Confidence in Angle Predictions for Clinical Decision Support

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Landmarks are used for clinical measurements, screening and to guide treatment decisions. In this work, we explore the clinical application of landmark-based angle measurements, with a particular aim of screening infants for Developmental Dysplasia of the Hip (DDH).
Our automated machine method uses a simple UNet++ architecture. The network is used to predict landmark heatmaps which represent the landmark localisation certainty. A Monte Carlo-like approach is then used to approximate an angle distribution from the landmark heatmaps. We propose a confidence metric from the derived angle distributions.
Multiple clinician annotations are combined and compared to the machine predictions. The machine-generated angle distribution is verified by confirming the correlation of the mean angle values and standard deviations per scan, between the multiple clinicians and the machine. The confidence scores correlate for the clinicians combined and the machine. The confidence of the machine strongly correlates with the sum of the confidence scores given by clinicians for each scan.
This work is the first to present a method for estimating the distribution of clinically relevant angles from predicted landmarks. Landmark-based angle confidence can establish robust methods and increase clinician trust in using automated or computer-aided methods.

Keywords: Angles · Heatmaps · Landmark detection · Probability Distribution · Ultrasound

1 Background

Clinical Relevance. Orthopaedic diseases such as Developmental Dysplasia of the Hip (DDH), Osteoarthritis (OAI), Adolescent Idiopathic Scoliosis (AIS) and knee deformities or instability, all require anatomical landmarks to be located in order to calculate diagnostic or treatment angles. Automated methods show reasonable performance in predicting angles when compared to ground truth annotations from one reviewer[10, 1, 7, 12].

Infant DDH is a condition in which the femoral head is not aligned with the acetabular socket. Early diagnosis is critical, as late diagnosis requires complex surgical procedures and prolonged hospital stays. Ultrasound (US) imaging is generally used for screening due to its accessibility. However, US images

are subjective and interpretation can be difficult [13]. The method pioneered by Dr. Reinhard Graf [4] assesses DDH severity. The Graf classification is determined from an anatomical angle (α), derived from manual landmark annotations (Fig. 1). Patients are screened as normal ($\alpha \geq 60°$) or abnormal ($\alpha < 60°$).

Previous methods for automating α angle measurements report within a reasonable range of clinical annotations (2.2°-3.8° of the ground truth α angle) [5, 2, 3]. None of these methods incorporates a probability distribution or a confidence measure in the reported angle measurement. This is needed to evaluate the angle range and to propose a confidence measure of the predicted angle.

Angle Estimation. Work to-date in Orthopaedics has calculated angles from the centre of predicted anatomical landmark annotations. A few methods have analysed the absolute difference of the machine predicted angles compared to clinicians [5, 7, 12]. These papers report the mean and standard deviation of the absolute difference in predicted angle across the dataset, but none have reported an angle distribution or angle confidence for each scan. Estimating an angle distribution can convey the probability of angle predictions. This can improve the robustness and trustworthiness of automated methods to promote clinical deployment by notifying when a machine-predicted angle has low confidence.

Landmark Localisation. Landmark 'heatmaps' are increasingly used to represent automated landmark predictions [9, 3, 2, 16]. Landmark 'heatmaps' allow for uncertainty in landmark placement to be quantified and visualised. Uncertainty metrics, such as the Expected Radial Error (ERE) [9], can be used to quantify the level of confidence in an individual landmark placement. Annotations are known to vary between clinicians (inter–rater variability). Landmark uncertainty accounts for inter–clinician variability in landmark placement [9, 14].

This work leverages landmark prediction uncertainty and extends this to estimate and validate an angle distribution that reflects inter–rater variability.

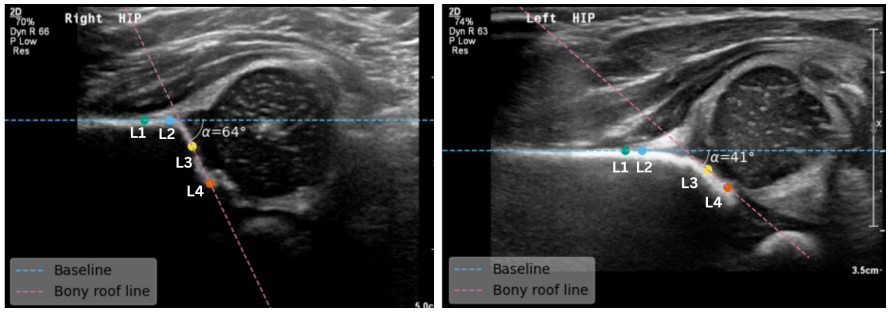

Fig. 1. The Graf Angle Calculation. Left: normal ($\alpha \geq 60°$); Right: abnormal ($\alpha < 60°$). Landmarks are used to calculate the Graf α angle: ilium points (L1, green and L2, blue), turning point (L3, yellow), lower limb point (L4, orange). The lines connect the landmarks to create the baseline (blue) and the bony roof line (purple).

Contributions

a. Estimate landmark-based angle distributions,
b. Validate the angle distribution method for machine-based predictions against multiple clinicians,
c. Assess the inter–rater angle variability for DDH which, in turn, informs the achievable precision of machine angle predictions,
d. Propose a confidence measure from the angle distribution and assess its correlation with clinician confidence.

2 Methods

2.1 Datasets

Main Dataset. A total of 1107 US scans were acquired from an Anonymous Hospital. The pixel size is 0.07×0.07 mm. This will be referred to as the *Main Dataset*. Landmarks were placed at four key anatomical points. The landmarks are used for the Graf α calculation (Fig. 1). Annotations in the *Main Dataset* were reviewed and confirmed by two additional expert clinicians. These annotations are referred to as the *Reference Expert*.

Data Subset. From the *Main Dataset*, 70 images were selected randomly. All images in this subset were annotated by 10 clinicians (n=700 annotations total). The 10 clinicians consisted of 5 Experts (consultants/attending) and 5 Non-Experts (registrar/residents). This will be referred to as the *Data Subset*.

Clinician Confidence. Clinicians were asked to give a score for their confidence in landmark placement for each scan. This score was from 1-5, with 1 meaning low confidence and 5 meaning high confidence. The sum of confidence across clinicians was used as a measure of clinician confidence for each scan.

Clinical Aggregated Landmarks. Post-processing was applied to the 10 Clinician annotations. We assumed that if we asked the clinician to place the same landmark multiple times, each precise placement would vary. To model this, the centre point of each landmark was used as the centre of a Gaussian distribution (σ=1). This represents the probability of the location of each landmark. We take a similar method for automated landmark generation (discussed in Section 2.2).

An aggregated heatmap of all 10 clinicians was generated by taking the sum across all annotations. The sum of the Gaussian heatmaps, for each landmark distribution was aggregated and then standardised $[0, 1]$ for comparisons with the machine method (see Figure 2). This is referred to as the Clinician Aggregated Landmark Distribution (CALD). This allows us to create a 'heatmap' that represents a probability distribution of the location of each landmark, illustrating the extent to which the reviewers agree. A wider spread of a landmark heatmap corresponds to weaker agreement, whereas a tight spread and bright colour correspond to stronger clinician agreement.

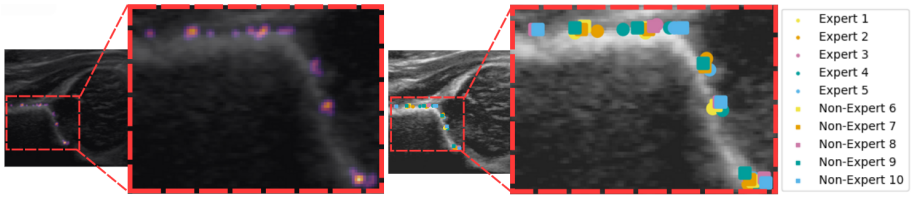

Fig. 2. Annotations and the Clinician Aggregated Landmark Distribution (CALD). The left, shows the combined heatmap from the sum of all 10 clinician Gaussian distributions (σ=1). The right shows the landmarks given by all 10 clinician annotations separately. Each side shows the entire image, with a section outlined in red dashed lines. The larger boxes outline a zoomed-in section.

Compliance with ethical standards The *Main Dataset* was retrospectively collected from images taken during routine screening at an Anonymous Hospital. A Philips EP1Q5G (L12-5 linear probe) US machine was used. Data was anonymised and stored securely in a Research Department. The study was approved by the required Anonymous Authorities and the Anonymous University. Predictive models ran on the Anonymous computational facilities.

2.2 Machine Landmark Predictions

The network input was an US image. A Gaussian heatmap (σ=1,using `scikit-image filter`[15]) was output for each landmark for training. The output of the model had 4 channels, one landmark per channel. This creates a per-pixel probability of the landmark location, which is referred to as a 'heatmap'. To exclude predictions near zero, any value below 0.05 in the heatmap was replaced with zero and normalised to $[0, 1]$.

Additional threshold post-processing was applied but first varied as a hyperparameter. This was done by taking a percentile of the total probability and setting anything outside that percentile to zero. This was done at multiple thresholds. Due to space limitations, we only report the landmark heatmap thresholded at 91%. This percentage is close to 100% due to the relative number of pixels that represent a landmark, compared to the total number of pixels in the image. Finally, these heatmaps are standardised $[0, 1]$.

All images were padded and resized to 512×352 in the data loader. The augmentation of images was done with speckle noise (factor$\leq$2), intensity shift (factor$\leq$1.0), scaling (factor$\leq$0.9), and translation ($x\leq$0.1 percentage of pixels, $y\leq$0.1 percentage of pixels). Three augmentations were randomly selected and applied to each image during data loading (`imgaug` [8]). A UNet++ [6] with a ResNet34 encoder was pre-trained on ImageNet data using `PyTorch` [11]. The decoder had five layers (256, 256, 256, 128, 64) each layer included batch normalisation, followed by a ReLU activation. Attention in the decoder was applied with Spatial and Channel 'Squeeze and Excitation Blocks'. The network was implemented using `Segmentation Models Pytorch` [6]. A spatial softmax func-

tion was applied to each channel for a probability-like distribution. The Negative Log Likelihood (NLL) loss function was used with L2 regularisation to optimise the model (batch size 6, epochs 20, and learning rate 0.005). The *Main Dataset* was split into training, validation and testing (70%:15%:15%). The class-balance was evenly distributed across all data splits (60.3% normal and 30.7% abnormal). The machine test set includes the same 70 scans as in the *Data Subset*.

2.3 Monte Carlo Angle Estimate

A Monte Carlo-like simulation was used to generate the angle probability distribution (10,000 iterations). In each iteration, a point was randomly selected from each landmark heatmap and then used to calculate α. The frequency of the α angles over all iterations was calculated. These values were binned (b=100) by α value to generate a histogram and rescaled between 0 and 1. The histogram represents the probability distribution of the α angles for each patient. This method was used to calculate the angle probability distributions for the CALD heatmaps and the machine heatmaps.

3 Evaluation

Mean and Standard Deviation. For each patient, the mean and standard deviation is calculated for both the CALD and the machine. A scatter plot is used to evaluate the correlation between the two. The absolute difference in α is measured between *Reference Expert* and the machine prediction. This is averaged for all patients, yielding an average absolute α difference. The average absolute α difference is calculated as well between the *Reference Expert* and the CALD. The Pearsons Correlation Coefficient (PCC) is reported for all comparisons.

Confidence Measures. The average of the average absolute α difference (see Section 3: Mean and Standard Deviation) between the *Reference Expert* and the CALD is used as a lower and upper bound around the mean. This is called the clinician mean-based interval. The area of the α distribution in this interval is defined as the confidence. The confidence is calculated for each patient for both: the machine and the CALD α distributions. This machine confidence calculated by the clinician mean-based bound, is compared to the sum of all of the clinician's confidence scores for each scan.

Similarly, the decision support confidence interval is generated using the Graf boundary. The area is calculated where the patient is normal ($\alpha \geq 60°$). This represents how confidently the α lays on the normal side of the decision boundary.

4 Results

Monte Carlo Angle Estimation Fig. 3 shows the output angle distribution for two different patients. It shows an example where one patient has larger landmark spread visually than another. As expected the wider spread in landmarks

results in a wider α probability distribution. This is reflected in the standard deviations and confidence values reported.

Mean and Standard Deviation. The average absolute α difference calculated between the CALD and the *Reference Expert* is 5.58°. When compared to the *Reference Expert*, the machine has a slightly lower α difference (3.76°).

The mean α for each patient predicted by the machine and the CALD are strongly correlated (r=0.90). The standard deviation of the machine and the CALD has a weak correlation (r=0.53). This is shown in Fig. 4.

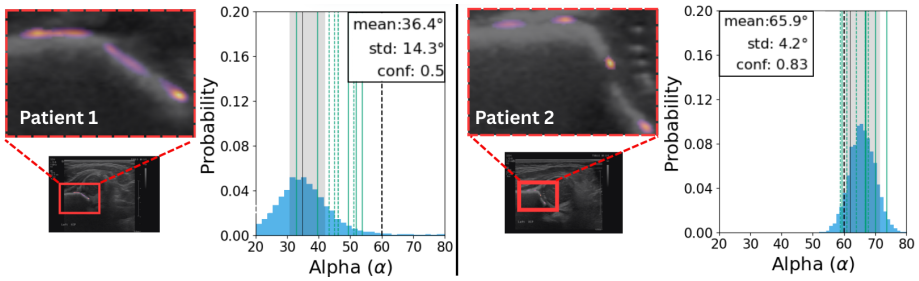

Fig. 3. Monte Carlo Angle Estimation. The right and left sections of the figure show examples of the α distributions for two unique patients. For each, an enlarged section of the patient scan with the machine-generated heatmaps. Associated α distribution is plotted for the resulting angles. All unique Expert mean α values are shown as vertical lines (green dashed for Non-Experts, green solid for Experts and black for the *Reference Expert*). A black dashed line illustrates the Graf decision boundary for abnormal vs. normal (60°). The grey area illustrates the clinician mean-based confidence interval (see Section 3). The mean, standard deviation (std) and confidence (conf) are reported.

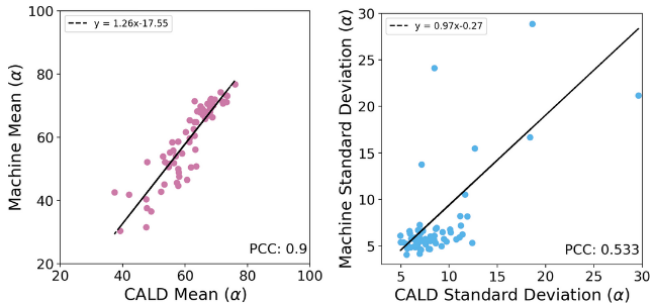

Fig. 4. Quantitative Results. The mean α and standard deviation are plotted for each patient. Left: comparison of the mean from the machine and the CALD. Right: comparison of the standard deviation of the machine and the CALD. All units are in α (°). The line of best fit is plotted for each graph. PCC are reported on each scatter plot.

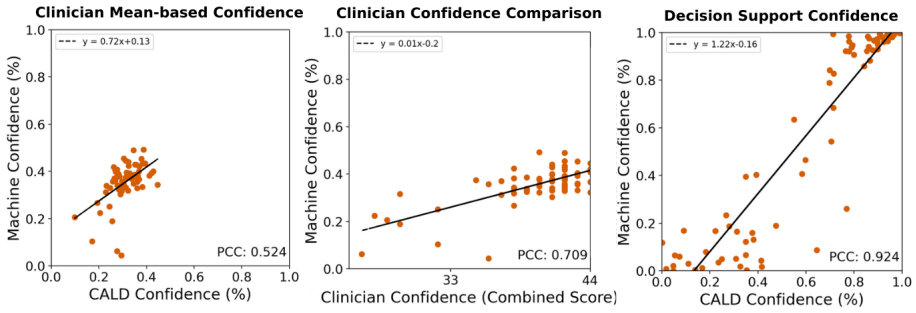

Fig. 5. Machine Confidence. Left: the clinician mean-based Machine confidence and CALD confidence. Centre: the sum of the clinician confidence scores against the clinician mean-based machine confidence. Right: the Graf decision boundary Machine confidence and CALD confidence.

Confidence Measures. There is a weak correlation (r=0.52, left of Fig. 5) between the machine confidence and the CALD confidence when using the clinician mean-based interval. There are three distinct outliers, where the machine predicts a confidence $< 0.1\%$. When examining the outliers, we see the landmark heatmaps overlap and are all extreme cases (severe dislocation/low α, Fig. 6).

There is a good correlation (r=0.7, centre of Fig. 5) between the Machine confidence predicted using the clinician mean-based interval and the reported clinician combined score. Outliers for the machine confidence in this plot (Fig. 5, left) are the same as the outliers identified in Fig. 6. A strong correlation exists (r=0.92, right of Fig. 5) between CALD confidence and machine confidence using the clinical decision boundary threshold (Fig. 6, interval visual).

5 Discussion and Conclusions

This work successfully leveraged landmark uncertainty to create a method for predicting angle probability uncertainty maps. This is illustrated in Fig. 3. This method is validated by the strong correlation in the mean α predicted by the CALD and the machine (r=0.9, Fig 4). The weaker correlation in standard deviation could be caused by the outliers consistent across all evaluation metrics. In Fig. 6 we see a large standard deviation. This is likely caused by the heatmaps of the landmarks overlapping (L1 and L2, Fig. 1), which results in a large variation of angles. Outliers with large distribution could have been caused by the class split available, or the visual heterogeneity within the abnormal class.

The α difference of the *Reference Expert*, compared to the α CALD, was 5.58° which suggests large inter–rater variability for DDH (see Section 4). This work is the first to report the variation across 10 clinicians and therefore helps to define a reference for the achievable machine precision needed for this task. The average absolute α difference in the machine predictions and the *Reference Expert* was smaller than the CALD (3.78°). This is likely because the machine was trained

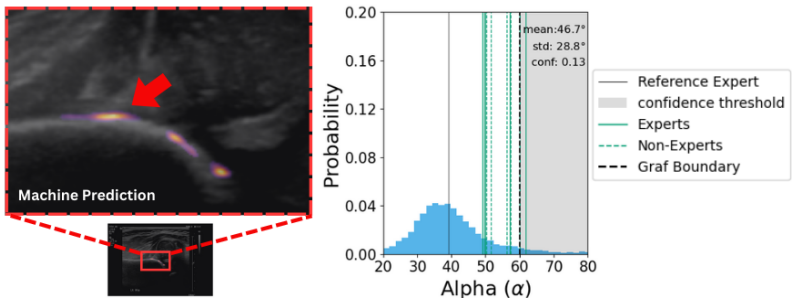

Fig. 6. Outlier Example. An example of a large standard deviation. Visual differences in the heatmap are overlaid on the ultrasound images (left). The associated angle distribution and standard deviation (right). The red arrow highlights overlapping heatmaps.

by the *Reference Expert*. Current methods in the literature report an average absolute α difference for the Graf method (2.2-3.8°) [5, 3], which are similar to our machine-generated predictions (3.78°). Although previous work has had a smaller difference, these are only relative to one clinical ground truth. This potentially suggests current single-reviewer-based models may be over-fitting. Future work can use the inter–rater variability outlined by this work as guidance for an achievable precision. We propose that, since machine-generated values of α have already exceeded clinical precision, there should be a move to focus on model generalisation and confidence metrics to flag difficult cases.

The proposed confidence metric is verified by the correlation between the total confidence score of clinicians and machine confidence (centre, Fig 5). The confidence measure is further verified by the strong correlation of machine confidence and CALD confidence when using the decision boundary as a reference (right, Fig. 5). Together, these results illustrate the potential use of the proposed angle confidence metric for clinical applications. The confidence metric developed in this paper can help promote trustworthy models by having the ability to report a confidence value that has been proven to correlate with clinicians. Future work must review clinical boundaries and define an angle confidence value that is clinically acceptable. Once these acceptable values are defined, this method can be deployed to create diagnostic safeguards by reporting uncertain scans.

Conclusion This work presents an automated method for computing possible angle ranges and probabilities from landmark heatmaps. We show that angle ranges predicted using this pipeline have a variability similar to that of clinical variability. Landmark-based angle screening with uncertainty metrics can help develop trust in automated methods for orthopaedic tasks by having a method for quantitative safeguards for flagging when the model is highly uncertain.

Acknowledgments. We are grateful for Dr. A Anonymous, Dr. B Anonymous and the Anonymous hospital for overseeing and ensuring all ethical standards were met.

We also thank them for securing the anonymised data and annotations used in this project. We extend gratitude to all clinicians involved in the annotation process.

References

1. Chen, B., Xu, Q., Wang, L., Leung, S., Chung, J., Li, S.: An automated and accurate spine curve analysis system. Ieee Access **7**, 124596–124605 (2019)
2. Chen, Y.P., Fan, T.Y., Chu, C.C., Lin, J.J., Ji, C.Y., Kuo, C.F., Kao, H.K.: Automatic and human level graf's type identification for detecting developmental dysplasia of the hip. Biomedical Journal **47**(2), 100614 (2024)
3. Clement, A., Singh, A., Perry, D., Voiculescu, I.: Improving automated ultrasound infant hip screening using an integrated clinical classification loss. In: Annual Conference on Medical Image Understanding and Analysis. Springer (July 2024)
4. Graf, R.: Fundamentals of sonographic diagnosis of infant hip dysplasia. Journal of Pediatric Orthopaedics **4**(6), 735–740 (1984)
5. Huang, B., Xia, B., Qian, J., Zhou, X., Zhou, X., Liu, S., Chang, A., Yan, Z., Tang, Z., Xu, N., et al.: Artificial intelligence-assisted ultrasound diagnosis on infant developmental dysplasia of the hip under constrained computational resources. Journal of Ultrasound in Medicine **42**(6), 1235–1248 (2023)
6. Iakubovskii, P.: Segmentation models pytorch. https://github.com/qubvel (2019)
7. Jo, C., Hwang, D., Ko, S., Yang, M.H., Lee, M.C., Han, H.S., Ro, D.H.: Deep learning-based landmark recognition and angle measurement of full-leg plain radiographs can be adopted to assess lower extremity alignment. Knee Surgery, Sports Traumatology, Arthroscopy **31**(4), 1388–1397 (2023)
8. Jung, A.B.: imgaug. https://github.com/aleju/imgaug (2018), [Online; accessed 01-Jan-2024]
9. McCouat, J., Voiculescu, I.: Contour-hugging heatmaps for landmark detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20565–20573 (2022). https://doi.org/10.1109/CVPR52688.2022.01994
10. McCouat, J., Voiculescu, I., Glyn-Jones, S.: Automatically diagnosing hip conditions from x-rays using landmark detection. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 179–182. IEEE (2021)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
12. Przystalski, K., Paleczek, A., Szustakowski, K., Wawryka, P., Jungiewicz, M., Zalewski, M., Kwiatkowski, J., Gądek, A., Miśkowiec, K.: Automated correction angle calculation in high tibial osteotomy planning. Scientific Reports **13**(1), 12876 (2023)
13. Rosendahl, K., Aslaksen, A., Lie, R., Markestad, T.: Reliability of ultrasound in the early diagnosis of developmental dysplasia of the hip. Pediatric radiology **25**, 219–224 (1995)
14. Thaler, F., Payer, C., Urschler, M., Stern, D.: Modeling annotation uncertainty with gaussian heatmaps in landmark localization. arXiv preprint arXiv:2109.09533 (2021)

15. Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: scikit-image: image processing in python. PeerJ **2**, e453 (2014)
16. Wyatt, J., Voiculescu, I.: Optimising for the unknown: Domain alignment for cephalometric landmark detection. arXiv preprint arXiv:2410.04445 (2024)