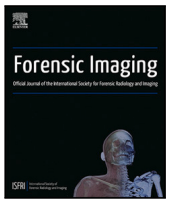
Forensic Imaging 40 (2025) 200619

Contents lists available at ScienceDirect

Forensic Imaging

journal homepage: www.sciencedirect.com/journal/forensic-imaging



Chronological age estimation for medico-legal expertise-based on sternoclavicular joint CT images using a deep neural network

Chang Sun [a,b], Yazdan Salimi [b], Isaac Shiri [b], Coraline Egger [c], Pia Genet [c,d], Habib Zaidi [b,e,f,g,*], Sana Boudabbous [h]

[a] Beijing University of Posts and Telecommunications, School of Information and Communication Engineering, 100876 Beijing, China
[b] Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland
[c] University Center of Legal Medicine Geneva, Geneva University Hospital and University of Geneva, Rue Michel-Servet 1, 1211 Geneva, Switzerland
[d] University Center of Legal Medicine, Lausanne, Lausanne University Hospital and University of Lausanne, Chemin de la Vulliette 4, 1000 Lausanne, Switzerland
[e] Geneva University Neurocenter, Geneva University, Geneva, Switzerland
[f] Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[g] Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark
[h] Geneva University Hospital, Division of Radiology, CH-1211, Geneva, Switzerland

ARTICLE INFO

Keywords:
Computed tomography
Age estimation
Deep learning
Age determination by skeleton
Sternoclavicular joints

ABSTRACT

The aim of this study was to develop and validate fully automated deep learning models to estimate chronological age from sternoclavicular CT images to help forensic age estimation and understand its limitations. A total of 742 whole-body CT and 164 pediatric chest-abdomen-pelvis CT scans (age: 1–60y, 437 m and 469f) were collected as a training dataset. A deep learning pipeline was implemented to segment the clavicle volume of interest, train an age estimation model, and finally fine-tune the network. The predictive performance of nine deep learning models was assessed and compared using 5-fold cross-validation. A transfer learning experiment was designed to evaluate the generalizability of the pre-trained models, using a fine-tuning group (age: 15–35y, 6 m and 4f) and a validation group (age: 16–35y, 6 m and 4f). Clinical age assessment based on clavicle bone was conducted on 5 thorax CT scans (4 m and 1f, age: 16–32y) and 5 sternoclavicular joint CT scans (unknown age) by one radiologist and two forensic pathologists. The intra- and inter-observer agreement of experts was assessed. A mean absolute error (MAE) of 4.23 ± 4.49 years, an area under the receiver operating characteristic (AUC) of 0.99 for age classification (> 14 years and > 18 years) and an accuracy of 0.97 for classification of ossification stages were achieved in the cross-validation. An MAE of 3.30 ± 3.58 years and an accuracy of 0.90 for ossification stage classification were achieved after fine-tuning. The three experts disagreed on the images that met the diagnostic requirements in 2 cases. Intra-observer agreement varied between experts. This study concluded that a fully automated deep neural network, employing a transfer learning strategy, exhibits potential for estimating chronological age from clavicular CT images.

1. Introduction

Age estimation plays a crucial role in various legal and social decision-making processes, such as criminal and civil proceedings, political decisions, asylum procedures, and sports competitions. In many countries, it is essential to demonstrate compliance with age limits when making legal decisions regarding procedural privileges or social benefits. In Switzerland, forensic pathologists conduct age estimation in criminal cases under the authorization of the Juvenile Court or the Public Prosecutor. They also perform these assessments in civil cases upon request from the State Secretariat for Migration (SEM). The rise in cross-border migration has resulted in a significant increase in the number of age estimations conducted by the SEM in Switzerland. In 2021, 528 age estimations were carried out, which is more than three times the number conducted in 2019 [1]. With the increasing phenomenon of migration, these assessments have become a more sensitive topic of interest due to their complexity and policy implications.

Forensic age estimation expertise follows the three-pillar principle recommended by the Arbeitsgemeinschaft für Forensische Altersdiagnostik (AGFAD) of the German Society of Legal Medicine [2–4]. These pillars include a clinical examination to assess puberty development

* Corresponding author at: Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).

https://doi.org/10.1016/j.fri.2024.200619
Received 14 September 2024; Accepted 4 December 2024
Available online 20 December 2024

and identify factors that may affect bone and dental age (such as illnesses, medication consumption, and malnutrition), a bone age estimation using X-rays of the hand and the wrist, and a dental age estimation involving an OPG (orthopantomogram) along with eventually a dental clinical examination. In cases where hand ossification is completed, it is recommended to conduct an additional CT examination of sternoclavicular joints. The closure of the medical epiphyseal joints of the clavicles occurs last among all the bones in the human body, thus making them particularly significant for forensic age diagnosis. Currently, the interpretation of the ossification status of the medial clavicular epiphysis is based on a classification by stages (5 stages with 3 subgroups for 2 of the 5 stages) [5–10] conducted by specialists with expertise and experience. However, it is worth noting that the literature on this topic is subject to debate, as the bone age-related factors (such as socioeconomic status) in the studied populations may not necessarily correspond to the individuals examined in the specific expertise under consideration, and the reproducibility of such examinations remains a concern.

In recent years, numerous studies have explored deep learning (DL)-based approaches for age estimation using various imaging techniques, including X-ray radiography, CT, MRI, and ultrasound images of the tooth, knee, wrist, and hand [11–16]. These fully automated or semi-automated DL-based methods offer a number of advantages, including high speed, reduced need for manual interpretation, and decreased human workload, thus enabling physicians to focus on rare and complex cases. However, challenges associated with DL methods include limited training samples, generalization to external datasets, and consistency with expert diagnoses.

The aim of this study was to develop a fully automated DL method for age estimation based on CT scans of the sternoclavicular joints within the medico-legal expertise setting. In addition, the study aimed to analyze the generalizability of the DL models through fine-tuning experiments and to assess the intra-observer and inter-observer variability of radiology and forensic pathology experts for ossification stage assessments of the medial clavicular epiphysis based on sternoclavicular joint CT images, and compare them with the fine-tuned DL model.

The main contributions of this work are as follows:

- To the best of our knowledge, this work is the first attempt to develop an automatic deep learning pipeline to facilitate and assist age estimation tasks based on clavicle CT images.
- We comprehensively collected CT images of individuals from different institutions with different voxel spacing to form a large dataset to train robust pre-trained models. Nine typical DL models were trained and compared.
- We performed a thorough evaluation of DL-based approaches for age estimation, age classification and the classification for the ossification stage of the medial clavicular epiphysis.
- We performed few-sample fine-tuning experiments on the pre-trained models to better understand the impact of sample size for fine-tuning on boosting prediction performance.
- We assessed the intra- and inter-observer reliability of three experts on 5 CT cases with standard protocol for bone age estimation and another 5 CT cases with slice thickness less than 1 mm, and then compared the performance to DL-based approaches.

The structure of this paper is as follows. In Section 2, we review studies related to age estimation based on medical images. In Section 3, we describe the details of building the CT dataset, the implementation of DL-based age estimation based on sternoclavicular joint CT images and the design of experiments. In Section 4, we present and discuss the experimental results. In Section 5, we conclude this work, discuss the limitations of this study and plans for future work.

2. Related work

This section presents a brief introduction to the development of age estimation approaches based on sternoclavicular joint CT images, age classification methods based on medial clavicular epiphysis and DL-based age estimation approaches using medical images.

2.1. Age estimation based on sternoclavicular joint CT images

In previous research on age estimation using medical clavicular epiphyseal based on CT images, Toutin et al. [17] established a regression model between age and bone mineral density (BMD) of the medial end of the clavicle and found MAE of 13.4 years for males and 13.1 years for females on 180 individuals (age between 15 to 100 years), respectively. Although this work revealed the correlation between BMD and bone age, the age estimation accuracy was relatively low. On the other hand, Shedge et al. [18] developed a regression model based on the degree of ossification of the medial clavicular epiphyseal evaluated by one observer on 350 Indian individuals (aged between 10.01 to 35.47 years), showing a MAE of 1.5 years and 1.1 years for male and female, respectively. Despite the promising results, this method has a shortage in under generalization and concerned reproducibility due to rigorous data cleaning and manual evaluation by expert physicians.

2.2. Age classification based on medial clavicular epiphysis

For medico-legal applications, 18 is the threshold age for criminal liability in many countries [19,20]. In some countries, children under the age of 14 at the time of the offense cannot be held criminally responsible and juveniles who commit crimes between the ages of 14 and 17 have more lenient sentences compared to adults [20]. In previous studies, the medial clavicular epiphyseal fusion stage was used to determine whether an individual is over the medicolegally critical thresholds of 18 years. For example, individuals classified as stages 4 and 5 based on CT scans of the medial clavicle epiphysis were observed to be above the age of 18 [21–24]. The reproducibility of staging clavicular development of forensic experts is of utmost importance for medical-legal age estimation applications. In addition to the utility of the fusion stage, Scendoni et al. [25] established a logistic regression model to determine the cut-off point for 18 years using the two features of MRI on medial clavicular epiphysis: REM-1 (ratio between the length of the whole epiphysis and the length of the metaphysis) and REM-2 (ratio between the length of epiphyseal-metaphyseal fusion and the length of the metaphysis). However, many factors, such as socioeconomic status [26,27] and nutritional status [28] are controversial issues in the general topic of age assessment based on epiphyseal ossification.

2.3. DL-based age estimation using medical images

Compared with traditional methods, the DL-based approach is more efficient, easier, faster, and cheaper because it is fully automated with acceptable performance and does not require any manual intervention. In recent years, a number of deep learning-based studies have made successful attempts at age prediction using radiological images of tooth [11,14], pelvis [15], cervical vertebral [16], knee [12] and hand [13]. For example, Atas et al. [11] used typical DL models, such as DenseNet [29] and MobileNetV2 [30], to estimate dental age from dental panoramic radiography (DPR) images and a modified network based on InceptionV3 [31] achieved a MAE of 3.13 years (age: 8–68 y). Demircioglu et al. [12] trained a ResNet-34 neural network to predict age from knee radiographs and achieved a MAE of less than one year (age: <21 y) and an AUC of 0.94 in age classification (< 18y). Hui et al. [13] developed a global-local feature fusion convolutional neural network to estimate bone age from hand X-ray images and achieved a MAE of less than one year (age: 4–18 y). Although deep learning proved to be a powerful method for age estimation with promising prediction accuracy, it is difficult to compare the performance of these methods. First, the use of radiological images of different body parts has a great impact on the prediction results of deep learning. In addition, the number and age distribution of the training and test samples, the health status of the participants, scanning modality and parameters will all affect the numerical results of the prediction effect, but they vary in different studies. Another issue of concern is the implementability of deep learning in clinical settings.

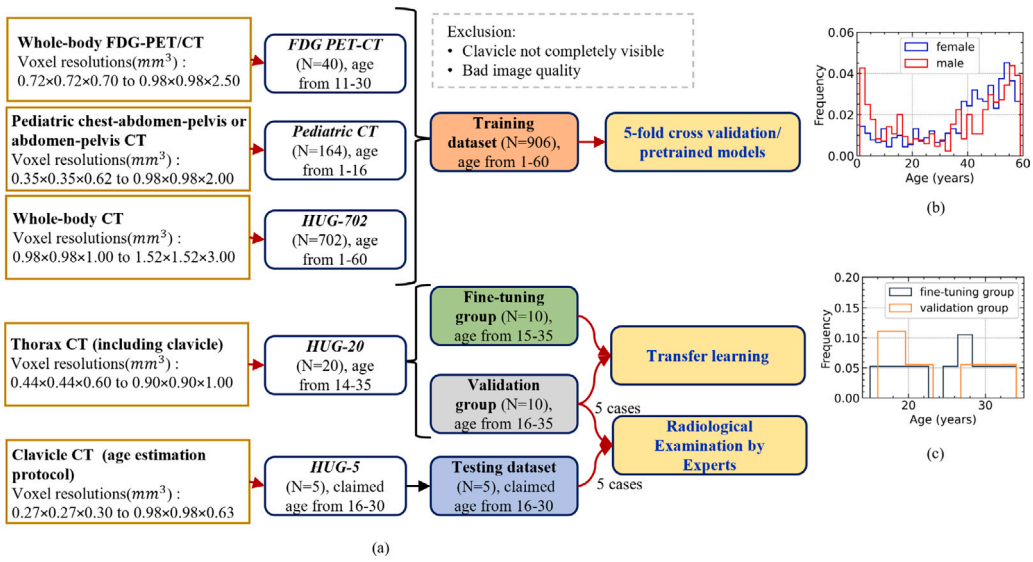


Fig. 1. Datasets. (a) Flowchart summarizing the enrollment of subjects in this study protocol. (b) Histogram of the chronological age of all individuals in the training dataset for cross validation. (c) Histogram of the chronological age of all individuals in HUG-20 cohort for transfer learning.

Table 1
Representation of the population group. Marker *: Real age estimation cases, chronological age unknown.

Name	Acquisition date	Manufacturer	Modality	Slice thickness (mm)	Num of parameters	Age range
FDG PET-CT	2014–2018	Siemens Biograph mCT	Whole-body CT	0.70 – 2.50	40	11y–30y
Pediatric CT	2004–2009	GE Lightspeed VCT/ Somatom Definition AS+/ GE Revolution	Pediatric chest-abdomen-pelvis CT	0.62 – 2.00	164	1y–16y
HUG-702	2018–2021	Siemens Biogragh mCT	Whole-body CT	1.00 – 3.00	702	1y–60y
HUG-20	2010–2011	Discovery CT750 HD	Thorax CT	0.60 – 1.00	20	14y–35y
HUG-5	2019–2021	Discovery CT750 HD	Clavicle CT	0.30 – 0.63	5	*

3. Materials and methods

3.1. Datasets

This is a retrospective study performed in collaboration between the radiology department of Geneva University Hospitals (HUG) and the University Center of Legal Medicine, Lausanne - Geneva (CURML). The project was approved by Geneva Canton ethical committee (CCER ID 2017-00922). Informed consent was waived owing to the retrospective nature of the study. As shown in Table 1, two cohorts (^{18}F-FDG PET-CT [32–34] and pediatric CT [35,36]) including 204 individuals from The Cancer Imaging Archive (TCIA), and three cohorts (HUG-702, HUG-20 and HUG-5) including 752 individuals from HUG were enrolled. Fig. 1(a) shows the enrollment flowchart. We constructed a training dataset for cross-validation and building pre-trained models, a fine-tuning group and a validation group for transfer learning experiments, and a testing dataset for radiological examination for experts.

The training dataset contains 906 CT scans (437 males and 469 females, mean age: 36.7 ± 18.4 years) from three cohorts: FDG PET-CT, pediatric CT and HUG-702 (Fig. 1(b)).

HUG-20 datasets (13 males and 7 females, mean age: 24.4 ± 6.1 years) were used for transfer learning experiments to test the generalizability of the pre-trained DL models (Fig. 1(c)). The HUG-20 dataset was acquired by covering the sternoclavicular area for other clinical indications other than forensic investigations and the patient claimed age according to the identification card was trusted as the chronological age. The acquisition and reconstruction parameters of HUG-20 were close to the routine parameters for forensic imaging in the HUG-5 dataset, however the average tube current was almost double.

HUG-5 (only males) were used for radiological examination conducted by the experts and as a test group for our model after fine tuning. HUG-5 dataset contains CT images acquired from cases referred for bone ossification estimation from the forensic medicine department, i.e., the protocol, acquisition, and reconstruction parameters were standardized for the specific aim of imaging which was bone age estimation. The detailed acquisition parameters are listed in supplementary Table S1.

3.2. Building and training networks

In this study, we map the CT images of the sternoclavicular joints to the continuous age variable. First, we cropped the medial clavicular epiphysis volume of interest (VOI) from clavicle bones. Then, we trained DL models in an end-to-end manner to automatically estimate the age of an individual from the VOI (Fig. 2).

In clavicle CT data, the left clavicle bones (closing to the medial end) were automatically cropped using a DL-based segmentation network based on a U-Net architecture [37]. The details of building and training the segmentation network are presented in the supplementary Table S2. We investigated two types of high-performing neural networks for age estimation: small models with fewer training parameters (ShuffleNet [38], ShuffleNet-V2 [39], MobileNet [40], MobileNet-V2 [30]) and large models with more training parameters (DenseNet121 [29], DenseNet169 [29], DenseNet201 [29], ResNet18 [41], and ResNet34 [41]). For each network, we performed 5-fold cross-validation. First, we split the training data into 5 equal sets. Then we trained a model on four sets and tested on the fifth set, which were unknown data to the model as they were not used for training. This process was repeated 5 times with a different set chosen for testing each

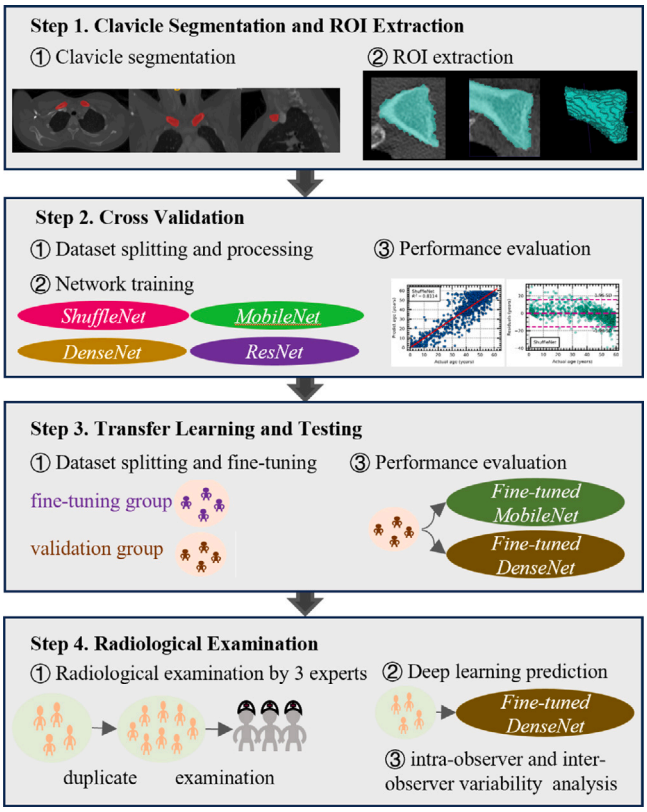


Fig. 2. An integrated workflow diagram (providing an outline of the whole process).

time. Finally, we got 5 models and the testing performance on every set. We averaged the testing results to get the final testing performance of the network. Data augmentation (random rotation by 90 degrees) was applied to all training samples. To adapt to the input size of the networks, we changed the convolution layers, down-sampling layers, and up-sampling layers to 3D implementation instead of 2D. In this study, all the training processes were performed on a server equipped with an NVIDIA GeForce GTX 2080Ti. The ShuffleNet and MobileNet family were completed under the PyTorch framework [42], whereas DenseNet and ResNet model family were retrieved from MONAI Model Zoo [43]. We choose the ADAM optimizer [44] with a momentum of 0.9, an initial learning rate of 1e-4, and a batch size set to 2. The networks were trained for a total of 100 epochs with the learning rate decays by half every 50 epochs. We validate the training models every two epochs and choose the model that had the best validation result as the final trained model.

3.3. Transfer learning experiments

The aim of this experiment was to test the generalization ability of pre-trained DL networks on new CT examples with acquisition parameters different from the training dataset. We divided the HUG-20 cohort into a fine-tuning group (10 cases) and a validation group (10 cases), then we conducted 10 transfer learning experiments on pre-trained MobileNet-V2 (from the small model family) and DenseNet169 (from the big model family). The pre-trained models were built by training on the whole training dataset using the above-mentioned parameters (optimizer, learning rate, etc.).

In the transfer learning experiments, firstly, we separated 10 cases randomly as the fixed validation group for fine-tuning evaluation on all experiments (0 to 10) and then used numbers of the other 10 cases group for fine-tuning. Specifically, in Experiment-0, we directly validated the pre-trained models on the validation group without any fine-tuning. In Experiment-1, 2, 3, 4, 5, 6, 7, 8, 9 and 10, we randomly selected 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 samples from the fine-tuning group, respectively, to fine-tune the pre-trained models and then tested them on the validation group. In each experiment from 1 to 9, where we had to choose a subset of data from 10 cases, we built 5 independent models fine-tuned on different selected data, and then reported the average testing results to avoid the randomness caused by the selection of different fine-tuning data. We choose the ADAM optimizer [44] with a momentum of 0.9, an initial learning rate of 1e-4, and a batch size set to 2. The networks were fine-tuned for a total of 5 epochs.

3.4. Radiological examination by experts

The aim of this experiment was to assess the intra-observer and inter-observer variability of the radiology and forensic pathology experts for ossification stage assessments of the medial clavicular epiphysis based on sternoclavicular joint CT images, and compared them with the fine-tuned DL model. Three experts were invited to participate in this study, one radiologist with 17 years of experience, and 2 forensic pathologists with 13 and 12 years of experience, respectively. The experts were independently provided with the entire CT series of the HUG-5 cohort and 5 cases of HUG-20 cohort and were asked to assess the stages of epiphyseal union both in right and left clavicles for each case in a blinded manner. To evaluate the intra-observer variability of the experts, these 10 cases were duplicated as 20 cases in the assessment, which the experts were not informed about.

3.5. Evaluation metrics

The age estimation performance was evaluated using four metrics: mean absolute error (MAE), mean error (ME), mean squared error (MSE), root mean square error (RMSE) and standard error of estimate (SEE).

For medico-legal applications in Switzerland, the Juvenile Court specifically handles criminal investigations involving individuals under the age of 18 at the time of the offense [19]. Additionally, in some other countries, the ages of 14 and 16 are also relevant for forensic applications [20]. To comprehensively assess the performance of deep learning models in age classification, we analyzed their ability to distinguish ages over 14, over 16, over 18, and over 30. We evaluated these models using metrics, such as precision, recall, F1-score, accuracy, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). The definitions of these metrics are explained as follows (taking the classification of over 18 as an example, where correctly classified adults are considered true positives).

- Precision refers to the proportion of correctly identified adults among the subjects who were actually adults.
- Recall measures the proportion of actual adults correctly identified by the model.
- F1-score represents the harmonic mean of precision and recall.
- Accuracy reflects the proportion of predictions that yield correct results.
- ROC plots sensitivity (recall) against 1-specificity (the proportion of actual minors identified as adults by the model) at different threshold settings.
- AUC, the area under the ROC curve, indicates the quality of the classifier, with a higher AUC indicating better performance.

Finally, $Accuracy_{oss}$ was used to analyze the classification performance for the ossification stage of the medial clavicular epiphysis. Based on the minimum and maximum ages of stages 1, 4, and 5 in the Kellinghaus study [6], and the minimum and maximum ages of stages 2a, 2b, 2c, 3a, 3b, and 3c in the Kellinghaus study [5], we converted the estimated and actual ages into estimated stages of ossification (ESO) and actual stages of ossification (ASO), respectively. Then we calculated $Accuracy_{oss}$:

$$Score_{oss} = \begin{cases} 1, ESO \bigcap ASO \neq \emptyset \\ 0, ESO \bigcap ASO = \emptyset \end{cases} \quad (1)$$

Table 2
MAE (in years) and standard deviation, Mean Error (in years) and standard deviation, MSE, RMSE (in years), SEE (in years), and the number of parameters of the deep learning models trained during cross-validation for age estimation.

Models	MAE	ME	MSE	RMSE	SEE	Number of parameters
ShuffleNet	5.81 ± 5.45	0.10 ± 7.97	63.53	7.97	7.98	0.94M
ShuffleNet-V2	6.56 ± 6.02	0.95 ± 8.85	79.29	8.90	8.91	1.30M
MobileNet	6.40 ± 6.42	−0.06 ± 9.06	82.16	9.06	9.07	3.30M
MobileNet-V2	4.74 ± 4.93	−0.48 ± 6.82	46.79	6.84	6.85	2.35M
DenseNet121	4.23 ± 4.49	0.16 ± 6.17	38.09	6.17	6.18	11.24M
DenseNet169	4.29 ± 4.58	0.02 ± 6.27	39.35	6.27	6.28	18.54M
DenseNet201	4.25 ± 4.48	0.31 ± 6.17	38.21	6.18	6.19	25.33M
ResNet18	4.88 ± 4.87	−0.17 ± 6.89	47.49	6.89	6.90	33.16M
ResNet34	4.78 ± 4.85	−0.06 ± 6.81	46.40	6.81	6.82	63.47M

Table 3
Precision, recall, F1 score and accuracy of the models trained during cross-validation for age classification.

Model	< 14 and ≥ 14 years				Model	< 16 and ≥ 16 years			
	Precision	Recall	F1	Accuracy		Precision	Recall	F1	Accuracy
ShuffleNet	0.9523	0.9866	0.9691	0.9482	ShuffleNet	0.9463	0.9890	0.9672	0.9460
ShuffleNet-V2	0.9320	0.9893	0.9598	0.9316	ShuffleNet-V2	0.9197	0.9890	0.9531	0.9217
MobileNet	0.9518	0.9773	0.9644	0.9405	MobileNet	0.9402	0.9699	0.9548	0.9261
MobileNet-V2	0.9813	0.9826	0.9820	0.9702	MobileNet-V2	0.9784	0.9918	0.9850	0.9757
DenseNet121	0.9748	0.9826	0.9787	0.9647	DenseNet121	0.9677	0.9863	0.9769	0.9625
DenseNet169	0.9788	0.9866	0.9827	0.9713	DenseNet169	0.9757	0.9890	0.9823	0.9713
DenseNet201	0.9723	0.9840	0.9781	0.9636	DenseNet201	0.9729	0.9849	0.9789	0.9658
ResNet18	0.9583	0.9826	0.9703	0.9504	ResNet18	0.9588	0.9890	0.9737	0.9570
ResNet34	0.9683	0.9799	0.9741	0.9570	ResNet34	0.9703	0.9849	0.9776	0.9636
Model	< 18 and ≥ 18 years				Model	< 30 and ≥ 30 years			
	Precision	Recall	F1	Accuracy		Precision	Recall	F1	Accuracy
ShuffleNet	0.9383	0.9383	0.9622	0.9394	ShuffleNet	0.9280	0.9806	0.9536	0.9350
ShuffleNet-V2	0.9067	0.9873	0.9453	0.9107	ShuffleNet-V2	0.9009	0.9854	0.9413	0.9162
MobileNet	0.9339	0.9760	0.9545	0.9272	MobileNet	0.9012	0.9741	0.9362	0.9096
MobileNet-V2	0.9630	0.9915	0.9771	0.9636	MobileNet-V2	0.9439	0.9806	0.9619	0.9471
DenseNet121	0.9696	0.9887	0.9791	0.9669	DenseNet121	0.9429	0.9887	0.9652	0.9515
DenseNet169	0.9682	0.9887	0.9887	0.9658	DenseNet169	0.9501	0.9854	0.9674	0.9548
DenseNet201	0.9682	0.9887	0.9784	0.9658	DenseNet201	0.9563	0.9903	0.9730	0.9625
ResNet18	0.9552	0.9915	0.9730	0.9570	ResNet18	0.9459	0.9903	0.9676	0.9548
ResNet34	0.9644	0.9929	0.9785	0.9658	ResNet34	0.9430	0.9903	0.9661	0.9526

Table 4
$Accuracy_{oss}$ of all tested models during cross-validation on the whole dataset, male, female individuals and patients with ages between 15 to 30 for ossification stage classification.

Models	All patients	Male patients	Female patients	Age between 15y to 30y
ShuffleNet	0.9305	0.9130	0.9468	0.7589
ShuffleNet-V2	0.8975	0.8650	0.9277	0.6786
MobileNet	0.9118	0.8970	0.9255	0.6607
MobileNet-V2	0.9636	0.9474	0.9787	0.8482
DenseNet121	0.9669	0.9588	0.9745	0.8839
DenseNet169	0.9735	0.9703	0.9766	0.9018
DenseNet201	0.9669	0.9588	0.9745	0.8482
ResNet18	0.9559	0.9405	0.9702	0.8393
ResNet34	0.9603	0.9474	0.9723	0.8304

$$Accuracy_{oss} = \sum \frac{Score_{oss}}{n} \quad (2)$$

The estimate was accurate if there was an overlap in the stage of ossification between ESO and ASO.

4. Results

4.1. Cross-validation performance

Cross-validation results showed that the DenseNet family outperformed other models with MAE ranging between 4.2 years and 4.3 years, and RMSE and SEE ranging between 6.1 years and 6.3 years (Table 2). MobileNet-V2 achieved the best balance between model size and performance. The mean error (bias, incorporating the direction of the error) of all models ranged from −1 to 1 year, indicating that they did not overestimate or underestimate age in the whole dataset. Figs. 3(a) and 3(b) depict the correlations and differences between the actual chronological age and the predict age. A good correlation coefficient (R^2=0.8871) was observed on DenseNet121.

For age classification, all DL models achieved AUC>0.9 on all age classification, especially, DenseNet169 achieved AUC=0.9945 for majority age classification, with precision=0.9682, recall=0.9887 (Table 3, Fig. 4). For classification of ossification stages of the medial clavicular epiphysis, Table 4 shows the $Accuracy_{oss}$ of DL models on the whole dataset, males, females, and individuals aged between 15 to 30 years. All models had $Accuracy_{oss}$ >86% on both male and female individuals. Especially, DenseNet169 had the highest $Accuracy_{oss}$ achieving a value of 97% (whole dataset), 97% (males), 98% (females), 90% (age between 15 to 30).

4.2. Performance on transfer learning experiments

Table 5 shows that there was an upward trend in the performance of DL models as more and more fine-tuned samples were included. For MobileNet-V2, Experiment-9 achieved the best performance in all metrics with MAE of 4.3 years. For DenseNet169, the performance of Experiment-10 was superior to other experiments with MAE of 3.3 years. Particularly, the MAE decreased by a factor of 2 for MobileNet-V2 (from 16.5 to 8.3 years) and from 16.8 to 6.8 years for DenseNet169 by using one fine-tuning sample for transfer learning (Experiment-1), demonstrating the power of pre-trained networks and transfer learning strategies. Furthermore, fine-tuning the pre-trained DenseNet169 network using 10 samples allows for a MAE of 3.3 years

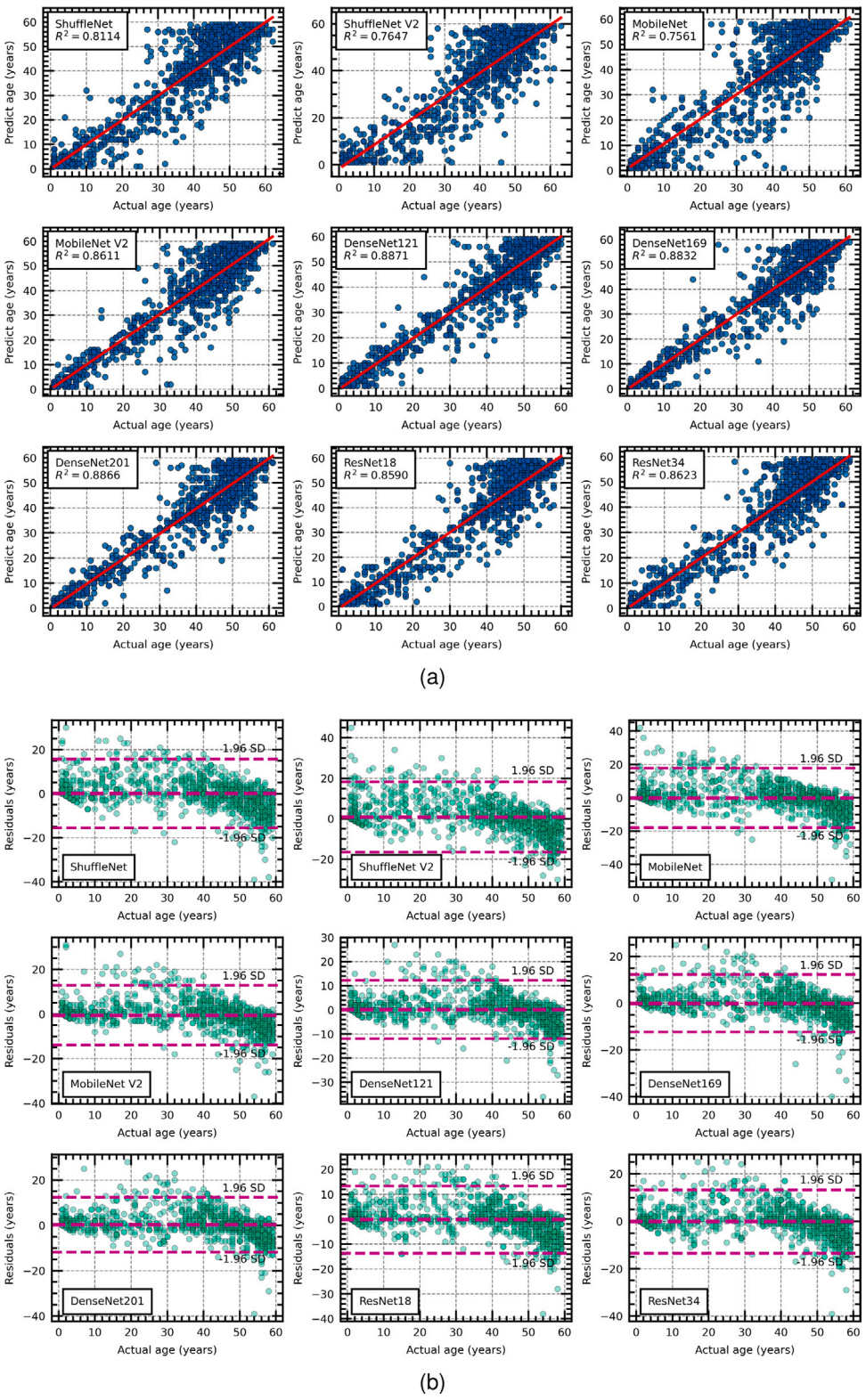


Fig. 3. Visualization results of the deep learning models trained during cross-validation. (a) Linear regression plots of actual age versus predicted age. (b) Bland-Altman plots show the difference between the estimated age of deep learning models and the ground truth age in cross-validation. The straight line marks bias whereas the dotted lines represent the 95% limits of agreement interval (bias±1.96×standard deviation).

for the external dataset, which is comparable to the results for the internal dataset (5-fold cross-validation). The MAE after fine-tuning was less than the MAE of the general model trained on 906 cases. One potential reason could be the better image quality we had in HUG-20 dataset. We used transfer learning since the number of high-quality images with the desired acquisition parameters were limited.

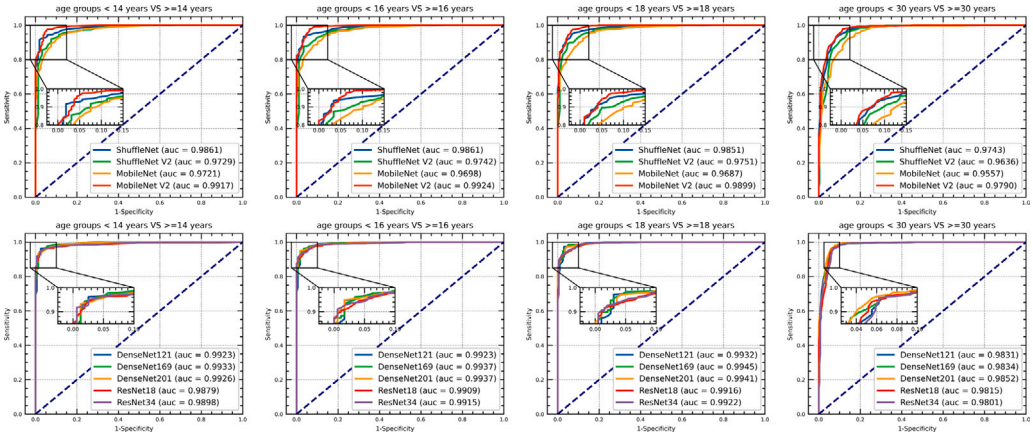


Fig. 4. ROC curves for separating the age groups < 14 and ≥ 14 years, < 16 and ≥ 16 years, < 18 and ≥ 18 years, and < 30 and ≥ 30 years of the deep learning models trained during cross-validation,.

Table 5
MAE (in years) and standard deviation, Mean Error (in years) and standard deviation, MSE, RMSE (in years), SEE (in years) and $Accuracy_{oss}$ of MobileNet-V2 and DenseNet169 in the transfer learning experiments.

Models	Experiments	MAE	ME	MSE	RMSE	SEE	$Accuracy_{oss}$
MobileNet-V2	Experiment-0	16.50 ± 7.94	12.50 ± 13.38	335.30	18.31	20.47	0.50
	Experiment-1	8.28 ± 7.42	−0.52 ± 11.11	123.60	11.12	11.35	0.60
	Experiment-2	6.72 ± 5.75	−1.68 ± 8.68	78.20	8.84	9.03	0.70
	Experiment-3	5.92 ± 5.38	−1.72 ± 7.82	64.04	8.00	8.17	0.80
	Experiment-4	6.00 ± 5.60	−2.68 ± 7.76	67.40	8.21	8.38	0.80
	Experiment-5	5.44 ± 4.89	−1.20 ± 7.22	53.52	7.32	7.47	0.80
	Experiment-6	5.14 ± 4.71	−0.18 ± 6.97	48.58	6.97	7.11	0.80
	Experiment-7	4.62 ± 4.83	−0.58 ± 6.66	44.66	6.68	6.82	0.80
	Experiment-8	4.94 ± 4.60	−2.22 ± 6.38	45.58	6.75	6.89	0.80
	Experiment-9	4.36 ± 4.29	0.00 ± 6.12	37.44	6.12	6.25	0.80
	Experiment-10	4.90 ± 5.75	−2.70 ± 7.06	57.10	7.56	8.45	0.70
DenseNet169	Experiment-0	16.80 ± 4.19	14.40 ± 9.61	299.80	17.31	19.36	0.50
	Experiment-1	6.80 ± 5.40	4.12 ± 7.64	75.40	8.68	8.86	0.90
	Experiment-2	5.34 ± 5.18	0.78 ± 7.40	55.34	7.44	7.59	0.80
	Experiment-3	5.28 ± 4.85	1.04 ± 7.09	51.40	7.17	7.32	0.80
	Experiment-4	4.66 ± 4.18	0.26 ± 6.25	39.18	6.26	6.33	0.80
	Experiment-5	4.90 ± 4.26	1.78 ± 6.25	42.18	6.49	6.63	0.90
	Experiment-6	4.22 ± 4.44	−1.18 ± 6.01	37.50	6.12	6.25	0.80
	Experiment-7	4.12 ± 4.33	−0.76 ± 5.92	35.68	5.97	6.10	0.80
	Experiment-8	4.14 ± 3.60	−0.78 ± 5.43	30.10	5.49	5.60	0.90
	Experiment-9	3.78 ± 3.72	−0.74 ± 5.25	28.10	5.30	5.41	0.90
	Experiment-10	3.30 ± 3.58	−0.70 ± 4.82	23.70	4.87	5.44	0.90

Table 6
Ossification stage classification results of three experts and DenseNet169. The marker "*" indicates that this case is uninterpretable for the expert. Cases with Case ID 1 to 5 come from the HUG-20 cohort. Cases with Case ID 6 to 10 come from the HUG-5 cohort. The first (left) and second (right) parenthesis show the physician decision on the 1st and 2nd review. (Left 1st/Right 1st) (Left 2nd/Right 2nd). The staging results of DL were derived from the predicted age of fine-tuned DenseNet169 based on the minimum and maximum ages of stages 1, 4, and 5 in the Kellinghaus study [6], and the minimum and maximum ages of stages 2a, 2b, 2c, 3a, 3b, and 3c in the Kellinghaus study [5]. The staging results of Ground truth are also transformed from the actual age in the same way.

Case ID	Observer #1	Observer #2	Observer #3	DL	Ground truth
1	(5/5)(5/5)	(4/4)(4/4)	(4/4)(4/4)	3b/3c/4	4/5
2	(3a/3c)(3a/3c)	(3c/3c)(*/*)	(*/*)(*/*)	2a/2b/2c/3a/3b/3c	2a/2b/2c/3a/3b
3	(5/5)(5/5)	(3c/3c)(4/*)	(3c/*)(4/*)	4/5	4/5
4	(3a/3b)(3a/3b)	(3c/3b)(3b/3b)	(3b/3b)(3b/3b)	3b/3c/4	2a/2b/2c/3a/3b/3c
5	(2b/3a)(2b/3a)	(3a/3a)(3a/3a)	(3a/3a)(3a/3a)	2a/2b/2c/3a/3b	1/2a/2b/2c/3a
6	(3c/3c)(3c/3c)	(3c/3c)(3c/3c)	(3c/3c)(3c/3c)	3b/3c/4	–
7	(4/5)(4/5)	(3c/3c)(3c/3c)	(3c/3c)(3c/3c)	2a/2b/2c/3a/3b	–
8	(3a/3a)(3a/3a)	(3a/3a)(3a/3a)	(3a/3a)(3a/3a)	3b/3c/4	–
9	(4/4)(3c/4)	(3c/3c)(3c/3c)	(3c/3c)(3c/3c)	2a/2b/2c/3a/3b	–
10	(5/5)(5/5)	(3c/3c)(3c/3c)	(3c/3c)(3c/3c)	3b/3c/4	–

4.3. Performance on real age estimation cases

Table 6 shows the results of three experts and DenseNet169 for the determination of the ossifications stage on 10 cases. The results of experts are presented as "(right clavicle in the first estimation/left clavicle in the first estimation) (right clavicle in the second estimation/left clavicle in the second estimation)". It can be seen that the reproducibility of Observer #1 is relatively high, but Observer #2 and #3 have a low reproducibility between the first and second estimation. The three experts did not agree on the images that met the diagnostic

requirements. Moreover, the first and second evaluations also showed inconsistent staging results. Some cases were interpretable for Observer #1 and Observer #2 in the first assessment but uninterpretable in the second assessment. Compared with the ground truth, Observer #1, #2 and #3 achieved an $Accuracy_{oss}$ of 100%/100%, 60%/80%, 60%/80% in the first/second assessment, respectively. DenseNet169 achieved 100% of $Accuracy_{oss}$.

5. Conclusion

To the best of our knowledge, this work is the first attempt to use deep learning to facilitate and to assist the age estimation task based on clavicle CT, making them fully automated, more accurate and more robust. Our model showed high prediction performance and could potentially be included as a reference to help radiologists and forensic pathologists to make decisions. A potential risk of using the pre-trained deep learning method is that the model might overfit the training data. Hence, we recommend using a fine-tuning strategy to mitigate the problem. The study's results could provide insights into the usefulness of DL methods in age assessment and contribute to the development of more accurate and reliable age estimation techniques. This study inherently bears several limitations: (1) Imbalanced classes: minors and adults do not make up an equal portion of the training dataset, i.e., the portion of minors is 21.8%, whereas the portion of subjects aged between 13 and 25 years is 11.8%. Hence, DL models cannot recognize data from minority classes easily. However, in the medico-legal setting, the classification of asylum seekers without valid identification documents is a challenge that requires careful consideration by legal authorities to prevent the misclassification of minors as adults. To solve this problem in the future, more sophisticated DL techniques can be utilized to address the class imbalance data problem. (2) In transfer learning experiments, the testing cohort was from the same institution as the training dataset, and the testing sample size was relatively small. To evaluate and improve the robustness of the model, more subjects from other institutions are needed. (3) The subjects from the real cases of age estimation are only males; however, all civil case requests were male in our practice. (4) Small sample size for radiological examination. (5) The 5 thorax CT cases from HUG-20 used for evaluating the intra-observer and inter-observer reproducibility are not specific to evaluate the development status of the clavicle, even if the slice thickness is less than 1 mm and the reconstruction kernel is almost the same as HUG-5.

In the future, more advanced networks can be tested in our proposed workflow to fully exploit the potential of the deep learning models. At the same time, the way to comprehensively evaluate the uncertainty of deep learning methods in forensic age estimation is also an important and promising research topic.

CRediT authorship contribution statement

Chang Sun: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Yazdan Salimi:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation. **Isaac Shiri:** Conceptualization, Methodology, Validation, Writing – review & editing. **Coraline Egger:** Data curation, Investigation, Visualization, Writing – review & editing. **Pia Genet:** Data curation, Formal analysis, Writing – review & editing. **Habib Zaidi:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Sana Boudabbous:** Conceptualization, Formal analysis, Methodology, Resources, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Habib Zaidi reports financial support was provided by the Swiss National Science Foundation

Acknowledgments

This work was supported by the Swiss National Science Foundation under grant SNSF 320030_231742. Chang Sun acknowledges the support from the China Scholarship Council (Grant No.202206470014) and the BUPT Excellent Ph.D. Students Foundation (CX2022203).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fri.2024.200619.

Data availability

The data used are not available owing to legal issues. Trained models and code will be made available on GitHub.

References

[1] SEM, Asile : statistiques de 2021, 2022, [Online]. Available: URL https://www.sem.admin.ch/sem/fr/home/sem/medien/mm.msg-id-87177.html.
[2] DGRM, Stellungnahme: forensische altersdiagnostik bei unbegleiteten minderjährigen flüchtlingen, 2018, [Online]. Available: URL https://www.dgrm.de/news/stellungnahmen.
[3] SSML, Forensische Altersdiagnostik, URL https://bit.ly/3C62nKI.
[4] A. Schmeling, C. Grundmann, A. Fuhrmann, H.-J. Kaatsch, B. Knell, F. Ramsthaler, W. Reisinger, T. Riepert, S. Ritz-Timme, F.W. Rösing, Criteria for age estimation in living individuals, Int. J. Legal Med. 122 (2008) 457–460.
[5] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, H. Pfeiffer, A. Schmeling, 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 (2010) 321–325.
[6] M. Kellinghaus, R. Schulz, V. Vieth, S. Schmidt, A. Schmeling, Forensic age estimation in living subjects based on the ossification status of the medial clavicular epiphysis as revealed by thin-slice multidetector computed tomography, Int. J. Legal Med. 124 (2010) 149–154.
[7] R. Schulz, Radiologische Untersuchungen zum zeitlichen Verlauf der Schlüsselbeinossifikation: Konsequenzen für die forensische Altersdiagnostik bei lebenden Personen (Ph.D. thesis), 2006.
[8] R. Schulz, M. Mühler, S. Mutze, S. Schmidt, W. Reisinger, A. Schmeling, Studies on the time frame for ossification of the medial epiphysis of the clavicle as revealed by CT scans, Int. J. Legal Med. 119 (2005) 142–145.
[9] D. Wittschieber, C. Ottow, R. Schulz, K. Püschel, T. Bajanowski, F. Ramsthaler, H. Pfeiffer, V. Vieth, S. Schmidt, A. Schmeling, Forensic age diagnostics using projection radiography of the clavicle: a prospective multi-center validation study, Int. J. Legal Med. 130 (2016) 213–219.
[10] D. Wittschieber, R. Schulz, V. Vieth, M. Küppers, T. Bajanowski, F. Ramsthaler, K. Püschel, H. Pfeiffer, S. Schmidt, A. Schmeling, The value of sub-stages and thin slices for the assessment of the medial clavicular epiphysis: a prospective multi-center CT study, Forensic Sci. Med. Pathol. 10 (2014) 163–169.
[11] I. Atas, C. Ozdemir, M. Atas, Y. Dogan, Forensic dental age estimation using modified deep learning neural network, 2022, arXiv preprint arXiv:2208.09799.
[12] A. Demircioglu, A.S. Quinsten, M. Forsting, L. Umutlu, K. Nassenstein, Pediatric age estimation from radiographs of the knee using deep learning, Eur. Radiol. 32 (7) (2022) 4813–4822, http://dx.doi.org/10.1007/s00330-022-08582-0.
[13] Q.L. Hui, C.L. Wang, J.W. Weng, M. Chen, D.X. Kong, A global-local feature fusion convolutional neural network for bone age assessment of hand X-ray images, Appl. Sci.-Basel 12 (14) (2022) 15, http://dx.doi.org/10.3390/app12147218.
[14] D. Milosevic, M. Vodanovic, I. Galic, M. Subasic, Automated estimation of chronological age from panoramic dental X-ray images using deep learning, Expert Syst. Appl. 189 (2022) 12, http://dx.doi.org/10.1016/j.eswa.2021.116038.
[15] L.Q. Peng, Y.C. Guo, L. Wan, T.A. Liu, P. Wang, H. Zhao, Y.H. Wang, Forensic bone age estimation of adolescent pelvis X-rays based on two-stage convolutional neural network, Int. J. Legal Med. 136 (3) (2022) 797–810, http://dx.doi.org/10.1007/s00414-021-02746-1.
[16] H. Seo, J. Hwang, Y.H. Jung, E. Lee, O.H. Nam, J. Shin, Deep focus approach for accurate bone age estimation from lateral cephalogram, J. Dental Sci. 18 (1) (2023) 34–43, http://dx.doi.org/10.1016/j.jds.2022.07.018.
[17] R. Toutin, M.F. Bilfeld, C. Raspaud, C. Bec, N. Telmon, F. Savall, F. Dedouit, Contribution of the use of clavicle bone density in age estimation, Int. J. Legal Med. 136 (4) (2022) 1017–1025, http://dx.doi.org/10.1007/s00414-021-02741-6.
[18] R. Shedge, T. Kanchan, P.K. Garg, S.G. Dixit, V. Warrier, P. Khera, K. Krishan, Computed tomographic analysis of medial clavicular epiphyseal fusion for age estimation in Indian population, Leg. Med. 46 (2020) 101735.

[19] Fedlex, Loi fédérale régissant la condition pénale des mineurs, URL https://www.fedlex.admin.ch/eli/cc/2006/551/fr#chap_1.
[20] Child rights international network, Country reports, URL https://home.crin.org/a2j-country-reports.
[21] D. Franklin, A. Flavel, CT evaluation of timing for ossification of the medial clavicular epiphysis in a contemporary Western Australian population, Int. J. Legal Med. 129 (3) (2015) 583–594, http://dx.doi.org/10.1007/s00414-014-1116-8.
[22] M.S. Gurses, N.T. Inanir, G. Gokalp, R. Fedakar, E. Tobcu, G. Ocakoglu, Evaluation of age estimation in forensic medicine by examination of medial clavicular ossification from thin-slice computed tomography images, Int. J. Legal Med. 130 (5) (2016) 1343–1352, http://dx.doi.org/10.1007/s00414-016-1408-2.
[23] T. Houpert, C. Rerolle, F. Savall, N. Telmon, P. Saint-Martin, Is a CT-scan of the medial clavicle epiphysis a good exam to attest to the 18-year threshold in forensic age estimation? Forensic Sci. Int. 260 (2016) 3, http://dx.doi.org/10.1016/j.forsciint.2015.12.007.
[24] S.U. Ramadan, M.S. Gurses, N.T. Inanir, C. Hacifazlioglu, R. Fedakar, S. Hizli, Evaluation of the medial clavicular epiphysis according to the Schmeling and Kellinghaus method in living individuals: A retrospective CT study, Leg. Med. 25 (2017) 16–22, http://dx.doi.org/10.1016/j.legalmed.2016.12.012.
[25] R. Scendoni, I.L.A. Ribeiro, M. Cingolani, A. Giovagnoni, M. Curzi, P. Fedeli, R. Cameriere, A new analytical cut-off point for determining 18 years of age using MRI on medial clavicular epiphysis, Leg. Med. 54 (2022) 6, http://dx.doi.org/10.1016/j.legalmed.2021.102010.
[26] A. Schmeling, R. Schulz, B. Danner, F. Rosing, The impact of economic progress and modernization in medicine on the ossification of hand and wrist, Int. J. Legal Med. 120 (2) (2006) 121–126, http://dx.doi.org/10.1007/s00414-005-0007-4.
[27] L. Meijerman, G.J.R. Maat, R. Schulz, A. Schmeling, Variables affecting the probability of complete fusion of the medial clavicular epiphysis, Int. J. Legal Med. 121 (6) (2007) 463–468, http://dx.doi.org/10.1007/s00414-007-0189-z.
[28] N.N. Win, B. Erok, The use of medial clavicular epiphysis ossification stages for bone age determination, Eur. J. Clin. Exper. Med. (4) (2021) 299–305.
[29] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708.
[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.
[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[32] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran, D. Rubin, A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions, Sci. Data 9 (1) (2022) 601.
[33] S. Gatidis, T. Kuestner, A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions) [Dataset], The Cancer Imaging Archive (2022).
[34] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, The cancer imaging archive (TCIA): maintaining and operating a public information repository, J. Digital Imaging 26 (2013) 1045–1057.
[35] P. Jordan, P.M. Adamson, V. Bhattbhatt, S. Beriwal, S. Shen, O. Radermecker, S. Bose, L.S. Strain, M. Offe, Pediatric chest/abdomen/pelvic CT exams with expert organ contours (pediatric-CT-SEG) (version 2) [dataset], The Cancer Imaging Archive (2021).
[36] P. Jordan, P.M. Adamson, V. Bhattbhatt, S. Beriwal, S. Shen, O. Radermecker, S. Bose, L.S. Strain, M. Offe, D. Fraley, Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours, Med. Phys. 49 (5) (2022) 3523–3528.
[37] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, pp. 234–241.
[38] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856.
[39] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, pp. 116–131.
[40] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[43] M.J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, MONAI: An open-source framework for deep learning in healthcare, 2022, arXiv preprint arXiv:2211.02701.
[44] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.