AUTOMATED LANDMARK DETECTION FOR ASSESSING HIP CONDITIONS: A CROSS-MODALITY VALIDATION OF MRI VERSUS X-RAY

Roberto Di Via,[1] *Vito Paolo Pastore,*[1] *Francesca Odone,*[1] *Siôn Glyn-Jones*[2] *and Irina Voiculescu*[3]*

[1]MaLGa Center, DIBRIS, University of Genoa, Italy
[2]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Oxford, UK
[3]Oxford University Department of Computer Science, UK

ABSTRACT

Many clinical screening decisions are based on angle measurements. In particular, FemoroAcetabular Impingement (FAI) screening relies on angles traditionally measured on x-rays. However, assessing the height and span of the impingement area requires also a 3D view through an MRI scan. The two modalities inform the surgeon on different aspects of the condition. In this work, we conduct a matched-cohort validation study (89 patients, paired MRI/x-ray) using standard heatmap regression architectures to assess cross-modality clinical equivalence. Seen that landmark detection has been proven effective on x-rays, we show that MRI also achieves equivalent localisation and diagnostic accuracy for cam-type impingement. Our method demonstrates clinical feasibility for FAI assessment in coronal views of 3D MRI volumes, opening the possibility for volumetric analysis through placing further landmarks. These results support integrating automated FAI assessment into routine MRI workflows.

***Index Terms*—** Femoroacetabular impingement, landmark detection, cross-modality, MRI, x-ray, deep learning

1. INTRODUCTION

Femoroacetabular impingement (FAI) is a pathomechanical hip disorder affecting 20–25% of the population, characterised by abnormal contact between the femoral head–neck junction and acetabular rim during motion [1]. Early identification is critical, as untreated FAI accelerates degenerative cartilage loss and predisposes to premature osteoarthritis [2]. FAI is the main cause of hip replacement. However, it is hard to diagnose at its early stages. Imaging is essential for early detection and treatment. Clinical assessment relies on two geometric parameters (Fig. 1): the α-angle, quantifying femoral head asphericity (cam morphology), and the lateral centre-edge (LCE) angle, measuring acetabular coverage (pincer morphology). Pathological thresholds are typically $\alpha{>}65^\circ$ and LCE$>40^\circ$ [3, 4].

Current clinical practice relies on anteroposterior (AP) pelvic x-rays, which have fundamental limitations from tissue density being projected into a single view plane. The foot-to-hip alignment and pelvic tilt can affect the plane in which the relevant angles are being measured, leading to a variability of up to 15° [5], sufficient to misclassify borderline cases. In this work, we investigate whether geometric FAI metrics can be reliably extracted from MRI alone. Whilst foot-to-hip misalignment can still be present in the coronal view, there are methods for correcting the misalignment, and measuring the angle in the correct plane. The first step towards this important 3D analysis is precisely showing that landmarks can be detected automatically in MRI slices.

*Correspondence to irina@cs.ox.ac.uk

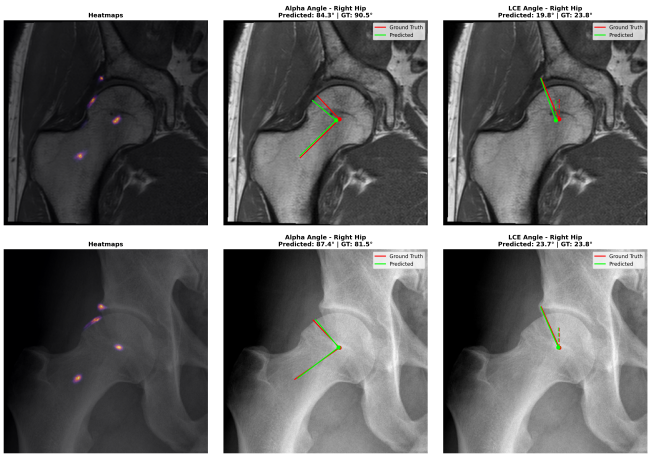

Fig. 1. FAI assessment from paired T1 MRI (top) and x-ray (bottom). The α-angle and LCE angle quantify femoral head asphericity and acetabular coverage, respectively. Our study confirms that automated landmark detection achieves equivalent accuracy across modalities, confirming the feasibility of MRI landmark detection.

Deep learning-based landmark detection have been used in diverse medical imaging tasks [6, 7, 8], with recent work demonstrating the feasibility of an automated FAI assessment from x-rays [9]. However, no prior work has validated automated FAI assessment on MRI or directly compared cross-modality performance on matched cohorts. MRI offers intrinsic advantages over x-ray, because the alignment of the hip can be corrected post-scan, and the true 3D nature of the impingement area can be assessed independently, either across multiple slices separately or, ideally, through a single volumetric assessment.

Contributions. (1) We present the first cross-modality validation of automated hip landmark detection between x-ray and MRI. Standard heatmap regression models achieve equivalent performance on T1-weighted MRI (2.98 MRI vs. 3.02 mm x-ray mean error, 87.5% diagnostic accuracy) on a matched patient cohort. This establishes coronal MRI as a successful modality for automated FAI assessment; (2) By extending measurements conventionally taken from x-rays to the corresponding slice in 3D MRI volumes, we enable seamless integration into clinical workflows; (3) We thereby establish a validated 2D foundation for future volumetric extensions towards full 3D geometric analysis and prediction.

2. METHOD

2.1. Problem Formulation

We formulate FAI angle computation as a landmark detection problem. For each hip, we localise four annotated anatomical keypoints: femoral head centre (FHC), neck-axis point (NA) along the femoral neck centreline, lateral acetabular edge (LAE), and lateral cam point (LCP) where the femoral head deviates from sphericity (see Fig. 2).

The α-angle is computed as the angle between the femoral neck axis (FHC→NA) and the cam deformity vector (FHC→LCP). Due to the LCP being hard to pinpoint – even by clinicians – α is notoriously difficult to calculate.

The easier LCE angle is calculated as the angle between the vertical axis and the acetabular coverage vector (FHC→LAE), quantifying lateral acetabular overhang. Angles are visualised in Fig 1.

2.2. Heatmap Regression Architecture

Following established practice in medical landmark detection [10, 11], we adopt a heatmap regression formulation. For each landmark $k \in \{1, \ldots, 4\}$, we generate a ground-truth Gaussian heatmap $H_k \in \mathbb{R}^{h \times w}$, centred at the annotated location (x_k, y_k) as $H_k(i,j) = \exp\left(-\frac{(i-y_k)^2+(j-x_k)^2}{2\sigma^2}\right)$, with σ=5 empirically providing the best trade-off between localisation precision and training stability. Encoder–decoder networks are trained to predict heatmaps $\hat{H}_k$ from input images $I \in \mathbb{R}^{512\times512}$ by minimizing the negative log-likelihood (NLL) over the four landmarks [12]. At inference, landmark coordinates are extracted as the spatial argmax of each predicted heatmap. For robustness, we employ test-time augmentation (TTA) by averaging heatmap predictions across different augmented views per test image, applying the stochastic transformations described in Sec. 3.1.

3. EXPERIMENTS

3.1. Dataset and Training Setup

We use a paired dataset of 89 patients who underwent AP pelvic x-ray and multiple hip MRI for FAI evaluation, collected at Oxford University as part of the FAIT trial [13]. This pathological cohort includes subjects with clinical and imaging evidence of FAI but without significant osteoarthritis or hip dysplasia. Data were acquired using the same protocol, across multiple UK sites, with some patients contributing multiple time points. Since acquiring such a rich dataset, with multiple imaging modalities, longitudinally for the same patient is logistically and clinically challenging, the resulting paired dataset remains relatively small.

For MRI, we only used T1-weighted coronal acquisitions consisting of 20 slices with 3.3 mm spacing. All images are resized/padded to 512×512 and min–max normalised to $[0, 1]$. All annotations are standardised to the middle slice (index 10), identified through pilot analysis of 20 subjects as the mid-acetabular coronal plane where all four landmarks (FHC, NA, LAE, LCP) are concurrently visible with maximal clarity. Future work will investigate automatic slice selection to eliminate manual standardisation. Example paired x-ray and MRI images with annotated landmarks are shown in Fig. 2. During training, stochastic augmentations include affine transforms (scale 0.95–1.05, translation $\pm5\%$, rotation $\pm10^\circ$, shear $\pm5^\circ$) and intensity jitter (brightness/contrast ±15–20%, gamma 0.85–1.15). The training:validation:test datasplit is 65:10:25 balanced across α-angle distributions given by Kolmogorov–Smirnov testing. This yields 105:17:40 images (57:8:24 patients).

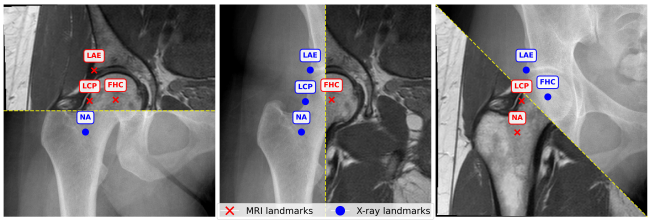

Fig. 2. Example of paired anteroposterior pelvic x-ray and corresponding T1-weighted MRI (slice 10) from the same subject, showing annotated landmarks: femoral head centre (FHC), neck-axis point (NA), lateral acetabular edge (LAE), and lateral cam point (LCP), illustrating their spatial correspondence across modalities.

We train a UNet++ with a ResNet18 encoder (ImageNet-initialised) using NLL loss. Optimisation uses AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-5}) with an ExponentialLR scheduler ($\gamma = 0.95$), batch size 4, and early stopping.

3.2. Evaluation Metrics

We evaluate performance across three hierarchical levels, and all metrics are computed independently for x-ray and MRI test sets. Experiment values are reported as mean ± standard deviation across three independent training runs with different random seeds.

Landmark localisation. We report (1) *Mean Radial Error* (MRE): the average Euclidean distance between predicted and ground-truth landmarks (mm); (2) *per-landmark MRE* for each keypoint [14]; and (3) *Success Detection Rate* at radius r (SDR@r): the percentage of landmarks localised within r mm.

Clinical angle assessment. We measure (1) Mean Absolute Error (MAE) between predicted and ground-truth angles; (2) the Intraclass Correlation Coefficient (ICC(2,1)), a two-way random effects model quantifying absolute agreement by partitioning variance into between-subject variability, between-rater systematic bias, and residual error through ANOVA decomposition of the paired measurements (model vs. clinician). ICC ranges from 0 (no agreement) to 1 (perfect agreement), with values <0.40 indicating poor, 0.40-0.59 fair, 0.60-0.74 good, and >0.75 excellent clinical reliability [15]. (3) the median absolute difference as a robust summary; and (4) Bland–Altman analysis [16] to decompose disagreement into systematic bias, proportional bias, and random error. For each subject, we compute the difference and mean between predicted and ground-truth angles, reporting: (i) mean bias (systematic over- or under-estimation), (ii) 95% limits of agreement (interval containing 95% of differences), and (iii) proportional bias via linear regression testing whether error magnitude depends on angle value.

Diagnostic performance. For cam-type impingement detection ($\alpha>65^\circ$), we report accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), together with the confusion matrix statistics.

4. RESULTS

4.1. Landmark Localisation Performance

Table 1 summarises localisation accuracy. Per-landmark metrics report mean ± standard deviation over three runs; the 'Overall' row shows per-run averages across landmarks, then aggregated across runs. MRI achieves an overall MRE of 2.98±0.23 mm, closely

Table 1. Landmark localisation accuracy for x-ray and MRI modalities. Mean and median radial errors (RE) are reported for each landmark. SDR@r indicates the success detection rate within r mm. ↓ indicates lower is better; ↑ indicates higher is better.

Landmark	Mean RE (mm) ↓		Median RE (mm) ↓	
	x-ray	MRI	x-ray	MRI
FHC	**1.07±0.08**	1.78±0.13	1.01±0.11	**0.84±0.14**
NA	2.35±1.02	**1.25±0.30**	1.61±0.17	**1.13±0.36**
LAE	3.18±0.64	**3.17±0.30**	2.81±0.74	**2.78±0.09**
LCP	**5.50±1.01**	5.72±0.76	**1.76±0.66**	2.55±0.64
Overall	3.02±0.10	**2.98±0.23**	**1.49±0.16**	1.62±0.29
SDR@2/3/4mm (%) ↑ x-ray: 62.5/79.2/85.4 **MRI:** 58.3/73.5/83.1				

matching that for x-rays (of 3.02±0.10 mm), demonstrating successful application of automated landmark detection to MRI. Individual landmark analysis reveals complementary strengths: x-rays excel at femoral head centre localisation (1.07 mm vs. 1.78 mm), likely due to higher bone-air contrast, while MRI shows superior neck-axis point detection (1.25 mm vs. 2.35 mm), potentially benefiting from improved soft-tissue delineation. Both modalities exhibit higher errors for the elusive lateral cam point (LCP: 5.50–5.72 mm mean), reflecting its clinical variability and frequent inter-observer disagreement; median errors (1.76–2.55 mm) are considerably lower, indicating that outliers in borderline FAI cases drive the means. Success detection rates at clinical thresholds (2–4 mm) remain similar for MRI 58–83% vs. x-ray 63–85%.

4.2. Angle Agreement and Diagnostic Accuracy

Table 2 summarises the agreement in angle measurement and cam screening results. Both modalities agree strongly on the LCE-angle (x-ray ICC: 0.82, MRI ICC: 0.73, excellent clinical significance) with small median errors (2.22° and 1.28°), indicating the geometric stability when localising the lateral acetabular edge. In contrast, the α-angle shows only moderate intra-modality agreement (ICC: 0.52 for x-ray, 0.41 for MRI; fair clinical significance) and larger MAEs (12.18°, 13.64°), comparable to the inter-observer variability among clinicians (≈ 0.45–0.56) [17], reflecting the inherent ambiguity of the cam deformity point. Median errors remain clinically acceptable (5.61°, 5.45°), suggesting that outliers inflate the mean disproportionately. Although continuous α-angle agreement differs modestly, both modalities achieve the same diagnostic accuracy for cam-type impingement (87.5%) considering the best runs. They have perfect specificity (100%) and equal sensitivity (54.6%), indicating that the remaining variability lies within clinically acceptable limits.

4.3. Bland–Altman Agreement Analysis

To complement the ICC results and assess continuous measurement agreement, we perform Bland–Altman analysis comparing predicted and ground-truth angles (Fig. 3). Each plot shows the mean of the two measurements on the x-axis and their difference on the y-axis; dashed blue line indicates mean bias and red ones mark the 95% limits of agreement (LoA). Points near zero bias and within the LoA indicate good agreement, while a regression trend denotes proportional bias. For the LCE-angle, MRI shows minimal bias (0.87°) and narrow LoA (±10.5°) without proportional bias ($p = 0.304$). In contrast, x-ray shows wider LoA (±21.6°) and proportional bias ($p < 0.001$), with errors increasing at higher acetabular coverage,

Table 2. Comparison of angle measurement agreement and diagnostic accuracy between x-ray and MRI. Results are reported as mean ± standard deviation across three different runs.

Metric	x-ray	MRI
LCE-angle agreement		
MAE (°) ↓	2.97±1.14	**2.96±0.52**
Median (°) ↓	2.22±0.42	**1.28±0.20**
ICC (2,1) ↑	**0.82±0.20**	0.73±0.14
α-angle agreement		
MAE (°) ↓	**12.18±1.63**	13.64±1.77
Median (°) ↓	5.61±1.41	**5.45±0.77**
ICC (2,1) ↑	**0.52±0.06**	0.41±0.15
Cam screening $\alpha > 65°$		
Accuracy (%) ↑	87.50	87.50
Sensitivity / Specificity (%) ↑	54.55 / 100.00	54.55 / 100.00
PPV / NPV (%)	100.00 / 85.29	100.00 / 85.29
TP / FP / TN / FN	6 / 0 / 29 / 5	6 / 0 / 29 / 5

suggesting more stable estimation from MRI. For the α-angle, both modalities underestimate clinician measurements (x-ray: $-2.68°$, MRI: $-7.39°$) with broad LoA (x-ray: $\pm 35.3°$, MRI: $\pm 38.3°$). X-ray errors remain magnitude-independent (p=0.518), whereas MRI shows proportional bias (p=0.010), reflecting slice-selection sensitivity. Despite these biases, both modalities achieve reliable cam-type classification (specificity 100%, sensitivity 54.6%), indicating that residual variability mainly arises from landmark ambiguity rather than modality differences.

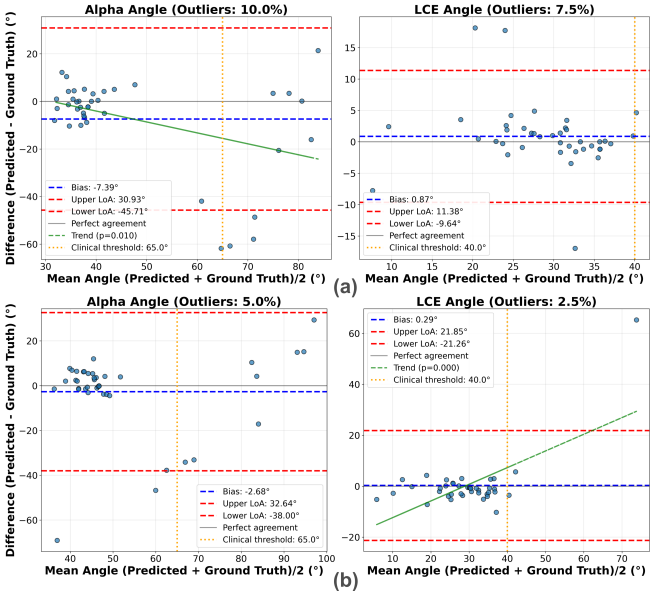

Fig. 3. Bland–Altman plots comparing predicted and ground-truth α- (left) and LCE-angles (right) for MRI (a) and x-ray (b).

4.4. Ablation Studies

Table 3 reports the effects of architecture choices and test-time augmentation (TTA) on the MRI results. To identify optimal network configurations, we compare established architectures (UNet++, UNet, DPT) with both convolutional (ResNet18, VGG16) and

Table 3. Ablation study of architecture–encoder combinations and test-time augmentation (TTA) on MRI performance, reporting mean radial error (MRE, mm), success detection rate at 2 mm (SDR@2 mm), and cam-type diagnostic accuracy (Cam Acc.).

Model	TTA	MRE (mm) ↓	SDR@2mm (%) ↑	Cam Acc. (%) ↑
UNet++ (ResNet18)	×	3.02 ± 0.28	56.46 ± 5.87	82.33 ± 3.79
UNet++ (ResNet18)	✓	$\mathbf{2.98 \pm 0.23}$	$\mathbf{58.44 \pm 4.86}$	$\mathbf{84.75 \pm 2.77}$
UNet++ (VGG16)	×	3.33 ± 0.03	56.25 ± 0.89	80.00 ± 7.07
UNet++ (VGG16)	✓	$\mathbf{3.29 \pm 0.36}$	$\mathbf{57.09 \pm 1.93}$	$\mathbf{80.17 \pm 4.41}$
UNet (MiT-B1)	×	4.11 ± 0.84	54.79 ± 1.44	66.67 ± 12.58
UNet (MiT-B1)	✓	$\mathbf{3.98 \pm 0.54}$	$\mathbf{57.33 \pm 0.91}$	$\mathbf{69.83 \pm 6.85}$
DPT (ResNet18)	×	4.49 ± 0.54	48.34 ± 3.79	67.50 ± 2.50
DPT (ResNet18)	✓	$\mathbf{4.44 \pm 0.19}$	$\mathbf{50.21 \pm 3.02}$	$\mathbf{70.17 \pm 6.56}$
DPT (MaxViT-Base)	×	3.91 ± 0.87	55.50 ± 0.71	73.75 ± 8.84
DPT (MaxViT-Base)	✓	$\mathbf{3.63 \pm 0.99}$	$\mathbf{55.94 \pm 1.33}$	$\mathbf{77.50 \pm 10.61}$

transformer-based (MiT-B1, MaxViT-Base) encoders of similar parameter counts (11–15M for lightweight, 119M for MaxViT-Base). All the encoders are ImageNet-initialised. Our objective is not architectural novelty but rather systematic validation of which existing designs transfer best to MRI-based FAI assessment. UNet++ with ResNet18 achieves optimal performance (MRE: 2.98±0.23 mm, mean cam accuracy: 84.75±2.77%), outperforming standard UNet (3.98 mm, 69.83%) and DPT variants (3.63–4.44 mm, 70–77.5%). Lighter encoders (ResNet18) outperform heavier alternatives, suggesting moderate-capacity networks with multi-scale skip connections are well-suited for this task. TTA consistently improves all architectures, providing 1.4–3.6 percentage point gains in cam accuracy while reducing variance. X-ray results show similar trends.

5. CONCLUSION & FUTURE WORK

Our study focuses on patients with FAI, whose diagnosis and management, including surgical planning, depend on an accurate two-dimensional and volumetric assessment of the impingement. By demonstrating that automated landmark detection on MRI achieves accuracy equivalent to that on x-rays, we establish that geometric measures of FAI can be extracted reliably from routine MRI scans. This represents an important first step towards supporting MRI as a suitable modality for automated FAI assessment.

Although AP x-rays remain the first-line tool for clinical screening due to their speed, low cost, and accessibility, they are not ideal for surgical planning. Our findings highlight the potential of MRI for quantitative geometric assessment in musculoskeletal analysis. Beyond the demonstrated equivalence, the volumetric nature of MRI offers opportunities for refined development of screening angles, particularly in patients with asymmetries or pelvic rotations, where measurements must be made relative to local axes.

Current limitations relate to analysing a single MRI slice, which underutilises volumetric information and disregards inter-slice coherence. The moderate α-angle reliability is due to the cam-point ambiguity and the modest size of the cohort.

Future work will extend the approach towards volumetric metrics averaged across slices or computed directly from 3D MRI volumes, exploring adjacent-slice training for data augmentation and incorporating uncertainty quantification to identify unreliable predictions.

This research was conducted retrospectively using human subject data from the FAIT study [13], which received ethical approval. The authors have no conflicts of interest to declare.

6. REFERENCES

[1] R. Ganz et al., "Femoroacetabular impingement: A cause for osteoarthritis of the hip," *Clinical Orthopaedics and Related Research*, vol. 417, pp. 112–120, 2003.

[2] R. Agricola et al., "Cam impingement of the hip: A risk factor for hip osteoarthritis," *Nature Reviews Rheumatology*, vol. 9, no. 10, pp. 630–634, 2013.

[3] C. R. Fraitzl et al., "Femoral head-neck offset measurements in 339 subjects: Distribution and implications for femoroacetabular impingement," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 21, no. 5, pp. 1212–1217, 2013.

[4] S. Kutty et al., "Reliability and predictability of the centre-edge angle in the assessment of pincer femoroacetabular impingement," *International Orthopaedics*, vol. 36, no. 3, pp. 505–510, 2012.

[5] J. C. Clohisy et al., "A systematic approach to the plain radiographic evaluation of the young adult hip," *J Bone and Joint Surgery*, vol. 90, no. Supplement 4, pp. 47–66, 2008.

[6] A. Clement et al., "Confidence in angle predictions for clinical decision support," in *MICCAI*, 2025.

[7] R. Di Via et al., "Is in-domain data beneficial in transfer learning for landmarks detection in x-ray images?," in *ISBI*, 2024, pp. 1–5.

[8] N. Patel et al., "A handful of data: Evaluating few-shot incremental landmark detection," in *ICIAP*, 2025.

[9] J. McCouat et al., "Automatically diagnosing hip conditions from x-rays using landmark detection," in *ISBI*, 2021, pp. 1273–1277.

[10] R. Di Via et al., "Self-supervised pre-training with diffusion model for few-shot landmark detection in x-ray images," in *WACV*, 2025, pp. 3886–3896.

[11] C. Payer et al., "Regressing heatmaps for multiple landmark localization using CNNs," in *MICCAI*, 2016, vol. 9901, pp. 230–238.

[12] J. McCouat et al., "Contour-hugging heatmaps for landmark detection," in *CVPR*, 2022, pp. 21544–21552.

[13] A. J. R. Palmer et al., "Protocol for the femoroacetabular impingement trial (FAIT): A multi-centre randomised controlled trial comparing surgical and non-surgical management of femoroacetabular impingement," *Bone & Joint Research*, vol. 3, no. 11, pp. 321–327, 2014.

[14] R. Di Via et al., "Are x-ray landmark detection models fair? A preliminary assessment and mitigation strategy," in *ICCV Workshops*, 2025.

[15] Domenic V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[16] Bland et al., "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, pp. 307–310, 1986.

[17] Ewertowski et al., "Automated alpha angle measurement on anteroposterior pelvic radiographs for cam morphology detection: validation against expert observers," *Skeletal Radiology*, vol. 51, no. 6, pp. 1177–1186, 2022.