




ALEXANDER TACK¹, BERNHARD PREIM², STEFAN ZACHOW³

Fully automated Assessment of Knee Alignment from Full-Leg X-Rays employing a "YOLOv4 And Resnet Landmark regression Algorithm" (YARLA): Data from the Osteoarthritis Initiative⁴

¹  0000-0002-2418-7629, corresponding author

²  0000-0001-9826-9478

³  0000-0001-7964-3049

⁴to appear in: Computer Methods and Programs in Biomedicine,

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: <http://www.zib.de>

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Abstract

Background and Objective: We present a fully automated method for the quantification of knee alignment from full-leg radiographs.

Methods: A state-of-the-art object detector, YOLOv4, was trained to locate regions of interests in full-leg radiographs for the hip joint, knee, and ankle. Residual neural networks were trained to regress landmark coordinates for each region of interest. Based on the detected landmarks the knee alignment, i.e., the hip-knee-ankle (HKA) angle was computed. The accuracy of landmark detection was evaluated by a comparison to manually placed ones for 180 radiographs. The accuracy of HKA angle computations was assessed on the basis of 2,943 radiographs by a comparison to results of two independent image reading studies (Cooke; Duryea) both publicly accessible via the Osteoarthritis Initiative. The agreement was evaluated using Spearman’s Rho, weighted kappa, and regarding the correspondence of the class assignment.

Results: The average deviation of landmarks manually placed by experts and automatically detected ones by our proposed ”YOLOv4 And Resnet Landmark regression Algorithm” (YARLA) was less than 2.0 ± 1.5 mm for all structures. The average mismatch between HKA angle determinations of Cooke and Duryea was $0.09 \pm 0.63^\circ$; YARLA resulted in a mismatch of $0.09 \pm 0.73^\circ$ compared to Cooke and of $0.18 \pm 0.67^\circ$ compared to Duryea. Cooke and Duryea agreed almost perfectly with respect to a weighted kappa value of 0.86, and showed an excellent reliability as measured by a Spearman’s Rho value of 0.98. Similar values were achieved by YARLA, i.e., a weighted kappa value of 0.83 and 0.87 and a Spearman’s Rho value of 0.98 and 0.98 compared to Cooke and Duryea, respectively. Cooke and Duryea agreed in 91% of all class assignments and YARLA did so in 90% against Cooke and 92% against Duryea.

Conclusions: YARLA yields HKA angles similar to those of human experts and provides a basis for an automated assessment of knee alignment in full-leg radiographs.

Keywords: Hip-knee-ankle angle, Varus, Valgus, Mechanical axes, Osteoarthritis, Deep learning

1 Introduction

Knee malalignment affects the distribution of loads across the joint in an unfavorable manner leading to increased contact pressure in the more heavily loaded regions [1]. Consequently, knee malalignment can be considered a risk factor for osteoarthritis and cartilage loss as well as a biomarker for assessing severity and progression [2, 3, 4].

As shown in Fig. 1, knee alignment is defined as the angle between the mechanical axes of the femoral and tibial bones. This angle has been termed the hip-knee-ankle (HKA) angle [5]. The mechanical axis of the femur is defined by a line from the center of the femoral head to the mid-condylar point between the cruciate ligaments (cf. Fig. 1). The mechanical axis of the tibia is defined as a line from the center of the tibial plateau to the center of the tibial plafond [6]. As a convention the HKA angle is expressed as the angular deviation from a straight angle of 180° . Varus deviations are expressed as negative angles and valgus deviations as positive ones.

Classically, the HKA angle was assessed in anterior-posterior radiographs in a manual or a semi-automated setting [7, 8, 5, 4, 9]. [10] proposed a computer-assisted, landmark-based method for the assessment of HKA angles from full-limb radiographs. In their semi-automated method the rater is guided to identify a set of landmarks. Afterwards, the HKA angle is derived automatically from the placed landmarks. The method of Sled et al. achieved an excellent accuracy, however requires manual input and trained image readers. In recent years there has been a general trend towards automated, deep learning-based methods for the analysis of medical images [11]. [12] presented an automated image analysis pipeline to predict the HKA angle from standard knee radiographs. Their proposed method utilizes random forest regression to predict landmarks at the outline of the knee bones. These landmarks are employed to estimate the HKA angle. The advantage of the method of Gielis et al. is that radiation exposure is minimized since only the knee is imaged instead of the full limb. However, this comes with the limitation of an average error of 1.8° as well as moderate intraclass correlation coefficients (ICCs) of 0.90 compared to manual readings from full-leg radiographs. [13] proposed a decentralized deep learning algorithm for HKA angle computation from full-leg radiographs. In a two-level approach 10 regions of interest (ROIs) are detected first using a convolutional neural network (CNN). Then, by utilizing these ROIs the coordinates of anatomical landmarks are computed employing a second CNN. The

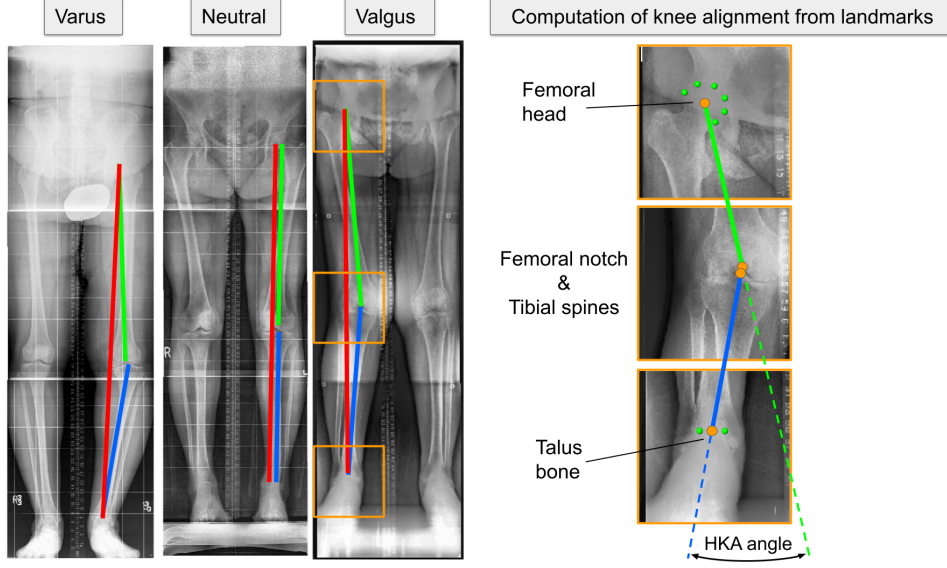


Figure 1: Left: Examples of legs with varus malalignment, neutral alignment, and valgus malalignment. The load-bearing axis is shown in red. Right: Computation of knee alignment based on landmarks illustrated by close-up images of the valgus leg. The mechanical axes of the femur (green line) and tibia (blue line) are computed based on landmarks for the hip, knee, and ankle. The center of the femoral head (orange circle) is derived from 6 landmarks placed at the boundary of the femoral head. The landmarks at the femoral notch and tibial spines are directly placed at the distinct anatomical regions. The center of the talus bone is derived from two landmarks defined at the superior medial and lateral edges of the talus. The HKA angle is the angle enclosed by the femoral mechanical axis and the tibial mechanical axis.

approach of Nguyen et al. is efficient with a run time of less than one second, but yields an average bias of -0.402° as well as a mismatch of more than 1.5° in 17.7% of the analyzed subjects compared to the ratings provided by radiologists. [14] presented a deep learning approach to assess the HKA angle from full-leg radiographs based on automated segmentations of the leg bones. A CNN similar to a U-net [15] was employed to segment the distal and proximal femur and tibia. The centers of these structures were used to compute the femoral and tibial mechanical axes in order to derive the HKA angle. The segmentations of hip, knee, and ankle were successful with a Dice similarity of up to 0.93. However, their method resulted in an average systematic bias of 0.49° as well as a mismatch greater than 1.5° in 10.83% of the analyzed subjects compared to the ratings of radiologists.

First automated approaches for HKA angle computation employing methods of machine learning were presented by [13], [14], and [12]. However, the proposed methods (i) often show a systematic bias compared to the measurements of radiologists, (ii) may result in deviations larger than 1.5° in a substantial amount of analyzed subjects. Additionally, the proposed studies evaluated a rather small set of images and merely assessed the accuracy of HKA angle computation – but did not analyze the accuracy of the underlying detection of anatomical landmarks.

The motivation for our study is to employ state-of-the-art methods of deep learning to determine the HKA angle from full-leg radiographs. We employ YOLOv4 [16], which is a fast object detection algorithm where the input image is subdivided and object detection is performed in each subregion. Such an approach is of advantage over methods which consecutively loop over all regions of the image, like in R-CNN [17] and might lead to increased accuracy and less run time. YOLOv4 uses the entire image during training and test time in order to implicitly encode contextual information about classes as well as their appearances and is highly generalizable and less likely to fail when applied to other domains or unexpected inputs [18]. After ROI detection using YOLOv4 we employ ResNets [19] to build deep

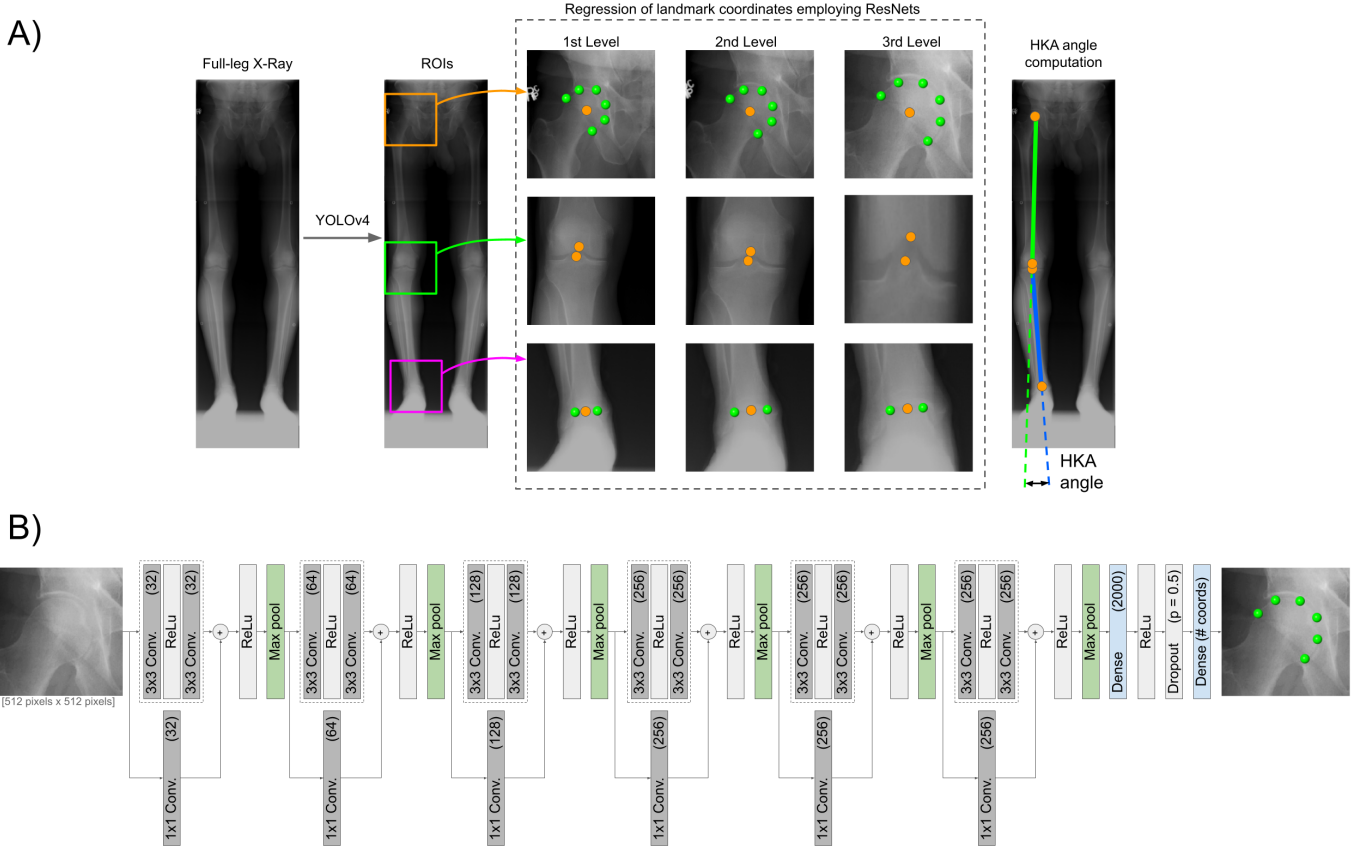


Figure 2: A) Pipeline of YARLA for computation of the HKA angle from full-leg X-Rays: The first step of the algorithm is YOLOv4 which locates ROIs in an image (hip: orange, knee: green, ankle: purple). In the second step three levels of ResNets are employed for the regression of landmark coordinates. The HKA angle is finally derived from the resulting two axes. B) Flow-chart of the ResNet architecture. Six residual blocks with projection shortcuts are employed. The number of filters is indicated in brackets. 3×3 convolutions are employed in each block. The last block is followed by a dense layer with 2000 nodes as well as a dense layer having as many nodes as the number of landmark coordinates.

CNNs for regression of landmark coordinates. We thoroughly evaluate the accuracy of our method on publicly available data from the Osteoarthritis Initiative (OAI)¹. The contributions of this study are (i) the transfer of YOLOv4 as a fast and reliable object detector for anatomical regions in full-leg radiographs, (ii) utilization of ResNets for precise landmark detection in these regions, and (iii) a thorough evaluation of the accuracy of landmark detection, HKA angle computation as well as of the class assignment (varus/neutral/valgus).

	OAI time point			
	v12	v24	v36	v48
Number of subjects	1,472	1,275	919	177
Subjects in common with v12	all	37	14	8
Subjects in common with v24	37	all	22	5
Subjects in common with v36	14	22	all	6
Subjects in common with v48	8	5	6	all
Sex (male; female)	671; 801	565; 710	349; 570	73; 104
Age [years]	61.86 \pm 9.08	63.67 \pm 9.22	64.27 \pm 8.94	64.58 \pm 9.1
BMI [kg/m ²]	29.75 \pm 4.83	28.36 \pm 4.97	28.31 \pm 4.94	28.57 \pm 5.59
Cooke: Legs measured	2,456	2,547	1,822	264
Duryea: Legs measured	2,858	2,521	1,828	352
Legs measured by both studies	2,942	2,549	1,838	354
Legs measured by Cooke only	84	28	10	2
Legs measured by Duryea only	486	2	16	90
Cooke: Average HKA angles	-1.37 \pm 3.86	-1.17 \pm 3.30	-0.95 \pm 3.02	-1.06 \pm 3.07
Duryea: Average HKA angles	-1.41 \pm 3.80	-1.24 \pm 3.33	-1.08 \pm 3.03	-1.27 \pm 3.05
Cooke: Alignment classes (Varus; Neutral; Valgus; NA)	1,025; 923; 416; 92	1,026; 1,114; 378; 29	638; 885; 293; 6	97; 125; 42; 0
Duryea: Alignment classes (Varus; Neutral; Valgus; NA)	1,234; 1,138; 486; 0	1,031; 1,130; 360; 0	660; 901; 267; 0	132; 173; 47; 0

Table 1: Demographics: In this study 3,843 X-Rays of the OAI database are analyzed that were acquired at four visits every 12 months (v12, v24, v36, and v48). At the different time points, mainly radiographs of different persons were taken. The majority of legs were assessed by both, Cooke and Duryea.

2 Methods

2.1 Full-Leg X-Rays from the OAI

3,843 full-leg X-Rays from the publicly available OAI database have been assessed. Detailed demographics are given in Table 1. Some X-Rays were excluded from this study due to an incomplete field of view ($N = 11$) and due to missing pixel size information in the DICOM metadata ($N = 3$). If the hip, knee, or ankle are outside of the field of view for one leg only, this leg is excluded from the study, but the contralateral one is still used ($N = 3$). See Supplementary Figure A1 for images and subject identifiers of all cases that were treated specially.

2.2 Knee alignment studies of Cooke and Duryea

Two independent image reading studies for measuring HKA angles in full-leg radiographs as provided by the OAI were independently conducted by Dr. Cooke and Dr. Duryea in two different image reading centers. The measurements of Cooke were funded by OAI and performed by OAISYS Inc.² with support of staff at Queen’s University (Kingston, Ontario) using the semi-automated tool Horizon Surveyor (OAISYS Inc., Perth, Canada). The measurements of Duryea were performed independently of OAI in the laboratory of Dr. Jeff Duryea at Brigham and Women’s Hospital in Boston, MA using custom software to guide the reader in placing the respective landmarks.³ Both, Cooke and Duryea, evaluated 6,965 of the available 7,683 legs. 124 remaining legs were evaluated by Cooke only and 594 by Duryea only.

¹<https://nda.nih.gov/oai/>

²www.oaisysmedical.com

³<http://www.spl.harvard.edu/pages/People/duryea>

2.3 Automated determination of HKA angles by employing YARLA

Our motivation is to develop an automated method for the assessment of the HKA angle in full-leg X-Ray images following the clinical workflow of radiologists. Hence, our approach is to locate the respective landmarks for each joint individually, to derive the mechanical axes, and finally to compute the HKA angle. For the hip, six landmarks are to be detected around the femoral head. These landmarks are placed manually equally distributed around the whole femoral head. For the knee and ankle, two distinct anatomical landmarks are to be detected, respectively (see Fig. 2). Several object detectors based on machine learning were proposed for an automated image analysis [20]. Many of them already show an excellent accuracy in object detection as well as a high accuracy in the determination of bounding boxes enclosing these objects. In our proposed YARLA we chose YOLOv4 for a detection of ROIs because it was shown to be very fast and precise on data from the "Microsoft COCO: Common Objects in Context" challenge [21] as well as data from PASCAL Visual Object Classes challenges (<http://host.robots.ox.ac.uk/pascal/VOC/>). In contrast to common object detection algorithms YOLOv4 gets the complete image as input, but processes it at once in order to achieve a low run time and high accuracy due to implicit encoding of contextual information. Its architecture is splitting the input image into a grid. If the center of an object of interest falls into a grid cell, this grid cell becomes responsible for detecting that object, i.e., to compute the respective class probabilities and bounding boxes. In our study, we transfer YOLOv4 to the medical domain. YOLOv4 is employed to detect ROIs in X-Ray images for the hip joint, the knee joint, and the ankle joint, respectively. Its architecture is kept as described in the original publications [18, 16].

It was shown that CNNs are well suited for accurate regression of landmark coordinates in medical images [22, 23, 24]. For classification and regression tasks, very deep CNNs often suffer from the vanishing gradient problem [25] as the gradient is back-propagated to earlier layers and the repeated multiplications result in an infinitely small gradient. ResNets [19] introduced so-called "identity shortcut connections" that skip one or more layers making it easier to train a deep CNN. The recent success of CNNs for landmark detection as well as the potential of better gradient flow of ResNets motivated us to employ them for landmark regression in each ROI (hip, knee, ankle) that was previously detected by employing YOLOv4. Within these regions the respective landmarks are located at the boundaries of the bones. Hence, a good contrast can be expected and the ResNets can be trained to detect these landmarks relying on strong gradients at the outlines of the bones as well as on local texture information. As shown in Fig. 2B, we employ ResNets with projection shortcuts consisting of six layers for regression of the landmark coordinates. The number of nodes of the final layer of the respective ResNet is equal to the number of landmark coordinates. The mean average error is employed as a loss function for training the ResNets and optimization is carried out using stochastic gradient descent (cf. Algorithm 1).

The input size of CNNs is restricted due to memory limitations. To allow our CNNs to focus on the relevant region around the respective landmarks with a maximum level of detail we employ an approach consisting of three levels of ResNets. In the first level the centers of the ROIs for each joint as detected by YOLOv4 are used for the computation of a surrounding region of $170\text{ mm} \times 170\text{ mm}$ in size. In the second level a $135\text{ mm} \times 135\text{ mm}$ region is extracted at the predicted location of the first stage. In the third and final level a region of $100\text{ mm} \times 100\text{ mm}$ in size for both the hip and the ankle, as well as a region of $100\text{ mm} \times 40\text{ mm}$ in size for the knee is computed with the centers at the respective second predictions. The sizes of the final regions are chosen according to the average sizes of adult joints and the initial sizes are chosen large enough to account for inaccuracies in the ROI detection by YOLOv4. Each region is resampled to 512×512 pixels and min-max normalized to an intensity range between zero and one before it is fed into the respective ResNet. The architecture of the ResNets is identical in each level.

After all landmarks have been detected by the ResNets, the resulting axes are computed from these landmarks. The center of the femoral head is computed by fitting a circle to the six landmarks of the hip and by minimizing the following equation:

$$\underset{\substack{r \\ x_c \\ y_c}}{\text{minimize}} E = \sum_{i=1}^n (d_i - r)^2, \quad (1)$$

with a circle of radius r and center (x_c, y_c) , and d_i being the distance of a landmark (x_i, y_i) from the center defined as

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} . \quad (2)$$

The femoral mechanical axis is defined from the femoral head \mathbf{c}_{hip} to the femoral knee landmark \mathbf{p}_{femur} . The center of the ankle is computed as the geometric mean of the two respective landmarks. The tibial mechanical axis is defined from the tibial knee landmark \mathbf{p}_{tibia} to the center of the ankle \mathbf{c}_{ankle} (cf. Fig. 1). Finally, the HKA angle is computed based on the two resulting axes as

$$\text{HKA angle} = \arctan \left(\frac{\det[\mathbf{c}_{ankle} - \mathbf{p}_{femur}, \mathbf{c}_{hip} - \mathbf{p}_{tibia}]}{(\mathbf{c}_{ankle} - \mathbf{p}_{femur}) \cdot (\mathbf{c}_{hip} - \mathbf{p}_{tibia})} \right) . \quad (3)$$

2.4 Experimental setup

Out of 3,843 X-Rays used in this study 900 X-Rays (OAI time point v12) are used as training, validation, and test data of YOLOv4 and the ResNets (60%, 20%, 20%). Manual landmarks are placed for both legs of all 900 X-Rays, i.e., six landmarks for the femoral head, two landmarks for the knee, and the ankle, respectively. In the following, these landmarks are termed LM_ZIB.

Our method is trained for the right leg only. The data is flipped in left-right direction to effectively double the amount of training data and to additionally reduce the variance within the data. Specifically, the X-Rays are flipped prior to training of YOLOv4 such that the patients' left legs appear similar to the right ones. All X-Ray images are resized to 512×1024 pixels and given as input into YOLOv4. The ROIs used for training YOLOv4 always have a size of $170 \text{ mm} \times 170 \text{ mm}$ centered at the geometric mean of the landmarks LM_ZIB for the respective joint. This size was empirically determined such that all details are covered by the ROIs.

For each level ResNets are trained using ROIs randomly placed around the center of the landmarks. To enhance the generalization ability and to cope with larger variation within the given image data, data augmentation is employed. The ROIs are translated (randomly up to $\pm 10\%$ of physical size of the ROI in every direction) and rotated (randomly up to $\pm 7^\circ$) as well as scaled (randomly between 75% and 125%).

2,943 X-Rays are neither used for training, validation nor testing. For these data HKA angle measurements are given by Cooke and Duryea but no manually placed landmarks. These 2,943 X-Rays are used for the evaluation of the HKA angles that are determined by YARLA and are in the following termed as "Angle_OAI" data.

2.5 Evaluation of the performance of YARLA

Table 2 shows the methods that are employed to evaluate the performance of YARLA. For all 360 legs contained in the testing data the mismatch between the landmarks computed by YARLA and the manually placed ones are evaluated. To investigate which regions are most important for the ResNet's computation of landmark coordinates occlusion heatmaps [26] are employed. A so-called "occluder" is moved over the respective ROI with a stride of 8. At each position the occluder sets the intensities of all pixels in a 64×64 pixels region to the mean image intensity to occlude the real image. The magnitude of change is evaluated for each position of the occluder and the most important regions for the ResNet's prediction are qualitatively assessed (see Fig. 3).

Additionally, the following five methods are employed for an evaluation of YARLA on both testing and Angle_OAI data: (1) Computation of non-parametric Spearman's Rho [27] to assess the agreement between HKA angles as determined by YARLA and the measurements of Cooke and Duryea. The Spearman's Rho measure is chosen since the HKA angle measurements are not normally distributed. (2) Generation of Bland-Altman plots [28] to investigate any systematic bias between YARLA results and the two studies, as well as to visualize the variance and to identify outliers.

In order to perform class assignments (varus/neutral/valgus), varus malalignment is defined as HKA angles ≤ -2 degrees, valgus malalignment as HKA angles $\geq +2$ degrees, and neutral alignment as any angle in between. Based on this definition of knee alignment, (3) the agreement of class assignment by YARLA is compared to the class assignments

of Cooke and Duryea. Further, (4) confusion matrices are plotted to analyze which classes were most often confounded with each other. Finally, (5) the weighted kappa [29] is computed to quantify agreement between YARLA and the two studies.

In order to investigate the quality of our manually placed landmarks, HKA angles and class assignments are determined using LM_ZIB for the testing data only. The conformity of class assignment is assessed using Spearman’s Rho and weighted kappa between the class assignments using LM_ZIB and those of Cooke, Duryea, and YARLA.

For a comparison to existing methods the ICC [30] and the proportion of errors being larger than 1.5° are computed for the Angle_OAI data.

Finally, we perform an ablation study to analyze the individual parts of our method. In the ablation study we evaluate the performance of YOLOv4 only and investigate the performance of adding up to three levels of ResNet landmark regression. To evaluate YOLOv4 only, the centers of the ROIs computed by YOLOv4 are utilized to compute the femoral and tibial mechanical axes. The HKA angle is derived from these axes. In further investigations one, two, or three levels of ResNets are employed to detect landmarks in the ROIs given by YOLOv4 (see Algorithm 1 and Fig. 2A).

YARLA output	Method for evaluation
Landmark location	Average distance
Landmark location	Occlusion heatmaps
HKA angle	Bland-Altman plots
HKA angle	Proportion of errors being $> 1.5^\circ$
HKA angle	Intraclass correlation coefficient (ICC)
HKA angle	Non-parametric Spearman’s Rho
Class assignments	Agreement
Class assignments	Confusion matrix
Class assignments	Weighted kappa

Table 2: Methods used for evaluating the performance of YARLA.

3 Results

In order to evaluate the quality of YARLA several components of the method were assessed, i.e., YOLOv4 as the employed object detector, preciseness of the located landmarks, HKA angle computations, and class assignments.

Success rate of YOLOv4

YOLOv4 detects the ROIs of the hip, knee, and ankle successfully for all legs contained in the testing data. For the Angle_OAI cases, YOLOv4 detects all regions successfully for 5,809 out of 5,818 legs (99,85% success rate, see Supplementary Figure B1 for images and subject identifiers of the nine X-Ray images for which YOLOv4 failed).

Location of the automatically detected landmarks

For the testing data, the difference between landmark positions as determined by YARLA and the manually located ones is on average 1.72 ± 1.00 mm for the center of the femoral head, 1.94 ± 1.33 mm for the distal femoral notch, 1.63 ± 1.29 mm for the tibial spines, and 1.54 ± 1.33 mm for the center of the talus bone at the ankle (Table 5).

Analysis of systematic bias and outliers

Bland-Altman plots were used to investigate if there is any systematic bias or if there are outliers in the HKA angles determined by Cooke, Duryea, or YARLA. Additionally, the HKA angles computed based on LM_ZIB are evaluated. LM_ZIB have an average mismatch of $0.12 \pm 0.6^\circ$ and $0.18 \pm 0.46^\circ$ to Cooke and Duryea, respectively. As shown in the Bland-Altman plots in Figure 4, the average disagreement between YARLA and LM_ZIB is $0.03 \pm 0.48^\circ$. The disagreement between YARLA and Cooke and Duryea is $0.13 \pm 0.65^\circ$ and $0.21 \pm 0.56^\circ$, respectively. Cooke and

Cooke vs. Duryea									
Cooke	Testing data			Cooke	Angle_OAI data				
		Duryea				Duryea			
		Varus	Neutral		Valgus		Varus	Neutral	Valgus
	Varus	99	7		0	Varus	2003	141	0
	Neutral	7	102		6	Neutral	154	2221	77
Valgus	0	5	53	Valgus	1	123	737		

YARLA vs. Cooke									
YARLA	Testing data			YARLA	Angle_OAI data				
		Cooke				Cooke			
		Varus	Neutral		Valgus		Varus	Neutral	Valgus
	Varus	95	3		0	Varus	1915	102	1
	Neutral	11	107		8	Neutral	255	2281	135
Valgus	0	5	50	Valgus	0	83	730		

YARLA vs. Duryea									
YARLA	Testing data			YARLA	Angle_OAI data				
		Duryea				Duryea			
		Varus	Neutral		Valgus		Varus	Neutral	Valgus
	Varus	125	3		0	Varus	2056	68	0
	Neutral	13	143		7	Neutral	235	2487	84
Valgus	0	3	58	Valgus	2	79	766		

Table 3: Confusion matrices for the testing as well as the Angle_OAI data.

Duryea had a mismatch of $0.07 \pm 0.57^\circ$. For the Angle_OAI data the mismatch is $0.09 \pm 0.73^\circ$ and $0.18 \pm 0.67^\circ$ between YARLA and the two studies, whereas Cooke and Duryea had a disagreement of $0.09 \pm 0.63^\circ$.

Spearman's Rho

In Table 4 statistical comparisons of the HKA angles between YARLA and the two studies are shown. For the testing data as well as the Angle_OAI data very strong correlations have been achieved. Spearman's Rho is greater than 0.98 for all four raters. With $p < 0.001$, significant correlations have been found in all cases.

Agreement of class assignment

The correctness of class assignment is equal to or higher than 90% for the testing data as well as the Angle_OAI data between YARLA and all other measurements (Table 4). The highest accuracy (93%) has been achieved between YARLA and LM_ZIB as well as Duryea (testing data). The lowest accuracy (90%) has been achieved between YARLA and Cooke (Angle_OAI data).

Confusion matrices

In Table 3 confusion matrices are shown for both testing and the Angle_OAI data. It can be seen for the Angle_OAI data that 255 knees with varus and 135 knees with valgus malalignment have been missclassified as neutral compared to Cooke as well as 235 and 84 compared to Duryea. The amount of legs in neutral alignment misclassified as malaligned is in tendency smaller (in total 185 misclassifications compared to Cooke and 147 misclassifications compared to Duryea).

Weighted kappa

Almost perfect agreement is found as measured by weighted kappa being higher than 0.80 between YARLA and the two other raters (Table 4). The highest kappa (0.88) is achieved between YARLA and the measurements employing

LM_ZIB as well as the ones of Duryea (testing data). The lowest kappa (0.83) is achieved between YARLA and Cooke (Angle_OAI data). The kappa decreases for YARLA vs. Cooke comparing the testing with the Angle_OAI data (0.85 vs. 0.83) and increases slightly for YARLA vs. Duryea (0.85 vs. 0.87).

Run time

ROI detection for both legs using YOLOv4 takes 0.6 seconds per X-ray on average. The application of the three ResNet levels takes 2.1 seconds on average for both legs. In total, the computation of the HKA angles for both legs in one X-Ray takes about 3 seconds on average.

Ablation study and comparison to other methods

The results of our ablation study and a comparison to four related methods is provided in Table 5. It can be seen that most performance measures are clearly improving in each level of ResNet landmark regression following the YOLOv4 ROI detection. The proportion of HKA angle errors larger than 1.5° is decreasing from 3.56% and 5.05% (YOLOv4 for Cooke and Duryea) to 3.38% and 1.82% using three levels of ResNet landmark regression.

4 Discussion

YARLA detects the joints of interest for almost all cases and achieves a good accuracy for locating the anatomical landmarks. Almost perfect agreement as measured by weighted kappa [31] and very strong correlations as measured by Spearman’s Rho [32] confirm the quality of our method. These evaluation methods as well as the conformity of class assignments are approximately in the range of disagreement between the two studies of Cooke and Duryea. Also, considering the related work, our proposed method achieves excellent results. The methods of both, [13] as well as [14], show a systematic mismatch between the automated results and the manual readings of around 0.5° (see Table 6). Our method shows only a slight deviation from the two studies by Cooke and Duryea. Moreover, stronger correlations are found using YARLA compared to [12], as well as clearly less bias. Also, in terms of the proportion of knees with a mismatch greater than 1.5° our method performs clearly better than the ones proposed by Nguyen et al. and Pei et al. (Table 6).

As shown in the occlusion heatmaps (Fig. 3) the ResNets focus on the image regions adjacent to the anatomical landmarks which are to be detected. Especially for the hip we noticed that the CNN learned the circular arrangement of the landmarks. It could be investigated in the future whether all six landmarks are needed or if fewer landmarks might also be sufficient.

In our evaluation of thousands of X-Ray images we did notice some limitations of YARLA. In nine cases YOLOv4 fails to identify all joint ROIs. These nine cases are shown in Supplementary Figure B1. This is probably due to an extremely low image contrast. Most of these cases have very dark regions around the knee and ankle or very bad contrast at the hip – whereas the other regions are displayed well. We argue that these X-Ray images should have been acquired in a better quality.

A strength of our method is that YOLOv4 indicates the certainty for all detected ROIs within the image allowing for quality assurance of the input X-Ray data (e.g. if the complete limb is covered). We believe our method performs better than the method of [13] since ResNets allow us to build deeper CNN consisting of 6 convolutional layers. Moreover, 3×3 convolutions are usually more efficient [33] than the 9×9 and 7×7 convolutions as employed by [13]. Each level of ResNet landmark regression is further improving the results since a higher level of detail is covered in the respective input images and the ResNet is explicitly guided to focus on the relevant region. Disadvantages of our method are that it is not trained end-to-end and that 9 ResNets need to be trained and executed. Moreover, our ResNet levels are completely independent and an iterative refinement of landmark positions could be beneficial [34]. As shown with the Bland-Altman plots (Fig. 4) the resulting mean disagreement as well as the standard deviation of HKA angle computations are consistent between the test data and the Angle_OAI data. The systematic mismatch between YARLA and the two studies is low, however, it can be seen that YARLA has the tendency to compute higher HKA angle values, i.e., valgus malalignment.

Spearman's Rho				
<i>Testing data</i>				
	Cooke	Duryea	LM_ZIB	YARLA
Cooke	—	0.99 (p < 0.001)	0.99 (p < 0.001)	0.98 (p < 0.001)
Duryea	0.99 (p < 0.001)	—	0.99 (p < 0.001)	0.99 (p < 0.001)
LM_ZIB	0.99 (p < 0.001)	0.99 (p < 0.001)	—	0.99 (p < 0.001)
YARLA	0.98 (p < 0.001)	0.99 (p < 0.001)	0.99 (p < 0.001)	—

<i>Angle_OAI data</i>			
	Cooke	Duryea	YARLA
Cooke	—	0.98 (p < 0.001)	0.98 (p < 0.001)
Duryea	0.98 (p < 0.001)	—	0.98 (p < 0.001)
YARLA	0.98 (p < 0.001)	0.98 (p < 0.001)	—

Accuracy of class assignment				
<i>Testing data</i>				
	Cooke	Duryea	LM_ZIB	YARLA
Cooke	—	0.92	0.92	0.90
Duryea	0.92	—	0.93	0.93
LM_ZIB	0.92	0.93	—	0.93
YARLA	0.90	0.93	0.93	—

<i>Angle_OAI data</i>			
	Cooke	Duryea	YARLA
Cooke	—	0.91	0.90
Duryea	0.91	—	0.92
YARLA	0.90	0.92	—

Weighted kappa				
<i>Testing data</i>				
	Cooke	Duryea	LM_ZIB	YARLA
Cooke	—	0.88	0.87	0.85
Duryea	0.88	—	0.89	0.88
LM_ZIB	0.87	0.89	—	0.88
YARLA	0.85	0.88	0.88	—

<i>Angle_OAI data</i>			
	Cooke	Duryea	YARLA
Cooke	—	0.86	0.83
Duryea	0.86	—	0.87
YARLA	0.83	0.87	—

Table 4: Evaluation of non-parametric Spearman's Rho, accuracy of class assignment, and weighted kappa. Agreement is computed for the Angle_OAI data between the automated HKA angle computations of YARLA and those of Cooke and Duryea. For the testing data, additionally, HKA angles were derived from our manually determined landmarks, LM_ZIB, and compared to the results of YARLA, Cooke, and Duryea.

Method	L2 landmark error [mm]	HKA angle error [Degree]	HKA angle error > 1.5°	HKA angle Spearman's Rho	ICC	HKA class as- signment weighted kappa	HKA class as- signment accuracy	Run time	Modality
Ours (YOLOv4)	H: 2.50 ± 2.49 FN: 4.21 ± 1.46 TS: 4.28 ± 1.80 A: 2.45 ± 3.15	C: 0.30 ± 0.70 D: 0.38 ± 0.69	C: 3.56% D: 5.05%	C: 0.98 D: 0.97	C: 0.97 D: 0.97	C: 0.82 D: 0.80	C: 0.89 D: 0.88	0.6 s	Full-leg X-Ray
Ours (YOLOv4 + 1 Level)	H: 2.39 ± 1.45 FN: 2.01 ± 1.34 TS: 1.91 ± 1.13 A: 2.55 ± 2.00	C: 0.24 ± 0.73 D: 0.32 ± 0.63	C: 3.85% D: 2.22%	C: 0.97 D: 0.98	C: 0.97 D: 0.98	C: 0.81 D: 0.84	C: 0.88 D: 0.90	1.3 s	Full-leg X-Ray
Ours (YOLOv4 + 2 Levels)	H: 2.09 ± 1.21 FN: 2.02 ± 1.33 TS: 1.86 ± 1.17 A: 2.05 ± 1.54	C: 0.20 ± 0.74 D: 0.29 ± 0.76	C: 3.42% D: 1.87%	C: 0.97 D: 0.98	C: 0.97 D: 0.97	C: 0.82 D: 0.85	C: 0.89 D: 0.91	2.0 s	Full-leg X-Ray
Ours (YOLOv4 + 3 Levels)	H: 1.72 ± 1.00 FN: 1.94 ± 1.33 TS: 1.63 ± 1.29 A: 1.54 ± 1.33	C: 0.09 ± 0.73 D: 0.18 ± 0.67	C: 3.38% D: 1.82%	C: 0.98 D: 0.98	C: 0.97 D: 0.98	C: 0.83 D: 0.87	C: 0.90 D: 0.92	2.7 s	Full-leg X-Ray

Table 5: Ablation study of the proposed method. To evaluate YOLOv4 on the testing data, the centers of the ROIs computed by YOLOv4 are used to derive the HKA angle. Moreover, the influence on the results of up to three levels of ResNet landmark regression following ROI detection by YOLOv4 is investigated. The L2 error is shown for the hip (H), femoral notch (FN), tibial spines (TS), and ankle (A). All other metrics are computed between our predictions and the measurements from Cooke (C) and Duryea (D), respectively. For reasons of comparability, additionally the intraclass correlation coefficient (ICC) is computed. The best results per column are highlighted in bold.

Method	L2 landmark error [mm]	HKA angle error [Degree]	HKA angle error > 1.5°	HKA angle Spearman's Rho	ICC	HKA class as- signment weighted kappa	HKA class as- signment accuracy	Run time	Modality
[10]	—	—	—	—	0.995	—	—	—	Full-leg X-Ray
[13]	—	-0.402 ± 0.68 (left leg)	17.7%	—	—	—	—	< 1 s	Full-leg X-Ray
[12]	—	1.8 ± 1.3	—	—	0.90	—	—	—	AP knee radio- graphs
[14]	—	-0.49 ± 0.75	10.83%	—	0.999	—	—	—	Full-leg X-Ray
Proposed method	H: 1.72 ± 1.00 FN: 1.94 ± 1.33 TS: 1.63 ± 1.29 A: 1.54 ± 1.33	C: 0.09 ± 0.73 D: 0.18 ± 0.67	C: 3.38% D: 1.82%	C: 0.98 D: 0.98	C: 0.97 D: 0.98	C: 0.83 D: 0.87	C: 0.90 D: 0.92	2.7 s	Full-leg X-Ray

Table 6: Comparison of our results to related work. The L2 error is shown for the testing data for the hip (H), femoral notch (FN), tibial spines (TS), and ankle (A) between the predictions of our method and our manual landmarks. All other metrics are computed for our method between our predictions and the measurements from Cooke (C) and Duryea (D), respectively. For reasons of comparability, additionally the intraclass correlation coefficient (ICC) is computed. The best results per column are highlighted in bold.

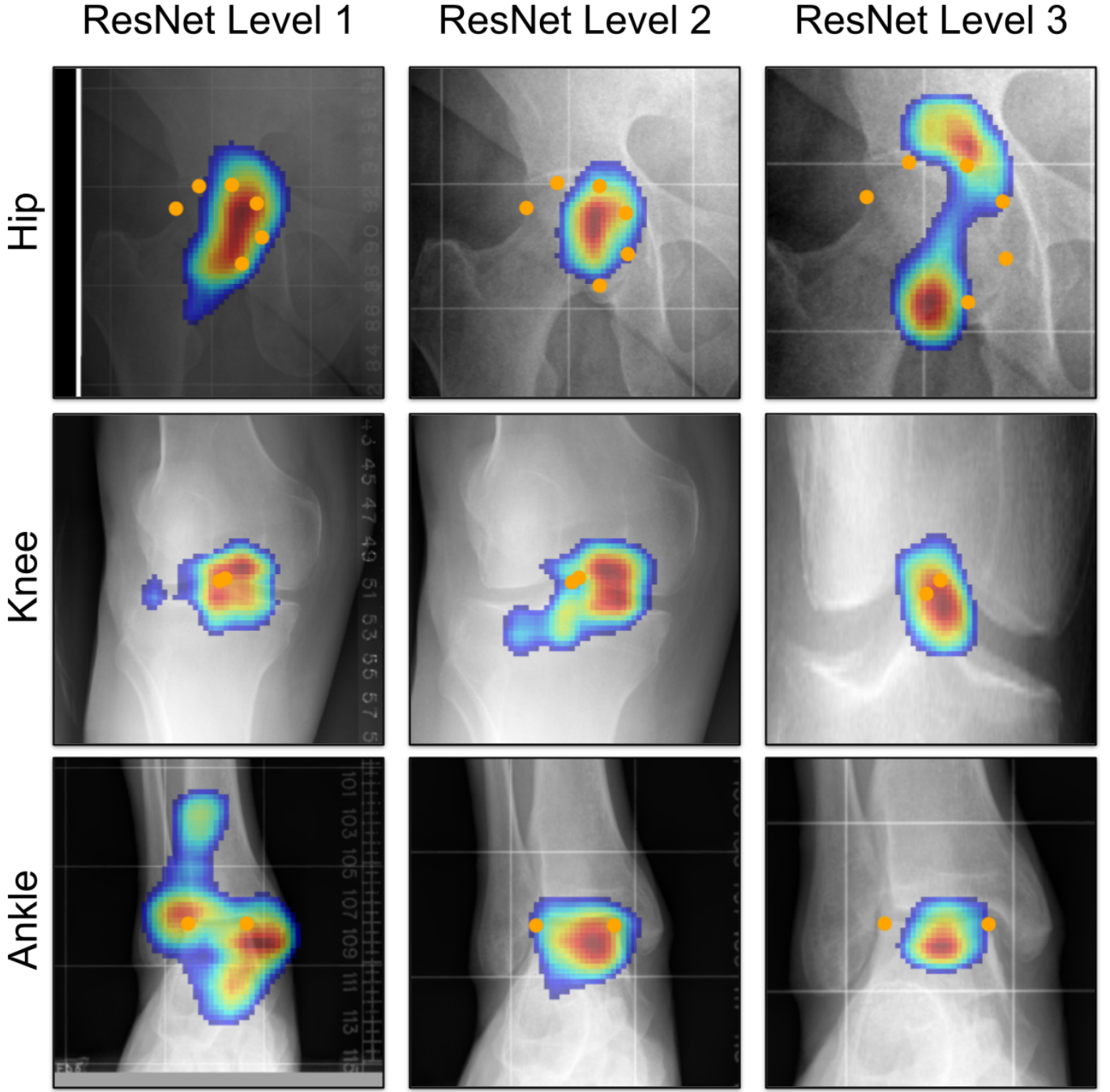
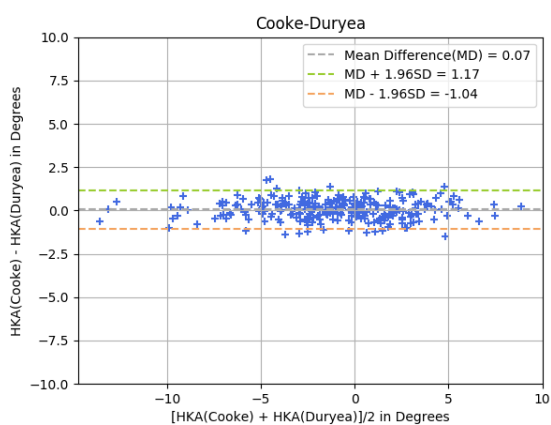
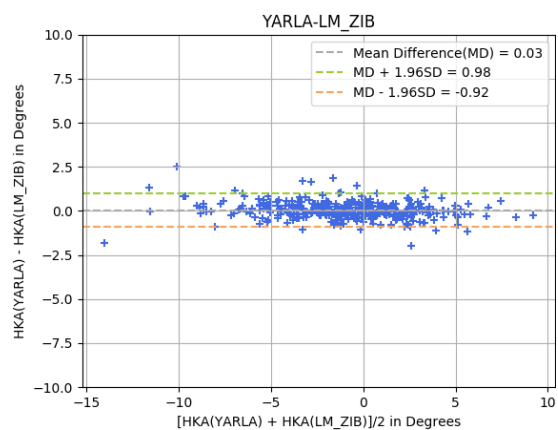


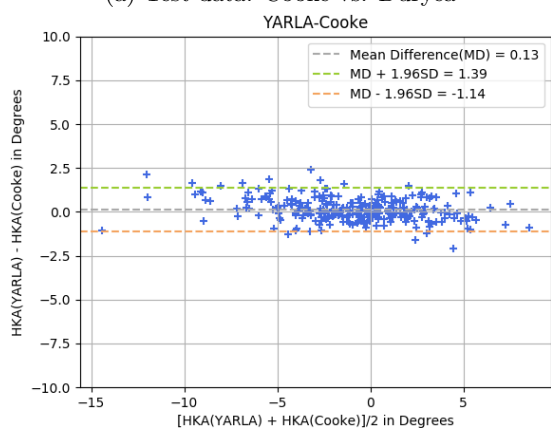
Figure 3: Occlusion heatmaps were computed for all levels of ResNet landmark regression. We utilized a so-called "occluder" of size 64×64 with mean image intensity. The occluder was moved over the respective X-Ray image with a stride of 8 pixels. The magnitude of change of the landmark coordinate prediction was evaluated at each position. The magnitudes of the occlusion heatmaps were normalized to $[0,1]$ and values lower than 0.7 were truncated.



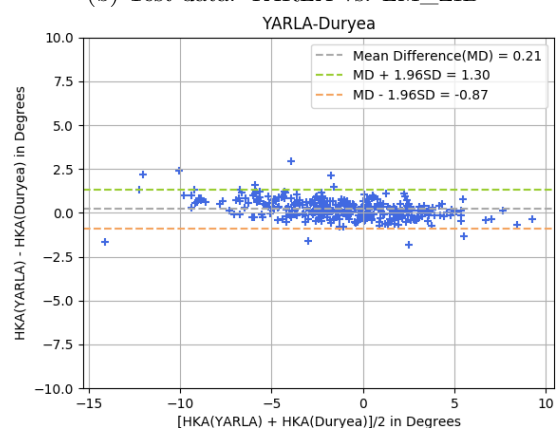
(a) Test data: Cooke vs. Duryea



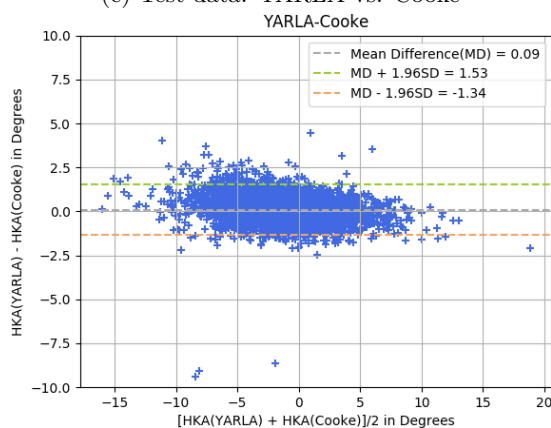
(b) Test data: YARLA vs. LM_ZIB



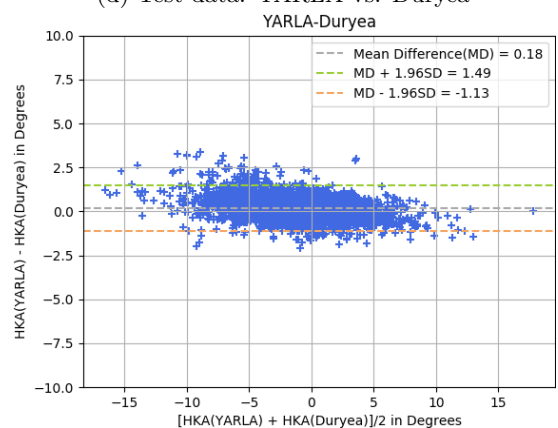
(c) Test data: YARLA vs. Cooke



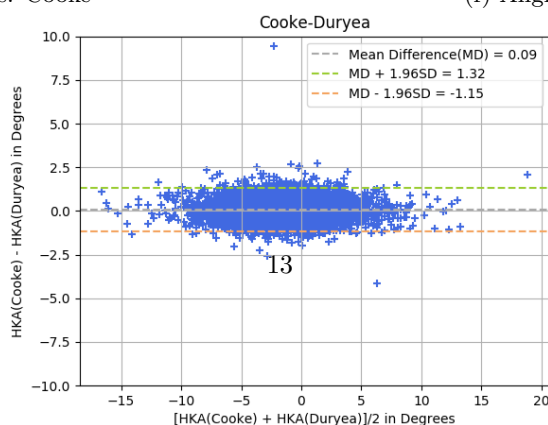
(d) Test data: YARLA vs. Duryea



(e) Angle_OAI data: YARLA vs. Cooke



(f) Angle_OAI data: YARLA vs. Duryea



(g) Angle_OAI data: Cooke vs. Duryea

Figure 4: Bland-Altman plots. For the Angle_OAI data Bland-Altman plots are shown comparing YARLA, Cooke, and Duryea. For the testing data, additionally, YARLA is compared against LM_ZIB.

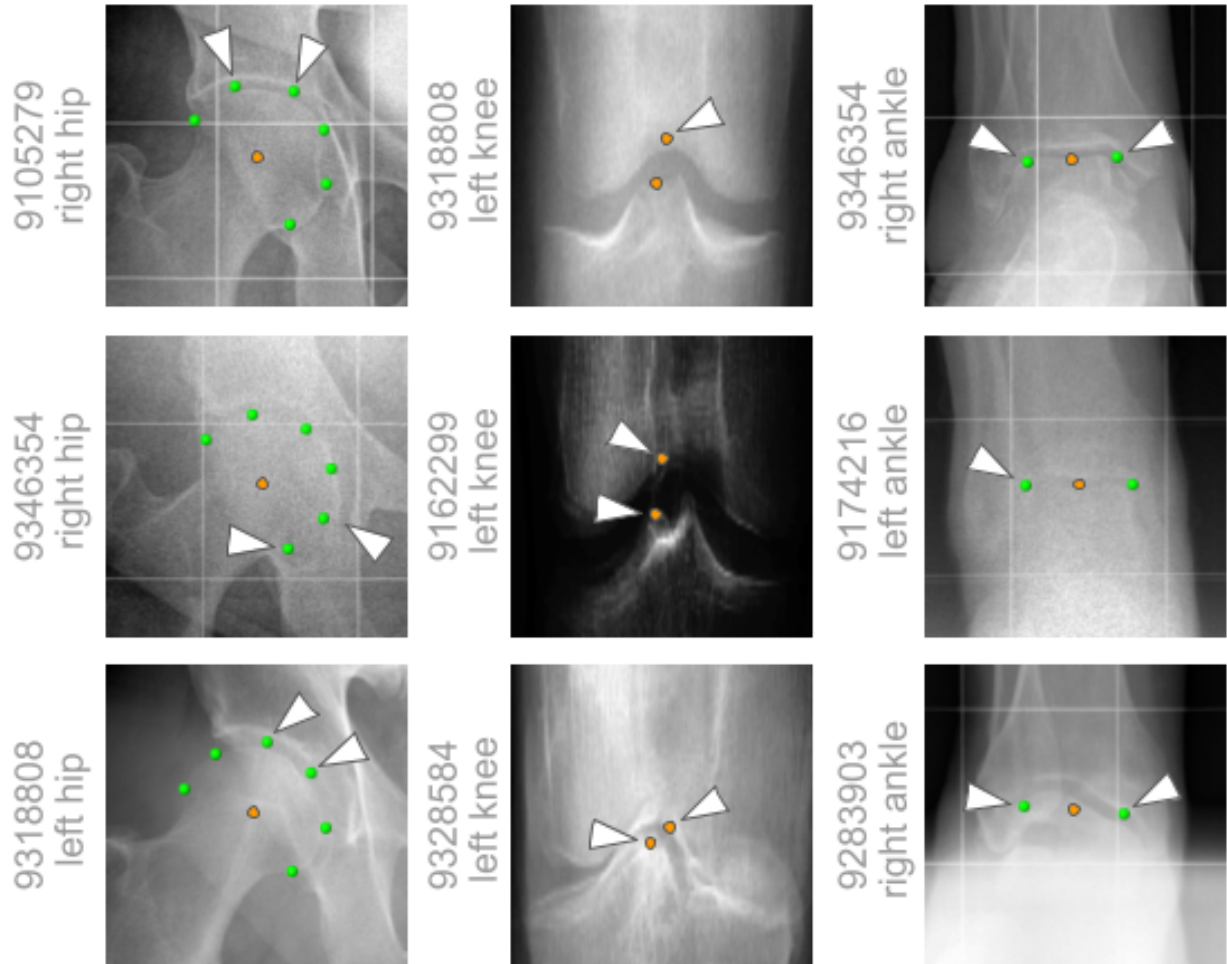


Figure 5: Examples for outliers. The first row contains examples which should actually be solved by YARLA, but show errors (white arrows). The second row shows examples of ROIs with bad image contrast which led to errors in the landmark detection. The last row shows examples in which unnatural shapes of the joint bones led to errors.

Algorithm 1 ResNets for regression of landmark coordinates for full-leg X-Rays

Require: ROIs for hip, knee, and ankle as given by YOLOv4.

```
1: procedure MULTI-LEVEL LANDMARK REGRESSION
2:   for  $ROI$  in [ $ROI_{hip}$ ,  $ROI_{knee}$ ,  $ROI_{ankle}$ ] do
3:      $c_0 \leftarrow$  compute center of  $ROI$ 
4:     // ResNet to regress landmarks  $lms_1 = (x_i, y_i, \dots, x_n, y_n)$ 
5:      $lms_1 \leftarrow$  RESNET( $c_0$ ,  $size_x=170$ ,  $size_y=170$ )
6:      $c_1 \leftarrow$  compute center of  $lms_1$ 
7:      $lms_2 \leftarrow$  RESNET( $c_1$ ,  $size_x=135$ ,  $size_y=135$ )
8:      $c_2 \leftarrow$  compute center of  $lms_2$ 
9:     if  $ROI\_TYPE$  is knee then
10:       $lms_3 \leftarrow$  RESNET( $c_2$ ,  $size_x=100$ ,  $size_y=40$ )
11:     else
12:       $lms_3 \leftarrow$  RESNET( $c_2$ ,  $size_x=100$ ,  $size_y=100$ )
13:     end if
14:   end for
15: end procedure
16: function RESNET( $c$ ,  $size_x$ ,  $size_y$ )
17:    $r \leftarrow$  extract  $size_x$  mm  $\times$   $size_y$  mm region in X-Ray around  $c$ 
18:    $r \leftarrow$  resample  $r$  to  $512 \times 512$  pixels (linear interpolation)
19:    $r \leftarrow$  min-max normalized intensities in  $r$  to  $[0,1]$ 
20:    $lms \leftarrow$  landmark regression by ResNet( $r$ )
21:   return  $lms$ 
22: end function
```

As an additional limitation, YARLA produces a few outliers (see Bland-Altman plots in Fig. 4 and examples shown in Fig. 5), usually due to bad image contrast (Fig. 5 middle row) or unexpected bone shapes (Fig. 5 lower row). Some outliers cannot be explained, though, and would be easily avoidable for a human (Fig. 5 upper row). Whereas more training data could improve the results, methodical changes to the landmark regression network could be considered as well. A different object detector, fully convolutional approaches for landmark detection, or end-to-end approaches could be employed. The end-to-end trained YOLOv4 shows a considerable error in landmark position as well as HKA angle computation. End-to-end trained approaches based on novel transformers [35] could improve the results due to attention mechanisms. However, these approaches require a lot of training data. It is common practice to increase the amount of training data using methods of data augmentation which is challenging for full-leg radiographs since global rotations of the image are not plausible and local deformations might influence the landmark position and thus the resulting HKA angle.

We consider this work a foundation for future artificial intelligence-based diagnosis of knee alignment. A clinical workflow could involve YARLA as a decision support for computer-aided diagnosis. In the first step YARLA computes anatomical landmarks for all joints determining the HKA angle. In the second step quality assurance by medical experts needs to be performed, i.e., confirmation of the landmark locations or modification of the locations via suitable software tools as well as agreement on the diagnosis of knee alignment proposed by YARLA. In future studies, the impact of YARLA could be investigated in terms of reliability and efficiency of knee alignment assessment from full-leg X-Rays. Moreover, a comparison with commercial products should be conducted⁴.

In conclusion, YARLA can be used for diagnosis of knee alignment from full-leg X-Rays, i.e., for computation of the hip-knee-ankle angle based on automated landmark detection. Our manual landmarks as well as the trained networks will be made publicly available with this publication to support future developments as well as an evaluation of the clinical value of YARLA (<https://pubdata.zib.de>).

Data availability

The manually placed landmarks as well as the python code are publicly available at <https://pubdata.zib.de>.

Compliance with ethical standards

Ethical approval: In this study we used data of the OAI database which is publicly available at <https://nda.nih.gov/oai/>. The institutional review board at each institute participated in the OAI study approved the protocol and consent form for the OAI study. Written informed consent was obtained prior to each clinic visit. Authorization for inclusion of the participant’s study data in public release datasets was part of the consent form.

Funding

The authors gratefully acknowledge the financial support by the German Federal Ministry of Education and Research (BMBF) – research network on musculoskeletal diseases, project Overload/PrevOP, grant no. 01EC1408B.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgements

The authors gratefully thank Henok Hagos Gidey for setting the foundation of this work in his Master’s thesis⁵. Also, we thank Henok Hagos Gidey and Loïc Dancelme for the creation of the manual landmarks LM_ZIB.

⁴<https://imagebiopsy.com/msk-platform/ib-lab-lama/>

⁵<https://opus4.kobv.de/opus4-zib/files/7126/HenokGideyMasterThesis.pdf>

The Osteoarthritis Initiative is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

References

- [1] OD Schipplein and TP Andriacchi. Interaction between active and passive knee stabilizers during level walking. *Journal of orthopaedic research*, 9(1):113–119, 1991.
- [2] Leena Sharma, Jing Song, Dorothy Dunlop, David Felson, Cora E Lewis, Neil Segal, James Torner, T Derek V Cooke, Jean Hietpas, John Lynch, et al. Varus and valgus alignment and incident and progressive knee osteoarthritis. *Annals of the rheumatic diseases*, 69(11):1940–1945, 2010.
- [3] Hamza Alizai, Frank W Roemer, Daichi Hayashi, Michel D Crema, David T Felson, and Ali Guermazi. An update on risk factors for cartilage loss in knee osteoarthritis assessed using MRI-based semiquantitative grading methods. *European radiology*, 25(3):883–893, 2015.
- [4] Derek T Cooke, Laurie Harrison, Bashir Khan, Allan Scudamore, and Ashraf M Chaudhary. Analysis of limb alignment in the pathogenesis of osteoarthritis: a comparison of Saudi Arabian and Canadian cases. *Rheumatology international*, 22(4):160–164, 2002.
- [5] T Derek V Cooke, Elizabeth A Sled, and R Allan Scudamore. Frontal plane knee alignment: a call for standardized measurement. *Journal of Rheumatology*, 34(9):1796–1801, 2007.
- [6] R Moyer, W Wirth, J Duryea, and F Eckstein. Anatomical alignment, but not goniometry, predicts femorotibial cartilage loss as well as mechanical alignment: data from the Osteoarthritis Initiative. *Osteoarthritis and cartilage*, 24(2):254–261, 2016.
- [7] TD Cooke, RA Scudamore, JT Bryant, C Sorbie, D Siu, and B Fisher. A quantitative approach to radiography of the lower limb. Principles and applications. *The Journal of Bone and Joint Surgery. British volume*, 73(5):715–720, 1991.
- [8] G Neumann, D Hunter, M Nevitt, LB Chibnik, K Kwok, H Chen, T Harris, S Satterfield, J Duryea, et al. Location specific radiographic joint space width for osteoarthritis progression. *Osteoarthritis and cartilage*, 17(6):761–765, 2009.
- [9] T Iranpour-Boroujeni, J Li, JA Lynch, M Nevitt, J Duryea, OAI Investigators, et al. A new method to measure anatomic knee alignment for large studies of oa: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 22(10):1668–1674, 2014.
- [10] Elizabeth A Sled, Lisa M Sheehy, David T Felson, Patrick A Costigan, Miu Lam, and T Derek V Cooke. Reliability of lower limb alignment measures using an established landmark-based method with a customized computer software program. *Rheumatology international*, 31(1):71–77, 2011.
- [11] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghahfoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

- [12] Willem Paul Gielis, Hassan Rayegan, Vahid Arbabi, Seyed Y Ahmadi Brooghani, Claudia Lindner, Tim F Cootes, Pim A de Jong, H Weinans, and Roel JH Custers. Predicting the mechanical hip–knee–ankle angle accurately from standard knee radiographs: a cross-validation experiment in 100 patients. *Acta orthopaedica*, pages 1–6, 2020.
- [13] Thong Phi Nguyen, Dong-Sik Chae, Sung-Jun Park, Kyung-Yil Kang, Woo-Suk Lee, and Jonghun Yoon. Intelligent analysis of coronal alignment in lower limbs based on radiographic image with convolutional neural network. *Computers in Biology and Medicine*, page 103732, 2020.
- [14] Yun Pei, Wenzhuo Yang, Shangqing Wei, Rui Cai, Jialin Li, Shuxu Guo, Qiang Li, Jincheng Wang, and Xueyan Li. Automated measurement of hip–knee–ankle angle on the unilateral lower limb x-rays using deep learning. *Physical and Engineering Sciences in Medicine*, pages 1–10, 2020.
- [15] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.
- [16] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2020.
- [17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [22] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer, 2016.
- [23] Julia MH Noothout, Bob D de Vos, Jelmer M Wolterink, Tim Leiner, and Ivana Išgum. Cnn-based landmark detection in cardiac cta scans. *arXiv preprint arXiv:1804.04963*, 2018.
- [24] Jupeng Li, Yinghui Wang, Junbo Mao, Gang Li, and Ruohan Ma. End-to-end coordinate regression model with attention-guided mechanism for landmark localization in 3d medical images. In *International Workshop on Machine Learning in Medical Imaging*, pages 624–633. Springer, 2020.
- [25] Hidenori Ide and Takio Kurita. Improvement of learning for cnn with relu activation by sparse regularization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2684–2691. IEEE, 2017.
- [26] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

- [27] Erich Leo Lehmann and Howard J D’Abrera. *Nonparametrics: statistical methods based on ranks*. Holden-day, 1975.
- [28] J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.
- [29] Joseph L Fleiss, Jacob Cohen, and Brian S Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin*, 72(5):323, 1969.
- [30] John J Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.
- [31] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [32] Haldun Akoglu. User’s guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.
- [33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [34] Yuanwei Li, Amir Alansary, Juan J Cerrolaza, Bishesh Khanal, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Fast multiple landmark localisation using a patch-based iterative network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 563–571. Springer, 2018.
- [35] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.