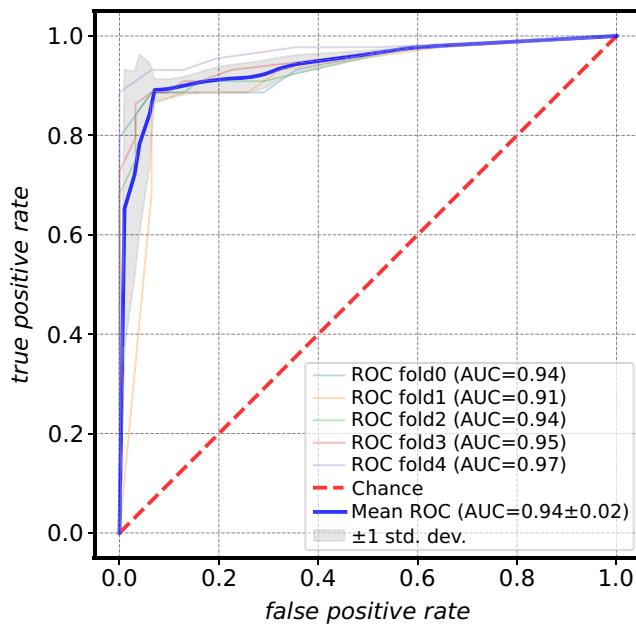
<sup>a</sup>Predict minors for the entire training set

<sup>b</sup>Repeated cross-validation for all five folds

<sup>c</sup>Machine learning algorithm

<sup>d</sup>Best cross-validation for all five folds

Regarding the classification task on the 18-year age limit, comparable results were attained using coronal or sagittal knee MRIs. The high AUC of 0.94 for the classifiers on sagittal data shows the capacity of this model to differentiate between adults and minors. A change of threshold can improve the amount of correctly classified minors which is a more favorable outcome in practice.



**Fig. 7** Receiver operating characteristic (ROC) curve for the best-performing model on majority classification ( $\geq 18$  years). The Random Forests Classifier (RFC) attained an accuracy of 90.9%, a sensitivity of 88.6%, a specificity of 94.2%, and an area under the ROC curve (AUC) of 94.4% averaged over five folds. The high mean AUC indicates that a shift of the threshold can further improve the sensitivity at the expense of the specificity (adapted from [37])

The inclusion of subject-related data, namely ossification stages and anthropometrics data, in the ML-based regressors and classifiers did not increase age estimation performance. In the classification task, it even slightly hurt the outcome. The limited contribution of anthropometric data when used in conjunction with MR images was also observed in [52]. Contrary, the inclusion of ossification stages positively contributed to the performance on age estimation in [21]. However, the data was not included in an MLA but used to pre-train a CNN, to achieve a better initialization of the network. Training a CNN on ossification stages rather than segmentation could be a viable alternative to the method developed in the current work.

### Comparison to similar studies

Similar studies in the field of automated age estimation are the works by Stern et al. [21, 27, 28, 59, 60] and Dallora et al. [30]. Both research groups have developed and analyzed methods for age estimation based on machine learning (incl. deep learning) as well. Further resemblance to the current work is the MRI modality, the number of samples, and the population, which included males in similar age ranges from 13 to 25 years (Table 3). While Dallora et al. also acquired MRIs of the knee, Stern et al. have focussed on different anatomical sites: hand, clavicle, and wisdom teeth.

From Table 3, it is visible that the models from the current work outperform the ones of Stern et al. and Dallora et al. in age regression on the basis of MAE. However, the differences can be caused by several factors such as a different age distribution, anatomical sites, sample number, pre-processing, and CNN architecture. The even lower MAE from a previous work [26] is less comparable because it was a proof of concept and no cross-validation

**Table 3** Comparison of age regression performance between the current work and other studies using automated and MRI-based methods. Techniques used include convolution neural networks (CNNs), machine learning algorithms (MLAs), gray-level variations (GLV), and principal component analysis (PCA). The metrics are presented in years and include the mean absolute error (MAE) and the standard deviation (SD)

Study	N	Gender	Age range (years)	Anatomical site	Technique	MAE ± SD
Pröve et al. [26] <sup>a</sup>	145	Male	14–21	Knee	CNN+MLA	0.48 ± 0.32
Current work <sup>b</sup>	185	Male	14–21	Knee	CNN+MLA	0.67 ± 0.49
Current work <sup>c</sup>	404	Male	13–21	Knee	CNN+MLA	0.79 ± 0.57
Dallora et al. [30]	221 (181)	Male (female)	14–21	Knee	CNN	0.79 <sup>d</sup>
Stern et al. [59]	56	Male	13–19	Hand	MLA	0.85 ± 0.58
Stern et al. [60]	132	Male	13–20	Hand	MLA	0.82 ± 0.56
Stern et al. [21]	103	Male	13–24	Hand, teeth, clavicles	CNN	1.14 ± 0.96
Stern et al. [27]	322	Male	13–25	Hand, teeth, clavicles	CNN	1.01 ± 0.74
Stern et al. [28]	328	Male	13–25	Hand	CNN	0.82 ± 0.65 <sup>e</sup>
Saint-Martin et al. [61]	80 (80)	Male (female)	8–25	Ankle	GLV+PCA	-

<sup>a</sup>Previous work as a proof of concept without *k*-fold cross-validation<sup>b</sup>Results of current work using *coronal* knee MRIs<sup>c</sup>Results of current work using *sagittal* knee MRIs<sup>d</sup>SD was not indicated in the paper, and the metric omits the MAE of 1.37 years for the 21-year-old age group<sup>e</sup>Metric only holds for subjects ≤ 18 years

was performed. A common result of these studies is the improvement in age regression through CNN *pre-training*. In the current work, the weights were initialized using the knowledge from the segmentation task. Stern et al. [21] pre-trained their CNN using information about the maturation of the growth plates by radiological assessment. In contrast, the models of Dallora et al. [30] were pre-trained on ImageNet [62], a large database of roughly 3.2 million images of animals, vehicles, etc.

Besides age regression, the studies mentioned above delivered results on the classification task as well. The metrics are markedly similar to the ones of the current

work (Table 4). Stern et al.'s higher specificities and lower sensitivities could be caused due to the larger amount of adults in their dataset. Nonetheless, the highlight is their excellent AUC of 0.98 in [27]. Dallora et al. has the highest sensitivity among all studies but does not provide values on accuracy and AUC. Another automated approach using MRI as modality which evaluated the classification task is the work by Saint-Martin et al. [61]. For their male dataset of 80 individuals, they obtained a low sensitivity of 69% but a very high specificity of 98%. In contrast to the current work, they investigated a different anatomical site (distal tibia) and had a wider age range (8–25 years).

**Table 4** Comparison of majority classification ( $\geq 18$  years) performance between the current work and other studies using automated and MRI-based methods. Besides accuracy, sensitivity, and specificity the area under the curve (AUC) represents the capacity of a model to discern between minors and adults. Metrics are presented in percentage. An overview of the population, datasets, and methods of the studies can be found in Table 3

Study	Minors	Accuracy	Sensitivity	Specificity	AUC
Current work <sup>a</sup>	49.3	89.7	88.2	92.3	92.0
Current work <sup>b</sup>	52.3	90.9	88.6	94.2	94.4
Dallora et al. [30]	47.1 <sup>c</sup>	-	90.4 <sup>c</sup>	95.6 <sup>c</sup>	-
Stern et al. [21]	42.7	91.3	88.6	93.2	-
Stern et al. [27]	41.6	90.7	82.1	96.8	98.0
Stern et al. [28]	-	-	-	-	95.7
Saint-Martin et al. [61]	37.5	-	69.0	98.0	-

<sup>a</sup>Results of current work using *coronal* knee MRIs<sup>b</sup>Results of current work using *sagittal* knee MRIs<sup>c</sup>Values for the male subjects

## Study limitations and practical implementation

The underlying population of this work was particularly *homogeneous*. Only Caucasian male individuals between 13 and 21 years of age with a middle to high socio-economic status were included. This makes comparison to other studies, e.g., Stern et al., more practicable since multiple factors influencing growth are largely mitigated. One limitation of the developed method is its suitability solely for this specific group. The application on other populations should be validated first. Notwithstanding, it is recommended to perform age estimation on homogeneous groups to attain more reliable results for insufficiently large datasets. Especially, to train separate models for males and females since the latter mature faster and have a different growth pattern [63–66].

Besides gender, the application of the method on individuals with lower socio-economic status has not been evaluated. EASO [10] and Schmeling et al. [67] state that in such a case the method would likely underestimate the subjects since lower socio-economic status generally delays skeletal maturation. This outcome would support the principle of “*in dubio pro reo*.” This hypothesis could be validated in the future by applying the method on other populations.

The age range in this work did not allow classification on the basis of other age limits such as 14 and 21 years which are relevant legal ages in specific fields. Nonetheless, the focus of this work was the 18-year age limit which is a crucial threshold to separate minors from adults in many countries. The selected age range made it possible to assess this age limit with sufficient margin of error.

A *uniform* age distribution is the optimal starting point for an unbiased age estimation. The coronal dataset was unbalanced but was countered by augmenting the knee MRIs of the subjects from the less populated age groups. A slight bias could still be present when comparing the regression results from DS<sub>cor</sub> with DS<sub>sag</sub> which was more uniformly distributed. However, the classification results were superior using DS<sub>sag</sub> which suggests that a more “favorable” distribution is not necessarily followed by better performance.

In addition to obtaining a uniform age distribution, the size of the dataset in medical applications is often limited due to laws and guidelines to protect the privacy of patients [68]. Although DS<sub>sag</sub> had over 400 three-dimensional MRIs of the knee, the roughly 40 cases per age group might not be sufficient to fully capture the variability of the group. Furthermore, data-driven methods such as deep learning require a large amount of data to learn a complex task given the high number of unknowns in the system. Cho et al. [68] concluded that 1000 samples per class in the training data

is necessary to correctly classify body parts in computed-tomography (CT) images using deep learning. Although it is a different field of study, it gives a perspective of how much data is necessary for data-driven models.

A final limitation of the current study is MRI as image modality due to the high cost in comparison to X-ray due to spacious facilities, equipment, medical technical assistants, and examination duration. In [69–78], a solution is proposed to accelerate acquisition times by reconstructing MRIs using undersampled data in *k*-space. Besides reducing the costs it can lower the stress to patients as well, which is particularly useful for young refugees with traumatic backgrounds. First promising applications of the acceleration technique in the area of age estimation have been analyzed in [71, 74, 75]. Other non-invasive techniques such as ultrasound (US) have yet to be evaluated for their suitability for age estimation, especially when used in combination with machine learning. A recent study by Herrmann et al. [79] proposed a growth plate classification system of the knee for age estimation using US volumes and compared the results with MRI.

## Conclusion

The current work presented a new method for automated age estimation of young individuals based on machine learning—notably CNNs—using MRIs of the knee in coronal and sagittal orientation. The three-step approach first pre-processed the MRIs (correction and standardization), subsequently extracted age-relevant structures (data reduction and simplification), and finally, estimated the chronological age based on the extracted data.

The computer-based method demonstrated its ability to both age regression and majority classification ( $\geq 18$  years) in a fully automated, non-invasive, and user-independent setting.

For the regression task, a combination of a CNN, i.e., AgeNet2D, and a tree-based machine learning algorithm and achieved an MAE of  $0.67 \pm 0.49$  years. Although this was the best-performing model and was evaluated similarly to other studies, it may not capture the full variability of the approach. Therefore, the specifically performed *repeated* stratified *k*-fold validation with a MAE of  $0.71 \pm 0.55$  years may provide a more realistic and reliable estimate of the model’s performance in practice.

For the classification task, AgeNet2D followed by a random forest classifier showed the highest values for sensitivity (88.6%) and specificity (94.2%) and comparable to similar studies. The *repeated* cross-validation results with a sensitivity of 88.4% and a specificity of 86.1% might offer a more realistic and reliable evaluation of the classifier’s performance.

The results of age estimation differed slightly for both imaging planes. While the regression was significantly better using coronal rather than sagittal MRI, the age distribution and size of the datasets were different. For a practical implementation of a similar approach, the imaging plane might have less influence than the factors mentioned above, especially when considering data-driven models that use deep learning, for example. Further analysis could be conducted in the future to evaluate this assumption.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This project is funded by the German Research Foundation (DFG), Project (SA 2530/6-1) and (JO 1198/2-1).

## Compliance with Ethical Standards

An ethical approval for this study was granted by the Ethics Committee of the Medical Association Hamburg (PV4527). The director of this study had full control of the data and the material submitted for publication.

**Conflict of interest** The authors declare that they have no conflict of interest.

**Informed consent** Written informed consent was obtained from all subjects in this study.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Schmeling A, Garamendi PM, Prieto JL, Landa MI (2011) Forensic age estimation in unaccompanied minors and young living adults. In: Forensic medicine - from old problems to new challenges, pp 1–34. InTech
- Kubilay S (2016) Ablauf des deutschen Asylverfahrens. Tech. rep. Bundesamt für Migration und Flüchtlinge (BAMF)
- Schmeling A, Kaatsch HJ, Marre B, Reisinger W, Riepert T, Ritz-Timme S, Rösing FW, Rötzscher K, Geserick G (2001) Empfehlungen für die Altersdiagnostik bei Lebenden im Strafverfahren. Rechtsmedizin 11:1–3
- Dvorak J (2009) Detecting over-age players using wrist MRI: science partnering with sport to ensure fair play. Br J Sports Med 43:884–885
- Schmidt S, Vieth V, Timme M, Dvorak J, Schmeling A (2015) Examination of ossification of the distal radial epiphysis using magnetic resonance imaging. New insights for age estimation in young footballers in FIFA tournaments. Science & Justice 55:139–144
- Tscholl PM, Junge A, Dvorak J, Zubler V (2016) MRI of the wrist is not recommended for age determination in female football players of U-16/U-17 competitions. Scand J Med Sci Sports 26:324–328
- European Asylum Support Office (2013) Age assessment practice in Europe. Tech. rep. December, European Asylum support office
- Britting-Reimer E (2015) Altersbestimmung in Deutschland und im Europäischen Vergleich. Jugendhilfe 53:88–95
- Fleischhauer J (2018) Minderjährige Flüchtlinge: Sind so kleine Hände. <https://www.spiegel.de/politik/deutschland/alterstests-bei-fluechtlingen-kolumne-von-jan-fleischhauer-a-1186171.html>
- European Asylum Support Office (2018) Practical guide on age assessment. Publications Office, Second edn
- Hagen M, Schmidt S, Rudolf E, Schmeling A (2020) Die Aus-sagekraft sozialpädagogischer Altersschätzungen im Vergleich zur forensischen Altersdiagnostik. Rechtsmedizin 30:233–240
- Greulich WW, Pyle SI (1959) Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci 238:393
- Tanner JM, Whitehouse R, Cameron N, Marshall WA, Healy MJR, Goldstein H (1983) Assessment of skeletal maturity and prediction of adult height (TW2 method). Academic Press 22:37
- Demirjian A, Goldstein H, Tanner JM (1973) A new system of dental age assessment. Human Biology 45(2):211–227
- Schulz R, Müller M, Reisinger W, Schmidt S, Schmeling A (2008) Radiographic staging of ossification of the medial clavicular epiphysis. Int J Legal Med 122(1):55–58
- Jopp E, Schröder I, Maas R, Adam G, Püschel K (2010) Proximale Tibiaepiphyse im Magnetresonanztomogramm: Neue Möglichkeit zur Altersbestimmung bei Lebenden? Rechtsmedizin 20:464–468
- Jiang J, Trundle P, Ren J (2010) Medical image analysis with artificial neural networks. Comput Med Imaging Graph 34:617–631
- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
- Setiono R, Liu H (1997) Neural-network feature selector. IEEE Trans Neural Netw 8:654–662
- 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 MR, Unal G, Wells W (eds) Medical image computing and computer-assisted intervention – MICCAI 2016, vol 9901 of Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 194–202
- Štern D, Kainz P, Payer C, Urschler M (2017) 61–69. In: Wang Q, Shi Y, Suk H-I, Suzuki K (eds) Machine learning in medical imaging. Springer International Publishing, Cham
- Lee H, Tajmir S, Lee J, Zissen M, Yesihwas BA, Alkasab TK, Choy G, Do S (2017) Fully automated deep learning system for bone age assessment. J Digit Imaging 30:427–441
- 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
- Iglovikov VI, Rakhlis A, Kalinin AA, Shvets AA (2018) Paediatric bone age assessment using deep convolutional neural networks. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, pp 300–308
- Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP (2018) Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 287:313–322
- Pröve PL, Jopp-van Well E, Stanczus B, Morlock MM, Herrmann J, Groth M, Säring D, Auf der Mauer M (2019) Automated segmentation of the knee for age assessment in 3D MR images using convolutional neural networks. Int J Legal Med 133:1191–1205

27. Štern D, Payer C, Giuliani N, Urschler M (2019) Automatic age estimation and majority age classification from multifactorial MRI data. *IEEE J Biomed Health Info* 23:1392–1403
28. Štern D, Payer C, Urschler M (2019) Automated age estimation from MRI volumes of the hand. *Med Image Anal* 58:101538
29. Li Y, Huang Z, Dong X, Liang W, Xue H, Zhang L, Zhang Y, Deng Z (2019) Forensic age estimation for pelvic X-ray images using deep learning. *Eur Radiol* 29:2322–2329
30. Dallora AL, Berglund JS, Brogren M, Kvist O, Ruiz SD, Dübbel A, Anderberg P (2019) Age assessment of youth and young adults using magnetic resonance imaging of the knee: A deep learning approach. *JMIR Medical Informatics* 7:e16291
31. Auf der Mauer M, Säring D, Stanczus B, Herrmann J, Groth M, Jopp-van Well E (2019) A 2-year follow-up MRI study for the evaluation of an age estimation method based on knee bone development. *Int J Legal Med* 133:205–215
32. Knußmann R (1992) Somatometrie. In: Martin R, Knußmann R (eds) *Anthropologie*. Gustav Fischer Verlag, pp 232–309
33. Martin R, Saller KF (1957) *Lehrbuch der Anthropologie: in systematischer Darstellung mit besonderer Berücksichtigung der anthropologischen Methoden: für Studierende, Ärzte und Forschungsreisende*. Stuttgart: Gustav Fischer Verlag
34. Hermanusse M, Lieberman LS, Schönfeld Janewa V, Scheffler C, Ghosh A, Bogin B, Godina E, Kaczmarek M, El-Shabrawi M, Salama EE, Rühli FJ, Staub K, Woitek U, Blaha P, Assmann C, van Buuren S, Lehmann A, Satake T, Thodberg HH, Jopp E, Kirchengast S, Tutkuviene J, McIntyre MH, Wittwer-Backofen U, Boldsen JL, Martin DD, Meier J (2012) Diversity in auxology: between theory and practice. Proceedings of the 18th Aschauer Soiree, 13th November 2010. *Anthropologischer Anzeiger; Bericht über die biologisch-anthropologische Literatur* 69(2):159–174
35. Jopp E, Schröder I, Püschele K, Hermanusse M (2012) Longitudinal shrinkage in lower legs: “negative growth” in healthy late-adolescent males. *Anthropol Anz* 69:107–115
36. Galić I, Mihanović F, Giuliodori A, Conforti F, Cingolani M, Cameriere R (2016) Accuracy of scoring of the epiphyses at the knee joint (SKJ) for assessing legal adult age of 18 years. *Int J Legal Med* 130:1129–1142
37. Auf der Mauer M (2020) Towards automated age estimation of young individuals - a new computer-based approach using 3D knee MRI. PhD thesis, Hamburg University of Technology
38. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC (2010) N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 29:1310–1320
39. Brieckle K, Hanebeck UD (2001) Template matching using fast normalized cross correlation. *Proc SPIE* 4387:95–102
40. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans Graph* 28:1
41. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention - MICCAI 2015*, vol 9351. Springer, Cham, pp 234–241
42. Clevert DA, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by Exponential Linear Units (ELUs). *arXiv:1511.07289*
43. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proceedings of Machine Learning Research* 37:448–456
44. Goodfellow I, Bengio Y (2016) *A courville, deep learning*. The MIT Press, Cambridge
45. 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
46. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin P-M, Larochelle H (2017) Brain tumor segmentation with deep neural networks. *Med Image Anal* 35:18–31
47. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M (2013) Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: *Advanced information systems engineering*. Springer, Berlin, pp 246–253
48. Lin M, Chen Q, Yan S (2013) Network in network. *arXiv:1312.4400*
49. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2015) Learning deep features for discriminative localization. *arXiv:1512.04150*
50. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: *3rd International conference for learning representations*, pp 1–15
51. Schmeling A, Dettmeyer R, Rudolf E, Vieth V, Geserick G (2016) Forensic age estimation: methods, certainty, and the law. *Deutsches Aerzteblatt Online* 113:44–50
52. De Tobel J, Fieuws S, Hillewig E, Phlypo I, van Wijk M, de Haas MB, Politis C, Verstraete KL, Thevissen PW (2020) Multi-factorial age estimation: a Bayesian approach combining dental and skeletal magnetic resonance imaging. *Forensic Sci Int* 306:110054
53. Krämer JA, Schmidt S, Jürgens KU, Lentschig M, Schmeling A, Vieth V (2014) Forensic age estimation in living individuals using 3.0T MRI of the distal femur. *Int J Legal Med* 128(3):509–514
54. Saint-Martin P, Rérolle C, Pucheu J, Dedouit F, Telmon N (2015) Contribution of distal femur MRI to the determination of the 18-year limit in forensic age estimation. *Int J Legal Med* 129(3):619–620
55. Vieth V, Schulz R, Heindel W, Pfeiffer H, Buerke B, Schmeling A, Ottow C (2018) Forensic age assessment by 3.0T MRI of the knee: proposal of a new MRI classification of ossification stages. *Eur Radiol* 28:3255–3262
56. Krämer JA, Schmidt S, Jürgens K-U, Lentschig M, Schmeling A, Vieth V (2014) The use of magnetic resonance imaging to examine ossification of the proximal tibial epiphysis for forensic age estimation in living individuals. *Forensic Science, Medicine, and Pathology* 10:306–313
57. Ottow C, Schulz R, Pfeiffer H, Heindel W, Schmeling A, Vieth V (2017) Forensic age estimation by magnetic resonance imaging of the knee: the definite relevance in bony fusion of the distal femoral- and the proximal tibial epiphyses using closest-to-bone T1 TSE sequence. *Eur Radiol* 27:5041–5048
58. De Tobel J, Bauwens J, Parmentier GIL, Franco A, Pauwels NS, Verstraete KL, Thevissen PW (2020) Magnetic resonance imaging for forensic age estimation in living children and young adults: a systematic review. *Pediatric Radiology* 50:1691–1708
59. Štern D, Ebner T, Bischof H, Grassegger S, Ehamer T, Urschler M (2014) Fully automatic bone age estimation from left hand MR images. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014* 17(Pt 2):220–227
60. Štern D, Urschler M (2016) From individual hand bone age estimates to fully automated age estimation via learning-based information fusion. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp 150–154
61. Saint-Martin P, Rérolle C, Dedouit F, Rousseau H, Rougé D, Telmon N (2014) Evaluation of an automatic method for forensic age estimation by magnetic resonance imaging of the distal tibial epiphysis - A preliminary study focusing on the 18-year threshold. *Int J Legal Med* 128:675–683
62. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database: 248–255

63. Marshall WA, Tanner JM (1969) Variations in pattern of pubertal changes in girls. *Arch Dis Child* 44(235):291–303
64. Marshall WA, Tanner JM (1970) Variations in the pattern of pubertal changes in boys. *Arch Dis Child* 45(239):13–23
65. Cole TJ, Rousham EK, Hawley NL, Cameron N, Norris SA, Pettifor JM (2015) Ethnic and sex differences in skeletal maturation among the birth to twenty cohort in South Africa. *Arch Dis Child* 100:138–143
66. Vucic S, de Vries E, Eilers PHC, Willemsen SP, Kuijpers MAR, Prahl-Andersen B, Jaddoe VWV, Hofman A, Wolvius EB, Ongkosuwito EM (2014) Secular trend of dental development in Dutch children. *Am J Phys Anthropol* 155:91–98
67. Schmeling A, Olze A, Reisinger W, König M, Geserick G (2003) Statistical analysis and verification of forensic age estimation of living persons in the Institute of Legal Medicine of the Berlin University Hospital Charité. *Leg Med* 5:S367–S371
68. Cho J, Lee K, Shin E, Choy G, Do S (2015) How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv:1511.06348*
69. Yang Y, Sun J, Li H, Xu Z (2016) Deep ADMM-Net for compressive sensing MRI. In: Advances in neural information processing systems
70. Wang S, Su Z, Ying L, Peng X, Zhu S, Liang F, Feng D, Liang D (2016) Accelerating magnetic resonance imaging via deep learning. In: IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE, p 2016
71. Terada Y, Tamada D, Kose K, Nozaki T, Kaneko Y, Miyagi R, Yoshioka H (2016) Acceleration of skeletal age MR examination using compressed sensing. *J Magn Reson Imaging* 44:204–211
72. Schlemper J, Caballero J, Hajnal JV, Price AN, Rueckert D (2018) A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans Med Imaging* 37:491–503
73. Zbontar J, Knoll F, Sriram A, Muckley MJ, Bruno M, Defazio A, Parente M, Geras KJ, Katsnelson J, Chandarana H, Zhang Z, Drozdza M, Romero A, Rabbat M, Vincent P, Pinkerton J, Wang D, Yakubova N, Owens E, Zitnick CL, Recht MP, Sodickson DK, Lui YW (2018) fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv:1811.08839*
74. 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
75. Neumayer B, Lesch A, Thaler F, Widek T, Tschauner S, De Tobel J, Ehammer T, Kirnbauer B, Boldt J, van Wijk M, Stollberger R, Urschler M (2019) The four-minute approach revisited: accelerating MRI-based multi-factorial age estimation. *Int J Legal Med* 134:1475–1485
76. Qin C, Schlemper J, Caballero J, Price AN, Hajnal JV, Rueckert D (2019) Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE Trans Med Imaging* 38:280–290
77. Knoll F, Hammernik K, Kobler E, Pock T, Recht MP, Sodickson DK (2019) Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magn Reson Med* 81:116–128
78. Mardani M, Gong E, Cheng JY, Vasanawala SS, Zaharchuk G, Xing L, Pauly JM (2019) Deep generative adversarial neural networks for compressive sensing MRI. *IEEE Trans Med Imaging* 38:167–179
79. Herrmann J, Säring D, Auf der Mauer M, Groth M, Jopp-van Well E (2020) Forensic age assessment of the knee: proposal of a new classification system using two-dimensional ultrasound volumes and comparison to MRI. *European Radiology*. <https://link.springer.com/article/10.1007%2Fs00330-020-07343-1>

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

## Affiliations

Markus Auf der Mauer<sup>1</sup>  · Elin Jopp-van Well<sup>2</sup> · Jochen Herrmann<sup>3</sup> · Michael Groth<sup>3</sup> · Michael M. Morlock<sup>4</sup> · Rainer Maas<sup>5</sup> · Dennis Säring<sup>1</sup>

Elin Jopp-van Well  
e.jopp@uke.de

Jochen Herrmann  
j.herrmann@uke.de

Michael M. Morlock  
morlock@tuhh.de

Rainer Maas  
rainermaas@gmx.de

<sup>1</sup> Medical and Industrial Image Processing, University of Applied Sciences of Wedel, Feldstraße 143, 22880 Wedel, Germany

<sup>2</sup> Department of Legal Medicine, University Medical Center Hamburg-Eppendorf (UKE), Butenfeld 34, 22529 Hamburg, Germany

<sup>3</sup> Section of Pediatric Radiology, Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf (UKE), Martinistr. 52, 20246 Hamburg, Germany

<sup>4</sup> Institute of Biomechanics M3, Hamburg University of Technology (TUHH), Denickestraße 15, 21073 Hamburg, Germany

<sup>5</sup> Radiologie Raboisen 38, Raboisen 38, 20095 Hamburg, Germany