# Automated Bone Age Assessment based on DXA scans for a variety of ethnicities using Deep Transfer Learning

MSc Thesis Biomedical Engineering - Medical Physics - BM51035

by

Ilva van Houwelingen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended on Thursday June 24, 2021 at 13:30 PM.

Student number: 4473965
Project duration: September 1, 2020 – June 24, 2021
Thesis committee:

| | |
|---|---|
| Dr. G. V. Roshchupkin, | Erasmus Medical Center Rotterdam, daily supervisor |
| Dr. F. M. Vos, | TU Delft, BME supervisor |
| Prof. Dr. W. J. Niessen, | Erasmus Medical Center Rotterdam & TU Delft |

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft Delft University of Technology

Erasmus MC University Medical Center Rotterdam

# Automated Bone Age Assessment based on DXA scans for a variety of ethnicities using Deep Transfer Learning

Ilva van Houwelingen[1], Fernando Rivadeneira[2], Frans M. Vos[1], Gennady V. Roshchupkin[2]

[1]Technical University of Delft,
[2]Erasmus University Medical Center Rotterdam

**Abstract**

A child's bone age is important for the diagnosis of a wide range of growth disorders. The most often used manual method for bone age assessment (BAA) consists of comparing hand-wrist radiographs with 'ground-truth' atlasses. This method is criticised for being time-invasive, prone to inter- and intra-observer variability and not applicable to the present-day multicultural population. Therefore, much research has been conducted in creating automated methods for BAA, using machine or deep learning (DL). Instead of using radiographs, dual-energy X-ray absorptiometry (DXA) scans could also be used for BAA, which have the benefit of a lower effective dose. This study focuses on two gaps in current research on automated BAA: developing an automated method for the use on DXA scans and incorporating ethnic information into the algorithm.

For this purpose, a DL network was constructed and pre-trained on a large data set of radiographs. Transfer learning was adopted to a data set containing DXA scans. The performance of four different models was measured in mean absolute difference (MAD) to observe the effect of adding gender and ethnic information as extra inputs. Final performance was measured on a lock box, which was kept aside during the entire training and tuning process. To gain more insight, regions important for the assessment by the automated model were being visualised using a modified version of Class Activation Mapping (CAM). Furthermore, a comparison was made with software created for automated BAA on radiographs.

Whether or not adding gender and ethnic information as extra inputs did not show a clear effect on the performance. The final performance on the lock box was an MAD of 6.8 months. The activation maps showed that the carpal region was the most important for the automated BAA. The comparison with the radiograph software showed it was not applicable on DXA scans and emphasised the need for a DXA-specific method.

This is the first study that developed an automated BAA method for the use on DXA scans rather than radiographs and the first that incorporates ethnic information inside the algorithm. An MAD of 6.8 months on a totally independent test set (lock box) is comparable with the inter-observer variability of manual BAA and performances reported for state-of-the-art automated BAA methods on radiographs. This method can contribute to reducing radiation exposure and time-intensiveness of the current BAA procedure.

**Index Terms**

Bone age, Skeletal maturation, DXA, Deep neural networks, Transfer learning

## I. INTRODUCTION

A child's bone age can be important information for pediatricians and endocrinologists. It can be different from chronological age, which is based on the date of birth of the child. The gap between bone age and chronological age can be a biomarker for a wide range of pediatric and endocrine disorders and a predictor for final adult height [1].

Current bone age assessment (BAA) methods rely mainly on manually comparing radiographs of the child's left hand to a predefined atlas or a detailed shape analysis of some bones of interest [2]. The most often used manual methods are the Greulich and Pyle (G-P) [3] and the Tanner-Whitehouse (TW) [4] methods. These procedures are very time-consuming for the pediatric endocrinologist or radiologist. King et al. reports an average assessment time of 1.4 minutes and 7.9 minutes for the G-P method and the TW method, respectively [5]. Furthermore, the quality of the assessment is highly dependent on the experience and skills of that particular health professional and therefore prone to inter- and intra-observer variability. Unsurprisingly, many studies have reported large standard errors between observers in the range of 5.4 to 10.0 months [6][7][8][9][10][11]. Moreover, there might be subtle

changes in bone maturation over a short time that are too small to be picked up by human visual inspection. These features can, however, play a significant role in BAA [12].

One way to overcome these drawbacks might be to automate the BAA using computerized techniques such as machine learning [13]. Especially deep learning (DL), where features are extracted automatically rather than manually, has shown to be a successful approach in a wide variety of medical imaging tasks [14].

To lower the radiation dose for the child, dual-energy X-ray absorptiometry (DXA) scans (Fig. 1) could be used as an alternative for BAA. DXA scans are currently used to measure bone mineral density for the diagnosis of osteoporosis and have the benefit of a 10-fold lower effective dose compared to conventional radiographs: 0.001 mSv versus 0.0001 mSv [15]. Limiting radiation exposure is always preferred in health care, but it is particularly important in children as their risk of developing radiation-induced complications over a lifetime is greater than in adults [16][17]. This is especially important if the scans are used for research purposes of a healthy population, rather than children with a suspected disorder. There have already been made some successful attempts to apply the G-P method on DXA scans [18][19][20][21]. However, to the best of our knowledge, there has not been attempted to develop automated methods for this.

Another issue to be aware of in BAA is the variability of skeletal maturation in different sexes and ethnicities [22][23][24]. It is widely known that on average, girls maturate faster than boys. Therefore, gender information is often incorporated in automated BAA methods. Furthermore, multiple studies have shown the difference in progression of skeletal maturation between different ethnic groups [25][23][26]. However, ethnic information is not yet widely incorporated in automated BAA methods [13][27].

This study has two aims: to study the feasibility of developing an automated DL algorithm for BAA based on DXA scans and incorporating demographic information into the algorithm. Conventionally, the use of DL requires a large amount of annotated data, which can be an important limitation in medical applications. A commonly used method to make use of the advantages of a DL network on a smaller data set is transfer learning. It uses pre-trained weights from a larger data set and fine-tunes the network on the (smaller) specific data set.

Accordingly, a large publicly available data set of hand-wrist radiographs labelled with bone age is used to pre-train a DL network. Next, different blocks of the network are frozen while other blocks are retrained on a data set containing DXA scans of children with various ethnic backgrounds from the population-based Generation R (GenR) study [28]. In addition to predicting a bone age, the most important regions of the hand for the assessment are being visualized and compared with features known to play an important role in (manual) BAA. Furthermore, a comparison is made with software developed for BAA on radiographs to show the need for a DXA-specific method. Finally, an approach is suggested to use the automated model in case of a data set with a different bone age range.

## II. RELATED WORK

### *A. Conventional machine learning algorithms on radiographs*

The first automated BAA methods made their entrance in the 1990s and started to develop since. Such techniques comprise semi-automated methods and later also fully automated methods using image processing or knowledge-based machine learning techniques [29][30][31][32][33]. Although some of these methods showed promising results in terms of reliability and reducing observer variability, there were some major drawbacks that contributed to the limited clinical use. First of all, the experiments in which the accuracy was determined were often based on small data sets and comparison in terms of accuracy was sometimes performed with chronological age instead of a ground-truth bone age. Furthermore, these methods were based on predetermined, hand-crafted features, limiting generalizability and finding other features with discriminative value. Finally, these methods were still time-consuming. The time complexity of such existing methodology is either due to the used image processing techniques that took up to minutes, or because there was still some manual intervention required, or both. The latter also makes a method less generalizable, less objective and more sensitive to inter- and intra-observer variability [31].

One system, called BoneXpert, which performed well enough to be made commercially available, is currently implemented in 150 hospitals [34][35]. The method has been validated in Northern-European children with

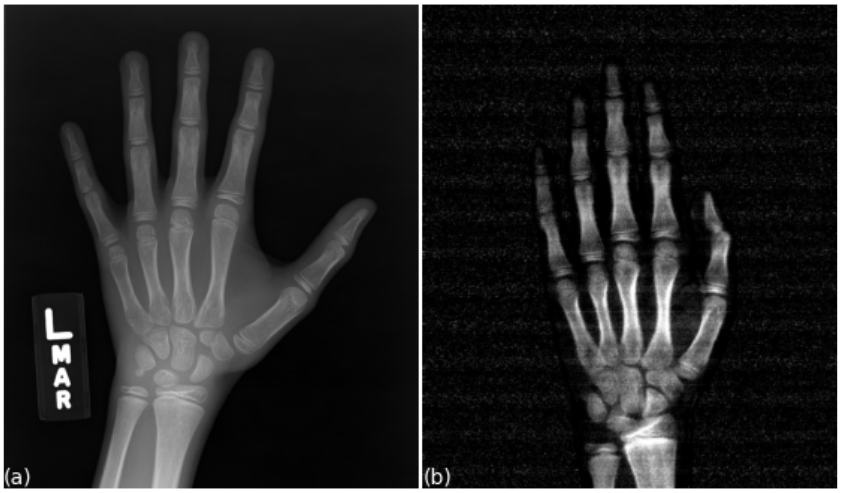


Fig. 1: Hand-wrist (a) radiograph [43] and (b) DXA scan [28] of two different children both with a bone age of 11 years.

short statures [36], precocious puberty [37], congenital adrenal hyperplasia [38] and children with a variety of ethnic backgrounds [8]. Although BoneXpert is commercially available, it still has its limitations. First of all, the predicted age is partly based on chronological age rather than solely on bone age [39][40]. Furthermore, the method does not use the carpal bones in the BAA, where the ossification pattern of these bones in younger children could be important [39][22]. Finally, it rejects low-quality images of which noise levels are too high [41]. Particularly, Zhang et al. reported a rejection of 4.5% (235 out of 5161) of individual bones [42]. Importantly, the software was developed for the use on radiographs and is not directly applicable on DXA scans.

### *B. Deep learning algorithms on radiographs*

A breakthrough in overcoming the limitations in generalizability of conventional machine learning techniques is the use of DL networks where features are automatically extracted rather than manually [44]. The first methods using deep learning started to develop from 2016, but most of these methods were used privately or with small to medium-sized data sets [45][46][39]. Furthermore, performances were not robust enough for practical applications.

Larson et al. used a large data set of 12611 radiographs and showed reasonable performance with their pre-trained convolutional neural network (CNN) based on residual learning [11].

In 2017, the Radiological Society of North America (RSNA) launched the Pediatric Bone Age Machine Learning Challenge [43]. For this challenge, they made the data set of Larson et al. [11] publicly available. The top five performing algorithms submitted to the challenge reported a mean absolute distance (MAD) below 4.6 months. One of the participating teams was that of the BoneXpert [34], which was the only one in the top five using conventional machine learning rather than DL.

With this large data set being publicly available, many new algorithms were created for the purpose of BAA [47][48][49][50][51][52][53]. The top performances on this data set are of Wang et al. [52] and Tang et al. [53], who report MAD values 4.0 and 2.3 months, respectively. The latter reference performed their evaluation on an unseen part of the training set rather than the original test set, which is important to note. Tang et al. [53] pre-trained their network on the RSNA data set and performed transfer learning on a different data set. Their final performance on a different data set was an MAD of 1.9 months, suggesting that transfer learning from the RSNA data set is a good approach for BAA.

Table 1 shows an overview of the performances of the aforementioned state-of-the-art methods on automated BAA based on radiographs.

TABLE 1: Performance of state-of-the-art methods in automated BAA based on hand-wrist radiographs

| Method | Data set | RMS | MAD |
|---|---|---|---|
| BoneXpert [34] | Digital Hand Atlas | 7.3 | |
| | RSNA [43] | | 4.4 |
| Chen [45] | Digital Hand Atlas | 13.2 | |
| Spampinato et al. [46] | Digital Hand Atlas | | 9.5 |
| Lee et al. [39] | Private | 10.6 | |
| Larson et al. [11] | RSNA | 7.6 | 6.0 |
| | Digital Hand Atlas | 8.8 | |
| RSNA Pediatric Bone Age Challenge [43] | RSNA | | Top-5 $< 4.6$ |
| Iglovikov et al. [47] | RSNA | | 4.9 |
| Wang et al. [52] | RSNA | | 4.0 |
| Tang et al. [53] | RSNA | | 2.3 |
| Human* [11] | RSNA | | 7.3 |

RMS = Root Mean Squared error, in months; MAD = Mean Absolute Difference, in months.
*Human performance is measured as the mean MAD of the four readings on the test set [11].

One main limitation of the RSNA data set is that it contains no information on ethnicity. Not incorporating ethnic information is known to be a gap in current work on automated BAA [13][27].

### C. Automated BAA based on DXA scans

To the best of our knowledge, there are no studies that attempted to automate BAA with the use of DXA scans instead of radiographs. One study comparing a DXA-based method with conventional radiographs for BAA, mentioned the use of Bone Age Assessment software. However, the age was still assigned by a DXA operator using reference values [54]. There are automated methods where DXA scans are used to predict other factors, for example hip fracture risk [55][56] and osteoporosis [57]. Such tasks are in essence similar to automated BAA based on DXA scans.

## III. METHOD

Inspired by Tang et al. [53], our study adopted transfer learning from the large RSNA radiograph data set [43] to a private data set containing DXA scans. Pre-training on the RSNA data set should assist the model with already pointing to features important for BAA that are similar in DXA scans. As such, this approach made use of the generalizability properties of a large data set and the specificity of the DXA data set for the task at hand. The features important for the prediction were visualised using Regression Activation Mapping (RAM). Next to this, a comparison was made with the commercial BoneXpert software that is designed for the use on radiographs [34]. Furthermore, experiments were performed to show how much training data is needed for reasonable performance if application of the model on a data set with a different age range is desired.

### A. Data sets and processing

The publicly available RSNA data set consists of 12611 hand-wrist radiographs labelled with bone age and gender [43]. Bone age was assessed by one reader making use of the G-P method [3]. The test set contains 200 radiographs labelled with gender and bone age. However, for this test set, bone age was determined by an average of six readings [43]. Bone age distribution of the training and test set is shown in figure 2 (left). As for the gender distribution, the data set consists of 5778 female and 6833 male samples. The test set is stratified with respect to gender (i.e., 100 samples for both classes). For the collection of this database informed consent was waived [11]. Image intensities were discretized through 8-bit grey levels. The images were stored in PNG format.

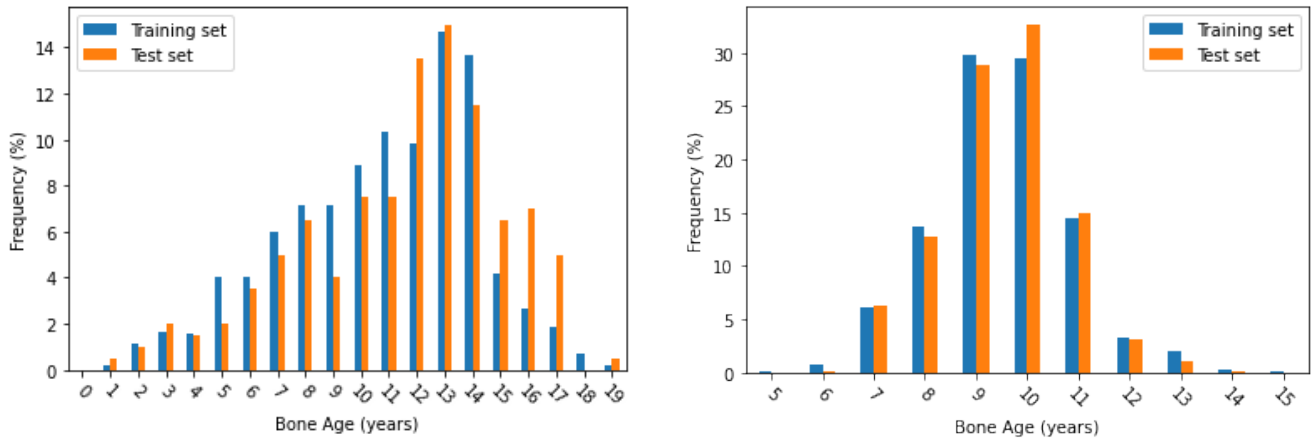


Fig. 2: Bone age distribution of RSNA [43] (left) and GenR focus 9 [28] (right) training and test set.

The DXA scans were retrieved from the Generation R Study: A multi-ethnic population-based prospective cohort study established at the Erasmus University Medical Center in Rotterdam. The goal of this study is to identify early causes of abnormal growth, development and health [28]. The data used for the purpose of this study was collected in the third cohort visit (mean age of $9.78 \pm 0.33$ years) with informed consent obtained for all participating children. DXA scans of the left hand-wrist area were obtained by a trained operator making use of an iDXA densitometer (General Electric, formerly Lunar Corp., Madison, WI, USA) and a standard protocol. Bone age was assessed by one trained observer using the G-P method [3]. 150 DXA scans were randomly selected for a blind re-assessment by the same observer. Our study included participants for which gender, ethnic and bone age information was available (N = 5310). The bone age distribution is displayed in figure 2 (right). Data was included from 2425 male and 2484 female samples. Concerning ethnicity, 4112 were classified as 'White', 463 as 'Black', 225 as 'Asian' and 109 as 'Hispanic' by the DXA operator. The obtained DXA scans were in MEB format but converted into JPG format during preprocessing that eliminated patient sensitive information.

For the purpose of transfer learning, the size of the images from both data sets had to be identical. During preprocessing the images were therefore resampled to a size of 500x500 while preserving aspect-ratio (i.e., zero-padding was used).

### *B. Network architecture*

The constructed base convolutional neural network (CNN) architecture is inspired by a study of Wang et al. [58] where they predict brain age based on magnetic resonance (MR) images. The initial architecture, along with its settings, are shown in figure 3. Inspired by some of the state-of-the-art methods on automated BAA based on radiographs [43][49][53][51], gender was included as a separate input to correct for the differences in skeletal maturation between sexes. A dropout layer was applied between the fully connected layers for regularization purposes.

### *C. Training and transfer learning*

The RSNA data set was randomly split into training and validation set with a ratio of 80:20 (see Supplementary Data A table 1: RSNA). The validation set was used to keep track of and limit overfitting (e.g., the model weights were only saved in case of improvement of validation loss) and tuning hyperparameters. During pre-training data augmentation was applied in the form of random rotations and translations (up to 10% of image size) for regularization purposes.

The loss function to optimize was the Mean Squared Error (MSE) loss, which is a common loss function in regression problems:

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

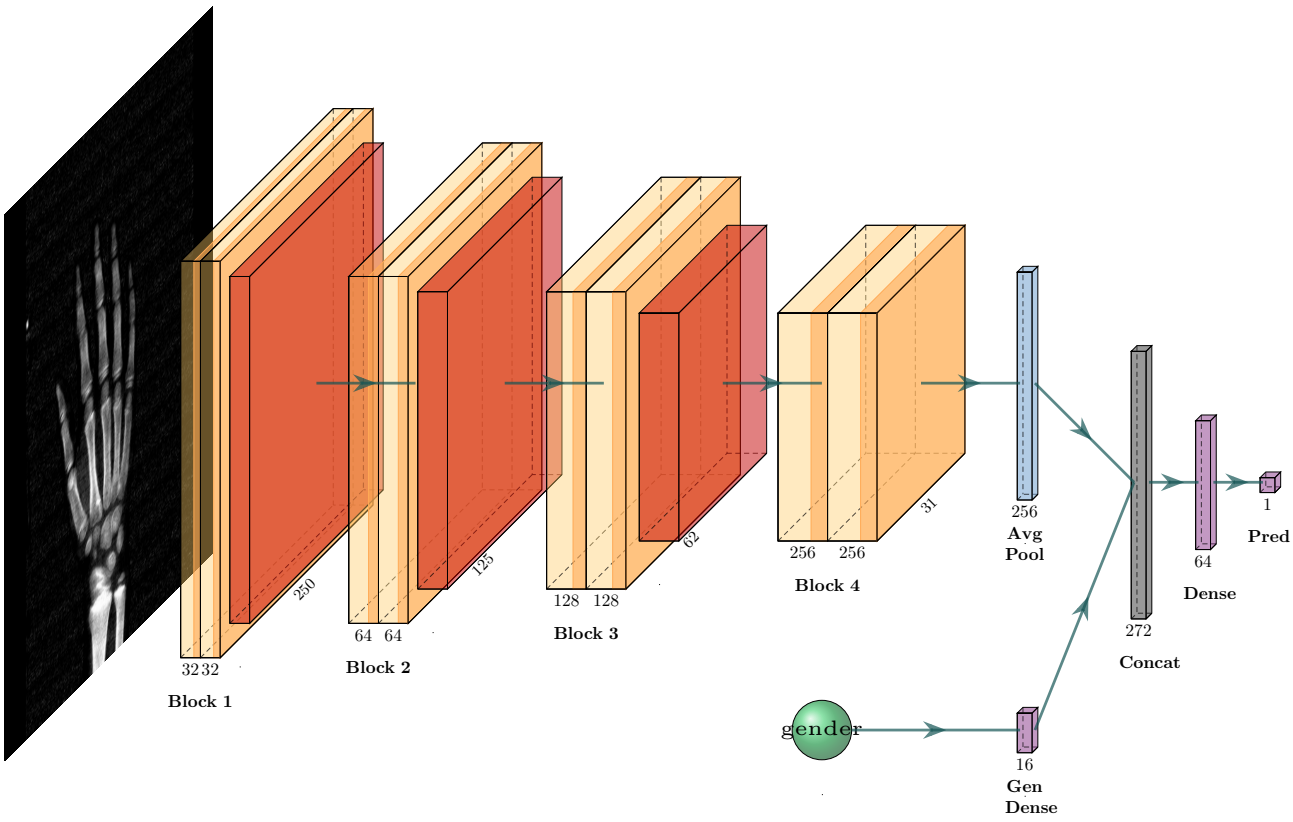


Fig. 3: Network architecture. It consists of three blocks of two convolutional layers each (including batch normalization and rectified linear unit (ReLU) activation) followed by a max-pooling layer. Block 4 consists of two convolutional layers (including batch normalization and ReLU activation) followed by a global average pooling layer. A binary number that represents gender (0=female, 1=male) is first passed along a fully connected layer before concatenated to the image information from the global average pooling layer. The fully connected layer following is used to propagate to a single regression output. A dropout layer is applied in between the last fully connected layer and the prediction layer.

During transfer learning, experiments were performed to find the optimal number of frozen blocks and fine-tuning the rest of the network (i.e., which weights are being updated during retraining and which are not) [59]. To this end, the Generation R data set was split up into training and validation set with a ratio of 75:15 and the performances of these measures were determined on a test set of 10% (see Supplementary Data A table 1: GenR preliminary experiments).

### *D. Final outcomes*

With the optimal transfer learning network, experiments were performed to determine the performance of four different network architectures:

1) the baseline network as shown in figure 3;
2) the baseline network without gender information;
3) the baseline network without gender information but with ethnic information, and
4) the baseline network with both gender and ethnic information.

Ethnic information was transformed in a numerical array using one-hot-encoding (i.e., a vector of size four comprised of zeros, except for one position labelled with one that is specific for a certain ethnicity). After that, it was incorporated similar to gender information passing through a dense layer of size 16 before concatenation.

The optimal network was found using 10-fold cross-validation and taking the mean of the 10 folds as performance measure. Next, final performance was measured on a lock box that was kept aside until the very last moment, following the approach of [60]. The model was retrained on the full data set (except the lock box) with a training/validation-ratio of 85:15 (see Supplementary Data A table 1: GenR final experiments). The lock box contained the 150 DXA scans of which a re-assessment was performed and was complemented with randomly selected other scans to gain a size of $\pm 10\%$ of the original data set. After random selection, the bone age, gender and ethnicity distributions were checked to guarantee a similar distribution in the lock box as in the training data. Final performance was reported in MAD for easy interpretability and comparison with automated BAA methods based on radiographs:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

For the final performance measure, the label of the re-assessed scans consisted of the first reading only. Next to this, the performance of the automated method was checked on the re-assessed scans, taking the average of the two readings as label.

Furthermore, a comparison was made with the bone age predictions on the lock box made with the BoneXpert software, which is designed for BAA based on radiographs [34].

### *E. Regression Activation Mapping*

To gain more insight in the predictions of the network, a modified version of Class Activation Mapping (CAM) [61] was used. The mapping makes use of the Global Average Pooling (GAP) layer in the network. Each neuron in this layer represents the spatial average of the feature maps from the convolutional layer prior to the GAP layer. Because the network only consists of fully-connected (dense) layers afterwards, the values of the GAP layer reflect the contribution of a feature map to the final prediction (if weighted with the weights from the fully-connected layers in between). More specifically, a feature map $g_k(i,j)$ from the last convolutional layer is mapped into a scalar $t_k$ by $t_k = \sum_{i,j} g_k(i,j)$ in the GAP layer (with $(i,j)$ being the spatial coordinates of a specific feature map $k$). The final prediction from the output layer $\hat{y}$ is the weighted sum of the output of the GAP layer using the weights from the fully-connected layers in between ($w_1$, $w_2$) and related to the Regression Activation Mapping (RAM) $G(i,j)$ as follows:

$$\hat{y} = \sum_{k=1}^{K}(w_{1,k} \bullet w_2 \sum_{i,j} g_k(i,j)) = \sum_{i,j} G(i,j)$$

The RAM created in this way is in essence the weighted sum of the feature maps from the last convolutional layer. The RAM was overlaid on the original DXA scan using a threshold to show the highest RAM values only.

The heat maps generated by RAM were compared with important regions known from literature.

A similar approach was conducted to find the relative contribution of the image, gender, and ethnic part of the network for the prediction. Instead of the GAP layer, the outputs of the concatenation layer were weighted with the weights of the fully-connected layers. The resulting vector consisted of an image part (indices 0-256), a gender part (indices 256-272), and an ethnic part (indices 272-288). The sum of each separate part divided by the sum of the entire vector was taken as the relative contribution.

### *F. Usability on new data set*

Because the age spread of the third cohort visit of the Generation R Study is not very large ($9.78 \pm 0.33$ years) [28], using this model to assess the bone age of children with a different age range might not be reliable. Therefore, additional experiments were performed to find out how many annotated DXA scans are needed for reasonable performance. This was done by taking 1%, 5%, 10%, 20%, and 50% of the GenR data for training. Performance (in MAD) was measured on the lock box.

## IV. RESULTS

The base network architecture as depicted in figure 3 had nearly 1.2M trainable parameters.

### A. Pre-training

Implementation of the network was done using Python 3.7.2, TensorFlow 1.14.0 and Keras 2.2.4 and experiments were conducted on a computer installed with an NVIDIA GeForce RTX 2080 GPU. The optimizer used to minimize the MSE loss was ADAM, with an initial learning rate of 0.001 and the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-08$, decay $= 0.0001$. Mini-batch size was set to 16. Dropout was applied with a rate of 0.1 between the final dense layer and the prediction layer. Training lasted for 150 epochs.

The final MAD on the RSNA test set was 7.7 ($\pm$6.2) months. From the predictions on the test set, one could find out the network did learn a certain pattern without any clear bias (see Supplementary Data B figure 1).

### B. Transfer learning

A flowchart of the inclusion of DXA data from the Generation R data set is provided in figure 4. The final hyperparameter optimization set contained 4909 scans and the lock box contained 544 scans. This hyperparameter optimization set was further divided into training, validation, and test set (see Supplementary Data A table 1: GenR preliminary experiments).

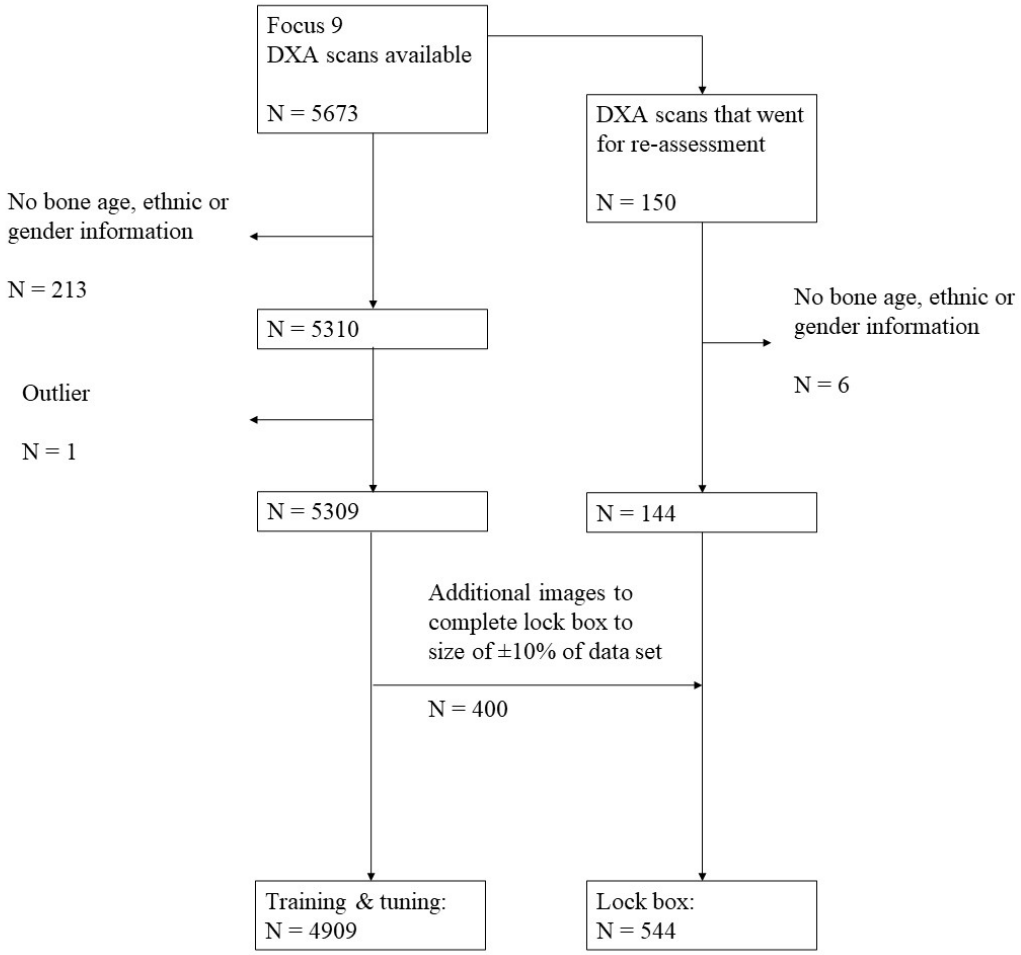


Fig. 4: Flowchart of available and used GenR [28] DXA data for training & tuning and lock box.

Similar settings as in pre-training were adopted for transfer learning, but initial learning rate and decay were reduced by a factor of ten, which is common practice in transfer learning to prevent overfitting. Training lasted for 50 epochs for determining the optimal number of blocks to freeze.

The results of the transfer learning are shown in table 2 (i.e., all models in this table make use of pre-trained weight initialization). Next to this, using randomly initialized weights resulted in an MAD of 7.0 ($\pm$5.4) months. This is not significantly different from the performance of the model using pre-trained weights with an MAD of 6.8 ($\pm$5.2) (p = 0.25). However, the visualization of the predictions shows a clear bias in the randomly initialized model compared to the pre-trained model (see Supplementary Data C figure 2). It can be hypothesized that random weight initialization takes longer to converge, but repeating the experiments for as long as 200 epochs did not show an improvement of the performance with an MAD of 7.1 ($\pm$5.6) and a similar bias as in Supplementary Data C figure 2.

TABLE 2: Transferred prediction results

| Frozen blocks | MAD (std) |
|---|---|
| None | 6.8 (5.2) |
| 1 | 7.1 (5.5) |
| 1,2 | 7.1 (5.4) |
| 1,2,3 | 7.8 (6.1) |
| 1,2,3,4 | 12.9 (9.9) |
| 2,3,4 | 7.2 (5.5) |
| 3,4 | 6.8 (5.3) |
| 4 | 6.8 (5.1) |
| Entire model* | 105.2 (75.5) |

MAD = Mean Absolute Difference, in months, std = standard deviation.
*Freezing the entire model means no retraining is done, i.e., the model for radiographs is used on DXA scans.

Using the network trained on radiographs without retraining on the DXA data set showed a poor performance with an MAD of 105.2 ($\pm$75.5) months, due to overestimation (see Supplementary Data C figure 3).
The other models, except for two, showed no significant difference in MAD value, with Bonferroni [62] correction applied for multiple testing: $p > \frac{0.05}{36} = 0.0014$. Only the models with frozen blocks 1, 2, and 3 and blocks 1, 2, 3, and 4 performed significantly worse (see Supplementary Data C figure 4). The choice with which model to continue was therefore also based on analysis of the predictions and learning curves. Analysis of the predictions comprised looking at the variance and if a bias was present. Low variance was preferred, but not too low as this can indicate overfitting. Logically, no appearance of bias was also preferred. A learning curve was considered to be more stable if the training loss was gradually going down and the validation loss did not differ too much between different epochs. With all these factors combined, the model where the weights of block 4 were frozen showed the best and most stable results and was therefore the model used for the final experiments.

### *C. Final outcomes*

From the learning curves of the transfer learning experiments, it was observed that overfitting started to occur at around 40 epochs (see Supplementary Data C figure 5). To limit the overfitting, the dropout rate was increased to 0.2 for the final experiments. Training lasted for 200 epochs. The final outcomes can be found in table 3.

From these results, we can observe that two models have equal best performance with an MAD of 6.9 months being the model including gender information and the model including both gender and ethnic information. However, these performances are not significantly better than the baseline model when tested against a Bonferroni [62] corrected p-value of $\frac{0.05}{6} = 0.0083$ (see Supplementary Data D figure 6).

The overall performances are worse than those of the transfer learning experiments which can be attributed to the fact that these results are the average of 10-fold cross-validation. This does, however, make these results more reliable and generalizable.

TABLE 3: Final prediction results of four different models

| Additional information in network architecture | MAD (std) |
|---|---|
| None | 7.1 (5.6) |
| Gender | 6.9 (5.4) |
| Ethnicity | 7.2 (5.6) |
| Gender & Ethnicity | 6.9 (5.4) |

MAD = Mean Absolute Difference, in months, std = standard deviation. Both are the averages of 10-fold cross-validation.

The mean errors of the different groups (i.e., males versus females and the different ethnic groups) got closer to each other when comparing the baseline model with the model including gender and ethnic information, indicating a reduction in bias originating from these factors (see Supplementary Data E table 2).

For the final predictions on the lock box, the architecture including both gender and ethnic information was used, resulting in an MAD of 6.8 (±5.4) months (Fig. 5, left) with an average assessment time of 32 ms per scan. From the figure, it might be observed that there is a small trend in overestimating bone age in the low bone age region and underestimating in the high bone age region. This could be attributed to the little amount of training data available in those regions (Fig. 2, right).

The re-assessed scans had an intra-class correlation of 89% [63]. Using the automated model on these scans, with the average of the two readings used as label, resulted in an MAD of 6.2 (±4.6) months (Fig. 5, right).

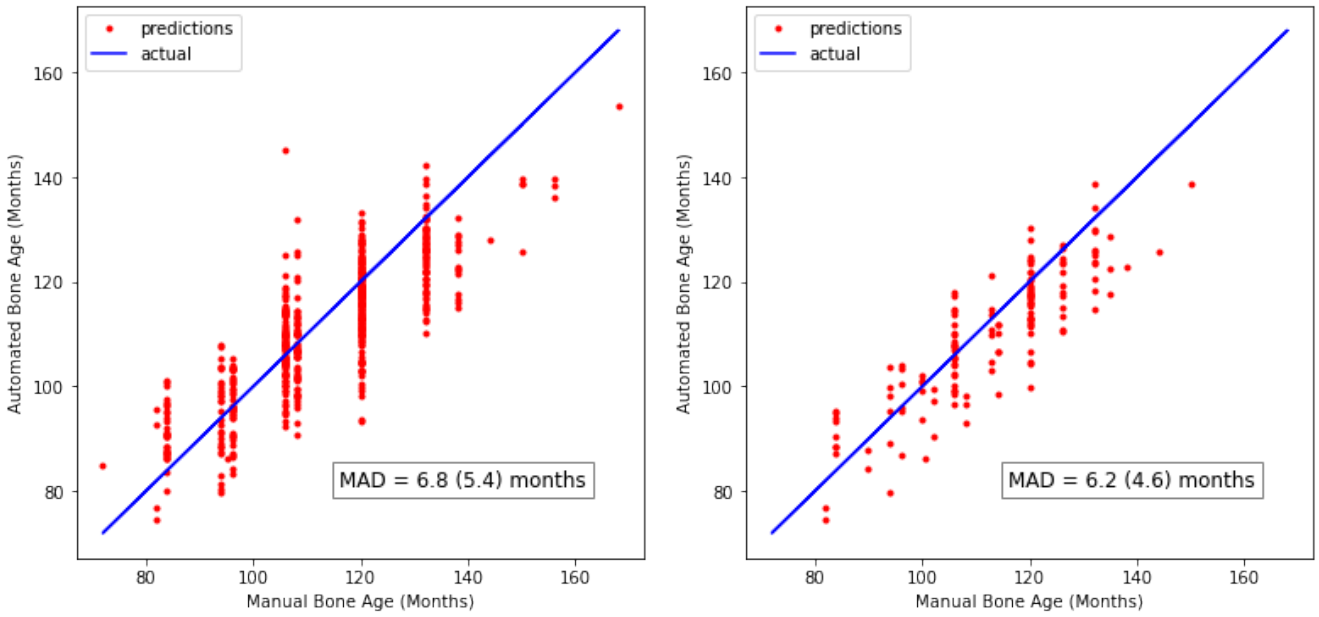


Fig. 5: Visualization of predictions of the final model on the lock box (left) and on the 144 re-assessed scans (right) where the average of the two readings is used as label.
MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

The BoneXpert predictions on the lock box resulted in a rejection rate of 27% (149/544) and an MAD of 15.0 (±10.3) months on the remaining scans (i.e., excluding rejected scans) (see Supplementary Data F figure 7). The predictions of the proposed DL model on the rejected scans showed an MAD of 7.3 (±5.8) months (see Supplementary Data F figure 8).

*D. Regression Activation Mapping*

For the predictions on the lock box, the relative contribution of the gender part varied between 0.02% and 4% and that of the ethnic part between 0.02% and 0.4%.

An example of a heat map showing the most important region for BAA using the automated model is shown in figure 6.

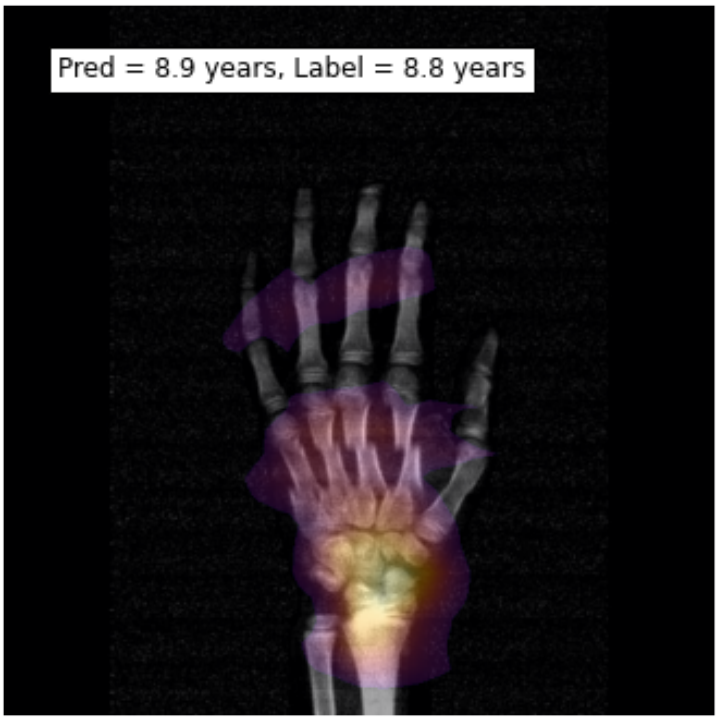


Fig. 6: DXA scan overlaid by a heat map calculated by Regression Activation Mapping (RAM) showing the most important regions of the hand for the BAA using the automated method. It also shows the prediction (Pred) from the network, together with the manually determined label.

A random selection of heat maps sorted by bone age group is shown in figure 7. Overall, one can observe that the heat maps point to specific bone regions of the hands, indicating that the network did learn to pick up specific features from the scans.

More specifically, it can be observed that for this network, the carpals seem to be the most determinant for the BAA. Figure 8 shows an example of the difference in bone maturation in the carpals between two participants with different bone ages. The secondary regions of interest are the metacarpophalangeal joints. From figure 7, it can be observed that with increasing bone age, the area indicating the most importance for the assessment increases. In the youngest group, the activation is mostly centered at the carpals, whereas in the oldest group, the activation is more spread out over the entire hand.

*E. Usability on new data set*

The results where only fractions of the data were used for training are collated in table 4. From these results and analysis of the predictions and learning curves, it was observed that from a fraction of 0.2 from the training data, or 982 in absolute number of scans, the results became stable and reasonable with an MAD of 8.0 ($\pm$6.4) months.

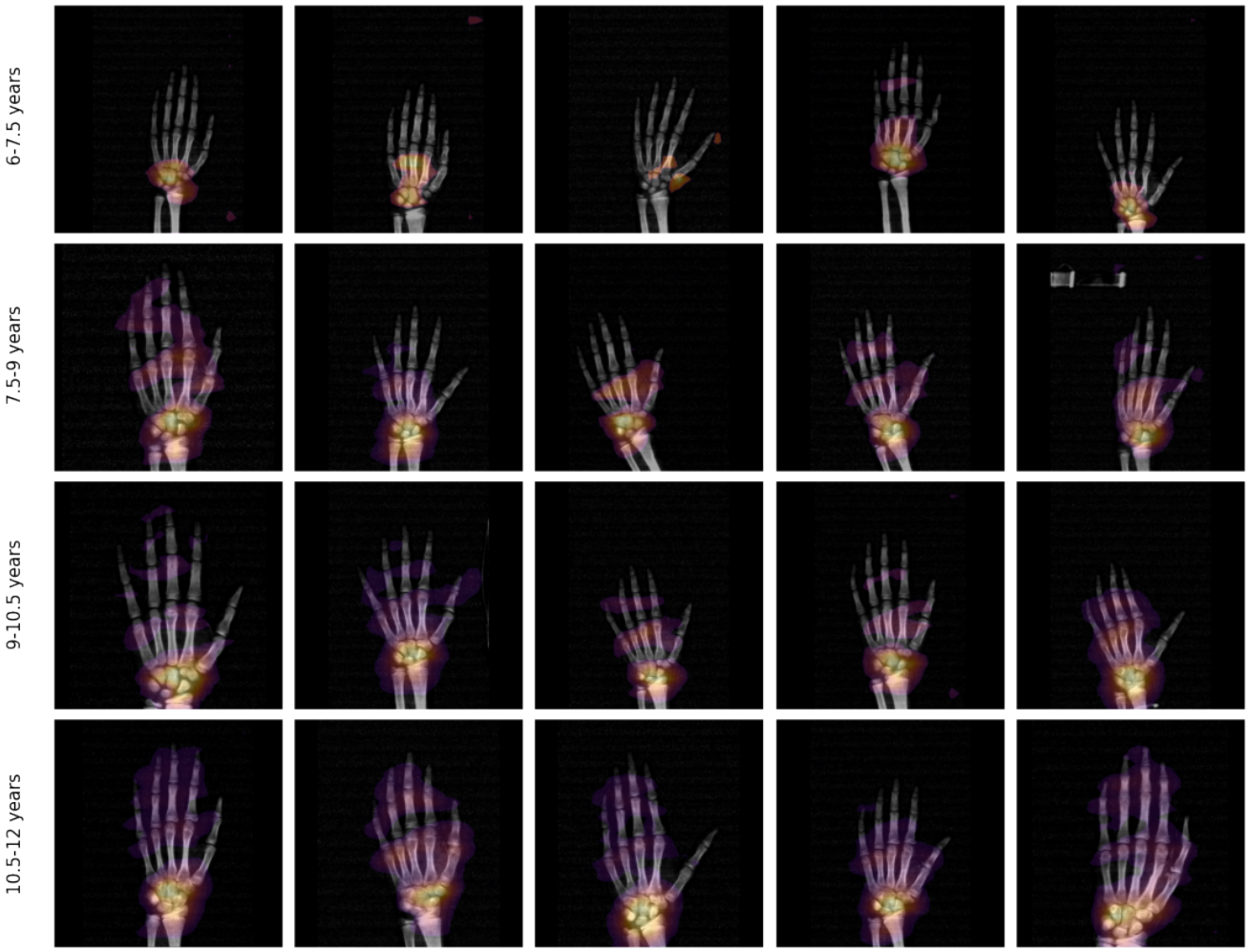


Fig. 7: Heat maps of scans representing different groups of bone age. The automated BAA has its primary focus on the carpals. The older the bone age group, the larger the area important for assessment becomes.

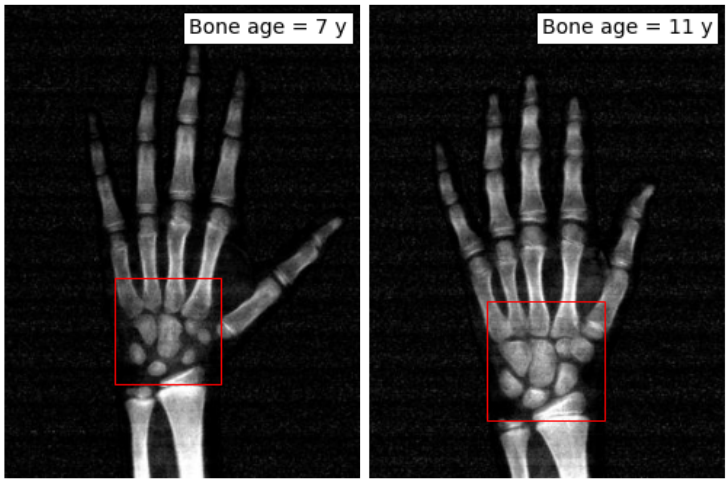


Fig. 8: DXA scans of two participants with different bone ages. The difference in bone maturation of the carpals (red box) is clearly visible and is the most important region for BAA by the automated method. y = years.

TABLE 4: Prediction results using a fraction of the training data

| Training data fraction | Absolute # of scans | MAD (std) |
|---|---|---|
| 0.01 | 49 | 14.8 (11.5) |
| 0.05 | 246 | 11.5 (8.9) |
| 0.1 | 491 | 8.4 (9.6) |
| 0.2 | 982 | 8.0 (6.4) |
| 0.5 | 2455 | 7.1 (5.4) |

MAD = Mean Absolute Difference, in months, std = standard deviation.

## V. DISCUSSION

This study describes the first method for automated BAA based on DXA scans. The chosen approach made use of transfer learning from a large data set of radiographs to another data set containing DXA scans labelled with bone age, gender, and ethnicity. The final network architecture included gender and ethnicity as extra input next to the DXA scans and resulted in an MAD of 6.8 months on the lock box. The comparison with software developed for BAA using radiographs showed the need for a method specifically created for the use on DXA scans. The heat maps showed that the carpal region was the most important region for the BAA by the proposed DL method.

The final result of an MAD of 6.8 months on the lock box is comparable with the performances of state-of-the-art automated methods and inter-observer variability on radiographs (Table 1). The assessment is performed in the order of milliseconds versus minutes when performing manual BAA [5]. The automated method outputs a continuous value rather than a categorical one, as is the case with the most commonly used manual method [3]. This makes it easier to accurately approximate the 'real bone age' and perform analyses because smaller differences can be picked up. The fact that the final performance is measured on a lock box which was kept aside during the entire project, enables to demonstrate the generalizability of the method. The improvement in the performance when the average of two readings is used as ground-truth label further signifies the generalization properties of the network (Fig. 5, right). Furthermore, the graphs of the predictions give more insight into the performance than only reporting single MAD values, which is often the only outcome in studies on automated BAA.

As regards the effects of including gender and ethnic information as a separate input in the network, the results are ambiguous. The relative contribution of the gender and especially the ethnic part of the network for prediction was low compared to the image part. Logically, the image part needs to play the most important role in the prediction, but increasing the contribution of the gender and ethnic information might lead to a larger reduction in bias originating from the different groups. This could be achieved by enlarging the fully-connected layers the additional information is passed through, before these are concatenated to the output of the GAP layer (i.e., the image information). However, it could also be that the addition of extra information as gender and ethnicity just has a limited effect on the performance of such a model. Future research could go deeper into explaining the difference in maturation between sexes and ethnicities into including this as separate input in an automated model.

The heat maps visualising the most activated areas for the BAA provide more insight in the predictions of the automated method. It was observed that the most important region seemed to be the carpal region (Fig. 6 & 7).

From literature on bone development and manual BAA, it is known that in the early bone ages the most weight is put on the distal interphalangeal joints, followed by the more proximal interphalangeal and the metacarpophalangeal joints and finally the carpals (see Supplementary Data G) [22]. These regions are indeed also used by our model, but in a different order of importance (Fig. 7).

In the later bone age stages, the distal and proximal interphalangeal and the metacarpophalangeal joints remain the most important regions followed by two very specific bones in the thumb and carpal area (see Supplementary Data G) [22]. Although the more distal parts of the hands seem to play a larger role in the later bone ages in

the automated method as well, the carpals remain the most important area for prediction (Fig. 7).

The commercially available BoneXpert software [34] does not use the carpal region for prediction. This finding could therefore also improve current BAA methods on radiographs.

Furthermore, we have shown the need for DXA specific software, because radiograph-based methods are not applicable on DXA scans.

One possible reason for rejection by the BoneXpert software [34] could be the apparent disconnection in the metacarpals visible on the image of figure 6. This type of disconnection was often observed in the DXA data set. Figure 6 shows that such an artifact hardly influences the heat map of the prediction by the proposed DL model. We hypothesize that registration of this artifact would therefore not seriously improve the performance and would only make the model less generalizable to a different DXA data set. Another cause of rejection by the BoneXpert software could be the lower resolution of DXA scans compared with radiographs [41] (Fig. 1). Because the proposed model was trained on the DXA scans, it implicitly learned to cope with the DXA scan resolution.

Additionally, we have shown that using the proposed DL network pre-trained on radiographs without retraining on the DXA data set resulted in a poor performance (Table 2 and Supplementary Data C figure 3). Consequently, an algorithm developed for the use on radiographs cannot be used on DXA scans directly without retraining.

From the transfer learning experiments, it also became clear that using pre-trained weight initialization over random weight initialization showed a better and more stable performance. This indicates the success of transfer learning for such an application.

Furthermore, when adopting transfer learning, freezing the last block showed the best and most stable results. We hypothesize that this is due to the deeper layers of the network being responsible for the high-level features. When using transfer learning from a radiograph data set for the purpose of BAA, these features are expected to point at the most determinant areas of the image for this specific task. In most transfer learning applications, the reverse can be observed. These are mostly based on large data sets containing natural images of simple objects (e.g., ImageNet [64]). In those situations, transfer learning benefits from the low-level features in the images, such as borders of objects. Therefore, often freezing the weights of the first layers rather than the last improves performance [59]. This might also be the reason why Tang et al. [53] only experimented with freezing the first few blocks, while freezing the last block(s) could have resulted in a better performance.

When using only part of the training data, stability of the performance started to occur at somewhat less than 1000 annotated scans (Table 4). Accordingly, in case of the application on a different bone age range than is used now (i.e., outside 7 to 13 years) the following approach can be used:

1) Manual labelling of 1000 scans;
2) Dividing this set into training, validation and test set;
3) Using the network with pre-trained weights to train on the training set and use the validation set to avoid overfitting;
4) Testing the performance on the test set;
5) In case of acceptable performance, using the automated model to assess the bone ages of the remaining scans.

We hypothesize that acceptable performance may be achieved even with less labelled data than $\pm$1000 scans in this case because the pre-training has now been performed on DXA scans (albeit from a different age range) rather than radiographs. Another approach could be to make use of a method called pseudo-labelling in semi-supervised learning [65]. However, some type of confidence measure is required for this approach to work.

*Limitations*

One of the main limitations of this study is the relatively small spread in bone age of the used DXA data set (Fig. 2, right). Using the method on a different bone age range might therefore not be reliable. To this end,

supported by the results in table 4, an approach was suggested where only a limited number of scans needs to be assessed manually and the remaining scans could be assessed using the automated method.

Another limitation is the bone age being labelled once by one observer. Despite the reported ICC of 89% on the re-assessed scans, a more powerful ground-truth label would be the average of multiple readings from different observers. However, the good results of figure 5 (right) show the generalization properties of the automated method, which can indicate issues related to inter-observer variability play a smaller role. At the same time, the manual approach is not a true gold standard, which is a known limitation in research on BAA.

Finally, the ethnicity variable is now classified based on visual interpretation of the DXA operator into 'White', 'Black', 'Asian' or 'Hispanic'. This classification resembles clinical practice and is a common and easy adoptable classification, making the method universally usable. However, a more reliable classification of ethnicity would be to look at the country of birth of the child and that of his or her parents. This could positively influence the performance of the model but might reduce the universal usability.

*Future work*

To overcome the limitation of not having a true gold standard for bone age, one approach could be to study the accuracy of the method to predict adult height, a common application of bone age [66]. Besides this, future work can consist of extending the training data with DXA scans of different age ranges to make the model reliable for the age spread of the entire childhood (i.e., 2 to 18 years). Finally, research can be conducted into integrating the model as part of a larger pipeline, combining multiple factors in the automated risk prediction of certain pediatric disorders, genetic analysis or adult height for which bone age is currently used in a manual way.

## VI. Conclusion

In this study, an automated BAA method was created for application on DXA scans rather than radiographs, which has not been done before. Yielding an MAD of 6.8 months on a totally independent test set (lock box), the performance is comparable with the inter-observer variability of manual BAA and the performances reported for state-of-the-art automated BAA methods on radiographs. To limit the radiation exposure in the examined children and improve time-effectiveness of the BAA procedure, the automated method can be used in clinical practice for an expected bone age range of 7 to 13 years. Inclusion of data from other bone ages can extend the use of the model in the future and can accelerate research on the topic of bone and growth development.

## References


[1] D. D. Martin, J. M. Wit, Z. Hochberg, L. Sävendahl, R. R. Van Rijn, O. Fricke, N. Cameron, J. Caliebe, T. Hertel, D. Kiepe, *et al.*, "The use of bone age in clinical practice–part 1," *Hormone research in paediatrics*, vol. 76, no. 1, pp. 1–9, 2011.

[2] A. M. Mughal, N. Hassan, and A. Ahmed, "Bone age assessment methods: A critical review," *Pakistan journal of medical sciences*, vol. 30, no. 1, p. 211, 2014.

[3] W. W. Greulich and S. I. Pyle, *Radiographic atlas of skeletal development of the hand and wrist*. Stanford university press, 1959.

[4] J. M. Tanner, R. Whitehouse, N. Cameron, W. Marshall, M. Healy, H. Goldstein, *et al.*, *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. London: Academic Press, 1983.

[5] D. King, D. Steventon, M. O'sullivan, A. Cook, V. Hornsby, I. Jefferson, and P. King, "Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods," *The British journal of radiology*, vol. 67, no. 801, pp. 848–851, 1994.

[6] S. Y. Kim, Y. J. Oh, J. Y. Shin, Y. J. Rhie, and K. H. Lee, "Comparison of the Greulich-Pyle and Tanner Whitehouse (TW3) methods in bone age assessment.," *Journal of Korean Society of Pediatric Endocrinology*, vol. 13, no. 1, pp. 50–55, 2008.

[7] R. Bull, P. Edwards, P. Kemp, S. Fry, and I. Hughes, "Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods," *Archives of disease in childhood*, vol. 81, no. 2, pp. 172–173, 1999.

[8] H. H. Thodberg and L. Sävendahl, "Validation and reference values of automated bone age determination for four ethnicities," *Academic radiology*, vol. 17, no. 11, pp. 1425–1432, 2010.

[9] G. F. Johnson, J. P. Dorst, J. P. Kuhn, A. F. Roche, and G. H. Dávila, "Reliability of skeletal age assessments," *American Journal of Roentgenology*, vol. 118, no. 2, pp. 320–327, 1973.

[10] A. F. Roche, C. G. Rohmann, N. Y. French, and G. H. Dévila, "Effect of training on replicability of assessments of skeletal maturity (Greulich-Pyle)," *American Journal of Roentgenology*, vol. 108, no. 3, pp. 511–515, 1970.

[11] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2018.

[12] N. Gorelik, J. Chong, and D. J. Lin, "Pattern recognition in musculoskeletal imaging using artificial intelligence," in *Seminars in musculoskeletal radiology*, vol. 24, pp. 38–49, Thieme Medical Publishers, 2020.

[13] A. L. Dallora, P. Anderberg, O. Kvist, E. Mendes, S. Diaz Ruiz, and J. Sanmartin Berglund, "Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis," *PloS one*, vol. 14, no. 7, p. e0220242, 2019.

[14] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[15] C. F. Njeh, T. Fuerst, D. Hans, G. M. Blake, and H. K. Genant, "Radiation exposure in bone mineral density assessment," *Applied Radiation and Isotopes*, vol. 50, no. 1, pp. 215–236, 1999.

[16] E. J. Hall, "Radiation biology for pediatric radiologists," *Pediatric radiology*, vol. 39, no. 1, p. 57, 2009.

[17] The International Commission on Radiological Protection, "ICRP publication 103," *Ann ICRP*, vol. 37, no. 2-4, pp. 1–332, 2007.

[18] D. Heppe, H. Taal, G. Ernst, E. Van Den Akker, M. Lequin, A. Hokken-Koelega, J. Geelhoed, and V. Jaddoe, "Bone age assessment by dual-energy X-ray absorptiometry in children: an alternative for X-ray?," *The British journal of radiology*, vol. 85, no. 1010, pp. 114–120, 2012.

[19] P. Płudowski, M. Lebiedowski, and R. S. Lorenc, "Evaluation of the possibility to assess bone age on the basis of DXA derived hand scans—preliminary results," *Osteoporosis international*, vol. 15, no. 4, pp. 317–322, 2004.

[20] H. Hoyer-Kuhn, K. Knoop, O. Semler, K. Kuhr, M. Hellmich, E. Schoenau, and F. Koerber, "Comparison of DXA scans and conventional X-rays for spine morphometry and bone age determination in children," *Journal of Clinical Densitometry*, vol. 19, no. 2, pp. 208–215, 2016.

[21] P. Braillon, A. Guibal, P. Pracros-Deffrenne, A. Serban, J. Pracros, and P. Chatelain, "Dual energy x-ray absorptiometry of the hand and wrist–a possible technique to assess skeletal maturation: methodology and data in normal youths," *Acta Paediatrica*, vol. 87, no. 9, pp. 924–929, 1998.

[22] V. Gilsanz and O. Ratib, *Hand bone age: a digital atlas of skeletal maturity*. Springer Science & Business Media, 2005.

[23] F. Ontell, M. Ivanovic, D. S. Ablin, and T. Barlow, "Bone age in children of diverse ethnicity.," *AJR. American journal of roentgenology*, vol. 167, no. 6, pp. 1395–1398, 1996.

[24] A. Zhang, J. W. Sayre, L. Vachon, B. J. Liu, and H. Huang, "Racial differences in growth patterns of children assessed on the basis of bone age," *Radiology*, vol. 250, no. 1, pp. 228–235, 2009.

[25] A. L. Creo and W. F. Schwenk, "Bone age: a handy tool for pediatric providers," *Pediatrics*, vol. 140, no. 6, 2017.

[26] R. T. Loder, D. T. Estle, K. Morrison, D. Eggleston, D. N. Fish, M. L. Greenfield, and K. E. Guire, "Applicability of the greulich and pyle skeletal age standards to black and white children of today," *American Journal of Diseases of Children*, vol. 147, no. 12, pp. 1329–1333, 1993.

[27] B.-D. Lee and M. S. Lee, "Automated bone age assessment using artificial intelligence: The future of bone age assessment," *Korean Journal of Radiology*, vol. 22, no. 5, p. 792, 2021.

[28] M. N. Kooijman, C. J. Kruithof, C. M. van Duijn, L. Duijts, O. H. Franco, M. H. van IJzendoorn, J. C. de Jongste, C. C. Klaver, A. van der Lugt, J. P. Mackenbach, *et al.*, "The Generation R study: design and cohort update 2017," *European Journal of Epidemiology*, vol. 31, no. 12, pp. 1243–1264, 2016.

[29] D. J. Michael and A. C. Nelson, "HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs," *IEEE transactions on medical imaging*, vol. 8, no. 1, pp. 64–69, 1989.

[30] J. Tanner and R. Gibbons, "Automatic bone age measurement using computerized image analysis," *Journal of Pediatric Endocrinology and Metabolism*, vol. 7, no. 2, pp. 141–146, 1994.

[31] M. Mansourvar, M. A. Ismail, T. Herawan, R. Gopal Raj, S. Abdul Kareem, and F. H. Nasaruddin, "Automated bone age assessment: motivation, taxonomies, and challenges," *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[32] S. Mahmoodi, B. S. Sharif, E. G. Chester, J. P. Owen, and R. Lee, "Skeletal growth estimation using radiographic image processing and analysis," *IEEE Transactions on Information Technology in Biomedicine*, vol. 4, no. 4, pp. 292–297, 2000.

[33] J. Seok, J. Kasa-Vubu, M. DiPietro, and A. Girard, "Expert system for automated bone age determination," *Expert Systems with Applications*, vol. 50, pp. 75–88, 2016.

[34] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," *IEEE transactions on medical imaging*, vol. 28, no. 1, pp. 52–66, 2008.

[35] D. Visiana, "BoneXpert for hospitals." `https://bonexpert.com/bonexpert-for-hospitals/`, 2019.

[36] D. D. Martin, D. Deusch, R. Schweizer, G. Binder, H. H. Thodberg, and M. B. Ranke, "Clinical application of automated Greulich-Pyle bone age determination in children with short stature," *Pediatric radiology*, vol. 39, no. 6, pp. 598–607, 2009.

[37] D. D. Martin, K. Meister, R. Schweizer, M. B. Ranke, H. H. Thodberg, and G. Binder, "Validation of automatic bone age rating in children with precocious and early puberty," *Journal of Pediatric Endocrinology and Metabolism*, vol. 24, no. 11-12, pp. 1009–1014, 2011.

[38] D. D. Martin, K. Heil, C. Heckmann, A. Zierl, J. Schaefer, M. B. Ranke, and G. Binder, "Validation of automatic bone age determination in children with congenital adrenal hyperplasia," *Pediatric radiology*, vol. 43, no. 12, pp. 1615–1621, 2013.

[39] H. Lee, S. Tajmir, J. Lee, M. Zissen, B. A. Yeshiwas, T. K. Alkasab, G. Choy, and S. Do, "Fully automated deep learning system for bone age assessment," *Journal of digital imaging*, vol. 30, no. 4, pp. 427–441, 2017.

[40] J. Seok, B. Hyun, J. Kasa-Vubu, and A. Girard, "Automated classification system for bone age X-ray images," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 208–213, IEEE, 2012.

[41] C. Booz, I. Yel, J. L. Wichmann, S. Boettger, A. Al Kamali, M. H. Albrecht, S. S. Martin, L. Lenga, N. A. Huizinga, T. D'Angelo, *et al.*, "Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method," *European radiology experimental*, vol. 4, no. 1, pp. 1–8, 2020.

[42] J. Zhang, F. Lin, and X. Ding, "Maturation disparity between hand-wrist bones in a Chinese sample of normal children: an analysis based on automatic BoneXpert and manual Greulich and Pyle atlas assessment," *Korean journal of radiology*, vol. 17, no. 3, pp. 435–442, 2016.

[43] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, *et al.*, "The RSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, 2019.
[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[45] M. Chen, "Automated bone age classification with deep neural networks," *Stanford University, USA, Technical Report*, 2016.
[46] C. Spampinato, S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi, "Deep learning for automated skeletal bone age assessment in X-ray images," *Medical image analysis*, vol. 36, pp. 41–51, 2017.
[47] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets, "Paediatric bone age assessment using deep convolutional neural networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 300–308, Springer, 2018.
[48] C. Liu, H. Xie, Y. Liu, Z. Zha, F. Lin, and Y. Zhang, "Extract bone parts without human prior: end-to-end convolutional neural network for pediatric bone age assessment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 667–675, Springer, 2019.
[49] X. Ren, T. Li, X. Yang, S. Wang, S. Ahmad, L. Xiang, S. R. Stone, L. Li, Y. Zhan, D. Shen, *et al.*, "Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph," *IEEE journal of biomedical and health informatics*, vol. 23, no. 5, pp. 2030–2038, 2018.
[50] X. Pan, Y. Zhao, H. Chen, D. Wei, C. Zhao, and Z. Wei, "Fully automated bone age assessment on large-scale hand X-ray dataset," *International Journal of Biomedical Imaging*, vol. 2020, 2020.
[51] F. Torres, C. González, M. C. Escobar, L. Daza, G. Triana, and P. Arbeláez, "An empirical study on global bone age assessment," in *15th International Symposium on Medical Information Processing and Analysis*, vol. 11330, p. 113300E, International Society for Optics and Photonics, 2020.
[52] D. Wang, K. Zhang, J. Ding, and L. Wang, "Improve bone age assessment by learning from anatomical local regions," *arXiv preprint arXiv:2005.13452*, 2020.
[53] W. Tang, G. Wu, and G. Shen, "Improved automatic radiographic bone age prediction with deep transfer learning," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, IEEE, 2019.
[54] J. Wang, Y. Sun, and T. Sanchez, "Assessing short term change in carpal bone age assessments by dual-energy X-ray absorptiometry," *Int J Radiol Imaging Technol*, vol. 4, p. 034, 2018.
[55] K. Yip and M. Husein, "Deep learning to predict hip fracture risk from clinical DXA-images," 2018.
[56] T. Nissinen, "Convolutional neural networks in osteoporotic fracture risk prediction using spine DXA images," *Computer Science*, 2019.
[57] D. Hussain and S.-M. Han, "Computer-aided osteoporosis detection from DXA imaging," *Computer methods and programs in biomedicine*, vol. 173, pp. 87–107, 2019.
[58] J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. de Bruijne, M. W. Vernooij, H. H. Adams, M. A. Ikram, W. J. Niessen, and G. V. Roshchupkin, "Gray matter age prediction as a biomarker for risk of dementia," *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 21213–21218, 2019.
[59] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
[60] M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble, "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data," *Neuroscience & Biobehavioral Reviews*, 2020.
[61] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization.," *CVPR*, 2016.
[62] C. E. Bonferroni, "Il calcolo delle assicurazioni su gruppi di teste," *Studi in onore del professore salvatore ortu carboni*, pp. 13–60, 1935.
[63] O. Grgic, E. Shevroja, B. Dhamo, A. G. Uitterlinden, E. B. Wolvius, F. Rivadeneira, and C. Medina-Gomez, "Skeletal maturation in relation to ethnic background in children of school age: The Generation R study," *Bone*, vol. 132, p. 115180, 2020.
[64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[65] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.
[66] H. H. Thodberg, J. Neuhof, M. B. Ranke, O. G. Jenni, and D. D. Martin, "Validation of bone age methods by their ability to predict adult height," *Hormone research in paediatrics*, vol. 74, no. 1, pp. 15–22, 2010.

# I. SUPPLEMENTARY DATA

## A. Data distribution

TABLE 1: Characteristics of used data sets

| | Mean bone age (std) | Male | Female | White | Black | Asian | Hispanic |
|---|---|---|---|---|---|---|---|
| RSNA [1] | | | | | | | |
| - Train | 10.6 (3.4) | 5462 | 4626 | - | - | - | - |
| - Val | 10.6 (3.4) | 1371 | 1152 | - | - | - | - |
| - Test | 11.0 (3.6) | 200 | 200 | - | - | - | - |
| GenR [2] preliminary experiments | | | | | | | |
| - Train | 9.4 (1.3) | 1869 | 1886 | 3154 | 355 | 160 | 86 |
| - Val | 9.4 (1.3) | 320 | 343 | 556 | 56 | 37 | 14 |
| - Test | 9.4 (1.3) | 236 | 255 | 402 | 58 | 28 | 9 |
| GenR [2] final experiments | | | | | | | |
| - Train | 9.4 (1.3) | 2065 | 2107 | 3503 | 190 | 389 | 90 |
| - Val | 9.3 (1.3) | 360 | 377 | 609 | 35 | 74 | 19 |
| - Lock box | 9.4 (1.2) | 254 | 290 | 426 | 49 | 25 | 8 |

Bone age is in years; std = standard deviation; GenR = Generation R.

## B. Final predictions on RSNA test data set

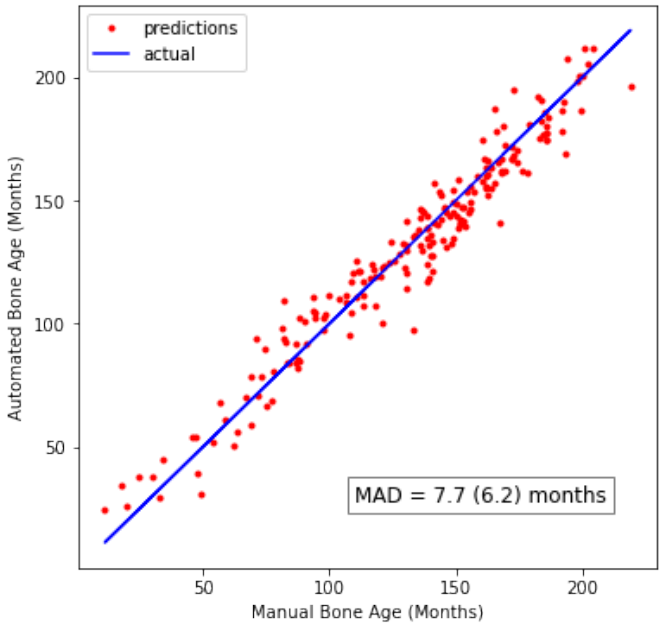


Fig. 1: Visualization of predictions on RSNA test data set showing a reasonable trend and no clear bias. MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

*C. Transfer learning experiments*

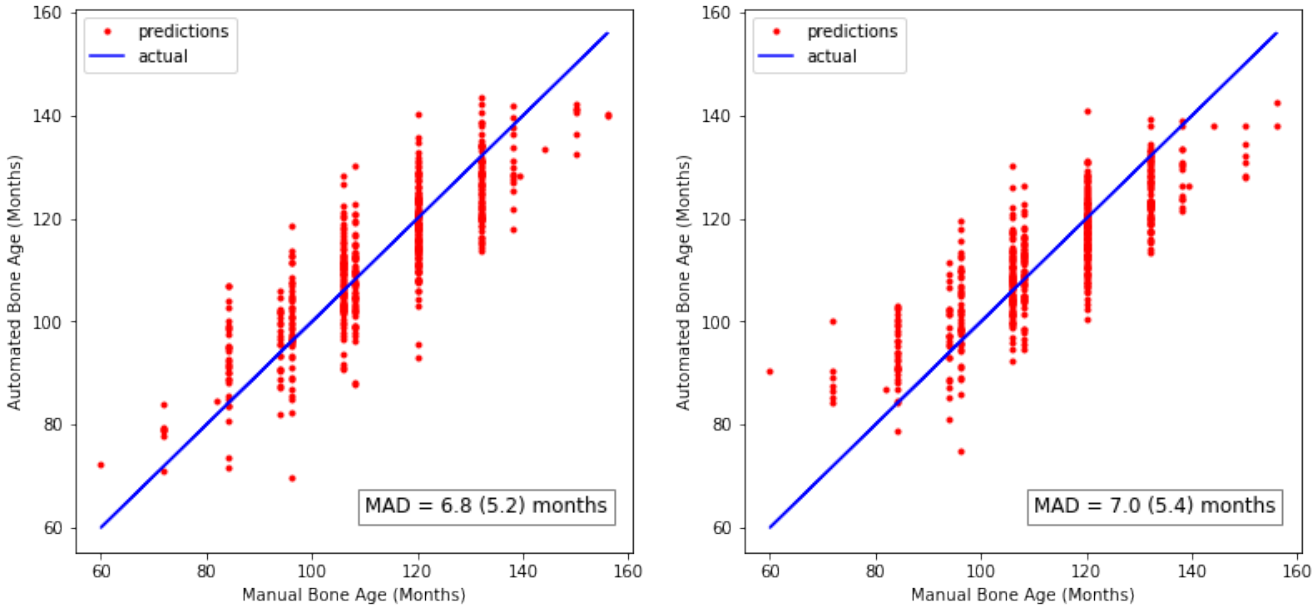


Fig. 2: Visualization of predictions of the network on the DXA test set with pre-trained weights (left) and randomly initialized weights (right) and no freezing. A bias can be observed in the model with randomly initialized weights. MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

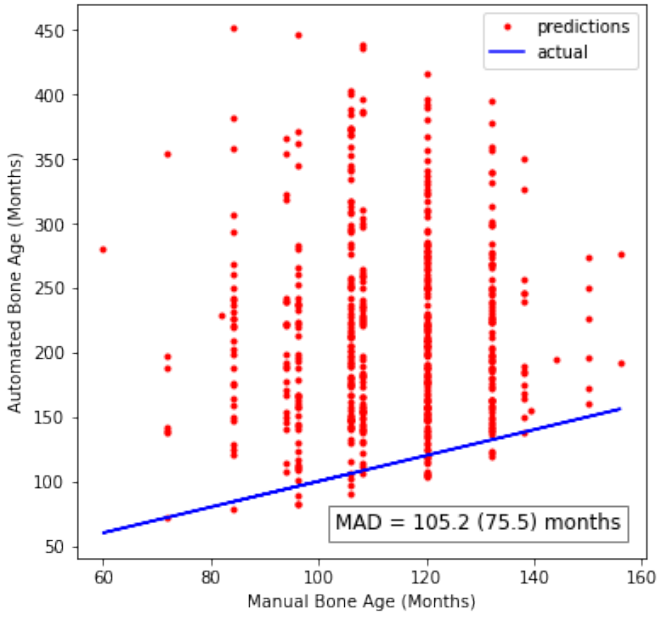


Fig. 3: Visualization of the predictions of the radiograph network on the DXA test set showing a poor performance due to overestimation, indicating a network trained on radiographs cannot be used without retraining on DXA scans.
MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

| | None | 1 | 1,2 | 1,2,3 | 1,2,3,4 | 2,3,4 | 3,4 | 4 | No re-training |
|---|---|---|---|---|---|---|---|---|---|
| None | | | | | | | | | |
| 1 | 0,110 | | | | | | | | |
| 1,2 | 0,156 | 0,865 | | | | | | | |
| 1,2,3 | **4,26E-05** | **0,001** | **3,03E-04** | | | | | | |
| 1,2,3,4 | **8,72E-35** | **3,16E-30** | **7,60E-32** | **1,66E-29** | | | | | |
| 2,3,4 | 0,111 | 0,767 | 0,689 | 0,008 | **6,95E-34** | | | | |
| 3,4 | 0,992 | 0,090 | 0,147 | **1,17E-05** | **1,95E-34** | 0,071 | | | |
| 4 | 0,908 | 0,019 | 0,080 | **2,14E-06** | **1,22E-33** | 0,081 | 0,893 | | |
| No re-training | **0,000** | **0,000** | **0,000** | **0,000** | **0,000** | **0,000** | **0,000** | **0,000** | |

Fig. 4: P-values of paired t-tests between performances of the same architecture with pre-trained weights, but a difference in frozen blocks in the transfer learning experiments. The numbers in the first row and first column indicate the frozen blocks in the model. Significance is tested against a Bonferroni [3] corrected p-value of 0.0014. Red cells with bold printed values represent a significant difference, whereas green cells with normal printed values represent no significant difference.

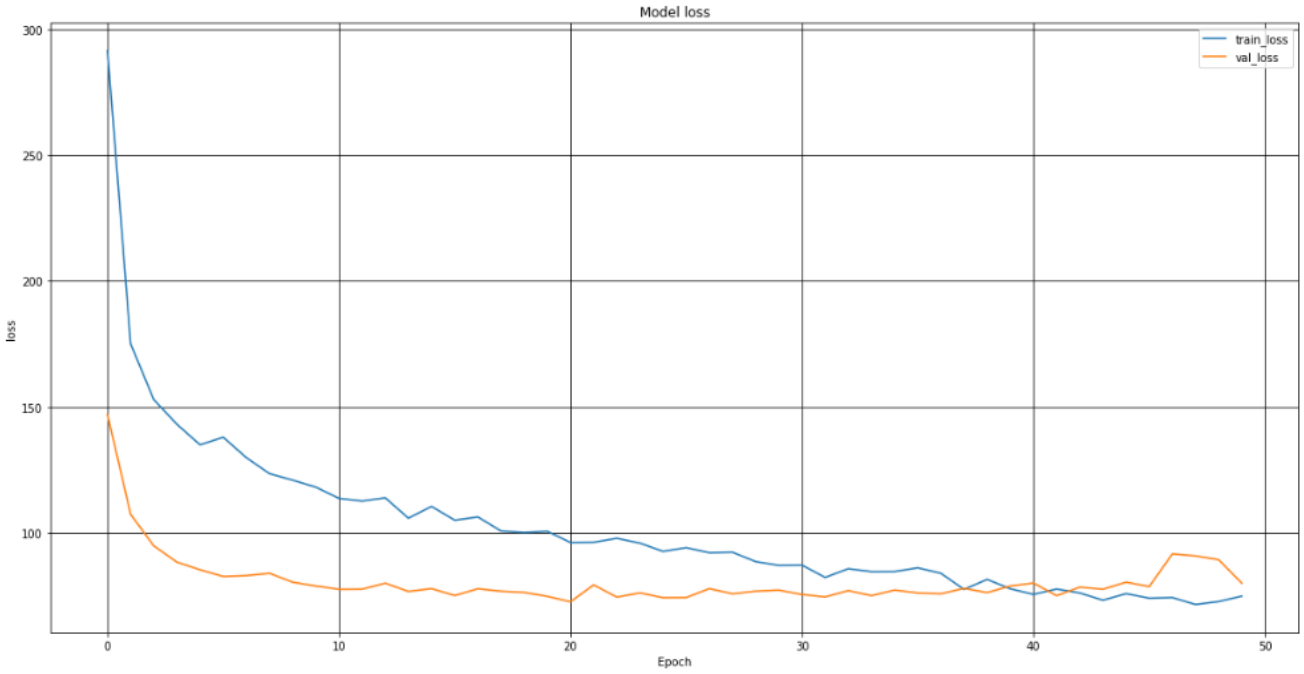


Fig. 5: Learning curve of the model when freezing block 4 during the transfer learning experiments.

*D. Final architecture experiments*

| | None | Gender | Ethnicity | Gender & Ethnicity |
|---|---|---|---|---|
| None | | | | |
| Gender | 0,028 | | | |
| Ethnicity | 0,266 | **1,10E-04** | | |
| Gender & Ethnicity | 0,027 | 0,815 | **1,47E-04** | |

Fig. 6: P-values of paired t-tests between performances of the same architecture, but a difference in additional input information next to the DXA scans. Significance is tested against a Bonferroni [3] corrected p-value of 0.0083. Red cells with bold printed values represent a significant difference, whereas green cells with normal printed values represent no significant difference.

*E. Bias reduction from gender and ethnicity*

TABLE 2: Mean deviation from zero error for different groups of two different models

| Group | **Base model** | | **Model with gender & ethnic information** | |
|---|---|---|---|---|
| | mean deviation* | T-stats (p-value) | mean deviation* | T-stats (p-value) |
| Males | -1.2 | -4.6 (0.000) | -0.8 | -3.0 (0.007) |
| Females | 0.6 | | 0.4 | |
| Whites | -0.3 | 1.0 (0.417) | -0.2 | 0.6 (0.621) |
| Blacks | 0.6 | | 0.2 | |
| Asians | -1.3 | | -0.8 | |
| Hispanics | 0.1 | | 0.1 | |

* from zero error averaged over 10 folds, in months, indicating bias based on groups.
T-stats (in months) are based on independent t-test between males and females and a one-way ANOVA between different ethnicities.

*F. Comparison with BoneXpert software*

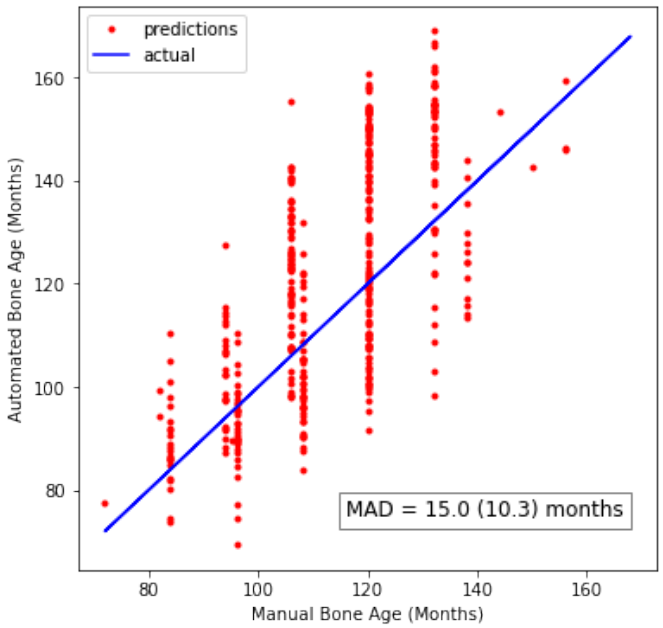


Fig. 7: Visualization of predictions of the BoneXpert software [4] on the lock box. The results are with exclusion of rejected scans (149/544).
MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

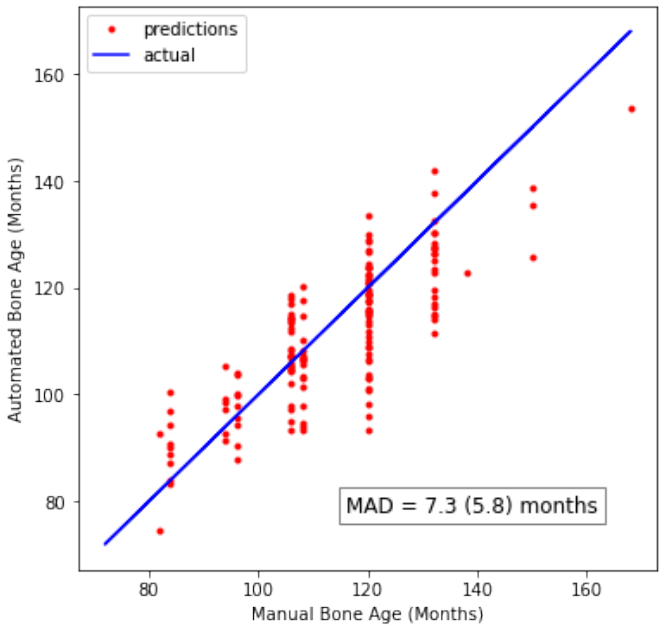


Fig. 8: Visualization of predictions of the proposed DL network on the 149 rejected scans by the BoneXpert software [4].
MAD = Mean Absolute Difference. Number in brackets represents the standard deviation.

*G. Important regions for BAA from literature*

In the pre-puberty stage (females: 2 years to 7 years of age, males: 3 years to 9 years of age) the BAA is primarily focused on the epiphyseal size of the phalanges. In this stage, the epiphyses increase in size and become as wide as the metaphyses (Fig. 9: left). The more distal the phalanges, the more weight is given in BAA. The development of the epiphysis of the ulna and all carpal bones (except for the pisiform) is also of interest in BAA for this stage, albeit to a lesser extent because of less reliability. [5]

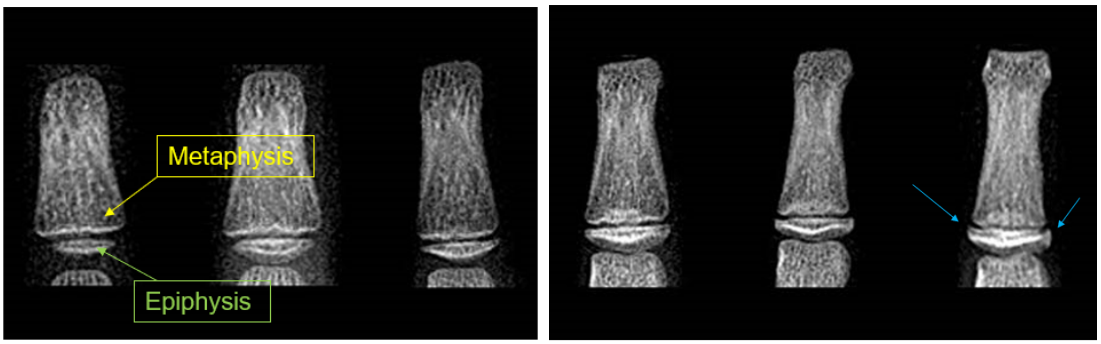


Fig. 9: Regions of interest in BAA from literature. Left: Primary focus of BAA in pre-puberty stage on progressive growth of the epiphyses until as wide as metaphyses. Right: Primary focus of BAA in early and mid-puberty stage on progressive growth of epiphyses until wider than metaphyses. Horn-like structures at both ends of the epiphysis are also of special interest, which are formed prior to epiphyseal fusion [5].

In the early and mid-puberty stage (females: 7 years to 13 years of age, males: 9 years to 14 years of age) the epiphyses grow even wider than the metaphyses and begin to cap the metaphyses by growing horny-like structures at both ends (Fig. 9: right). Furthermore, the pisiform and the ulnar sesamoid of the thumb become visible (Fig. 10). However, the latter indicators are less reliable. [5]


Fig. 10: Secondary focus of BAA in early and mid-puberty. The ulnar sesamoid of the thumb and the pisiform become visible during this stage. NB: radiograph is of 30-year old man and only depicted for the purpose of showing the location of specific bones [6].

## REFERENCES SUPPLEMENTARY DATA


[1] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, *et al.*, "The RSNA pediatric bone age machine learning challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, 2019.

[2] M. N. Kooijman, C. J. Kruithof, C. M. van Duijn, L. Duijts, O. H. Franco, M. H. van IJzendoorn, J. C. de Jongste, C. C. Klaver, A. van der Lugt, J. P. Mackenbach, *et al.*, "The Generation R study: design and cohort update 2017," *European Journal of Epidemiology*, vol. 31, no. 12, pp. 1243–1264, 2016.

[3] C. E. Bonferroni, "Il calcolo delle assicurazioni su gruppi di teste," *Studi in onore del professore salvatore ortu carboni*, pp. 13–60, 1935.

[4] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The BoneXpert method for automated determination of skeletal maturity," *IEEE transactions on medical imaging*, vol. 28, no. 1, pp. 52–66, 2008.

[5] V. Gilsanz and O. Ratib, *Hand bone age: a digital atlas of skeletal maturity*. Springer Science & Business Media, 2005.

[6] O. Kose, F. Guler, A. Turan, K. Canbora, and S. Akalin, "Prevalence and distribution of sesamoid bones of the hand. a radiographic study in Turkish subjects," *Int. J. Morphol*, vol. 30, no. 3, pp. 1094–9, 2012.