Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research Paper

Multicentric development and validation of a multi-scale and multi-task deep learning model for comprehensive lower extremity alignment analysis

Nikolas J. Wilhelm [a,b,*], Claudio E. von Schacky [d], Felix J. Lindner [c], Matthias J. Feucht [e,g], Yannick Ehmann [c], Jonas Pogorzelski [c], Sami Haddadin [b], Jan Neumann [d], Florian Hinterwimmer [a], Rüdiger von Eisenhart-Rothe [a], Matthias Jung [f], Maximilian F. Russe [f], Kaywan Izadpanah [f], Sebastian Siebenlist [c], Rainer Burgkart [a], Marco-Christopher Rupp [c]

[a] Department of Orthopedics and Sports Orthopedics, Klinikum rechts der Isar, School of Medicine, Munich, Germany
[b] Munich Institute of Robotics and Machine Intelligence, Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany
[c] Department of Orthopedic Sports Medicine , Klinikum rechts der Isar, School of Medicine, Munich, Germany
[d] Department of Radiology, Klinikum rechts der Isar, School of Medicine, Munich, Germany
[e] Department of Orthopedics and Trauma Surgery, Medical Center, Faculty of Medicine, Albert-Ludwigs-University of Freiburg, Freiburg, Germany
[f] Department of Radiology, Medical Center, Faculty of Medicine, Albert-Ludwigs-University of Freiburg, Freiburg, Germany
[g] Orthopedic Clinic Paulinenhilfe, Diakonie-Hospital, Stuttgart, Germany

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Lower extremity
Mechanical alignment
Deep learning
Object detection
Multiscale
Multitask

ABSTRACT

Osteoarthritis of the knee, a widespread cause of knee disability, is commonly treated in orthopedics due to its rising prevalence. Lower extremity misalignment, pivotal in knee injury etiology and management, necessitates comprehensive mechanical alignment evaluation via frequently-requested weight-bearing long leg radiographs (LLR). Despite LLR's routine use, current analysis techniques are error-prone and time-consuming. To address this, we conducted a multicentric study to develop and validate a deep learning (DL) model for fully automated leg alignment assessment on anterior–posterior LLR, targeting enhanced reliability and efficiency. The DL model, developed using 594 patients' LLR and a 60%/10%/30% data split for training, validation, and testing, executed alignment analyses via a multi-step process, employing a detection network and nine specialized networks. It was designed to assess all vital anatomical and mechanical parameters for standard clinical leg deformity analysis and preoperative planning. Accuracy, reliability, and assessment duration were compared with three specialized orthopedic surgeons across two distinct institutional datasets (136 and 143 radiographs). The algorithm exhibited equivalent performance to the surgeons in terms of alignment accuracy (DL: 0.21 ± 0.18° to 1.06 ± 1.3° vs. OS: 0.21 ± 0.16° to 1.72 ± 1.96°), interrater reliability (ICC DL: 0.90 ± 0.05 to 1.0 ± 0.0 vs. ICC OS: 0.90 ± 0.03 to 1.0 ± 0.0), and clinically acceptable accuracy (DL: 53.9%–100% vs OS 30.8%–100%). Further, automated analysis significantly reduced analysis time compared to manual annotation (DL: 22 ± 0.6 s vs. OS; 101.7 ± 7 s, $p \leq 0.01$). By demonstrating that our algorithm not only matches the precision of expert surgeons but also significantly outpaces them in both speed and consistency of measurements, our research underscores a pivotal advancement in harnessing AI to enhance clinical efficiency and decision-making in orthopaedics.

1. Introduction

Lower extremity osteoarthritis (OA) is a prevalent cause of musculoskeletal disability, with a prevalence of 3754.2/100,000 and a 9.3% increase since 1990, contributing to 4.4% of the global health burden [1]. Lower extremity malalignment is a biomechanical condition significantly influencing the etiology of various musculoskeletal lower extremity pathologies, leading to the development of knee joint OA [2,3]. These pathologies encompass (osteo-)chondral defects [4], meniscal tears [5], ligamentous insufficiency [6,7], and patellofemoral instability [8]. Additionally, there is growing evidence that biomechanical optimization of leg alignment can predict clinical outcomes for numerous non-surgical and surgical treatment options, including brace therapy, knee joint preserving surgery [9–13], and knee replacement [14].

A comprehensive preoperative analysis of lower extremity alignment on anterior–posterior (a.p.) long leg radiographs (LLR) is a crucial

[*] Corresponding author at: Department of Orthopedics and Sports Orthopedics, Klinikum rechts der Isar, School of Medicine, Munich, Germany.
E-mail address: nikolas.wilhelm@tum.de (N.J. Wilhelm).

https://doi.org/10.1016/j.artmed.2024.102843
Received 30 May 2023; Received in revised form 11 March 2024; Accepted 11 March 2024
Available online 14 March 2024

clinical exam required for diagnosis, quantification, decision making and surgery planning for lower extremity malalignment [9–14]. Consequently, it is a frequently ordered radiographic exam in musculoskeletal care. Despite the clinical consensus on a standardized approach to alignment assessment [3,15], the complex analysis involving over ten radiographic parameters remains time-consuming and prone to inaccuracies when performed by clinical providers [15–19]. Previously reported intra- and interreader reliabilities for this analysis vary from excellent to poor [15,16,18]. High accuracy is vital for subsequent treatment decisions, as inconsistencies may impact clinical outcomes post-surgery, highlighting the need to address this factor in the increasingly personalized care of lower extremity musculoskeletal injuries.

While in line with current macrotrends, multiple clinical specialties have leveraged the potential of artificial intelligence branches such as machine learning (ML) and deep learning (DL), clinically applicable solutions specific to muscoluskeletal care and orthopedics have been relatively underrepresented; despite musculoskeletal pathologies accounting for a substantial part of the global health burden. However, in the frequently required and tedious but clinically highly relevant task of radiographic lower extremity alignment analysis, there seems significant potential for fully automated machine learning (ML) algorithms, given that reader experience [16,18] and technical support tools [15,17,19] have been identified as factors limiting precision and consistency; in particular as ML algorithms have demonstrated to outperform human raters in comparable clinical use cases [20–23]. The potential of ML algorithms lies in enhancing accuracy, reliability, and expediting preoperative leg alignment analysis. While previous studies have reported high accuracy of DL models in predicting single alignment parameters [24,25], the main limitation is their clinical applicability to only isolated parameters, as a comprehensive alignment analysis as required in the clinical practice of musculoskeletal care providers relies on multiple relevant alignment parameters to generate meaningful results that can be leveraged for clinical decision making [3,15].

From a technical standpoint, algorithms trained to perform clinically relevant analyses on radiographs of the lower extremity used segmentation methods to obtain leg length measurements [26], single lower extremity alignment parameters [24], or hip joint detection [27]. In contrast, other studies used landmark detection methods to obtain alignment analysis of the spine [28] or knee joint [25,29–32]. As such, there is a significant potential for a combination of these approaches to improve the reliability and accuracy of automatic measurements on a radiograph, analogous to ensemble learning in ML [33]. Previous studies have successfully used multitasking methods in similar clinical use cases, resulting in high precision [27]. However, as information technology devices utilized in healthcare facilities typically face limitations by computing power, there is a critical unmet clinical need for lean and effective DL architecture to not compromise on the comprehensivess and accuracy of a lower extremity analysis.

In this study, we present a DL system capable of performing fully automated, comprehensive leg alignment assessment on anterior–posterior (a.p.) LLR. We demonstrate the efficient utilization of knowledge between master and expert networks during evaluation to maximize accuracy, surpassing single-scale approaches while avoiding hardware limitations. Comparing key performance indicators such as accuracy, reliability, and evaluation time with specialized orthopedic surgeons (OS) in a multicenter validation study, we reveal that a fully automated, comprehensive leg alignment analysis based on the DL model achieves clinical-level OS performance. Moreover, the developed DL algorithm considerably outperforms specialized human raters in processing time. These findings underscore the potential of state-of-the-art DL models to augment orthopedic providers' capabilities in managing lower extremity pathologies for high-volume, critical tasks demanding precision and reliability.

2. Material and methods

This Institutional Review Board-approved study (460/21s) was conducted in accordance with institutional privacy policies. Patients that had received radiographic evaluation prior to corrective surgery for lower extremity malalignment at the senior authors' institution between 01/2014 and 01/2021 were retrospectively included. Conventional preoperative weight-bearing a.p. LLRs were required for inclusion. Unconsolidated fractures, metal implants or hardware overlying the contours of the cortical bone, as well as inadequate radiographic quality due to severe malrotation or incomplete visualization of the bony structures were defined as exclusion criteria. The data were split on patient level 60%, 10%, and 30% for training, validation, and hold-out testing, respectively. An additional test dataset with a size equal to the internal dataset was acquired from an external institution (University of Freiburg) to serve as an external validation of the DL model.

2.1. Radiographic acquisition

Two or three preoperative weight-bearing a.p. radiographs were acquired, depending on the height of the patient. The overlapping radiographs were merged to obtain a full LLR. A ruler and a reference sphere served as length reference

2.2. Dataset annotation for image analysis

Landmark segmentation and annotation were performed on all patients' LLRs. An internal validation of the annotation protocol was conducted using 50 randomly selected images reviewed by three experienced OS (M.C.R., J.P., M.J.F.). Labels and segmentations were created by one OS (F.L.) using 3D Slicer (version 4.11, Slicer Community, open source, slicer.org) and verified by a second OS (M.C.R.). Disagreements were resolved by a third OS (M.J.F.). These annotations served as the baseline reference for training. Landmark annotations were made at anatomic locations relevant for orthopedic surgical planning, following deformity analysis principles and preoperative analysis requirements. As depicted in Fig. 1, segmentations and landmark placements were performed to define femoral head center, femoral and tibial anatomic axes, femoral and tibial joint lines, and talar joint line. Segmentation was conducted on the femoral head using a best-fit circle approach, the distal femur excluding non-weight-bearing structures, the entire proximal tibia distinguishing anterior and posterior structures, and the fibular head. The boundaries were defined to include anatomic structures necessary for planning TKA or alignment corrective osteotomy (Fig. 1).

2.3. Advanced deep learning techniques for leg alignment analysis

The multi-level approach adopted for automatic detection and labeling of various landmarks and segmentations in a single image focused on leg anatomy. Fig. 2 illustrates the formulated architecture which partitions the entire leg image anatomically into nine distinct objects (hip, femur trochanter, femur shaft, femur condyles, tibia eminence, tibia joint line, tibia shaft, ankle, and reference sphere), each with segmentation, landmarks, bounding box, and class.

To limit the input area for a single annotation network, we utilize an upstream recognition step. This step identifies sub-objects within the entire image, thereby streamlining the task and bypassing hardware constraints. It is the main recognition network's duty to designate these categorized image regions. This procedure employs a network based on an RCNN ResNet-101, pre-trained with the Microsoft Common Objects in Context dataset [34]. The loss function comprises the aggregate of cross-entropy classification losses $L_{cls,i}$ and intersection over union (IoU) losses $L_{box,i}$ for each object i, as outlined by Girshick [35]. To

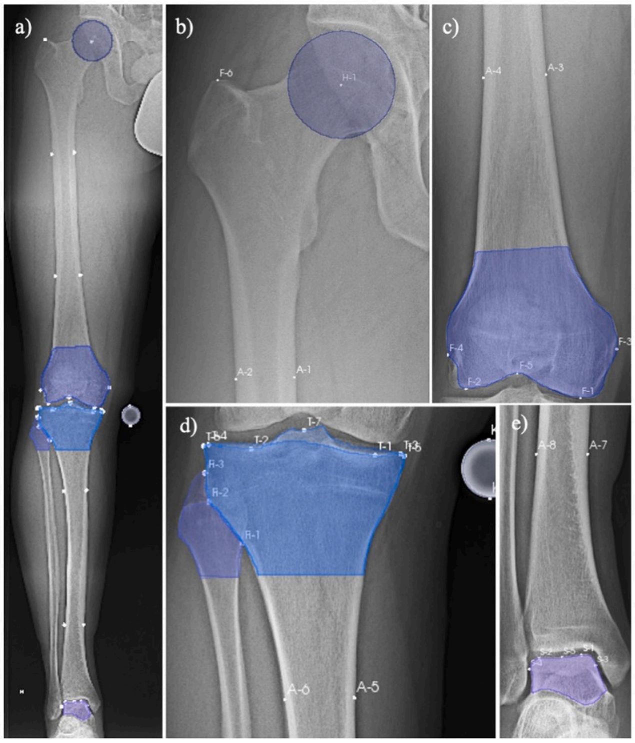

Fig. 1. Annotation and segmentation of clinically significant landmarks on standard anteroposterior hip-knee-ankle radiograph using 3D Slicer (v4.11, Slicer Community, open-source, slicer.org): (a) Radiograph annotations overview. (b) Hip region: H-1 marks femoral head center; F-6 indicates greater trochanter tip; segmentation covers largest femoral head diameter. (c) Distal femur: F-1, F-2 establish joint line; F-3, F-4 delineate condyles borders; F-5 identifies notch center; segmentation covers weight-bearing distal femur. (d) Proximal tibia and fibula detail: T-3, T-4 define proximal tibial joint line; T-5, T-6 indicate medial and lateral proximal tibia borders; T-5 marks intercondylar eminence midpoint; K-1, K-2 represent 25-mm reference body diameter; segmentation includes entire proximal tibia and fibula, differentiating anterior and posterior structures. (e) Ankle joint: S-5 denotes talar surface center; S-3, S-4 indicate talus borders; joint line established by S-1, S-2; segmentation involves proximal talus. Points A-1 through A-8 identify cortical borders at various elevations, delineating anatomic axes.

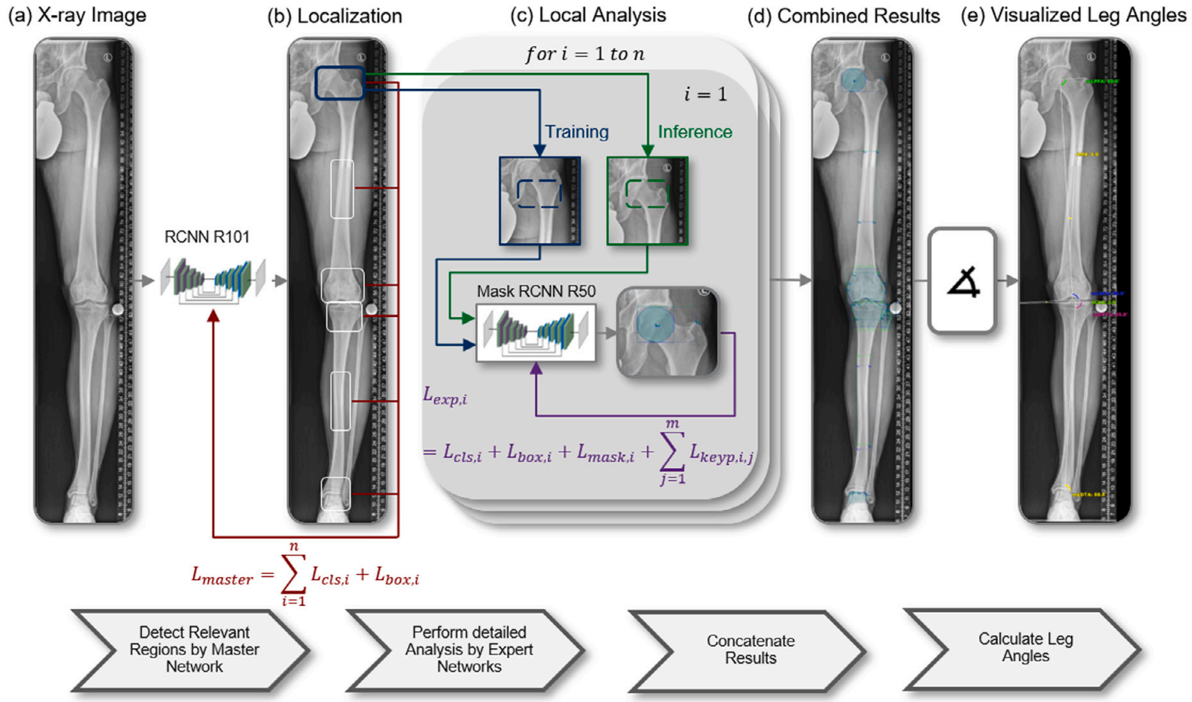
(a) X-ray Image
(b) Localization
(c) Local Analysis
(d) Combined Results
(e) Visualized Leg Angles
RCNN R101
$for\ i = 1\ to\ n$
$i = 1$
Training
Inference
Mask RCNN R50
$L_{exp,i}$
$= L_{cls,i} + L_{box,i} + L_{mask,i} + \sum_{j=1}^{m} L_{keyp,i,j}$
$L_{master} = \sum_{i=1}^{n} L_{cls,i} + L_{box,i}$
Detect Relevant Regions by Master Network
Perform detailed Analysis by Expert Networks
Concatenate Results
Calculate Leg Angles

Fig. 2. Comprehensive lower extremity alignment analysis via deep learning algorithm: (a) Utilization of standard weight-bearing anteroposterior radiographs for hip, knee, and ankle evaluation. (b) Detection network processes radiographs, downscaled for memory conservation and optimizing $L_{cls,i}$ and $L_{box,i}$ for the network heads during training, while specialized networks examine detected regions at full resolution. (c) Prior to employing expert networks, the image area optimized for inference aligns the object with the average location in the training dataset. Expert network training is further extended by segmentation ($L_{mask,i}$) and landmark identification ($\sum_{j=1}^{m} L_{keyp,i,j}$) for the network heads. (d) Projection of expert networks' data onto the hip-knee-ankle radiograph. (e) Result: Automated leg alignment assessment on radiographs.

ensure the GPU memory usage stays below 2 GB, we downscale the images prior to processing them through the main network.

Each image region is then directed to expert networks based on their category, utilizing the original full image resolution. The primary image analysis, comprising landmark placement, bounding boxes, segmentation, and classification, is conducted by the expert multitasking network. All network heads are assigned equal weightage for loss functions, considering their equal impact on the backbone [34]. Each specialized network employs a Mask-R CNN ResNet-50 architecture pre-trained with the COCO dataset and fine-tuned with the training dataset for 30,000 iterations, with implementations based on PyTorch 1.12 and the detectron2 v0.6 library [36].

In order to optimize the proposed pipeline's performance, an adaptive method is introduced. This leverages the master network's knowledge and employs distinct strategies for training and inference as depicted in Fig. 2(c). During training, random segments of the entire X-ray image are selected to ensure variance in image representation while maintaining sufficient object presence and minimizing utilized image area. Therefore each expert network is trained on an individually created training dataset.

During inference, the main detection network's positional information is leveraged to minimize the task complexity faced by the expert network. This is accomplished by positioning the detected object at the location where it was on average in the training dataset used by the individually trained expert network. If the main detection network fails to detect the object during inference, the object's position is roughly predetermined using the statistical data from the main network's training dataset, and the expert network is tasked with annotation within this expanded search area.

The final result of the DL-based image analysis is obtained by projecting the integrated annotations from specialized networks, along with the main detector network's positional information, onto the complete X-ray image (see Fig. 2(d)). In-depth information on each network's implementation can be accessed through the online training parameters and source code: https://github.com/NikonPic/AlignmentNet and in Appendix A. The obtained results form the foundation for subsequent post-processing steps aimed at enhancing overall performance, encompassing segmentation, bounding box, class, and landmark locations of each object identified by the DL algorithm.

2.4. Optimization of landmark accuracy through annotation integration and local edge filtering

To enhance the precision of angle calculations, we introduce a process that integrates additional intermediate steps, tailored to the specific anatomy. This approach merges information from bounding boxes, segmentation, and landmarks to augment the result's accuracy. For instance, landmarks of the tibial and femoral condyles and the ankle joint are carefully positioned on the predefined segmentation edge. Through the consolidation of segmentation and detected landmarks, the accuracy of the final landmarks can be amplified. We calculate these final landmarks as follows:

$$\mathrm{mark}_{\mathrm{final}} = k \cdot \mathrm{mark}_{\mathrm{orig}} + (1-k) \cdot \mathrm{mark}_{\mathrm{seg}\perp}. \tag{1}$$

Where $\mathrm{mark}_{\mathrm{final}}$ represents the landmark defined for angle analysis, $\mathrm{mark}_{\mathrm{orig}}$ is the original landmark as defined by the specialist network, and $\mathrm{mark}_{\mathrm{seg}\perp}$ is the perpendicular projection of the original landmark onto the segmentation. The coefficient k, determined to be 0.6, is the optimally chosen parameter that minimizes angular loss in the validation dataset. This value was empirically found to provide the most accurate balance between the initial and projected landmark positions for precise angle analysis.

To further optimize the accuracy of critical landmarks, specifically those situated on the convex contour of the femoral condyles (F-1 and F-2), we employ a local edge filter. These landmarks are especially prone to errors but are essential for determining clinical parameters like JLCA and mLDFA. We necessitate the use of an edge filter due to the conservative behavior of the landmark detector during training, which is driven by the loss function for landmark placement. The application of the local edge filter is visualized in Fig. 3.

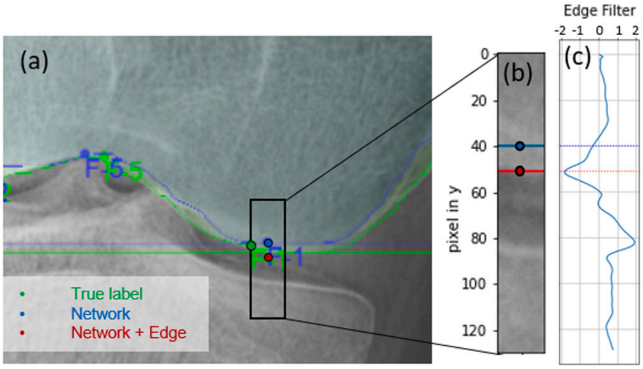

Fig. 3. Local edge filter for accurate anatomical localization of F-1 on the femoral condyle: (a) Relevant X-ray image section displaying physician's annotations in green, network's annotations in blue, and redefined F-1 point using local edge filter in red. (b) Image section serving as input for the edge filter. (c) Edge filter output, illustrating network markers in blue and the minimum in red. The redefined red point, while farther from the OS's annotation, provides superior alignment parameter calculation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The network's label serves as an anchor for the local edge analysis. A window expands around this area (as seen in Fig. 3(a) black box or (b)). The intensity of the selected area is averaged, converting the image into a one-dimensional intensity vector as:

$$\mathrm{intensity}_y = \sum_{x=-10}^{10} \mathrm{intensity}_{x,y} \tag{2}$$

By applying a digital zero-phase filter based on [37] to the acquired intensity and identifying the filtered signal's minima, we ascertain and apply the new y-position for the landmark to redefine the F-1 or F-2 point, as shown in Fig. 3(c).

2.5. Conducted ablation studies

To optimize the DL model's accuracy, the presented methods were evaluated and compared in detail. Initially, various DL architectures were assessed on the internal test dataset. The selected benchmarks included:

1. Single-scale approach (SC) that directly evaluates entire images without a localization step.
2. Multiscale approach (MS) that incorporates the master network but centers the local object in image subregions.
3. Proposed approach combining multiscale approach and simulated training scenario for the expert network (MS-TrainSim).

These models were compared with the annotations of OS1 on the internal test dataset, particularly analyzing and comparing the root mean square error (RMSE) of landmarks across all objects. The architecture exhibiting the lowest RMSE in landmark detection served as the foundation for further analysis.

A subsequent ablation study assessed the advantages of anatomical landmark optimization filters. The DL model landmarks were compared with the combination of the DL model with local edge filters for points F-1 and F-2. Landmark accuracy and angular accuracy for mLDFA and JLCA were compared, and the approach with the lowest error for JLCA and mLDFA was utilized for further analysis.

2.6. Lower extremity alignment analysis

The final automatic landmark detection algorithm generated numerical output for lower limb alignment, including relevant parameters such as mLPFA, mLDFA, JLCA, mMPTA, mLDTA, AMA, mFA-mTA, and limb length [3,38].

The mLPFA is the lateral angle between a line from the femoral head's center to the greater trochanter's tip and the femur's mechanical axis. The mLDFA is the lateral angle between the distal femoral knee joint line and the femur's mechanical axis. The JLCA is the angle between the distal femoral knee joint line and the proximal tibial knee joint line. The mMPTA is the medial angle between the proximal tibial knee joint line and the tibia's mechanical axis. The mLDTA is the lateral angle between the tibia's mechanical axis and the talar joint line. The AMA is the angle between the mechanical and anatomic axes of the femur. The mechanical tibiofemoral angle is the angle between the femur's and tibia's mechanical axes. MAD is the distance between the Mikulicz line and the tibial joint line's center, while limb length is determined using the Mikulicz line's length.

The DL model was trained to recognize and measure the reference sphere's diameter for accurate length calibration. If the reference sphere was absent, the ruler was used for calibration using the latest OCR version by Ooms [39]. The developed model's final workflow for an example radiograph is shown in Fig. 4. Further details on calculating each angle based on the determined landmarks can be found in the GitHub repository https://github.com/NikonPic/AlignmentNet.

2.7. Evaluation of performance

To evaluate the performance of the fully automated alignment analysis, reference measurements were performed by three different experienced human raters for both the internal and external test data sets.

For the internal dataset, manual landmark annotation was performed using 3D Slicer (version 4.11, Slicer Community, open source, slicer.org), and subsequent alignment analysis parameters were calculated using the principles of deformity analysis [3] as clinically required for the [38] planning method. In addition, a comprehensive leg alignment analysis of OS was performed using the state-of-the-art US Food and Drug Administration (FDA)-approved mediCAD orthopedic planning software (version mediCAD Classic, Knee 2D, version 6.0; Hectec GmbH) according to the software instructions to obtain the outcome parameters generated by the DL algorithm. While OS1 (J.P.) and OS2 (Y.J.E.) were two specialist OS employed in the specialized tertiary lower extremity deformity correction department at the senior author's institute that performed measurements for the purpose of this study, OS3 reflects the attending-level surgeon approved measurements that were utilized for the preoperative lower alignment analysis.

To evaluate inter- and intrarater reliability, segmentation and landmark detection and reference measurement by OS1 and OS2 were performed twice at four-week intervals in 30 randomly selected patients.

2.8. Performance metrics

The fully automated landmark detection accuracy was gauged using the root mean square error (RMSE) and Sørensen-Dice coefficient as per [40]. This assessment compared bounding box placement and segmentation accuracy between manual annotations by an orthopedic surgeon (OS1) and our deep learning (DL) model on the internal test dataset.

For final performance evaluation, the DL model's clinical parameters were contrasted with the average of human measurements (OS1, OS2, and OS3) on both internal and external test datasets. Discrepancies between the DL model's predictions and these ground truth measurements were quantified to assess alignment parameter accuracy.

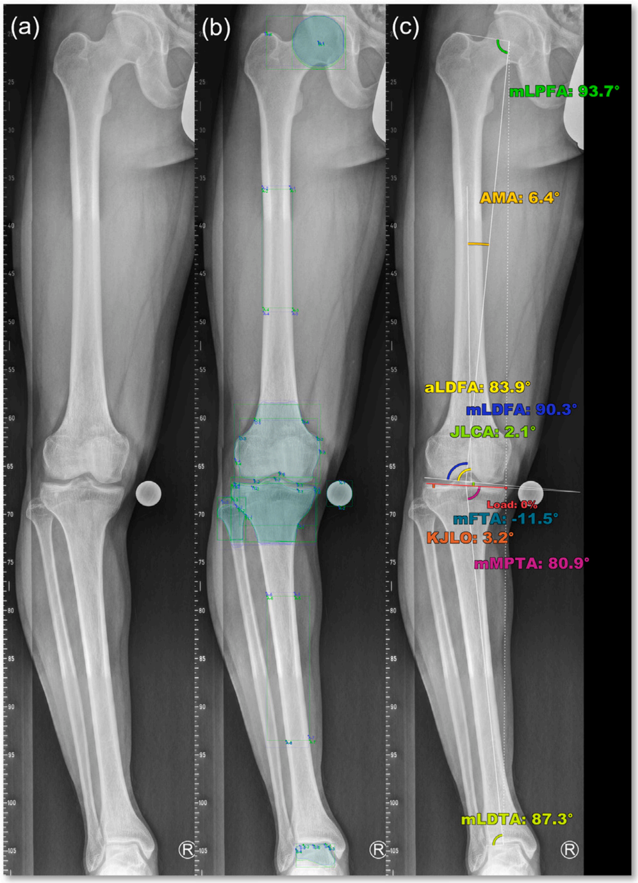

Fig. 4. Detailed algorithm workflow: (a) Analysis conducted on a standard weight-bearing anterior–posterior hip-knee-ankle radiograph. (b) DL model's predicted annotations and segmentation in blue, with ground truth annotations in green for comparison. (c) Post-processing visualization of alignment parameters derived from landmark annotations for clinical application. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Interrater reliability was examined by comparing intraclass coefficients (ICC) between individual human raters and between the ground truth and DL model parameters (ICC3K; average fixed rater). Intrarater reliability, for 30 cases in the external dataset, involved repeating measurements by OS1, OS2, and the DL model after a 30-day period, with ICC comparisons made between each rater's repeated measurements.

Finally, to gauge the clinical reliability and safety of the DL algorithm, we compared the percentage of DL algorithm parameters within a clinically acceptable range against human raters (OS1 vs. OS2; OS1 vs. OS3; OS2 vs. OS3). Clinically acceptable error thresholds were defined as over 2° for angle measurements [15,41], 2 mm for MAD, and 5 mm for limb length [41,42].

2.9. Statistical analysis

Statistical analysis was performed using the python pingouin package [43]. Categorical variables are reported as counts and percentages, while continuous variables are reported as mean ± standard deviation. To evaluate the accuracy of landmark annotation between OS and the DL system, the root mean square error (RMSE) was calculated. To evaluate the accuracy of segmentations between OS and the DL system, the root mean square error was calculated via the Sørensen-Dice coefficient. For reliability assessment, ICC values of ≥0.9 were defined as excellent, ≥0.75 as good, ≥0.5 as moderate, and ≤0.5 as poor [44]. The normality of continuous variables was assessed using the Shapiro–Wilk test. According to the respective distribution, continuous

Table 1
Patient characteristics. Continuous variables are presented as mean ± standard deviation (range); Categorical variables are presented as count and percentage.

	Overall (n = 594)	Train (n = 399)	Valid (n = 59)	Test (n = 136)	External (n = 143)
Age [years]	41.1 ± 13.2	41.6 ± 13.0	46.8 ± 12.1	37.4 ± 13.4	40.7 ± 11.4
Left	388 (65.3%)	296 (74.2%)	0 (0.0%)	92 (67.6%)	72 (50.4%)
Female	182 (30.6%)	117 (29.3%)	16 (27.1%)	49 (36.0%)	36 (25.2%)

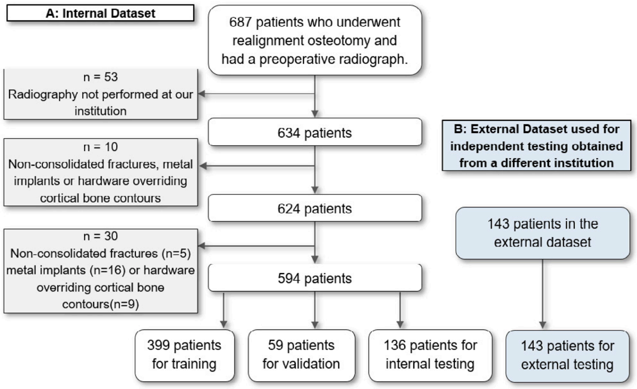
A: Internal Dataset
687 patients who underwent realignment osteotomy and had a preoperative radiograph.
n = 53
Radiography not performed at our institution
634 patients
n = 10
Non-consolidated fractures, metal implants or hardware overriding cortical bone contours
624 patients
n = 30
Non-consolidated fractures (n=5) metal implants (n=16) or hardware overriding cortical bone contours(n=9)
594 patients
B: External Dataset used for independent testing obtained from a different institution
143 patients in the external dataset
399 patients for training
59 patients for validation
136 patients for internal testing
143 patients for external testing

Fig. 5. Flowchart illustrating the patient population for this study after considering inclusion criteria, exclusion criteria for the internal (A) and external dataset (B).

variables were compared using t-tests or Mann–Whitney U tests. The significance level was set at $p < 0.05$.

3. Results

Between 01/2014 and 01/2021, review of the institutional database identified 687 patients who had a preoperative weight-bearing a.p. LLR. Accounting for inclusion and exclusion criteria (Fig. 5), 594 the remaining patients (mean age 41.1 ± 13.2 years, 182 females, 388 left leg) were included. The demographic characteristics of the study population are summarized in Table 1. At the senior author's institution, a reference sphere was present on 580 radiographs, whereas a ruler was present on the remaining 14 radiographs. Accordingly, patient radiographs were randomly divided into training (n = 399, 60%), validation (n = 59, 10%), and test (n = 136, 30%) data sets.

3.1. Performance in landmark detection and segmentation

To calculate all relevant angles, the DL model must first identify the necessary anatomical structures and landmarks, learning to detect these from the ground truth annotations provided by the expert. High agreement was observed between the predicted landmarks by the DL model and the annotated landmarks, enabling accurate angle calculation. Table 2 presents the detailed results for the accuracy of each individual object and landmark. For length measurements, a reference object, such as a reference sphere or a ruler, must be identified within the X-ray. The algorithm successfully detected the reference sphere in 100% (132 of 132) of cases where it was present and performed measurements on the ruler in 100% (4 of 4) of the remaining cases.

The mean RMSE across all landmarks of each object ranged from 0.48 mm ± 1.0 mm (Sphere) to 7.1 mm ± 4.55 mm ($\text{femur}_{\text{shaft}}$). The Sørensen-Dice coefficient for bounding box placement ranged from 0.92 ± 0.04 ($\text{tibia}_{\text{shaft}}$) to 0.97 ± 0.01 ($\text{femur}_{\text{trochanter}}$) and for segmentation between 0.89 ± 0.2 (ankle) and 0.97 ± 0.01 ($\text{femur}_{\text{trochanter}}$).

Table 2
Root Mean Square Errors (RMSE) in mm of the averaged landmark detections and Dice Score Coefficient for Bounding Box placement (Dice BBox) and Segmentation (Dice Seg) for the individual objects on the internal test dataset. Continuous variables are presented as mean ± standard deviation (range) and data listed in brackets are 95% CIs.

	RMSE	Dice BBox	Dice Seg
Hip	0.6 ± 0.3	0.97 [0.97,0.97]	0.97 [0.97,0.97]
$\text{Fem}_{\text{troch}}$	0.9 ± 0.5	0.97 [0.97,0.97]	0.97 [0.97,0.97]
$\text{Fem}_{\text{shaft}}$	7.0 ± 4.6	0.90 [0.89,0.91]	–
Fem_{cond}	1.9 ± 2.7	0.96 [0.96,0.96]	0.97 [0.97,0.97]
Tib_{emin}	1.1 ± 1.1	0.96 [0.96,0.96]	0.95 [0.93,0.97]
Tib_{jl}	1.4 ± 1.4	0.96 [0.96,0.96]	0.96 [0.96,0.96]
$\text{Tib}_{\text{shaft}}$	5.5 ± 3.5	0.92 [0.91,0.93]	–
Ankle	0.8 ± 1.6	0.93 [0.91,0.95]	0.89 [0.89,0.89]
Sphere	0.5 ± 1.0	0.95 [0.92,0.98]	0.95 [0.92,0.98]

Table 3
Assessment accuracy of alignment parameters measured by the mean deviation from the average OS rating for the internal and external test dataset. The highest accuracy for each parameter is highlighted in **bold**. Continuous variables are presented as mean ± standard deviation. TP, tibial plateau.

Internal test dataset (n = 136)				
	OS1-OS2	OS1-OS3	OS2-OS3	$\text{AI-OS}_{\text{mean}}$
Leg length [mm]	**6.91 ± 7.99**	–	–	9.31 ± 9.22
Load on TP [%]	2.58 ± 12.36	**0.76 ± 0.71**	1.1 ± 1.72	1.73 ± 6.18
MAD [mm]	2.35 ± 1.33	1.66 ± 1.17	1.32 ± 1.86	**1.16 ± 0.68**
mLPFA [°]	0.82 ± 1.13	1.47 ± 1.23	1.38 ± 1.28	**0.69 ± 0.71**
AMA [°]	**0.29 ± 0.27**	0.34 ± 0.22	0.32 ± 0.26	0.36 ± 0.52
mLDFA [°]	**0.48 ± 0.56**	0.66 ± 0.61	0.58 ± 0.55	0.5 ± 0.71
JLCA [°]	**0.56 ± 0.56**	0.83 ± 0.74	0.69 ± 0.62	0.76 ± 0.93
mMPTA [°]	**0.39 ± 0.39**	0.88 ± 1.7	0.91 ± 1.65	0.67 ± 0.61
mFTA [°]	**0.13 ± 0.14**	0.32 ± 0.25	0.32 ± 0.24	0.16 ± 0.13
KJLO [°]	**0.45 ± 0.76**	–	–	0.65 ± 0.63
mLDTA [°]	**0.5 ± 0.51**	0.77 ± 0.69	0.53 ± 0.45	0.94 ± 0.86
External test dataset (n = 143)				
	OS1-OS2	OS1-OS3	OS2-OS3	$\text{AI-OS}_{\text{mean}}$
Leg length [mm]	**4.24 ± 2.89**	15.38 ± 27.47	16.53 ± 27.7	7.58 ± 9.31
Load on TP [%]	**0.82 ± 0.65**	1.22 ± 1.59	1.37 ± 1.69	1.1 ± 1.1
MAD [mm]	**0.93 ± 0.72**	1.58 ± 3.01	1.76 ± 3.04	1.86 ± 1.38
mLPFA [°]	**0.86 ± 0.72**	1.69 ± 1.92	1.72 ± 1.96	1.06 ± 1.3
AMA [°]	0.28 ± 0.21	**0.23 ± 0.19**	0.32 ± 0.25	0.35 ± 0.37
mLDFA [°]	**0.45 ± 0.74**	0.48 ± 0.53	0.55 ± 0.76	0.45 ± 0.52
JLCA [°]	**0.48 ± 0.71**	0.74 ± 0.61	0.83 ± 0.86	0.59 ± 0.59
mMPTA [°]	**0.37 ± 0.31**	0.68 ± 0.78	0.74 ± 0.81	0.46 ± 0.57
mFTA [°]	**0.21 ± 0.16**	0.23 ± 0.26	0.25 ± 0.29	0.21 ± 0.18
KJLO [°]	**0.42 ± 0.51**	–	–	0.42 ± 0.59
mLDTA [°]	**0.48 ± 0.46**	0.91 ± 1.1	0.95 ± 1.06	0.95 ± 1.13

3.2. Performance of DL algorithm compared to manual reference measurements

The DL model calculates the relevant alignment parameters based on the definitions provided in Section 2.6 and underwent both internal and external testing. For the internal dataset, the mean differences in the angular alignment parameters between the OS ranged from 0.13° ± 0.14° (mFTA, OS1-OS2) to 1.47 ± 1.23 (mLPFA, OS1-OS3). In comparison, the mean differences between the DL model and the mean measurement of the OS ranged from 0.16° ± 0.14° (mFTA) to 0.94° ± 0.86° (mLDTA).

In the validation of the performance on an external test dataset, the mean differences in the angular alignment parameters assessed between the OS ranged from 0.21° ± 0.16° (mFTA, OS1-OS2) to 1.72° ± 1.96° (mLPFA, OS2-OS3). In the comparison, mean differences between the DL model and the mean measurement of the OS, these ranged from 0.21° ± 0.18° (mFTA) to 1.06° ± 1.3° (mLPFA). Detailed information on the performance is presented in Table 3.

Table 4
Interreader reliability as quantified by intraclass correlation (ICC) values on the internal and external test dataset. Data listed in brackets are 95% CIs. TP, tibial plateau.

Internal test dataset (n = 136)				
	OS1-OS2	OS1-OS3	OS2-OS3	AI-OS$_{mean}$
Leg length	**0.99 [0.99, 0.99]**	–	–	0.99 [0.97, 0.99]
Load on TP	0.92 [0.89, 0.95]	**1.0 [0.99, 1.0]**	0.99 [0.98, 1.0]	0.98 [0.97, 0.99]
MAD	0.99 [0.8, 1.0]	1.0 [0.95, 1.0]	**1.0 [0.99, 1.0]**	1.0 [0.98, 1.0]
mLPFA	0.98 [0.98, 0.99]	0.95 [0.74, 0.98]	0.95 [0.79, 0.98]	**0.99 [0.99, 0.99]**
AMA	**0.97 [0.95, 0.98]**	0.95 [0.68, 0.98]	0.95 [0.62, 0.99]	0.93 [0.9, 0.95]
mLDFA	**0.99 [0.98, 0.99]**	0.96 [0.93, 0.97]	0.96 [0.94, 0.98]	0.98 [0.97, 0.99]
JLCA	**0.92 [0.87, 0.95]**	0.87 [0.79, 0.92]	0.92 [0.87, 0.95]	0.73 [0.62, 0.81]
mMPTA	**0.99 [0.99, 0.99]**	0.81 [0.7, 0.88]	0.82 [0.71, 0.88]	0.97 [0.96, 0.98]
mFTA	**1.0 [1.0, 1.0]**	1.0 [0.98, 1.0]	1.0 [0.99, 1.0]	**1.0 [1.0, 1.0]**
KJLO	**0.96 [0.94, 0.97]**	–	–	0.95 [0.92, 0.97]
mLDTA	**0.99 [0.99, 0.99]**	0.98 [0.95, 0.99]	0.99 [0.98, 1.0]	0.97 [0.96, 0.98]
External test dataset (n = 143)				
	OS1-OS2	OS1-OS3	OS2-OS3	AI-OS$_{mean}$
Leg length	**1.0 [1.0, 1.0]**	0.92 [0.87, 0.95]	0.92 [0.86, 0.95]	0.99 [0.97, 0.99]
Load on TP	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**
MAD	**1.0 [1.0, 1.0]**	0.99 [0.99, 0.99]	0.99 [0.99, 0.99]	1.0 [0.95, 1.0]
mLPFA	**0.99 [0.99, 0.99]**	0.95 [0.89, 0.97]	0.95 [0.91, 0.97]	0.98 [0.97, 0.98]
AMA	0.97 [0.94, 0.98]	**0.98 [0.97, 0.98]**	0.96 [0.9, 0.98]	0.94 [0.91, 0.96]
mLDFA	0.97 [0.95, 0.98]	**0.98 [0.97, 0.99]**	0.96 [0.95, 0.97]	0.98 [0.97, 0.99]
JLCA	**0.9 [0.87, 0.93]**	0.86 [0.8, 0.9]	0.79 [0.7, 0.85]	0.9 [0.85, 0.93]
mMPTA	**0.99 [0.99, 1.0]**	0.97 [0.95, 0.98]	0.96 [0.94, 0.97]	0.98 [0.98, 0.99]
mFTA	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**
KJLO	**0.97 [0.96, 0.98]**	–	–	0.96 [0.95, 0.97]
mLDTA	**0.99 [0.99, 1.0]**	0.97 [0.96, 0.98]	0.97 [0.96, 0.98]	0.97 [0.95, 0.98]

3.3. Interreader reliability

Details on the interreader reliability of the internal and external test datasets are provided in Table 4. For the internal test dataset, interreader reliability was good (0.81) to excellent (1.0) between OS for all angular measurements, while it was excellent (0.92–0.99) for measurements involving absolute distances. Comparable to human performance, interreader reliability between the DL model and ground truth measurements of the OS was moderate (0.73) to excellent (1.0) for all angular measurements, while it was excellent (0.98–0.99) for measurements involving absolute distances.

For the external test dataset, interreader reliability was moderate (0.79) to excellent (1.0) between OS for all angular measurements, while it was excellent (0.92–1.0) for measurements involving absolute distances. Similar to human performance, interreader reliability between the DL model and ground truth measurements of the OS was excellent (0.9–1.0) for all angular parameters, and excellent (0.99–1.0) for measurements involving absolute distances.

3.4. Intrarater reliability

To determine intrarater reliability, the measurement was performed twice with Medicad on the external dataset for n = 30 cases and compared. The same cases were also repeated with the algorithm, which always gave the same results due to its deterministic behavior. Intratrater reliability results are shown in Table 5.

3.5. Determination of clinically acceptable accuracy

Detailed results on the clinically acceptable accuracy are presented in Table 6.

In the internal test dataset, the rate of clinically acceptable agreement between OS1, OS2 and OS3 in the assessment of the respective alignment parameters ranged between 13.6% and 100% of the cases. The rate of clinically acceptable agreement between the DL model and the ground truth measurements by OS1, OS2 and OS3 in the assessment of the respective alignment parameters ranged between 32.8% and 100% of the cases.

Table 5
Intrarater reliability as quantified by intraclass correlation (ICC) values on the internal test dataset for OS1 and OS2. Data listed in brackets are 95% CIs. The AI outperforms both OS, due to her deterministic behavior and scores **1.0 [1.0, 1.0]** for all parameters. TP, tibial plateau.

	OS1-OS1	OS2-OS2
Leg length	**1.0 [1.0, 1.0]**	**1.0 [1.0, 1.0]**
Load on TP	0.99 [0.98, 0.99]	0.95 [0.9, 0.98]
MAD	**1.0 [0.99, 1.0]**	0.99 [0.92, 1.0]
mLPFA	0.97 [0.93, 0.98]	0.98 [0.96, 0.99]
AMA	0.98 [0.92, 0.99]	0.95 [0.49, 0.99]
mLDFA	0.98 [0.97, 0.99]	0.99 [0.96, 0.99]
JLCA	0.94 [0.87, 0.97]	0.83 [0.65, 0.92]
mMPTA	0.98 [0.97, 0.99]	0.98 [0.96, 0.99]
mFTA	1.0 [0.99, 1.0]	0.99 [0.93, 1.0]
KJLO	0.99 [0.97, 0.99]	0.97 [0.95, 0.99]
mLDTA	0.95 [0.89, 0.98]	0.99 [0.98, 1.0]

In the external test dataset, the rate of clinically acceptable agreement between OS1, OS2 and OS3 in the assessment of the respective alignment parameters ranged between 31% and 100% of the cases. The rate of clinically acceptable agreement between the DL model and the ground truth measurements by OS1, OS2 and OS3 in the assessment of the respective alignment parameters ranged between 54% and 100% of the cases.

3.6. Analysis of cases with worst performance

The cases with the worst performance on the external data set compared with the surgeons' averaged measurements (OS$_{mean}$) are shown in Fig. 6. In case (a), the mLPFA is misdetermined by 8.5° (97.8° DL versus 89.3° OS$_{mean}$). In case (b), the angle of the mLDFA is too low by 2.9° and the JLCA by 5.8°. In case (c), the AMA is too low by 3.2° compared with the OS. Similarly, in case (d), the KJLO and mMPTA are too high by 6.1° and 6°, respectively. Finally, in case (e), the mLDTA is too high by 10.6°.

3.7. Comparison of time required for analysis

Mean time required of human raters for a full comprehensive analysis of the alignment using MediCAD version 6.0 was 103 ± 8 s for

Table 6
Clinically acceptable accuracy according to clinically relevant tolerance margins on the internal and external test dataset. Values represent the percentage of individual cases, clinically acceptable agreement was achieved. TP, tibial plateau.

Internal test dataset				
	OS1-OS2	OS1-OS3	OS2-OS3	AI-OS$_{mean}$
Leg length (tol = 5 mm)	**56.0**	–	–	32.8
Load on TP (tol = 2%)	**94.4**	15.2	14.4	90.4
MAD (tol = 2 mm)	41.6	20.0	26.4	**87.2**
mLPFA (tol = 2°)	93.6	13.6	13.6	**96.0**
AMA (tol = 2°)	**99.2**	16.8	16.8	98.4
mLDFA (tol = 2°)	**97.6**	55.2	56.8	96.8
JLCA (tol = 2°)	82.4	53.6	50.4	**91.2**
mMPTA (tol = 2°)	**100.0**	56.8	56.0	96.0
mFTA (tol = 2°)	**100.0**	59.2	59.2	**100.0**
KJLO (tol = 2°)	**97.6**	–	–	94.4
mLDTA (tol = 2°)	**96.8**	15.2	16.8	88.8
External test dataset				
	OS1-OS2	OS1-OS3	OS2-OS3	AI-OS$_{mean}$
Leg length (tol = 5 mm)	**68.5**	31.5	30.8	53.9
Load on TP (tol = 2%)	**95.1**	72.7	69.9	86.0
MAD (tol = 2 mm)	**98.6**	73.4	74.1	62.9
mLPFA (tol = 2°)	**90.9**	71.3	74.8	88.8
AMA (tol = 2°)	**100.0**	95.8	95.8	99.3
mLDFA (tol = 2°)	**97.9**	94.4	94.4	97.2
JLCA (tol = 2°)	97.9	93.0	92.3	**98.6**
mMPTA (tol = 2°)	**100.0**	93.0	93.0	98.6
mFTA (tol = 2°)	**100.0**	96.5	95.8	**100.0**
KJLO (tol = 2°)	**98.6**	–	–	97.2
mLDTA (tol = 2°)	**99.3**	86.7	86.7	89.5

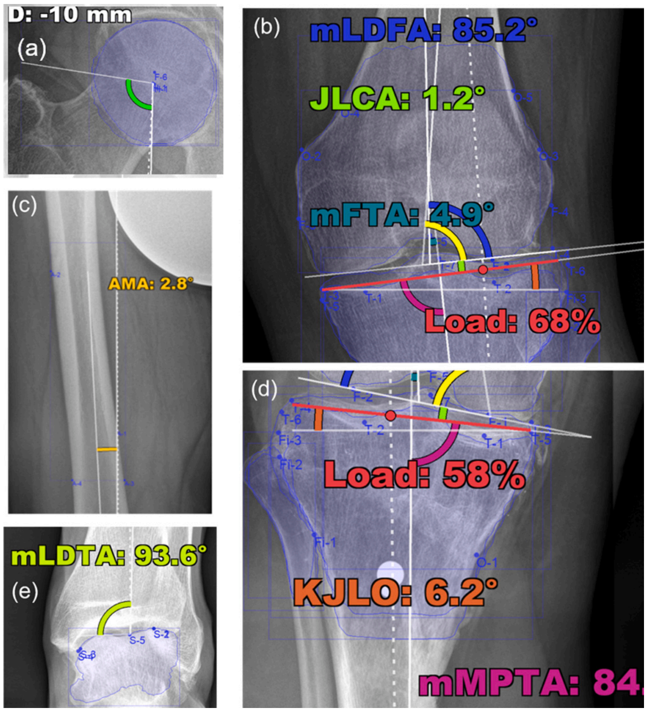

Fig. 6. Worst-Performing cases on external test dataset: Displaying the lowest performance for (a) mLPFA, (b) mLDFA and JLCA, (c) AMA, (d) KJLO and mMPTA, and (e) mLDTA, with DL algorithm annotations (segmentations and landmark detections) in blue and corresponding angles and parameters in their respective colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

OS1 and 98 ± 6 s for OS2. The processing time for the DL model for a full comprehensive analysis per radiograph at inference was 22 ± 0.5 s, utilizing a consumer-grade personal computer (Nvidia GTX 2070 maxQ) as a simulated clinical resource environment. This was significantly faster ($p \leq 0.01$) than OS1 or OS2 on the internal test dataset. Similarly, on the external test dataset, the mean time required of human raters for a full comprehensive analysis of the alignment using MediCAD version 6.0 was 105 ± 7 s for OS1 and 101 ± 7 s for OS2, while it was 22 ± 0.6 s for the DL model ($p \leq 0.01$). As such, the processing time of the DL model was more than four times faster than OS1 and OS2.

4. Discussion

This study aimed to develop and externally validate a DL model for the autonomous, comprehensive assessment of leg alignment using a.p. LLR radiographs through a multitasking approach. The resulting DL model performed leg alignment analysis, incorporating comprehensive parameters essential for clinical decision-making, with a level of precision, inter-reader reliability, and clinically acceptable accuracy matching that of specialized OS. Furthermore, the algorithm notably outperformed human raters in intra-rater reliability and processing time. With degenerative knee joint pathologies on the rise, the demand for lower extremity analysis using LLR is increasing. The DL model developed in this study holds significant potential to streamline workflow, bolster diagnostic confidence and accuracy, and reduce processing time.

Despite the adoption of ML and DL advances by numerous clinical specialties to improve quality of care and workflows in resource-constrained environments, orthopedic musculoskeletal care has seen a dearth of clinically applicable solutions [45]. Although AI applications in orthopedic care have primarily concentrated on risk stratification, clinical outcome prediction [46], and enhancing radiographic diagnostic accuracy for certain pathologies [47], there is a lack of meaningful applications that tangibly improve orthopedic providers' clinical workflows. While previous publications have striven to expedite standardized lower extremity alignment analysis [24,25,29,30,41,48], the majority have focused solely on measuring a single alignment parameter, such as the mFTA angle [24,25,29,30,48], AMA [49], or limb length [26]. While these analyses may serve as proof of concept, they offer limited utility in clinical decision-making and surgical planning contexts, such as lower extremity deformity correction [15,50] and knee arthroplasty [51].

4.1. Technical innovation

From a technical perspective, this study's model uses a master network to guide multiple expert networks, employing a multi-algorithmic approach. These networks perform multi-task feature selection like segmentation and landmark detection, optimizing accuracy and reliability while reducing memory use. The master network improves performance by sharing knowledge with the expert networks during inference and applying local edge detection to error-prone landmarks.

This stands in contrast to previous solutions for related clinical use cases that only utilized single-task features such as landmark detection [25,28–30] or segmentation features [24–26,29,30]. This study's development approach effectively harnesses the power of ensemble learning in ML [33]. The distinct technologies used in the multitasking approach outperformed previously reported results in terms of landmark detection accuracy, with previously reported L2 errors ranging from 1.63 ± 1.29 mm (tibial spines) to 1.7 ± 1.0 mm (femoral head) [25], and segmentation tasks, with Sørensen-Dice coefficients ranging from 0.97 ± 0.09 (femoral) to 0.96 ± 0.11 (tibial) [24] or 0.82 (hip) to 0.93 (knee) [48].

4.2. Evaluating alignment accuracy

In evaluating the accuracy of clinically relevant alignment parameters as a key performance indicator for potential applicability in orthopedic care, the accuracy of the developed DL model generally performed on the level of different specialist OS. Notably, using a multitasking approach, the algorithm outperformed previously reported studies such as Simon et al. [41], which documented accuracies of 0.39° for mFTA, 0.96° for mLDFA, and 1.07° for mMPTA. Comparing the systematic bias of the performance, the results align well with those obtained by Schock et al. [24] with mean differences between raters and AI of −0.04° to 0.01° for mFTA and −0.31° to 0.3° for AMA and Tack et al. [25] with 0.13 ± 0.65°–0.21 ± 0.56° for mFTA, but with substantially lower SDs compared with these studies, indicating higher accuracy on average. Moreover, the developed DL model outperformed the results published by Pei et al. [48], reporting a systematic bias of 0.49° for mFTA, Gielis et al. [29] with 1.8° for mFTA, and Gielis et al. [29] with −0.40° for mFTA.

4.3. Interreader and intrarater reliability

Evaluation of the algorithm's consistency and reliability by comparing the interrater reliability between the human specialist raters and the developed DL model showed good (JCLA) to excellent (all other paramters) interrater reliability for the external test dataset and was comparable to the interreader reliability of specialized OS using a FDA-approved digital planning program [15]. More specifically, the ICCs between the mean observer and DL model in the present study are higher for load, mLPFA, mLDFA, mMPTA, mFTA, while they are lower for the leg length and JCLA measurements compared with human raters. Interrater reliability between human raters and the DL model was higher in the present study compared with previously published studies using high-performance algorithms on internal test datasets, with reported ICCs for the mFTA of 0.99 [0.99, 1.0] [24] or 0.97 [25] and ICCs for AMA of 0.87 [0.83, 0.9] – 0.89 [0.86, 0.92] [24,52]. Compared to publications reporting on the reliability of DL algorithms on external datasets, the developed DL model showed higher ICCs compared to the human raters than previous studies, reporting ICCs of 0.99 for the mFTA, 0.87 [0.84–0.89] for mLDFA, and 0.93 [0.91–0.94] for mMPTA. Regarding intrarater reliability, the algorithm showed excellent results due to its deterministic behavior, aligning closely with the algorithm published by Simon et al. [41]. Relative to Mitterer et al. [53], our algorithm was slightly outperformed in JLCA (0.95 vs. 0.9) and AMA (0.96 vs. 0.94), equaled in mLDFA (0.99 vs. 0.98), mLDTA (0.97 vs. 0.97), and mMPTA (0.97 vs. 0.98), but edged ahead in MAD (0.99 vs 1.0), HKA (0.98 vs. 1.0), and mLPFA (0.93 vs. 0.98), also providing additional measurement parameters.

4.4. Clinically acceptable accuracy

Measurement reliability within a clinically acceptable safety margin is a critical parameter for assessing the clinical value of a fully automated tool like the developed DL algorithm. The algorithm demonstrated high clinical accuracy, with a notable percentage of its measurements falling within the clinically tolerable safety margin compared to human raters. Specifically, this ranged from 89%–100% for the internal dataset and 88%–100% for the external dataset, exceeding 90% agreement in all measurements crucial for orthopedic decision making, such as mMPTA, mLDFA, JLCA, AMA, and mFTA. This performance is on par with human assessments, which varied between 82%–100% for internal data and 71%–100% for external data. In comparison, prior studies reported clinically acceptable agreement of only 82.3% within a 1.5° margin [30] or 90% agreement in class assignment (varus<−2; valgus>2) of 90°–92° [25] for mFTA. However, the DL algorithm introduced here showed a clinical accuracy of 100% within a 2° margin. When compared with previously published fully automated commercial solutions, our DL algorithm exhibited higher clinical accuracy across all evaluated angles [41].

4.5. Length calibration

Calibration, though a minor manual task, is crucial for accurate implant sizing and cut planning in orthopedic surgery [15,50,51]. Our DL algorithm effectively performs two calibration methods, achieving a 100% success rate, depending on the available reference objects in the radiograph. Despite less precision in absolute distances compared to angular measurements, the algorithm's accuracy and reliability were clinically acceptable, paralleling specialized OS. Challenges in achieving precise calibration mainly stemmed from minor errors due to small reference objects like spheres or rules, impacting measurements such as MAD, implant size, or osteotomy gap opening. This aligns with previous studies [41] and reflects the inherent difficulty in calibration rather than algorithmic limitations. For parameters significantly affected by calibration, like limb length discrepancies, clinical relevance lies in the difference between sides, aiding healthcare providers in maintaining physiological limb length during surgeries [51].

4.6. Processing time

Harnessing the true power of DL applications for clinical tasks, the processing time for time consuming clinical tasks can be significantly and substantially reduced and automatically be performed in the background prior to human evaluation. With a processing time of 22 ± 0.6 s, the DL model in this study significantly and substantially outperformed specialized OS with advanced FDA-approved digital planning software by a factor of 4.6 (excluding the time required to open the image and save the analysis results) and by a factor of 8.6 compared with the literature. In addition, the models outperform commercially available AI models for leg alignment analysis on a consumer-grade laptop, demonstrating the computational advantage of a multitasking approach [41]. Considering the multitude of parameters, the performance is on par with other single-task models with similar performance designed to measure single alignment parameters, ranging from 3–7 s [24]. The fact that these findings were obtained on a consumer grade laptop demonstrates the potential for a local deployment in a clinical information technology environment with limited computing power, while not compromising on speed and accuracy of the analysis.

4.7. Susceptibility to errors

However, as the analysis of the individual radiographs with the worst performance shows, the DL algorithm is not error-free. In particular, in the presence of anatomical abnormalities, such as a surgical revision situation and/or high-grade osteoarthritis, the performance of DL models suffers. Furthermore, in the presence of risk factors indicative of suboptimal radiographic quality of LLRs, such as flexion of the knee and deviated position of the patella or proximal fibula suggestive of lower extremity malrotation, human review is required to prevent treatment decisions from being made based on incorrect alignment parameters [54].

4.8. Clinical potential

In terms of its clinical potential, we consider a highly comprehensive, accurate, reliable and fast algorithm such as the DL model developed to be of potential substantial benefit to the clinical radiologic and orthopedic workforce upon clinical deployment. With a precision, consistency, and clinical accuracy on a level of experienced OS, a solution such as this may increase the reliability as well as yield the potential of substantial time saving by sourcing out the time intensive process of a comprehensive leg alignment analysis. While at the current point in time, human input may be required to review and confirm the product of the fully automated assessment, solutions such as the DL model developed unlock potential for increases in efficacy as well as savings of time, resources, and cost.

4.9. Limitations

There were several limitations to this study. First, as the performance of a DL model was compared to a ground truth, the results are limited by the human performance. By including measurements of highly trained OS as well as measurements performed in a realistic clinical environment, efforts were made to obtain high quality as well externally valid ground truth measurements, yet the individual human performance is an inherent limitation of this methodology. Second, the quality of the radiograph is a factor significantly limiting precision of the DL model developed. Especially in case of suboptimal quality limited by incomplete depiction of relevant landmark, malrotation or flexion of the lower extremity, the fully automated assessment requires human rater review. Third, as the DL algorithm is primarily designed to assist in therapeutical decision making and preoperative planning, radiographs including hardware overriding the cortical bones were a priori excluded from training.

5. Conclusions

The developed DL model allowed for a comprehensive analysis of leg alignment on a.p. LLR with precision, reliability, and robustness comparable to that of OS, not failing on a single image during internal and external validation. Furthermore, by significantly and substantially outperforming human raters in terms of processing time for assessment as well as repeated measurement reliability, the DL model developed yields potential to accelerate and enhance clinical practice. This highlights, how a state-of-the-art DL model could augment the abilities of orthopedic providers in managing lower extremity pathologies in a high-volume, critical tasks that demands a high degree of precision and reliability.

Code availability

Detailed analysis on the developed DL model and information on calculating each angle based on the determined landmarks can be found in the GitHub repository https://github.com/NikonPic/AlignmentNet.

CRediT authorship contribution statement

Nikolas J. Wilhelm: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Claudio E. von Schacky:** Supervision, Project administration, Investigation, Conceptualization. **Felix J. Lindner:** Validation, Formal analysis, Data curation. **Matthias J. Feucht:** Validation, Supervision. **Yannick Ehmann:** Validation, Data curation. **Jonas Pogorzelski:** Validation, Supervision. **Sami Haddadin:** Validation, Supervision, Formal analysis. **Jan Neumann:** Validation, Supervision. **Florian Hinterwimmer:** Validation, Methodology. **Rüdiger von Eisenhart-Rothe:** Validation, Project administration, Formal analysis. **Matthias Jung:** Validation, Formal analysis. **Maximilian F. Russe:** Validation, Formal analysis. **Kaywan Izadpanah:** Validation, Formal analysis. **Sebastian Siebenlist:** Validation, Supervision, Investigation, Formal analysis. **Rainer Burgkart:** Visualization, Validation, Supervision, Resources, Project administration, Formal analysis. **Marco-Christopher Rupp:** Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sebastian Siebenlist reports a relationship with Arthrosurface that includes: consulting or advisory. Sebastian Siebenlist reports a relationship with Medi Bayreuth that includes: consulting or advisory.

Acknowledgments

The authors would like to thank Fritz Seidl, MA, for language editing this manuscript. M.J. was supported by the Berta-Ottenstein-Programme for Clinician Scientists, Faculty of Medicine, University of Freiburg.

Appendix A. Neural network implementation details

Dataset preparation

The main detection network, trained on Lower Limb Radiographs (LLRs), outputs classes and bounding boxes for each object. The expert network is trained on specially prepared datasets. For each dataset, the relative positions of the objects in the LLRs are determined using top and bottom limits b_i, t_i. The overall image range is selected so that range = [bottom, top] with bottom = $\min(b_i)$ and top = $\max(t_i)$ for $i = 1, \dots, n$, where n is the number of images. This ensures that each local dataset contains the targeted object. Average object widths, heights, and centers are recorded for each dataset to enhance inference accuracy.

Preprocessing

Preprocessing was an integral part of our data processing pipeline for both the internal and external datasets. This preprocessing primarily involved downscaling the images to make them suitable for processing by our RCNN model. The downscaling approach was governed by the Detectron2 framework's configurations, where the shorter side of each image was dynamically resized. For training, this size was randomly selected from a predefined list ranging between 640 and 800 pixels, while for testing, a fixed size of 800 pixels for the shorter side was used. The longer side of the images was correspondingly resized, capped at a maximum of 1333 pixels, to maintain the aspect ratio.

Architecture and inference process

A detailed visualization of the image analysis procedures employed is presented in Fig. 7. Initially, the complete lower limb radiograph (LLR) image is subjected to preprocessing, preparing it for the primary detection network. This network, a Region-based Convolutional Neural Network (RCNN) with a ResNet101 backbone, is adapted from the architecture described in [55]. Equipped with a Feature Pyramid Network (FPN), it generates a nuanced representation of the image's features. The Region Proposal Network (RPN) within the RCNN framework is tasked with generating preliminary object and bounding box predictions. To refine these predictions and mitigate overlaps, Non-Maximum Suppression (NMS) by Neubeck and Van Gool [56] is applied, filtering based on object scores. Following this, Region of Interest (RoI) Align, as proposed in [34], is utilized to accurately map the predicted regions to their corresponding objects using the extracted features, culminating in the box head's final predictions of object classes and bounding boxes.

Subsequent to the network's output of predictions, encompassing class and bounding box coordinates $(class, bbox)$, a specialized expert network is selected for more in-depth analysis. The center coordinates (x, y) of the identified bounding box are computed to extract a specific subimage from the X-ray. This selection is meticulously aligned to coincide with the training data's statistical mean, ensuring that the detected center (x, y) matches the average center $(x_{\text{mean}}, y_{\text{mean}})$ observed during training. The alignment process extends to the dimensions of the subimage, with the width and height adjusted to match the average dimensions width = $\text{width}_{\text{mean}}$ and height = $\text{height}_{\text{mean}}$ derived from the localized dataset used for training the expert network.

In the final stage, this expert network, a Mask R-CNN as referenced in [34], conducts a thorough analysis of the chosen subimage. Differing

Table 7
Network parameters of RCNN-101 and Mask-RCNN-50 networks utilized for lower extremity alignment analysis.

Parameter	RCNN-101	Mask-RCNN-50
Batch size	1 image per batch	1 image per batch
Base learning rate	0.00025	0.00025
Training iterations	30,000	30,000
RoI per image	512	512
Training input size range	640–800 pixels	640–800 pixels
Maximum input dimension	1333 pixels	1333 pixels
Anchor sizes	[32, 64, 128, 256, 512]	[32, 64, 128, 256, 512]
Non max surpression thresholds	0.7 (training), 0.5 (testing)	0.7 (training), 0.5 (testing)
ROI box head pooler resolution	7	7
Mask activation	–	Specific to object
Convolutional layers in mask head	–	4 layers
Pooler resolution for ROI mask head	–	14

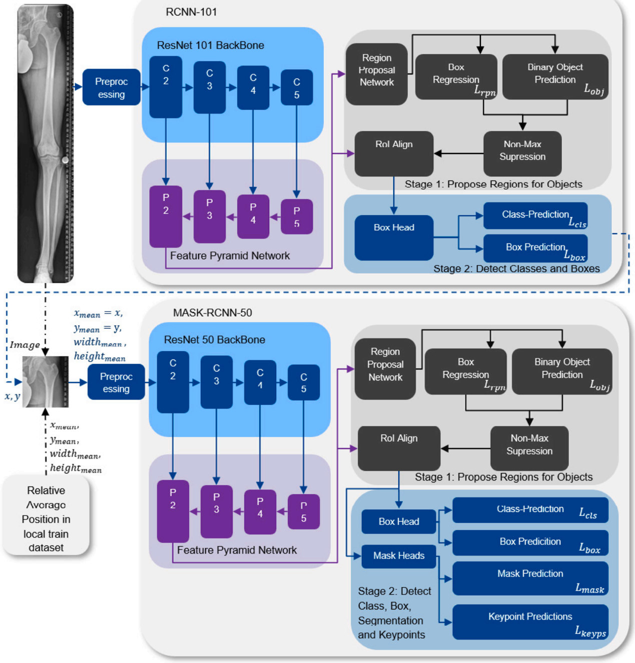

Fig. 7. Detailed view of the network architectures and their interplay during inference. The master RCNN101 network first analyzes the X-ray image, predicting classes and bounding boxes. For each detected object, the respective expert is selected based on class, and center coordinates (x, y) are determined. A subimage is then chosen based on average values from the expert network's training dataset, positioning the object at the average training location. The expert Mask RCNN50 network subsequently performs detailed analysis, extracting final class, bounding box, segmentation, and keypoints. Key training losses are indicated at respective stages where they primarily contribute to the network's learning process.

from the initial RCNN with ResNet101, the expert network's architecture includes additional network heads for segmentation and keypoint detection for each object, besides the standard class and bounding box heads. This Mask R-CNN's enhanced capabilities allow for more detailed and specific analysis, particularly beneficial for complex or overlapping anatomical structures in the radiograph.

Training losses

The utilized Mask R-CNN is trained with a multi-task loss function based on He et al. [34], which combines several different loss components. The total loss is a sum of these individual losses:

- **Classification Loss** (L_{cls}): This is the softmax loss over the object classes, including a background class.

  $$L_{cls} = -\log(p_u)$$

  where p_u is the softmax probability for the true class u.
- **Bounding Box Regression Loss** (L_{box}): This loss is applied to the predicted bounding box coordinates. It is typically a Smooth L1 loss between the predicted bounding box offsets and the true offsets.

  $$L_{box} = \text{Smooth}_{L1}(t_u - v)$$

  where t_u are the predicted offsets for bounding box and v are the ground truth offsets, and Smooth L1 loss is defined as:

  $$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases}$$
- **Mask Loss** (L_{mask}): This is a per-pixel binary cross-entropy loss, used for the mask prediction branch.

  $$L_{mask} = -\frac{1}{m^2}\sum_{1}^{m^2}[y_{ij}\log(\hat{y}_{ij}) + (1 - y_{ij})\log(1 - \hat{y}_{ij})]$$

  where m is the dimension of the mask, y_{ij} is the ground truth and $\hat{y}_{ij}$ is the predicted mask at pixel (i, j).
- **Keypoint Loss** (L_{keyp}): This loss is used for keypoint detection tasks. It is a per-keypoint binary cross-entropy loss.

  $$L_{keyp} = -\frac{1}{Km^2}\sum_{k=1}^{K}\sum_{u=1}^{m^2}[y_{kij}\log(\hat{y}_{kij}) + (1 - y_{kij})\log(1 - \hat{y}_{kuv})]$$

  where K is the number of keypoints, m is the dimension of the mask, y_{kij} is the ground truth and $\hat{y}_{kij}$ is the predicted probability at location (i, j) for keypoint k.
- **Objectness Loss** (L_{obj}): This is a binary cross-entropy loss for the objectness score predicted by the RPN.

  $$L_{obj} = -[y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})]$$

  where y is the binary indicator (0 or 1) if the anchor is an object, and $\hat{y}$ is the predicted objectness score.
- **RPN Box Regression Loss** (L_{rpn}): This is a Smooth L1 loss applied to the bounding box predictions of the RPN.

  $$L_{rpn} = \text{Smooth}_{L1}(t_i - v_i)$$

  where t_i are the predicted offsets for the RPN bounding boxes and v_i are the ground truth offsets.

The total loss for each expert Mask R-CNN is a weighted sum of these losses. With equal weights, the expert network's total loss is:

$$L_{expert} = L_{cls} + L_{box} + L_{mask} + L_{keyp} + L_{obj} + L_{rpn}$$

For the master RCNN model, which does not require segmentation or keypoint detection, the loss simplifies to:

$$L_{master} = L_{cls} + L_{box} + L_{obj} + L_{rpn}$$

Table 8
Mean Root Mean Square Errors (RMSE) of the landmark detections for the individual objects and the different approaches on the internal test dataset. The approaches are divided into the singlescale (SC), multiscale (MS), and multiscale training simulation (MS-TrainSim) approaches. Continuous variables are presented as mean ± standard deviation (range).

Object [mm]	SC	MS	MS-TrainSim
Hip	6.6 ± 24.8	0.6 ± 0.4	**0.6 ± 0.3**
Femur$_{trochanter}$	6.9 ± 29.4	1.2 ± 1.5	**0.9 ± 0.5**
Femur$_{shaft}$	55.5 ± 105.3	17.5 ± 11.0	**7.1 ± 4.5**
Femur$_{condyles}$	6.9 ± 42.4	1.9 ± 2.0	**1.7 ± 1.6**
Tibia$_{eminence}$	12.5 ± 61.6	**1.0 ± 1.1**	1.1 ± 1.1
Tibia$_{joint\ line}$	6.9 ± 42.4	**1.3 ± 1.0**	1.3 ± 1.3
Tibia$_{shaft}$	95.4 ± 139.5	5.8 ± 4.1	**5.5 ± 3.5**
Ankle	1.1 ± 1.9	1.7 ± 3.5	**0.9 ± 1.7**
Sphere	8.2 ± 49.5	0.5 ± 1.0	**0.4 ± 1.0**
Average	20.9 ± 73.9	3.0 ± 6.0	**1.9 ± 2.9**

Table 9
Ablation study for anatomical landmark optimization of F-1 and F-2 on the internal test dataset. Displayed are the Root Mean Square Errors (RMSE) for F-1 and F-2 in mm, as well as the accuracy of the angles JLCA and mLDFA in deg for the plain label by the object detector (AI) and the combination of object detector and local edge filter (AI + Edge).

	AI	AI + Edge
F-1 [mm]	**1.6 ± 1.4**	1.7 ± 4.1
F-2 [mm]	**2.0 ± 1.8**	2.0 ± 1.9
JLCA [°]	1.1 ± 1.0	**0.8 ± 1.0**
mLDFA [°]	1.3 ± 1.1	**0.5 ± 0.7**

Network parameters

This paragraph delves into the specifics of the neural network architectures utilized in our study for leg alignment analysis using radiographic images. It focuses on detailing the RCNN-101, used as the primary detection network, and the Mask-RCNN-50, employed as specialized subnetworks for nuanced analysis. Essential optimization parameters are summarized in Table 7.

Appendix B. Ablation study for optimal architecture determination

When determining the optimal architecture of the DL model to attain the most accurate landmark accuracy, the three architectures Singlescale (SC), Multiscale (MS) and Multiscale Training Simulation (MS-TrainSim) are compared. The results of the average landmark accuracy are shown in Table 8 for the internal test dataset.

The MS TrainSim achieved the highest average landmark accuracy (1.9 mm ± 2.9 mm) and substantially outperformed the SC architecture (20.9 mm ± 73.9 mm) as well as moderately outperformed the MC architecture (3.0 mm ± 6.0 mm). Therefore, the MS TrainSim approach was used as the basis for all further analyses.

Appendix C. Ablation study for anatomical landmark optimization

The results for the ablation study regarding the utility of employing edge filters are shown in Table 9. While the accuracy of the landmark detection decreased slightly (F-1: 1.6 mm ± 1.4 mm vs. 1.7 mm ± 4.1 mm, F-2: 2.0 mm ± 1.8 mm vs. 2.0 mm ± 1.9 mm), the accuracy of the affected angles JLCA and mLDFA increased moderate to significantly (JLCA: 1.1° ± 1.0° vs. 0.8° ± 1.0°, mLDFA: 1.3° ± 1.1° vs. 0.5° ± 0.7°). As such, the edge filter, while moving the landmark further away from its target on average, is more accurate for its anatomical position. Therefore, the landmark filter used for F-1 and F-2 was activated for further analyses.

Table 10
Comparison between the Dice scores for object detection of the main detection network (RCNN-101) and the final results of the specialized networks (RCNN101 + MASK-RCNN50).

Object	Dice score (Specialized)	Dice score (Main)
Hip	**0.97 ± 0.02**	0.93 ± 0.2
Femur$_{troch}$	**0.97 ± 0.01**	0.96 ± 0.09
Femur$_{shaft}$	**0.9 ± 0.06**	0.88 ± 0.17
Femur$_{cond}$	**0.96 ± 0.02**	0.96 ± 0.09
Tibia$_{emin}$	**0.96 ± 0.02**	0.94 ± 0.12
Tibia$_{jl}$	**0.96 ± 0.02**	0.94 ± 0.15
Tibia$_{shaft}$	**0.92 ± 0.04**	0.89 ± 0.17
Ankle	**0.93 ± 0.09**	0.89 ± 0.22
Sphere	0.95 ± 0.15	**0.96 ± 0.09**

Appendix D. Performance of the RCNN-101

The efficacy of our master RCNN-101 network is evaluated using the Dice score metric, which measures the overlap between the network's detected objects and the ground truth labels on our internal test dataset. To demonstrate the effectiveness of our two-stage approach, we compare the Dice scores of the master RCNN-101 network with those achieved using the subsequent expert network. Table 10 presents this comparative analysis.

Our results indicate that the combined methodology, where the general RCNN-101 is used for object localization and the specialized MASK-RCNN for detailed object detection, generally surpasses the performance of the standalone RCNN-101 in most tasks. The Dice scores for the general RCNN-101 range from 0.89 to 0.96, while the specialized MASK-RCNN shows an improvement with scores ranging from 0.90 to 0.97. Notably, the sole exception to this trend is in Sphere detection, where the standalone RCNN-101 marginally outperforms the two-stage process.

References

[1] Safiri S, Kolahi A-A, Smith E, Hill C, Bettampadi D, Mansournia MA, Hoy D, Ashrafi-Asgarabad A, Sepidarkish M, Almasi-Hashiani A, Collins G, Kaufman J, Qorbani M, Moradi-Lakeh M, Woolf AD, Guillemin F, March L, Cross M. Global, regional and national burden of osteoarthritis 1990–2017: a systematic analysis of the Global Burden of Disease Study 2017. Ann Rheum Dis 2020;79(6):819–28. http://dx.doi.org/10.1136/annrheumdis-2019-216515.

[2] Liu X, Chen Z, Gao Y, Zhang J, Jin Z. High tibial osteotomy: Review of techniques and biomechanics. J Healthc Eng 2019;2019.

[3] Paley D. Principles of deformity correction -. Berlin, Heidelberg: Springer; 2014.

[4] Brown ML, McCauley JC, Gracitelli GC, Bugbee WD. Osteochondritis dissecans lesion location is highly concordant with mechanical axis deviation. Am J Sports Med 2020;48(4):871–5.

[5] Hwang B-Y, Kim S-J, Lee S-W, Lee H-E, Lee C-K, Hunter DJ, Jung K-A. Risk factors for medial meniscus posterior root tear. Am J Sports Med 2012;40(7):1606–10.

[6] Wang Y-L, Yang T, Zeng C, Wei J, Xie D-X, Yang Y-H, Long H-Z, Xu B, Qian Y-X, Jiang S-d, Lei G-H. Association between tibial plateau slopes and anterior cruciate ligament injury: A meta-analysis. Arthroscopy 2017;33(6):1248–59.e4.

[7] Webb JM, Salmon LJ, Leclerc E, Pinczewski LA, Roe JP. Posterior tibial slope and further anterior cruciate ligament injuries in the anterior cruciate Ligament–Reconstructed patient. Am J Sports Med 2013;41(12):2800–4.

[8] Imhoff FB, Funke V, Muench LN, Sauter A, Englmaier M, Woertler K, Imhoff AB, Feucht MJ. The complexity of bony malalignment in patellofemoral disorders: femoral and tibial torsion, trochlear dysplasia, TT–TG distance, and frontal mechanical axis correlate with each other. Knee Surg Sports Traumatol Arthrosc 2020;28(3):897–904.

[9] Ackermann J, Merkely G, Arango D, Mestriner AB, Gomoll AH. The effect of mechanical leg alignment on cartilage restoration with and without concomitant high tibial osteotomy. Arthroscopy 2020;36(8):2204–14.

[10] Cao Z, Mai X, Wang J, Feng E, Huang Y. Unicompartmental knee arthroplasty vs high tibial osteotomy for knee osteoarthritis: A systematic review and Meta-Analysis. J Arthroplasty 2018;33(3):952–9.

[11] Liu JN, Agarwalla A, Gomoll AH. High tibial osteotomy and medial meniscus transplant. Clin Sports Med 2019;38(3):401–16.

[12] Lutz PM, Winkler PW, Rupp M-C, Geyer S, Imhoff AB, Feucht MJ. Complex patellofemoral reconstruction leads to improved physical and sexual activity in female patients suffering from chronic patellofemoral instability. Knee Surg Sports Traumatol Arthrosc 2021;29(9):3017–24.

[13] Yamaguchi KT, Cheung EC, Markolf KL, Boguszewski DV, Mathew J, Lama CJ, McAllister DR, Petrigliano FA. Effects of anterior closing wedge tibial osteotomy on anterior cruciate ligament force and knee kinematics. Am J Sports Med 2018;46(2):370–7.

[14] Sappey-Marinier E, Batailler C, Swan J, Schmidt A, Cheze L, MacDessi SJ, Servien E, Lustig S. Mechanical alignment for primary TKA may change both knee phenotype and joint line obliquity without influencing clinical outcomes: a study comparing restored and unrestored joint line obliquity. Knee Surg Sports Traumatol Arthrosc 2021.

[15] Schröter S, Ihle C, Mueller J, Lobenhoffer P, Stöckle U, van Heerwaarden R. Digital planning of high tibial osteotomy. Interrater reliability by using two different software. Knee Surg Sports Traumatol Arthrosc 2013;21(1):189–96.

[16] Matos MB, Faria JLRd, Pavão DM, Sandt ML, Pigozzo BCdA, Filho PGTdS, Albuquerque RPe. Evaluation of intraobserver and interobserver reliability of mechanical axis alignment measure of the lower limb through the panoramic radiograph in patients in the preoperative and postoperative periods of total knee arthroplasty. Open J Orthop 2020;10(09):221–33.

[17] Marx RG, Grimm P, Lillemoe KA, Robertson CM, Ayeni OR, Lyman S, Bogner EA, Pavlov H. Reliability of lower extremity alignment measurement using radiographs and PACS. Knee Surg Sports Traumatol Arthrosc 2011;19(10):1693.

[18] Schmale GA, Bayomy AF, O'Brien AO, Bompadre V. The reliability of full-length lower limb radiographic alignment measurements in skeletally immature youth. J Child Orthop 2019.

[19] Specogna AV, Birmingham TB, DaSilva JJ, Milner JS, Kerr J, Hunt MA, Jones IC, Jenkyn TR, Fowler PJ, Giffin JR. Reliability of lower limb frontal plane alignment measurements using plain radiographs and digitized images. J Knee Surg 2004;17(4):203–10.

[20] Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, Zhang Z, Nitche N, Lacave E, Pourchot A, Felter A, Lassalle L, Regnard N-E, Feydy A. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: A multicenter cross-sectional diagnostic study. Radiology 2021;300(1):120–9.

[21] Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, Zaid M, McGill KC, Patel R, Sohn JH, Wright A, Darger BF, Padrez KA, Ozhinsky E, Majumdar S, Pedoia V. Automatic hip fracture identification and functional subclassification with deep learning. Radiol: Artif Intell 2020;2(2):e190023.

[22] Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skelet Radiol 2019;48(2):239–44.

[23] Wu JT, Wong KCL, Gur Y, Ansari N, Karargyris A, Sharma A, Morris M, Saboury B, Ahmad H, Boyko O, Syed A, Jadhav A, Wang H, Pillai A, Kashyap S, Moradi M, Syeda-Mahmood T. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. JAMA Netw Open 2020;3(10):e2022779.

[24] Schock J, Truhn D, Abrar DB, Merhof D, Conrad S, Post M, Mittelstrass F, Kuhl C, Nebelung S. Automated analysis of alignment in Long-Leg radiographs by using a fully automated support system based on artificial intelligence. Radiol: Artif Intell 2021;3(2):e200198.

[25] Tack A, Preim B, Zachow S. Fully automated assessment of knee alignment from Full-Leg X-Rays employing a "YOLOv4 and resnet landmark regression algorithm" (YARLA): Data from the osteoarthritis initiative. Comput Methods Programs Biomed 2021;205:106080.

[26] Zheng Q, Shellikeri S, Huang H, Hwang M, Sze RW. Deep learning measurement of leg length discrepancy in children based on radiographs. Radiology 2020;296(1):152–8.

[27] von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, Posadzy M, Foreman SC, Nevitt MC, Link TM, Pedoia V. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 2020;295(1):136–45.

[28] Yeh Y-C, Weng C-H, Huang Y-J, Fu C-J, Tsai T-T, Yeh C-Y. Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs. Sci Rep 2021;11(1):7618.

[29] Gielis WP, Rayegan H, Arbabi V, Ahmadi Brooghani SY, Lindner C, Cootes TF, de Jong PA, Weinans H, Custers RJH. Predicting the mechanical hip-knee-ankle angle accurately from standard knee radiographs: a cross-validation experiment in 100 patients. Acta Orthop 2020;91(6):732–7.

[30] Nguyen TP, Chae D-S, Park S-J, Kang K-Y, Lee W-S, Yoon J. Intelligent analysis of coronal alignment in lower limbs based on radiographic image with convolutional neural network. Comput Biol Med 2020;120:103732.

[31] Steele JR, Jang SJ, Brilliant ZR, Mayman DJ, Sculco PK, Jerabek SA, Vigdorchik JM. Deep learning phenotype automation and cohort analyses of 1, 946 knees using the coronal plane alignment of the knee classification. J Arthroplasty 2023;38(6):S215–21.e1. http://dx.doi.org/10.1016/j.arth.2023.02.055.

[32] Wang J, Hall TA, Musbahi O, Jones GG, van Arkel RJ. Predicting hip-knee-ankle and femorotibial angles from knee radiographs with deep learning. Knee 2023;42:281–8. http://dx.doi.org/10.1016/j.knee.2023.03.010.

[33] Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems: first international workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 proceedings 1. Springer; 2000, p. 1–15.

[34] He K, Gkioxari G, Dollár P, Girshick RB. Mask R-CNN. 2017, CoRR abs/1703.06870. URL: http://arxiv.org/abs/1703.06870, arXiv:1703.06870.

[35] Girshick RB. Fast R-CNN. 2015, CoRR abs/1504.08083. URL: http://arxiv.org/abs/1504.08083, arXiv:1504.08083.

[36] Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R. Detectron2. 2019, https://github.com/facebookresearch/detectron2.

[37] Gustafsson F. Determining the initial states in forward-backward filtering. IEEE Trans Signal Process 1996;44(4):988–92. http://dx.doi.org/10.1109/78.492552.

[38] Miniaci A, Ballmer FT, Ballmer PM, Jakob RP. Proximal tibial osteotomy. Clin Orthop Relat Res 1989;NA;(246):250???259.

[39] Ooms J. tesseract: Open source OCR engine. 2023, https://docs.ropensci.org/tesseract/ (website) https://github.com/ropensci/tesseract (devel).

[40] Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologiske skrifter, Munksgaard in Komm.; 1948, URL: https://books.google.de/books?id=rpS8GAAACAAJ.

[41] Simon S, Schwarz GM, Aichmair A, Frank BJH, Hummer A, DiFranco MD, Dominkus M, Hofstaetter JG. Fully automated deep learning for knee alignment assessment in lower extremity radiographs: a cross-sectional diagnostic study. Skelet Radiol 2021;51(6):1249–59.

[42] Knutson GA. Anatomic and functional leg-length inequality: a review and recommendation for clinical decision-making. Part I, anatomic leg-length inequality: prevalence, magnitude, effects and clinical significance. Chiropr Osteopat 2005;13:11.

[43] Vallat R. Pingouin: statistics in Python. J Open Source Softw 2018;3(31):1026. http://dx.doi.org/10.21105/joss.01026.

[44] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(2):155–63.

[45] Ramkumar PN, Pang M, Polisetty T, Helm JM, Karnuta JM. Meaningless applications and misguided methodologies in artificial intelligence–related orthopaedic research propagates hype over hope. Arthrosc: J Arthrosc Relat Surg 2022;38(9):2761–6. http://dx.doi.org/10.1016/j.arthro.2022.04.014.

[46] Martin RK, Ley C, Pareek A, Groll A, Tischer T, Seil R. Artificial intelligence and machine learning: an introduction for orthopaedic surgeons. Knee Surg Sports Traumatol Arthrosc 2021;30(2):361–4. http://dx.doi.org/10.1007/s00167-021-06741-2.

[47] Ko S, Pareek A, Ro DH, Lu Y, Camp CL, Martin RK, Krych AJ. Artificial intelligence in orthopedics: three strategies for deep learning with orthopedic specific imaging. Knee Surg Sports Traumatol Arthrosc 2022;30(3):758–61. http://dx.doi.org/10.1007/s00167-021-06838-8.

[48] Pei Y, Yang W, Wei S, Cai R, Li J, Guo S, Li Q, Wang J, Li X. Automated measurement of hip-knee-ankle angle on the unilateral lower limb X-rays using deep learning. Phys Eng Sci Med 2020;44(1):53–62.

[49] von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Gassert FG, Foreman SC, Gassert FT, Jung M, Jungmann PM, Russe MF, Mogler C, Knebel C, von Eisenhart-Rothe R, Makowski MR, Woertler K, Burgkart R, Gersing AS. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. Radiology 2021;204531. http://dx.doi.org/10.1148/radiol.2021204531.

[50] Feucht MJ, Winkler PW, Mehl J, Bode G, Forkel P, Imhoff AB, Lutz PM. Isolated high tibial osteotomy is appropriate in less than two-thirds of varus knees if excessive overcorrection of the medial proximal tibial angle should be avoided. Knee Surg Sports Traumatol Arthrosc 2020;29(10):3299–309. http://dx.doi.org/10.1007/s00167-020-06166-3.

[51] Tanzer M, Makhdom AM. Preoperative planning in primary total knee arthroplasty. J Am Acad Orthop Surg 2016;24(4):220–30.

[52] Pagano S, Müller K, Götz J, Reinhard J, Schindler M, Grifka J, Maderbacher G. The role and efficiency of an AI-powered software in the evaluation of lower limb radiographs before and after total knee arthroplasty. J Clin Med 2023;12(17):5498. http://dx.doi.org/10.3390/jcm12175498.

[53] Mitterer JA, Huber S, Schwarz GM, Simon S, Pallamar M, Kissler F, Frank BJH, Hofstaetter JG. Fully automated assessment of the knee alignment on long leg radiographs following corrective knee osteotomies in patients with valgus or varus deformities. Arch Orthop Trauma Surg 2023.

[54] Ahrend M-D, Baumgartner H, Ihle C, Histing T, Schröter S, Finger F. Influence of axial limb rotation on radiographic lower limb alignment: a systematic review. Arch Orthop Trauma Surg 2021.

[55] Ren S, He K, Girshick RB, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. 2015, CoRR abs/1506.01497. URL: http://arxiv.org/abs/1506.01497, arXiv:1506.01497.

[56] Neubeck A, Van Gool L. Efficient non-maximum suppression. In: 18th international conference on pattern recognition. ICPR'06, 2006, p. 850–5. http://dx.doi.org/10.1109/ICPR.2006.479.