scientific reports

Check for updates

OPEN

Enhancing automatic landmark localization in X-ray images using combined segmentation and regression models: application to lower limb alignment assessment

Ashkan Zarghami[1,5], Sebastián Amador Sánchez[1,2,5]✉, Philippe Van Overschelde[3] & Jef Vandemeulebroucke[1,2,4]

Manual landmark detection in lower limb medical imaging is time-consuming and error-prone. Recently, authors have proposed automatic landmark detection methods based on image segmentation, coordinate regression, or a combination of both to aid clinicians. While the latter approach shows promising results, detailed optimization of its design choices, including the integration strategy and hyperparameter tuning, remains unexplored. This study investigates the optimal approach to combining image segmentation and coordinate regression, focusing on selecting suitable network architectures and optimizing their configurations. We contrasted two methods for training the network: end-to-end training and cascading the subnetworks, and assessed the optimal architecture for each strategy. For landmark segmentation, we compared U-Net and Swin-UNETR models, and for coordinate regression, we assessed VGG-16, ResNet-50, and Swin-B. Performance was evaluated in detecting eight landmarks in each leg of a complete lower-limb X-ray and in examining their influence on the clinical task of measuring lower-limb malalignment. Swin-UNETR slightly outperformed U-Net, with a lower Euclidean distance error (1.74 mm [0.98 mm] versus 1.75 mm [1.06 mm]) and significantly fewer false positives (160 vs. 502). For coordinate regression, VGG-16 in the end-to-end configuration achieved the highest accuracy (2.44 mm [2.12 mm]) and proved optimal for lower limb alignment assessment, with 97.88% of estimations falling within 1.5° of the hip–knee–ankle angle ground truth. Combining Swin-UNETR with VGG-16 within an end-to-end framework yields the most accurate and robust performance for automated lower-limb landmark detection and alignment assessment.

Keywords Lower limb alignment, Landmark detection, Localization, Segmentation, Coordinate regression

Recent studies reveal a prevalence of hip (8.55%), knee (24%), and ankle (1%) osteoarthritis worldwide[1–3]. Lower limb alignment plays a crucial role in musculoskeletal health since any coronal malalignment of the lower limb is considered a risk factor for osteoarthritis (OA)[4]. Lower-limb coronal malalignment is commonly described as varus or valgus based on the deviation of the mechanical axis of the lower extremity from neutral alignment.

To confirm malalignment, full lower-limb (FLL) radiographs are used, as they are considered the gold standard[5]. On an FLL image, the mechanical axis is defined as the line connecting the center of the femoral head to the center of the ankle joint. In varus malalignment, this axis shifts medially relative to the knee joint center, resulting in increased load distribution in the medial tibiofemoral compartment. In valgus alignment, the mechanical axis shifts laterally, increasing load in the lateral compartment[4].

To perform a complete malalignment test[6], in addition to the mechanical axis, complementary axes are defined over the FLL image. Through them, measurements that allow malalignment assessment are computed [5]. Figure 1 illustrates the anatomical landmarks and axes definitions used for these measurements.

[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [2]IMEC, Kapeldreef 75, 3001 Leuven, Belgium. [3]MoveUP, Ghent 9000, Belgium. [4]Department of Radiology, Universitair Ziekenhuis Brussel, Laarbeeklaan 101, 1090 Brussels, Belgium. [5]Ashkan Zarghami and Sebastián Amador Sánchez are contributed equally to this work. ✉email: sebastian.amador.sanchez@vub.be

Despite technological advances, landmark annotation remains a manual process performed by clinicians. As a result, malalignment assessment is time-consuming and prone to variability in alignment-quality evaluation, as interpretations can differ with physicians' expertise[7].

Arthroplasty is one of the most common surgical treatments for alleviating pain and restoring function in patients with end-stage osteoarthritis of the lower limb[8]. While pain relief is the primary goal, achieving neutral mechanical alignment, typically defined as a hip–knee–ankle (HKA) angle close to 0°, remains essential, as it contributes to the longevity of the prosthesis[9,10] and is a significant predictor of the risk of long-term aseptic wear and loosening[11,12].

Artificial intelligence (AI) applications for assessing and managing lower limb malalignment in orthopedics have gained significant attention in recent years due to their potential to enhance healthcare delivery and improve diagnostic accuracy[13]. Recent studies demonstrate the use of AI for the automatic evaluation of lower limb alignment[14–18]. Nguyen et al.[14] utilized a cascaded regression approach to locate essential anatomical landmarks required for malalignment evaluation. Pei et al.[15] employed three independent U-Net-based neural networks to segment the hip, knee, and ankle joints from full lower limb (FLL) X-rays to measure the hip-knee-ankle angle.

Moon et al.[16] segmented bone regions from full-leg radiographs, enabling automated measurement of lower extremity alignment. Sanchez et al.[17] introduced an approach that combines image segmentation with coordinate regression, enhancing lower limb alignment assessment, which is crucial for orthopedic diagnostics and treatment planning. Kim et al.[18] designed a ResNet-based deep learning model for automated measurement of the hip-knee-ankle (HKA) angle on lower limb radiographs, achieving high accuracy.

In contrast to the methodologies that rely on segmenting regions of interest or performing regression to estimate the level of malalignment, Sanchez et al.[17] proposed a segmentation-guided coordinate regression (SGR) architecture to improve the robustness of landmark segmentation. Landmarks are first segmented, and the probability segmentation map is processed along with the input image to perform coordinate regression. Their proposed network topology compensated for missed or ambiguous segmentations, achieving superior accuracy compared to standard coordinate regression. That said, the approach increases the complexity of the architecture and training, leaving ample room for hyperparameter optimization that was previously unexplored.

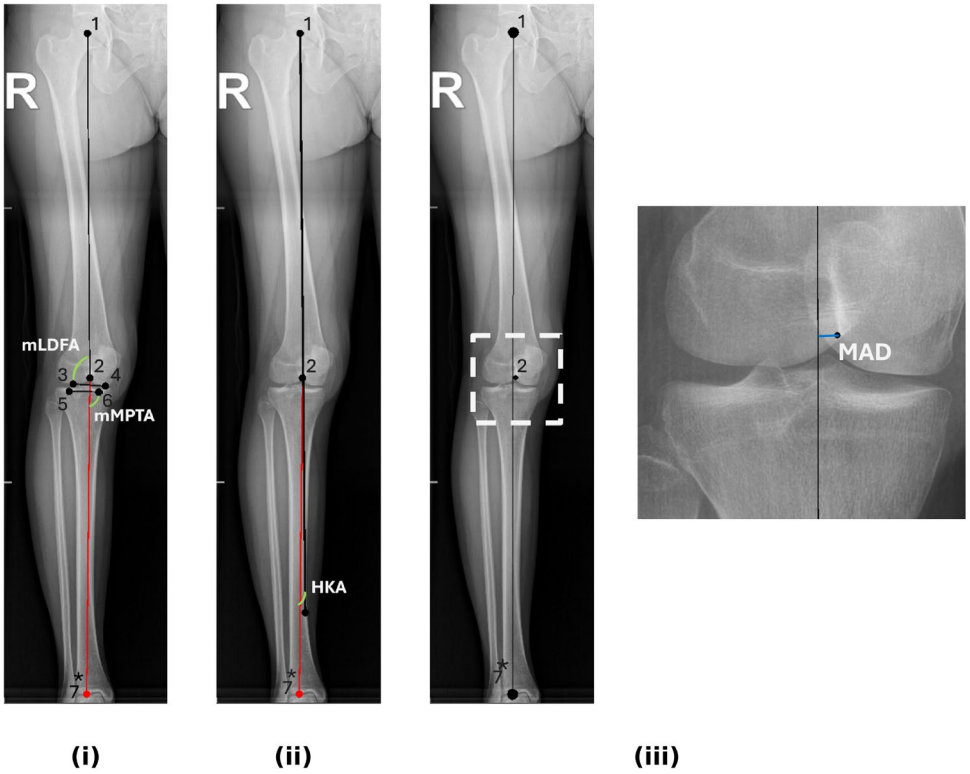

Fig. 1. Angular and distance measurements used in lower limb alignment assessment. (**i**) Mechanical lateral distal femoral angle (mLDFA): angle between the mechanical axis of the femur, defined by the line connecting the femoral head center to the knee center (1 → 2), and the tangent to the distal femoral condyles (3 → 4). Mechanical medial proximal tibial angle (mMPTA): angle between the mechanical axis of the tibia, defined by the line connecting the knee center to the ankle center (2 → 7*), and the tangent to the tibial plateau (5 → 6). (**ii**) Hip–knee–ankle angle (HKA): angle between the mechanical axis of the femur (1 → 2) and the mechanical axis of the tibia (2 → 7*). (**iii**) Mechanical axis deviation (MAD): perpendicular distance from the knee center (2) to the mechanical axis of the lower limb, defined by the line connecting the femoral head center to the ankle center (1 → 7*). Here, 7* denotes the ankle joint center, computed as the midpoint of two ankle landmarks.

This study builds on the segmentation-guided coordinate regression framework introduced in[17] and uses the same annotated dataset, while focusing on benchmarking alternative segmentation architectures, regression backbones, and training strategies.

In this study, we aimed to enhance landmark detection by systematically analyzing and benchmarking the segmentation-guided coordinate regression architecture. Our main contributions are the following. First, we evaluated two segmentation architectures to identify the most effective for guiding landmark detection. Second, we compared end-to-end training with a cascaded approach, where the segmentation and regression networks are trained separately. This analysis helps determine whether joint optimization improves performance or if independent training yields comparable accuracy. Third, we investigated various regression architectures coupled to each training scheme to understand their impact on localization performance, thereby identifying the most suitable topology for refining landmark predictions across diverse anatomical structures and imaging scenarios.

Materials and methods

Landmark segmentation

We contrasted two network architectures: the U-Net[19] and Swin-UNETR[20]. U-Net represents a robust and widely validated fully convolutional approach, providing a reliable baseline due to its effectiveness in capturing spatial features at multiple scales. Swin-UNETR, on the other hand, incorporates transformer blocks into its encoder, enabling it to leverage global context through self-attention mechanisms, which may potentially improve segmentation accuracy by modeling long-range dependencies. Swin-UNETR was included as a representative transformer-based segmentation architecture to enable a controlled comparison with U-Net within the same implementation framework, rather than as an exhaustive benchmark against all possible 2D transformer-based alternatives.

For both models, we used the implementations available in the Medical Open Network for AI (MONAI) deep learning framework, applying them to our 2D X-ray landmark segmentation task and modifying only the input and output layers to align with our specific task requirements. In this context, "landmark segmentation" refers to the prediction of landmark-centered probability maps rather than delineation of full anatomical structures. Each landmark was represented during training by a binary circular mask centered on the annotated point, and the predicted masks were later converted into landmark coordinates by thresholding and centroid extraction. We trained independent models with randomly initialized weights for each anatomical region of interest, with the number of output channels corresponding to the number of landmarks present in the target image. The hip model had one channel, the knee model five, and the ankle model two. During evaluation, to derive coordinates from the segmentation model outputs, we applied a probability threshold (t) of 0.5 to binarize the predicted probability maps. Next, we computed the centroid of each binary mask and took these as the corresponding coordinates. Next, we converted pixel coordinates to physical space measured in millimeters, using the DICOM pixel spacing values embedded in each radiograph.

Regression head assessment

We investigated three models for the coordinate regression section: VGG-16, ResNet-50, and Swin-B. VGG-16[21] is particularly effective at capturing hierarchical features. Meanwhile, ResNet-50[22] was selected for its ability to enable the training of deeper architectures through residual connections, which help mitigate vanishing gradient issues. Deeper networks can model more complex spatial relationships, potentially improving the accuracy of regressing anatomical landmarks from medical images. The Shifted Window (Swin) transformer[23] (baseline configuration: Swin-B) was included for its ability to capture fine-grained spatial information, which could make it effective for landmark detection.

Each regression model had pretrained (ImageNet) weights. Additionally, we modified the input layer of each network to accept the concatenation of 3-channel X-rays and the probability maps generated by the segmentation models, where the number of probability map channels corresponds to the number of landmarks in each region. Thus, the network took four input channels for the hip, eight for the knee, and five for the ankle. The output of each network corresponded to the total x- and y-coordinates to be estimated. Similar to image segmentation, on inference, we converted the estimated pixel coordinates to physical points.

Since the pretrained backbones were originally designed for RGB natural images, each grayscale X-ray was replicated across three identical channels before concatenation with the segmentation probability maps; the radiographs were therefore not treated as true RGB images.

End-to-end versus Cascade networks

We compared two training strategies to identify the most effective one for landmark detection. The first was the end-to-end strategy proposed by Sanchez et al.[17], termed segmentation-guided coordinate regression. In this strategy, a segmentation model initially estimates landmark positions by generating probability maps centered on the target location. The previous output is concatenated with the original X-ray image and input to a regression network to estimate the x- and y-coordinates of the landmarks. In the strategy proposed by Sanchez et al., both models are trained end-to-end, meaning that the entire network architecture is optimized simultaneously; see Fig. 2.

On the other hand, the cascade strategy integrates a separately trained segmentation with a regression model (see Fig. 3). In a first phase, a network is trained to detect the desired landmarks by segmenting masks centered at the desired position. Once the segmentation network is trained, probability maps are generated for each X-ray. In a second phase, X-rays are concatenated with the inferred probability maps, and a coordinate regression network is independently trained.

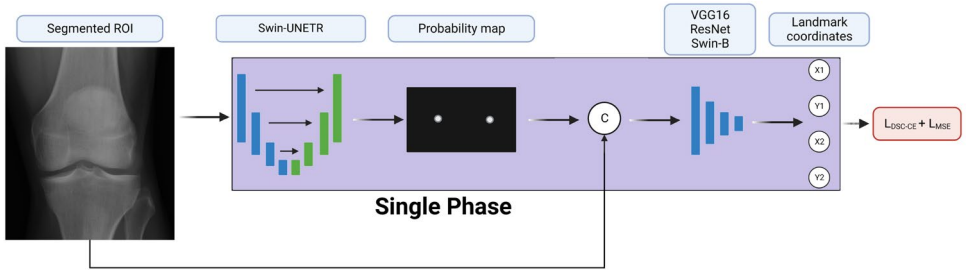

Fig. 2. End-to-end landmark detection approach. This methodology combines an image segmentation model (*e.g.*, a Swin-UNETR) and a regression network to estimate landmark coordinates. In this approach, both models are trained jointly.

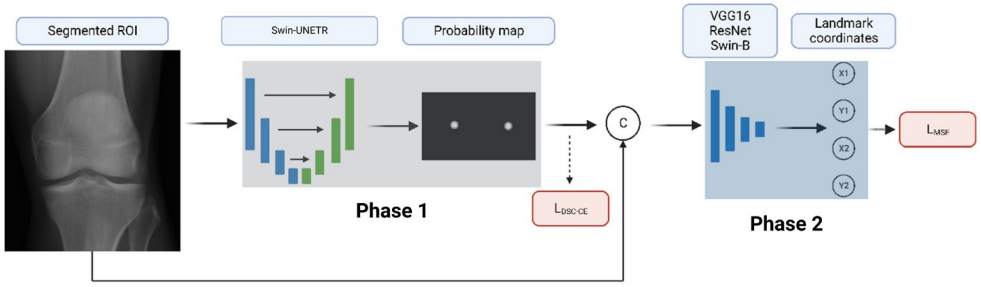

Fig. 3. Graphical representation of the cascade approach. In the first phase, an image segmentation model (*e.g* a Swin-UNETR) is trained to detect a set of landmarks. Then, probability maps indicating the possible location of the landmarks are generated using the X-rays. In the second phase, the probability maps are concatenated with the X-ray images and used to train a regression network.

To provide a controlled and reproducible comparison between the assessed strategies, we trained all the described networks for 100 epochs using the Adam optimizer with a constant learning rate of 10^{-5}. This unified training setup was adopted to provide a controlled comparison across architectures and training strategies, although it may not represent the optimal hyperparameter configuration for every model family. For the end-to-end training strategy, the total loss is expressed as follows:

$$L_{total} = L_{Sgm} + L_{Reg}. \tag{1}$$

For the segmentation branch of the end-to-end strategy (L_{Sgm}) and the segmentation network of the cascade approach, we employed a uniform weighted combination of Dice (L_{Dsc}) and Cross-entropy (L_{CE}) as the loss function[24]:

$$L_{\text{Dsc-CE}} = L_{\text{Dsc}} + L_{\text{CE}}, \tag{2}$$

$$L_{\text{Dsc}} = 1 - \frac{2TP}{2TP + FP + FN}, \tag{3}$$

$$L_{\text{CE}} = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right). \tag{4}$$

where *TP*, *FP*, and *FN* correspond to the true positive, false positive, and false negative classified pixels, respectively. For L_{CE}, $\hat{y}$ and y represent the predicted and ground truth value, respectively. Meanwhile, for the regression branch of the end-to-end scheme (L_{Reg}) and the coordinate regression of the cascade, we utilized the mean squared error (MSE) loss function:

$$L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} \|\text{p}_i - \hat{\text{p}}_i\|_2^2 \tag{5}$$

where n is the number of samples, p_i denotes the ground-truth landmark coordinates, and $\hat{\text{p}}_i$ denotes the estimated landmark coordinates.

Training and network implementations were performed in Python 3.8.6 using PyTorch 1.7.1, with MONAI 0.7.0 integrated for medical imaging tasks. The training was conducted on Google Colab using the embedded Nvidia Tesla P100 GPU.

Variable	Category	Value
Gender	Male [%]	34.06
	Female [%]	56.47
	Unknown [%]	9.47
Age	Years [median (IQR)]	64 (18)
Presence of external devices	Hip (one side) [%]	10.01
	Hip (both sides) [%]	4.13
	Knee (one side) [%]	35.15
	Knee (both sides) [%]	13.93
	Ankle (one side) [%]	2.39
	Ankle (both sides) [%]	0.11
Presence of "R" marker	Obstructing marker [%]	7.73
KL grade	0 [%]	44.29
	1 [%]	33.02
	2 [%]	10.88
	3 [%]	6.80
	4 [%]	5.01

Table 1. Descriptive characteristics of the dataset. Continuous variables are reported as median (IQR), and categorical variables as percentages.

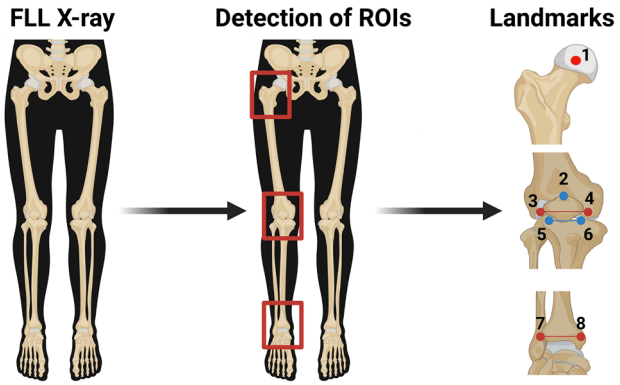

Fig. 4. Bounding boxes were placed to define and crop the hip, knee, and ankle regions. A total of eight landmarks were annotated on each leg side of each FLL X-ray and were considered the ground-truth positions.

Dataset

A single radiologist annotated a dataset of 919 X-rays, including pre-operative and post-operative scans retrospectively obtained. The dataset was randomly selected from the Hip and Knee Unit clinic at AZ Maria Middelares Hospital, representing a single-center cohort. All subjects had previously signed an informed consent form authorizing the collection and use of their data for research purposes. All radiographs were acquired using a Siemens Luminos dRF Max fluoroscopy system. During acquisition, subjects were positioned 1.50 m from the X-ray source, with collimation set to 41.8 cm $\times$ 41.8 cm and tube voltage set to 90 kV. The system automatically adjusted exposure parameters for patients with larger body mass to ensure adequate image quality. In post-processing, up to four individual images were stitched to generate the final full lower limb radiograph. The demographic and radiographic characteristics of the dataset are summarized in Table 1, including age, sex distribution, presence of external devices, laterality markers, and Kellgren–Lawrence grades (KL). These bounding boxes were manually defined to extract regions of interest (ROIs) around each joint, ensuring consistent anatomical coverage while reducing irrelevant background and enabling standardized preprocessing prior to model training, as shown in Fig. 4. These landmarks correspond to anatomically defined points at the hip, knee, and ankle that are used to derive the lower-limb alignment measurements described in Section "Lower limb malalignment evaluation". The dataset and landmark annotation protocol used in this study are the same as those introduced in[17]. We cropped the images using the annotated bounding boxes and used 80% of the images for training and validation, and 20% for testing. During training and validation, we employed a batch size of 8 for all models, with images being resampled to a fixed size of 512×512 pixels.

For the segmentation task, we generated binary circular masks with a radius of 15 pixels, centered on the desired landmark positions. This value was selected as a moderate mask size to provide sufficient spatial support for the segmentation-guided regression pipeline while preserving localized supervision around each landmark.

In our framework, the segmentation output serves as an intermediate probability map that guides the regression network rather than directly defining the final landmark coordinates.

The selected radius of 15 pixels was based on prior work[25], in which multiple mask sizes were evaluated and medium-sized masks were found to provide the best trade-off between segmentation learnability and localization precision.

Lower limb malalignment evaluation

To evaluate the clinical relevance of precise landmark detection, we measured four lower-limb alignment metrics (MAD, mLDFA, mMPTA, and HKA) using the annotated landmarks and compare these measurements with those obtained using the examined landmark-detection methods. In our implementation, see Figs. 1 and 4, the mechanical axis deviation (MAD)[26] is the distance between the lower extremity's mechanical axis ($1 \rightarrow \text{center between } 7 \text{ and } 8$) and the knee joint's center (2).

The deviation from the normal of the lower extremity alignment was evaluated by the joint orientation angles: the mechanical proximal tibial angle (mMPTA) and lateral distal femoral angle (mLDFA)[26]. The mMPTA was defined as the angle formed between the mechanical axis of the tibia ($\text{center between } 7 \text{ and } 8 \rightarrow 2$) and the tibial plateau line ($5 \rightarrow 6$). Meanwhile, the mLDFA was defined as the angle between the mechanical axis of the femur ($1 \rightarrow 2$) and the femur chondyle line ($3 \rightarrow 4$). Additionally, we measured the hip-knee-ankle angle (HKA), which is defined as the angle between the mechanical axes of the femur and tibia in the coronal plane. In this study, the evaluation focuses on the absolute error between predicted and reference HKA values rather than on directional classification. Although varus and valgus alignment can be defined clinically based on the direction of mechanical axis deviation, such a sign convention was not explicitly incorporated in this work.

Metrics

Landmark detection accuracy of each approach was assessed using the Euclidean distance, which is calculated as follows:

$$\text{Distance}(\mathrm{p}_i, \hat{\mathrm{p}}_i) = \sqrt{(p_{i,x} - \hat{p}_{i,x})^2 + (p_{i,y} - \hat{p}_{i,y})^2}. \tag{6}$$

where $(p_{i,x}, p_{i,y})$ and $(\hat{p}_{i,x}, \hat{p}_{i,y})$ correspond to the ground-truth and estimated x-y coordinates of landmark *i*, respectively. Additionally, we compared the approaches by counting the number of detections above 10 mm, which we considered erroneous. This threshold was used as an operational criterion to identify large localization failures and quantify outlier detections, rather than as a strict clinical cutoff. To complement this analysis, we also report Successful Detection Rate (SDR) values across multiple thresholds. In parallel, we evaluated the proposed methods using the Successful Detection Rate (SDR), which is the ratio of estimated landmarks that fall within a reference threshold. SDR is computed as follows:

$$\text{SDR}_z = \frac{n_d}{n} \times 100\% \tag{7}$$

In (7), n_d denotes the number of landmarks whose Euclidean distance error is below the threshold z and n denotes the total number of evaluated landmarks. Regarding the alignment evaluation, we quantified the error between the ground truth values and those derived from the estimated landmark positions via the mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{m}_i - m_i| \tag{8}$$

where $\hat{m}_i$ represents the alignment metric value computed from the estimated landmark positions, m_i denotes the corresponding value computed from the ground-truth annotated landmarks, and n is the number of test samples.

Results

First, we compared the performance of the U-Net and Swin-UNETR models. Table 2 shows the achieved median Euclidean distance error and interquartile range (IQR) for each network type. From Table 2, we observe a superior performance from the Swin-UNETR model after the exclusion of false positives. In terms of the distance error, the difference between the Swin-UNETR network and the U-Net is marginal (1.74 mm and 1.75 mm, respectively). However, the Swin-UNETR yields significantly fewer false positives (160 vs. 502). Given the superior results achieved with the Swin-UNETR architecture, we employed this network in the end-to-end vs. cascade comparison.

Network	Euclidean distance [mm]	Missed detection	False positives
U-Net	1.75 (1.06)	4	502
Swin-UNETR	1.74 (0.98)	4	160

Table 2. Metrics for each image segmentation network for landmark detection. We report the median Euclidean distance, IQR, missed detections, and false positives.

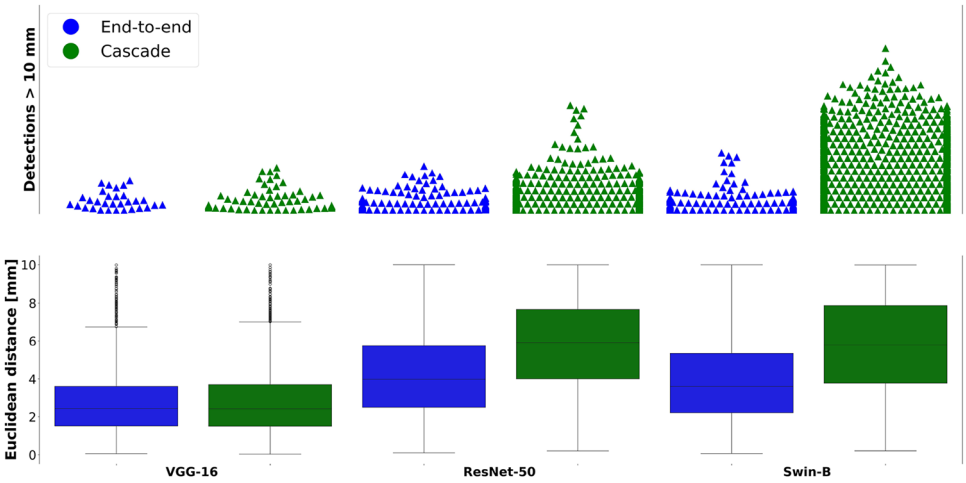

Fig. 5. Euclidean distances and detections above 10 mm for the end-to-end and cascade approaches using different regression heads.

Method	End-to-end	Cascade
VGG-16	**2.44 (2.12)**$^{*+}$	2.45 (2.31)
ResNet-50	**4.17 (3.51)**	7.95 (6.68)
Swin-B	**3.76 (3.52)**	17.81 (21.11)

Table 3. Landmark detection accuracy achieved for each training strategy combined with different regression heads. We report the median Euclidean distance error [mm] and IQR. Bold numbers indicate statistically significant differences (Wilcoxon test, $\alpha = 0.05$) between the end-to-end and cascade methods. Within the end-to-end approaches: * indicates statistically significant differences for VGG-16 vs. ResNet-50 (Wilcoxon test, $\alpha = 0.016$ after Bonferroni correction). $^{+}$ indicates statistically significant differences for VGG-16 vs. Swin-B (Wilcoxon test, $\alpha = 0.016$ after Bonferroni correction)

Figure 5 and Table 3 group the results over the eight targeted landmarks and show the distribution and numerical values of the Euclidean distance for the two training strategies using different regression heads. Figure 5 shows the distribution of the Euclidean distance, where the detections above 10 mm have been filtered and are displayed independently. Conversely, Table 3 describes the results obtained from the complete test data. From Fig. 5, we observe that VGG-16 is the optimal regression network. Regardless of the training modality (end-to-end or cascade), VGG-16 achieves lower Euclidean distance errors, leading to fewer detections above 10 mm. For ResNet-50 and Swin-B, a trade-off occurs in terms of median Euclidean distance error. Although ResNet-50 in the cascade configuration outperforms Swin-B, in the end-to-end approach, Swin-B surpasses ResNet-50. Altogether, the end-to-end approach performs better than the cascade one.

Regarding detections above 10 mm, Fig. 5 shows a notable performance difference between the training schemes and regression networks. Out of 2944 landmarks, we observe that VGG-16, in the end-to-end approach, yielded fewer (31) of this type of detection than the cascade approach (61). Meanwhile, the ResNet-50 produced 150 on the end-to-end configuration and 1046 on the cascade mode. The Swin-B architecture achieved 171 detections above 10 mm using the end-to-end approach and 2132 using the cascade. Table 3 confirms that the VGG-16 had the best performance compared to ResNet-50 and Swin-B since a significantly lower median Euclidean distance error (Wilcoxon test $\alpha = 0.05$) was achieved with this configuration.

To complement the quantitative evaluation, Fig. 6 provides a qualitative comparison of landmark localization performance across different backbone networks and training strategies. Consistent with the quantitative results, the end-to-end configuration produces predictions that are more closely aligned with the ground-truth landmarks than the cascade approach. Among the evaluated backbones, VGG-16 shows more consistent and accurate landmark localization, particularly in challenging cases such as prosthetic joints and low-contrast anatomical regions.

Figure 7 displays the SDR for each training and architecture setup, where all targeted landmarks are grouped together. The VGG-16 shows superior performance, achieving the highest SDR at all distances. At 3 mm, both VGG-16-based methods already achieve an SDR above 60%, a condition not observed for the ResNet-50 and Swin-B approaches. Additionally, at 6 mm, VGG-16 models reach SDR values of around 90%, a level that ResNet-50 and Swin-B cannot match even at 10 mm. Overall, in terms of SDR, VGG-16 is the best network, followed by Swin-B and ResNet-50, both on the end-to-end configuration.

In addition to previous findings, Fig. 7 shows a clear difference between end-to-end and cascade training for the ResNet-50 and Swin-B approaches. End-to-end training consistently outperforms cascade training at

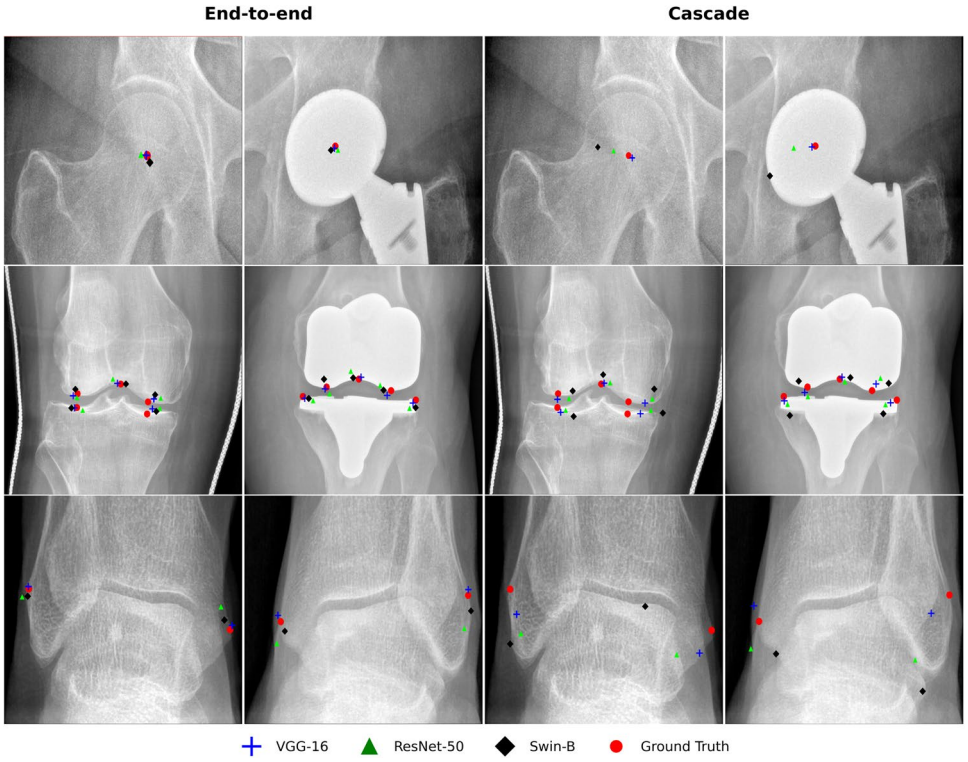

Fig. 6. Qualitative comparison of landmark predictions using end-to-end and cascade approaches. Example X-ray images from the hip, knee, and ankle are shown with overlaid landmark predictions from different backbone networks (VGG-16, ResNet-50, and Swin-B) together with the ground-truth annotations. The examples include both native and prosthetic joints and illustrate the visual differences in landmark localization between the two approaches.

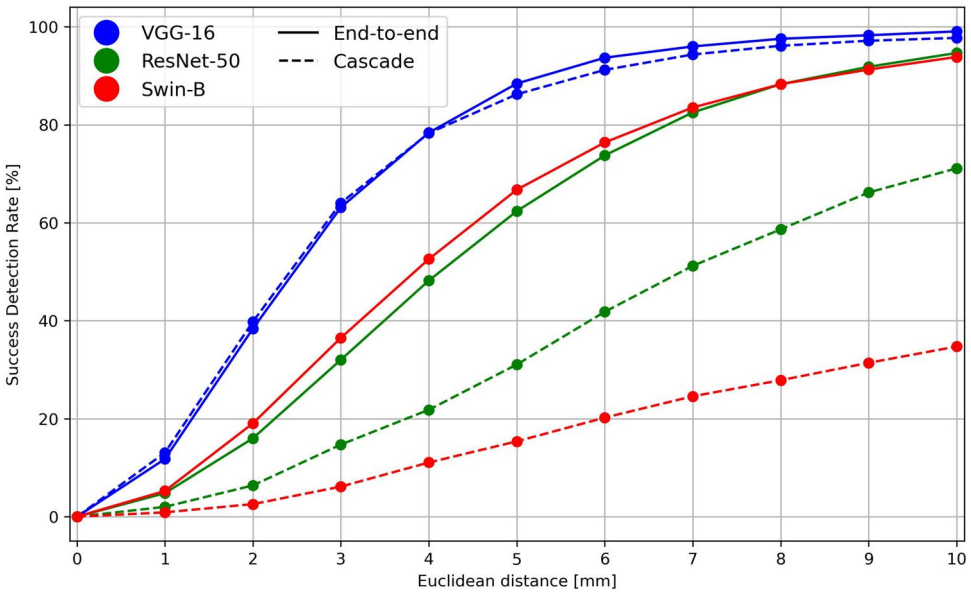

Fig. 7. The successful detection rate for the investigated training schemes and regression models.

Method	MAD [mm]	mLDFA [°]	mMPTA [°]	HKA value [°]	HKA error [°] $< 1.5°$
End-to-end					
VGG-16	$\mathbf{1.88 \pm 1.47}$	1.91 ± 1.39	1.10 ± 1.31	$\mathbf{0.47 \pm 0.37}$	**97.88**
ResNet-50	4.11 ± 2.90	2.57 ± 1.72	1.37 ± 1.35	1.1 ± 0.77	72.42
Swin-B	2.74 ± 2.18	2.34 ± 1.61	1.23 ± 1.33	0.71 ± 0.55	89.83
Cascade					
VGG-16	2.21 ± 1.93	$\mathbf{1.78 \pm 1.24}$	$\mathbf{0.94 \pm 1.09}$	0.56 ± 0.48	94.49
ResNet-50	6.13 ± 4.75	2.32 ± 1.63	1.73 ± 1.43	1.77 ± 1.28	47.03
Swin-B	7.37 ± 5.33	2.68 ± 1.90	1.81 ± 1.58	2.29 ± 1.53	35.16

Table 4. Mean absolute error values for each of the assessed alignment metrics. Bold numbers represent the approaches with the best performance.

all evaluated threshold levels. However, for the VGG-16 architecture, this difference is almost negligible and varies with the threshold level. Below 4 mm, cascade training yields higher SDR values compared to end-to-end training. Conversely, above a 4 mm threshold, end-to-end training achieves higher SDR values than the cascade approach.

To better understand landmark-dependent performance, we also conducted a landmark-wise analysis. Although Fig. 7 aggregates all landmarks, landmark-wise errors reveal a clear dependence on landmark type (Supplementary Table S2). For the best-performing configuration (VGG-16 end-to-end), the femoral landmark (HoF) achieved the lowest mean localization error (1.50 ± 1.02 mm), while the knee joint-line landmarks (Right/Left-FC and Right/Left-TP) exhibited higher mean errors (3.89–3.95 mm) with larger variability. In general, the cascade strategy amplified errors for the most challenging landmark types; for example, Swin-B in cascade mode produced markedly larger errors for several knee-related landmarks (e.g., Left-FC 30.54 ± 6.73 mm and Left-TP 34.70 ± 8.14 mm), indicating that these landmarks are more sensitive to error propagation and are harder to localize reliably. Overall, consistently higher mean errors and larger SDs for these landmarks indicate lower localization reliability and a higher likelihood of large-error outliers compared with more distinctive landmarks such as HoF.

In addition to continuous error measures, we report the percentage of cases with HKA error below $1.5°$ as an interpretable indicator of agreement with the reference measurements. This threshold was used as a practical tolerance to reflect small angular deviations and to facilitate comparison across models, rather than as a strict clinical decision boundary. Table 4 shows the mean absolute error achieved on the alignment metrics between the metrics computed using the ground truth landmarks and those achieved by employing the estimated landmark positions. Table 4 indicates that accurate landmark detection leads to correct alignment measurements, as evidenced by the VGG-16 model, which achieves lower error across all evaluated metrics. As shown in Table 4, we observe a trade-off in the alignment metrics' accuracy. The end-to-end approach using the VGG-16 network achieves the best results for MAD, HKA, and percentage of HKA measurements below $1.5°$. However, the cascade training scheme is superior for measuring mLDFA and mMPTA. The performance of VGG-16 is followed by that of Swin-B and ResNet-50, both in the end-to-end configuration. However, both are suboptimal, as errors in specific metrics are twice those of the VGG-16 network. This difference is more apparent in the percentage of HKA errors below $1.5°$, where the gap compared to Swin-B is 8.05%, and compared to ResNet-50 is 25.46%. From a clinical perspective, these results indicate that improvements in landmark localization are meaningful insofar as they translate into more reliable lower-limb alignment measurements, including MAD, mLDFA, mMPTA, and HKA. Because these measurements are routinely used in deformity assessment and preoperative planning, the proposed automation may support clinical workflow by reducing manual measurement burden and improving consistency across evaluations.

Discussion

In terms of image segmentation network performance, our results show that Swin-UNETR can detect anatomical landmarks as reliably as the traditional U-Net, while yielding fewer false positives. This outcome is particularly interesting given that transformer-based models such as Swin-UNETR are generally considered to require large-scale pretraining or extensive datasets to achieve optimal performance[20,27].

In our experiments, Swin-UNETR achieved nearly identical Euclidean distance errors to U-Net, but generated substantially fewer false positives (see Supplementary Table S1). This indicates that Swin-UNETR not only detects landmarks accurately but also focuses its predicted probability maps more precisely around the true anatomical locations. These results are consistent with reports by Hatamizadeh et al.[20], Dosovitskiy et al.[28], and Chen et al.[29] that the hierarchical Swin Transformer encoder, which utilizes shifted window-based self-attention, can effectively integrate local details and global context, thereby improving discrimination between true and false detections in medical image segmentation tasks.

When evaluating the coordinate regression architectures paired with the training modalities, VGG-16 proved to be the most accurate. In terms of Euclidean distance error, SDR, and alignment assessment, VGG-16 outperformed ResNet-50 and Swin-B. We hypothesize that the metrics achieved by VGG-16 result from its architecture, the choice of hyperparameters, and the nature of the task: estimating landmark coordinates. Both VGG-16 and ResNet-50 consist of stacked convolutional layers; however, VGG-16 includes three FC layers at

the end of the network that can be employed to learn the mapping from encoded image features to landmark coordinates better. Conversely, ResNet-50 has only one FC layer after the final convolutional block, which might not be sufficient to learn a pixel-coordinate representation. Furthermore, the VGG-16 architecture processes the entire region of interest, whereas Swin-B analyzes patches of the same image. This patch-wise approach could be hindering the Swin-B network from properly learning the mapping from image to coordinates.

Regarding the evaluation of end-to-end versus the cascade training scheme, we found that end-to-end algorithms generally outperformed their cascade counterparts, except for the VGG-16 model, where the difference was marginal. We theorize this superior performance arises from two main reasons: (1) the end-to-end architecture includes more parameters in its design and (2) the simultaneous optimization of both segmentation and regression models. Having more parameters enables the end-to-end method to learn the mapping from images to coordinates better. Additionally, due to the concurrent optimization, segmentation errors can be corrected. This does not occur in the cascade approach, where previously (erroneously) inferred segmentations could mislead the regression network and hamper optimal landmark detection.

As mentioned earlier, the performance of the VGG-16 configuration varied with the training scheme and the threshold used in the SDR analysis. We hypothesize that this threshold-dependent behavior may be related to the quality of segmentation produced during the initial phase of the cascade approach. If these segmentation maps are well-centered, they can provide strong spatial cues for precise regression. However, if segmentation deviates from the true landmark, the regression stage lacks a mechanism to correct it, since both stages are optimized independently. These findings emphasize the benefit of joint optimization when combining segmentation and regression tasks. Particularly in challenging clinical scenarios with noisy or ambiguous landmarks, end-to-end configurations appear to offer more resilience and consistency in final predictions.

Finally, the landmark-wise analysis in Supplementary Table S2 suggests that localization difficulty is landmark-dependent, with the knee joint-line landmarks (FC and TP) exhibiting higher mean errors and greater variability than the femoral landmark (HoF) and the ankle malleoli in most settings. This pattern likely reflects anatomical and projection-related differences across regions. The femoral head is a relatively large and distinctive structure, which may facilitate more stable localization. In contrast, the femoral condyles and tibial plateaus are defined along more complex and overlapping bony contours on AP full-length radiographs, making the corresponding knee landmarks more sensitive to small deviations in localization. Although ankle landmarks are often more sharply delineated, the malleoli can still be sensitive to the selected architecture and training scheme, as reflected by increased errors in certain configurations (Supplementary Table S2). These findings indicate that the observed performance differences are not only architecture-dependent, but also related to the anatomical distinctiveness and radiographic appearance of the target landmarks.

From a clinical perspective, the reported Euclidean and angular errors suggest that the best-performing configuration can provide lower-limb alignment measurements that are generally close to the reference annotations. In particular, the low HKA, MAD, mLDFA, and mMPTA errors indicate that improved landmark localization translates into more reliable alignment assessment, which is relevant for deformity evaluation and preoperative planning. However, even small landmark-localization errors may propagate to the derived measurements, and therefore the clinical significance of these deviations should be interpreted with caution. In addition, although the achieved SDR values indicate robust overall localization performance, clinical applicability depends not only on high average accuracy but also on reliability in difficult cases such as low-contrast regions, overlapping anatomy, and prosthetic joints. Finally, because all annotations were performed by a single radiologist, the present results reflect consistency relative to one expert reference rather than inter- or intra-observer variability. Future work should therefore include reproducibility analysis across multiple raters, as well as methodological refinements beyond architecture choice, such as stronger data augmentation, improved hyperparameter optimization, and validation on broader external datasets.

In summary, this study highlights several observations. First, within the landmark-segmentation framework, the Swin-UNETR architecture demonstrated improved performance relative to U-Net, as reflected by a lower incidence of failure detections. Second, end-to-end training showed advantages over the cascade approach, potentially because joint optimization allows partial mitigation of segmentation-related errors that would otherwise propagate through a sequential pipeline. Finally, among the evaluated regression heads, VGG-16 yielded the most favorable results, suggesting that its feature extraction characteristics may be well suited to this task. Overall, these findings provide insight into how training strategy and regression-head selection influence the performance of segmentation-guided landmark detection pipelines for lower-limb radiographic analysis.

Limitations & future work

The study's limitations include the dataset's size and diversity, which affect its generalizability. Annotation accuracy also impacts performance, making inter-annotator agreement analysis important. In addition, real-world validation in clinical settings is needed, along with further optimization of model efficiency.

A further limitation of this study is that all compared architectures were trained under the same hyperparameter configuration in order to maintain a controlled and reproducible evaluation protocol. Although this choice facilitates direct comparison across model families and training strategies, it may not reflect the optimal training regime for each architecture, particularly for transformer-based and residual models, which may require architecture-specific learning rates, schedules, regularization strategies, or longer training. Therefore, the reported results should be interpreted as comparative performance under a unified experimental setup rather than as the best achievable performance of each individual model. Future work should include systematic hyperparameter optimization tailored to each architecture and should also evaluate additional models while extending the analysis to broader anatomical regions beyond the hip, knee, and ankle to improve generalizability.

Future work should also include external validation on publicly available datasets, such as OAI or MOST, to improve reproducibility and enable broader comparison with previously published methods.

Conclusion
Accurate landmark detection (median error of 2.44 mm with an IQR of 2.12 mm) and lower limb alignment assessment are obtained when combining a Swin-UNETR model and a VGG-16 network following an end-to-end optimization. This arrangement demonstrated better performance than its cascade counterpart, even when paired with other models in the regression phase.

Data availability
No datasets were generated or analysed during the current study.

Code availability
The code used in this study is currently not publicly available.

Received: 14 January 2026; Accepted: 16 April 2026
Published online: 22 April 2026

References
1. Fan, Z. et al. The prevalence of hip osteoarthritis: A systematic review and meta-analysis. *Arthr. Res. Therapy* **25**(1), 51 (2023).
2. Herrera-Pérez, M. et al. Ankle osteoarthritis aetiology. *J. Clin. Med.* **10**(19), 4489 (2021).
3. Culvenor, A. G. et al. Prevalence of knee osteoarthritis features on magnetic resonance imaging in asymptomatic uninjured adults: A systematic review and meta-analysis. *Br. J. Sports Med.* **53**(20), 1268–1278 (2019).
4. Han, X. et al. Association between knee alignment, osteoarthritis disease severity, and subchondral trabecular bone microarchitecture in patients with knee osteoarthritis: a cross-sectional study. *Arthritis Res.Therapy* **22**, 1–11 (2020).
5. Luís, N. M. & Varatojo, R. Radiological assessment of lower limb alignment. *EFORT Open Rev.* **6**(6), 487–494 (2021).
6. Paley, D. *Principles of deformity correction* (Springer, Berlin, Heidelberg, 2014).
7. Vaishya, R., Vijay, V., Birla, V. P. & Agarwal, A. K. Inter-observer variability and its correlation to experience in measurement of lower limb mechanical axis on long leg radiographs. *J. Clini. Orthopaed. Trauma* **7**(4), 260–264 (2016).
8. Matassi, F., Pettinari, F., Frasconà, F., Innocenti, M. & Civinini, R. Coronal alignment in total knee arthroplasty: A review. *J. Orthop. Traumatol.* **24**(1), 24 (2023).
9. Fang, D. M., Ritter, M. A. & Davis, K. E. Coronal alignment in total knee arthroplasty: Just how important is it?. *J. Arthroplasty* **24**(6), 39–43 (2009).
10. Berend, M.E., Ritter, M.A., Meding, J.B., Faris, P.M., Keating, E.M., Redelman, R., Faris, G.W., & Davis, K.E. The chetranjan ranawat award: Tibial component failure mechanisms in total knee arthroplasty. *Clin. Orthopaed. Relat. Res.®* **428**, 26–34 (2004)
11. Collier, M. B., Engh, C. A. Jr., McAuley, J. P. & Engh, G. A. Factors associated with the loss of thickness of polyethylene tibial bearings after knee arthroplasty. *JBJS* **89**(6), 1306–1314 (2007).
12. Liau, J.-J., Cheng, C.-K., Huang, C.-H. & Lo, W.-H. The effect of malalignment on stresses in polyethylene component of total knee prostheses-a finite element analysis. *Clin. Biomech.* **17**(2), 140–146 (2002).
13. Myers, T. G. et al. Artificial intelligence and orthopaedics: An introduction for clinicians. *JBJS* **102**(9), 830–840 (2020).
14. Nguyen, T. P. et al. Intelligent analysis of coronal alignment in lower limbs based on radiographic image with convolutional neural network. *Comput. Biol. Med.* **120**, 103732 (2020).
15. Pei, Y. et al. Automated measurement of hip-knee-ankle angle on the unilateral lower limb x-rays using deep learning. *Phys. Eng. Sci. Med.* **44**, 53–62 (2021).
16. Moon, K.-R., Lee, B.-D. & Lee, M. S. A deep learning approach for fully automated measurements of lower extremity alignment in radiographic images. *Sci. Rep.* **13**(1), 14692 (2023).
17. Sanchez, S.A., Van Overschelde, P., & Vandemeulebroucke, J. Segmentation-guided coordinate regression for robust landmark detection on X-rays: Application to automated assessment of lower limb alignment. *IEEE Access* (2024)
18. Kim, Y.-T. et al. Hka-net: clinically-adapted deep learning for automated measurement of hip-knee-ankle angle on lower limb radiography for knee osteoarthritis assessment. *J. Orthop. Surg. Res.* **19**(1), 777 (2024).
19. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., & Schnabel, J.A. Left-ventricle quantification using residual u-net. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, 371–380 (Springer, 2019).
20. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., & Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, 272–284 (Springer, 2021).
21. Simonyan, K.,& Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556 (2014)
22. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021)
24. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T.,& Wirkert, S., et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint* arXiv:1809.10486 (2018)
25. Sanchez, S.A., Zarghami, A., Van Overschelde, P., & Vandemeulebroucke, J. Optimization and benchmarking of image segmentation for improved landmark detection in lower limb x-rays and accurate coronal plane alignment of the knee classification. *IEEE Access* (2025)
26. Watanabe, Y., Takenaka, N., Kinugasa, K., Matsushita, T. & Teramoto, T. Intra-and extra-articular deformity of lower limb: Tibial condylar valgus osteotomy (TCVO) and distal tibial oblique osteotomy (DTOO) for reconstruction of joint congruency. *Adv. Orthop.* **2019**(1), 8605674 (2019).
27. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A.,& Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357 (PMLR, 2022).
28. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929 (2020)
29. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., & Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* arXiv:2102.04306 (2021)

Author contributions
Conceptualization, A.Z. and S.A.S.; methodology, A.Z. and S.A.S.; validation, A.Z. and S.A.S.; formal analysis, A.Z. and S.A.S.; investigation, A.Z.; data curation, P.V.O.; writing–original draft preparation, A.Z. and S.A.S.;

writing–review and editing, A.Z., S.A.S., J.V.; All authors have read and agreed to the published version of the manuscript.

Funding
The presented job was funded by the Innoviris grant [BHF / 2020-RDIR-6a] of the Brussels-Capital Region, as part of the project ANTICIPATE on Augmented Intelligence in Orthopaedics Treatments.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval and consent to participate
We did not apply for ethical committee approval for this study. The retrospective and anonymized data used in this study was collected in previous studies. In those studies, patients signed an informed consent form, permitting the use of their data in future studies.

Consent for publication
Not applicable.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-49750-2.

Correspondence and requests for materials should be addressed to S.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

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