Improving Automated Ultrasound Infant Hip Screening using an Integrated Clinical Classification Loss

Allison Clement[1,2][0000−0003−0339−3674], Abhinav Singh[1,2][0000−0002−7329−6792], Daniel Perry[1,2,3][0000−0001−8420−8252], and Irina Voiculescu[1][0000−0002−9104−8012]

[1] Computer Science Department, Oxford University, Oxford, UK
allison.clement@cs.ox.ac.uk
[2] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Oxford, UK
[3] Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, UK

Abstract. Angle measurements in medical imaging tasks primarily rely on landmark localisation. Developmental dysplasia of the hip (DDH) is a disorder amongst infants where the hip does not form properly such that the femoral head is poorly located within the socket. The Graf method is an ultrasound-based screening technique in which lines are drawn through anatomical landmarks to characterise the deformity. The most important feature in the Graf method is an angle termed 'alpha', which is used to define severity and guide treatment decisions.

In the Graf method, ultrasound images are annotated by clinicians to determine the optimal treatment. However, the subjective nature of ultrasound image interpretation results in significant intra– and inter–observer variability in measuring alpha.

Deep learning has shown reasonable performance in predicting the Graf class compared to clinicians. However, these automated methods lack the evaluation of geometric criteria (landmark detection and angle measurements) and classification metrics. Until now, no work has evaluated the effect of incorporating the clinical classification within the loss function.

To ensure the clinical adoption of automated methods, it is important to replicate the clinical pipeline. This paper shows improved performance by adding a weighted class into the loss function for most metrics. This work illustrates the importance of considering all metrics in the clinical pipeline to determine the best methods. The developed methods for evaluating geometric criteria can be applied to other angle-based medical imaging classification tasks.

Keywords: Angle Predictions · Landmark Detection · Classification · Ultrasound · Combined Loss

1 Introduction

1.1 Clinical Background

Developmental dysplasia of the hip (DDH) is a common childhood disease where the 'ball and socket' joint of the hip does not form properly, such that the femoral ball is poorly located within the acetabular socket. If diagnosed in the neonatal period, it can be treated with a simple harness (brace). However, if diagnosed after three months of age, it often requires multiple complex surgical procedures. There is considerable debate on the optimal method of DDH detection [15]. In many parts of Europe ultrasound screening is undertaken in all infants, while in other countries (i.e. the UK and US), ultrasound is only undertaken on high-risk children.

The most widely agreed-upon method of ultrasound assessment for DDH was designed by Prof. Reinhard Graf (the Graf method). It is a technique where anatomical landmarks are used to derive two diagnostic angles; alpha and beta. This method determines the depth of the bony socket (alpha) and cartilaginous coverage of the femoral head (beta) [17]. These angles, alongside a clinician's judgment, determine the clinical classifications that dictate treatment strategies. A recent consensus by UK surgeons has supported the use of the alpha angle to primarily guide treatment [1].

The Graf method stipulates the precise anatomical location of each landmark to improve objectivity in the assessment of scans. However, in clinical medicine, there are often significant cognitive biases [8]. When placing landmarks clinicians actively consider how the annotations affect patient treatment. Clinicians readily admit to altering the position of landmarks to fit what they deem the best clinical action. Their inference is likely based on other clinical parameters (such as the patient's age or family history) that allow them to consciously influence borderline cases, i.e. those that fall on the cusp of two different treatment strategies.

1.2 Automated Methods

The identification of key anatomical landmarks to determine angles is used in the decision-making process for several orthopaedic diseases. Deep learning methods, commonly U-Nets, have reasonable performance in the ability to predict angles from landmarks, compared to expert clinicians in orthopaedic tasks [13, 2, 9].

For the diagnosis of DDH, the clinical workflow and how the landmark localisation affects the generation of Graf angles are shown in Figure 1. While previous studies have looked at automating Graf measurements for diagnosis, most have not expressed the pixel probability of each landmark [16, 11, 5, 6]. Landmark localisation has been performed on radiographs to create predictions with non-Gaussian point distributions in the medical image [12]. There has been little work in ultrasound diagnostics to compare localisation metrics from landmarks generated from pixel-wise probability outputs [4].

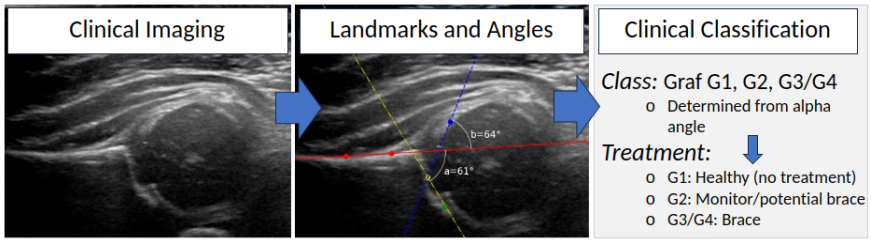

Fig. 1. The clinical pipeline for determining the Graf classification, based on angle measurements derived from landmarks. This example shows a normal (Graf G1) hip. The left image shows the raw data. The centre image shows landmark annotations and lines drawn to intersect these landmarks. Landmarks (Red, Yellow, Blue and Green) represent specific anatomical points (listed in Figure 2). The lines in the centre image create the clinical angles, alpha (a) and beta (b), used for classifying the Graf type and sub type, respectively. The rightmost image summarises how the Graf type is used to guide treatment decisions

1.3 Gaps in the literature

It is important to compare all metrics in the DDH clinical pipeline to understand the best-performing automated method. Previously work has evaluated model accuracy in predicting the final Graf class (92.3-97.7%) [16, 3] and reported mean absolute error of angle metrics (1-3.4°) alongside the percent of alpha values within 5°(93%) [6]. However, there has been no consideration for angle error agreement in the clinical pipeline and how this influences disease classification.

Evaluation of model performance has focused on Accuracy, rather than Recall (sensitivity). [3, 11, 6] Recall is critical for clinicians to understand if the model is unable to identify abnormal patients. Cases where a scan is abnormal but the model predicts it to be normal i.e. 'false negatives' are of high risk and consequences for the child.

Recent work has reported the landmark localisation errors, the angle errors, and all classification metrics [4]. However, this work did not evaluate how the angle differences related to the final decision (Graf classification).

The cognitive bias in placing landmarks mentioned in Section 1.1 may affect the final classifications. Angles with small differences, e.g. 1°, may fall on opposite sides of a class boundary. This means that, although a model may be within a degree, the resultant output is a different Graf class. This complex relation between landmark placement and output class could be captured by placing more weight on the final classification in the loss function.

No work to date for angle predictions in medical imaging has incorporated the final output (angle prediction and class) into the loss functions to help the model learn the relation between landmarks. Most published work, except for Hu et al, has used standard loss functions for model training. [3, 11, 6] Hu et al experimented with different loss functions that incorporated landmarks, bony rim, and shape loss but did not add the final output to the loss [6]. This relation,

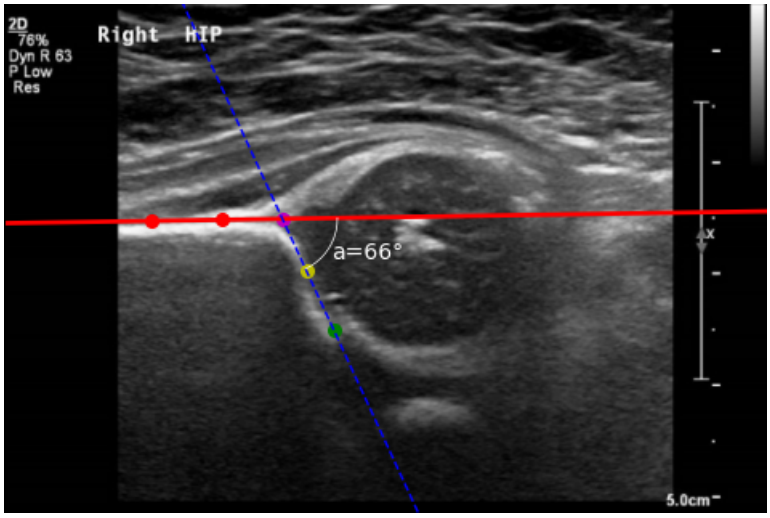

Fig. 2. An example of clinically determining the alpha angle (a). Landmarks are placed by clinicians: two ilium points (red), the turning point (yellow), and the lower limb point (green). The lines show the baseline (red, full line) and cartilage roof line (blue, dashed line) created by drawing through the relevant landmarks.

captured in the Graf angle calculation, may be influential when training the model but is currently not weighted highly in image-alone methods. Its addition could improve all geometric criteria during inference.

1.4 Contributions

We proposed a method that accounts for non-Gaussian probability distributions, such as heatmaps, to express model uncertainty to clinicians. The main contributions of this work were to improve current methods for automating the Graf method by:

1. *Including landmark probability distribution (heatmaps) and Graf class in loss functions,*
2. *Evaluating both geometric outputs (landmarks and angles) and Graf classification results.*

2 Research Methods

2.1 Datasets

In this work, based on the recommendations from the British Society for Children's Orthopaedic Surgery (BSCOS), the alpha angle was used to classify the

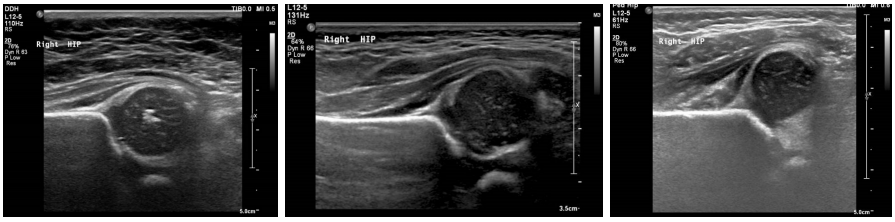

Fig. 3. Example ultrasound scan for each Graf Class. Showing G1 (left), G2 (center), and G3/4 (Right).

Table 1. Graf Classification from Experts. Showing the class, the associated alpha angle range, and the Graf class description.

Class	Alpha	Description
G1	$\geq 60°$	Normal: Discharge Patient
G2	≥ 43 and $<60°$	Borderline: Clinical Review ± Brace
G3&4	$<43°$	Abnormal: Brace

patient as normal, borderline, or abnormal, as seen in Table 1. Ultrasound scans were collected from Alder Hey Children's Hospital (n=516 scans of left and right hips, from 397 unique patients).

Training and validation of an automated method to replicate the clinical pipeline requires both landmark annotations and a final Graf diagnosis given to the patient. Landmark annotations were completed on all 516 scans. Of these 516 annotated scans, 186 were annotated by two reviewers, with one reviewer annotating the images twice (a=3 annotations). These annotations have been considered the *Clinical Ground Truth*. The Graf classification assigned to the patient by the Alder Hey Children's Hospital was known and referred to as the *Gold Standard* as it was used to guide the treatment. Additionally, an External Dataset of 50 ultrasound scans (m=50 patients) was collected from the Royal National Orthopaedic Hospital. This was referred to as the External Dataset. The number of scans per class in each dataset are reported in Table 2. Ultrasound Scan examples for each class are shown in Figure 3.

Ethical Considerations The primary dataset was collected from Alder Hey Children's Hospital as part of routine screening. Scanning was performed using a Philips EP1Q5G (L12-5 linear probe) ultrasound machine. All data (images and summary metadata) was anonymised prior to use for research and stored securely in the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS). Where higher processing power was required, advanced computational facilities (ARC- Advanced Research Computing) within the University of Oxford were used. The study was approved by the University of Oxford and the NHS Health Research Authority (HRA). The External Data set was obtained from the Royal National Orthopaedic Hospital. This data followed

Table 2. Number of scans in each class for both datasets. The number reported with percentage relative to the entire set.

	G1	G2	G3/4
Primary Dataset	184 (35.7%)	227(44.0%)	105 (20.3%)
External Dataset	13 (26.0%)	30 (60.0%)	7 (14.0%)

the same anonymisation protocols, registration requirements, and data storage procedures as the primary dataset.

2.2 Model Architecture

The model took raw images as inputs, with the annotation landmarks and Graf class as outputs. The problem was modeled as a classification task for each image pixel. This allowed the algorithm to output a heatmap of where each landmark was predicted which, crucially, was not restricted to a Gaussian distribution. Models outputs that were more spread meant a less confident prediction.

To do this, a UNet++ [7] with a ResNet34 encoder was pre-trained on ImageNet data using `PyTorch` [14]. The decoder had five layers which were set channel sizes of 512, 256, 256, 128, and 64. Layers were batch normalized, followed by an activation function (ReLU). These methods were implemented using `Segmentation Models Pytorch` [7]. Each output channel corresponded to single landmarks. During training a Gaussian distribution was applied to each channel (σ=6) and the softmax function was applied to get the individual channel's probability-like distribution.

Augmentation on every image was applied using the `imgaug` library [10]. The augmentation included rotation (factor$\leq$1), intensity shift (factor$\leq$0.5), scaling (factor$\leq$0.2), and translation ($x \leq 0.05, y \leq 0.1$).

A Negative Log Likelihood (NLL) loss function was used to train the network, with L2 regularisation to improve generalisation. NLL was chosen due to the success in previous work developed to predict medical imaging landmarks as predicted heatmaps [12].

The baseline method was trained with the image loss alone. This loss was a NLL loss function comparing the input Gaussian heatmaps with the predicted heatmaps. The function for the baseline image loss is below:

$$J_{NLL}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} * log(\hat{y}_{ij}) \tag{1}$$

where, the model was parameterised by θ, y_{ij} was the target Gaussian heatmap, $\hat{y}_{ij}$ was the predicted heatmap. Additional parameters included C, which summed over the number of channels, and N which was the number of samples in the batch. Landmarks were determined by taking the hottest point in the heatmap of each channel. Using these landmarks the alpha angle, α, was calculated.

Further experimentation was performed on adding classes to the loss function and described in Section 3.

3 Experiments

The model was trained as a classification of pixel location task. All models used 516 unique scans (selecting only one reviewer within the 186 that were repeated), split into a 70:15:15 training, validation, and testing ratio, respectively.

To combine NLL image loss with the output class, Mean Squared Error (MSE) was used. NLL was not used, as this would require a probability of an output class. MSE was chosen to directly compare the output angles. To do this the NLL image loss was added with a class loss using MSE and a weighted MSE to compare the predicted angles.

To do this, the models were trained with the following loss function adaptations:

1. **Mean Squared Error Class Loss (NLL-MSECL)**

$$J_{msecl}(\theta) = (1-\gamma) \times J_{NLL}(\theta) + (\gamma) \times \sum_{i=1}^{N} (\alpha_i - \hat{\alpha}_i)^2 \tag{2}$$

where $\hat{\alpha}$ was the angle predicted by the landmarks from the model, and α was the true alpha from the center of clinician annotated landmarks, γ was a variable for weighting the image and class loss.

2. **Weighted Mean Squared Error Class Loss (NLL-wMSECL)**

$$J_{msecl}(\theta) = (1-\gamma) \times J_{NLL}(\theta) + (\gamma) \times \sum_{i=1}^{N} \frac{(\omega_{class\alpha_i} * \alpha_i - \omega_{class\hat{\alpha}_i} \times \hat{\alpha}_i)^2}{\omega_{class\hat{\alpha}_i}} \tag{3}$$

where ω was the weight for the alpha class.

To evaluate the effect of the weights, γ was calculated and compared for a few values in a specified range $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$. The models were first trained and validated. The models were then run on the held-out test set and additionally tested on the External Dataset. Metrics were reported for the held-out test set and External Dataset. The weighting of each Graf class was chosen to push weights towards more severe cases. This was done because clinicians would rather incorrectly predict a more severe case rather than incorrectly classify an abnormal patient as normal (false negative). Weights were used as 1, 2, and 4 for Graf classes G1, G2, and G3/4, respectively.

4 Evaluation Metrics

To evaluate the success of the predictions generated by the model: landmark localisation metrics, angle difference metrics, and overall classification of disease were assessed.

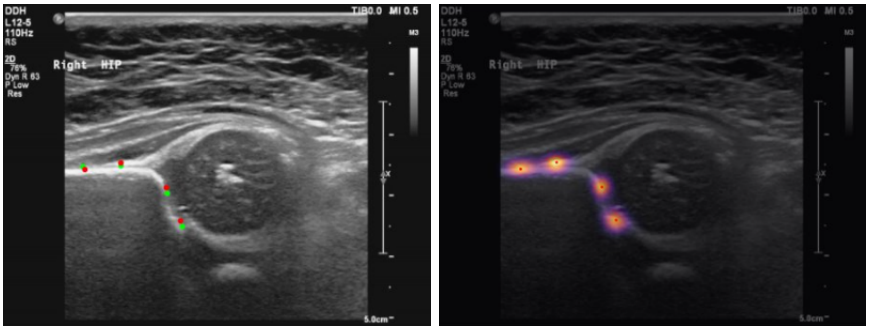

Fig. 4. On the left, an ultrasound with ground truth landmarks (green) and predicted model outputs (red) from the NLL-trained model. On the right, the outputs as a heatmap, where red points were selected as each 'hottest point'.

4.1 Landmark Metrics

Once landmarks were predicted by the network, the *hottest point* (highest value) in each channel was selected as the final predicted position. We followed common metrics used in the literature to ensure the baseline landmarks showed reasonable performance (showing similar values, or improved metrics) to existing methods.

To evaluate the performance of the landmark localisation, Mean Radial Error (MRE) was first calculated. The MRE was the average Euclidean distance between landmarks predicted by the model and ground truth. The Successful Detection Rate (SDR) was also calculated as the percentage of landmarks that fell within a specific threshold of the MRE as commonly used in the literature [12, 4].

4.2 Angle Agreement

Predicted landmarks were used to calculate the alpha angle. The predicted angle by the model for the test set was compared to the Clinical Ground Truth values (CGT, the difference in angles calculated between the set of examples with three reviewers (n=186) and the Gold Standard from Alder Hey Children's Hospital). The Gold Standard used was the diagnosis that was given to the patient as part of their treatment at the hospital. The percent 'agreement' was the difference in class outcome for these angle predictions. This illustrated how often the angles predicted agreed on the same class prediction. For example, although a predicted alpha angle could be one degree different than a ground truth value, this difference could shift the overall Graf class predicted by the model.

4.3 Classification Metrics

Once the angle was calculated the disease-specific class, Graf class, was computed. To compare, we evaluated the results by comparing the output model

Table 3. Successful detection rates (SDR), in pixels. Literature methods, compared to image-based loss and models trained with additional class loss. Gamma was reported showing the gamma which performed the best, all γ values in Supplementary S1 Materials.

	SDR (5 pix)	**SDR (10 pix)**	**SDR (20 pix)**	**SDR (40 pix)**
NLL	**81.7**%	**96.1**%	99.2%	99.7%
NLL+MSECL (γ=0.8)	78.1%	**96.1**%	**99.7**%	**100**%
NLL+wMSECL (γ=0.2)	80.6%	95.0%	97.5%	99.4%
External (NLL-wMSECL γ = 0.6)	62.9%	76.7%	85.0%	97.5%

Accuracy to the Accuracy of the human variability for similar landmark identification tasks. Accuracy, Precision, and Recall were all evaluated. This was because Accuracy alone only illustrates the general model performance and not its ability to avoid false negatives (classifying the patient into a Graf class which was less severe). This was critical in this application to ensure doctors would not send patients home who required monitoring or bracing.

The comparisons were done between G1 compared to G2 and G3&4 as well as between G1 and G2 compared to G3&4. This was due to the clinical relevance of the separation of these groups. Some clinicians wanted to focus on identifying and discharging patients with normal scans (G1), whereas other clinicians wished to prioritise the abnormal cases (G3/4).

Additionally, as seen in Figure 3, it was more difficult to differentiate between normal (G1) and borderline (G2) than to identify an abnormal scan G3&4. Separating the model's performance into the two groupings described above allowed our methods to be compared to the difficulties faced by clinicians in separating normal and borderline scans.

5 Results

The best-performing model, and γ where appropriate, for each metric was reported.

5.1 Landmark Metrics

The output heatmaps and generated landmarks can be seen in Figure 4.

Landmark metrics were reported with the best γ values for SDR and MRE. The SDR was found to stay within similar ranges when class loss was added (Table 3). All were found to be higher than values reported in the literature for similar tasks.

Table 4. Mean Radial Error (MRE) in pixels and mm. NA-Not Available. See all values reported in Supplementary Table S2

	MRE (pixels)	**MRE (mm)**
NLL	5.14±1.14	0.36±0.09
NLL+MSECL (γ =0.8)	**5.14±0.86**	0.36±0.09
NLL+wMSECL (γ =0.6)	5.57 ± 1.43	0.39 ± 0.1
External (NLL γ = 0.8)	11.99 ± 8.13	NA

These values were reported in pixel size and mm to allow for comparison across datasets as well as scans. This was important to understand as the resolution of the image greatly affects the model's ability to identify features. Additionally, pixel size was valuable for understanding the clinical anatomy and application relevance. The pixel size of the dataset was reported to be 0.07mm by 0.07mm. The pixel size of the External Dataset was not available and thus, only pixel MRE was reported.

The NLL image-alone method was found to have the best performance when compared to all other methods. MRE between predicted and landmark locations were reported (Table 4).

5.2 Angle Agreement

Angle agreement, explained in Section 4.2 was reported in Table 5. This table compared the ground truth ranging between three human reviewers to the Gold Standard (See Section 4.2 for description). This was called the Clinical Ground Truth (CGT) variability. Within the percentage of the angle, whether the class agreed (A) or disagreed (D) was reported.

All model output angle measurements were comparable to the CGT variability (hereafter, this was referred to as the clinical variability). Some models outperformed the agreement to the Gold Standard from Alder Hey Children's Hospital. The model's performance was able to predict much more reliably than the ground truth reviewers, for all models, within 1 ° to 2 °.

The External Dataset showed superior performance, when compared to the Clinical Ground Truth values and the testing set, within 1° and 2°. Additionally, the image-alone almost halved this value (91.7%), reducing the angle within 1° to be 47.9% (Shown in Supplementary Materials).

5.3 Classification Metrics

Classification metrics for G1 was compared to G2 and G3&4 in Table 6, and for Classes G1 and G2 compared to G3&4 Table 7.

Table 5. Clinical Ground Truth ranges from human experts (comparing three different annotations) compared to the predictions made across various loss functions. This showed the percentage of angles found within these specified thresholds, along with the percentage that agreed or disagreed. Where A was Agree and D was Disagree.

	$\alpha \leq$**10°**	$\alpha \leq$**5°**	$\alpha \leq$**2°**	$\alpha \leq$**1°**
CGT	90.0% (A:72.5%,D:18.4%)	64.8% (A:56.9%,D:7.9%)	33.2% (A:31.9%,D:1.2%)	18.2% (A:17.8%,D:1.5%)
NLL	**98.6%** (A:77.8%,D:20.8%)	91.7% (A:72.2%,D:19.4%)	79.2% (A:61.1%,D:18.0%)	73.6% (A:55.6%,D:18.0%)
NLL-MSECL (γ=0.4)	95.8% (A:66.7%,D:29.2%)	91.7% (A:66.7%,D:25.0%)	79.2% (A:54.2%,D:25.0%)	75.0% (A:50.0%,D:25.0%)
NLL -wMSECL (γ=0.6)	91.7% (A:68.1%,D:23.6%)	87.5% (A:65.3%,D:22.2%)	81.9% (A:59.7%,D:22.2%)	75.0% (A:52.8%,D:22.2%)
External (NLL-wMSECL γ=0.6)	95.8% (A:60.4%,D:35.4%)	**95.8**% (A:60.4%,D:35.4%)	**93.7%** (A:58.3%,D:35.4%)	**91.7**% (A:56.2%,D:35.4%)

The model performance was the same, if not better, than CGT values. The only metric that was outperformed when comparing to clinicians was the Precision for determining class G1 compared to G2/3/4 (Table 7). Weighting the class (explained in Section 3) with a γ=0.6 increased the Precision to be above clinical performance, although reducing the Recall.

6 Discussion

This work demonstrates that it is necessary to employ a suite of metrics which take into account both the geometric properties of the final output and the final classification. Since the DDH literature [3, 11, 5, 6, 4] considers only individual metrics, for us to compare against existing work, we need to examine each of the clinical workflow outputs. Additionally, we show comparable performance to clinical expert variability for the same tasks.

Landmarks Metrics. Our reported pixel values in Tables 3 and 4 follow the metric and thresholds of other landmark detection work [12]. Since this literature uses different data, a direct comparison is not warranted. However, the values are similar (ranges around 76.7%−96%).

Conversely, the DDH literature reports ranges of $0.38-0.98mm$ within 50% of predictions [4]. A fair comparison is against our SDR for 10 pixels ($0.7mm$), where all of our methods achieved landmark detection that is much higher than 50% (75.7-96.1%).

Within the variations of our algorithm, the image-alone loss method performed best for landmark detection when comparing all other proposed loss

Table 6. Clinical Ground Truth (CGT) compared to the Gold Standard from the hospital, compared to current methods for G1 vs. G2 or G3&4. Comparing Accuracy, Precision, Recall/True Positive Rate (TPR)

	Accuracy	**Precision**	**Recall(TPR)**
CGT	74.6±4.6%	98.2±1.2%	69.3±6.5%
NLL	**90.2**%	94.4%	**73.9**%
NLL-MSECL (γ=0.6)	86.1%	88.2%	65.2%
NLL-wMSECL γ=0.6	87.5%	**100**%	60.9%
External (NLL-wMSECL) γ=0.2	87.5%	75%	60%

Table 7. Clinical Ground Truth (CGT) compared to the Gold Standard from the hospital, compared to current methods for G1 or G2 vs. G3&4. Comparing Accuracy, Recall/true positive rate (TPR), and Precision

	Accuracy	**Precision**	**Recall(TPR)**
CGT	82.3%±2.8%	81.4%±3.8%	44.8±5.3%
NLL	86.1%	96.1%	85.9%
NLL-MSECL (γ = 0.6)	87.5%	92.9%	91.23%
NLL-wMSECL (γ = 0.6)	80.6%	87.7%	87.7%
External (NLL)	**93.7**%	**97.6**%	**95.4**%

functions (Table 4). This is perhaps because the model with image-only loss is trying to optimise solely for predicting landmarks. The networks of the same size, but with added classification metrics, must optimise both landmark and angle metrics simultaneously. Although image-alone loss performs best for MRE, it does not perform best for all metrics.

With the additional angle information in the loss, the model may have learned a representation which also encompassed the relation between the landmarks and the angles. This is perhaps why we see the image-alone method slightly outperforming the other methods.

Angle Agreement. The proposed angle metric is the first to compare angle agreement. Although algorithms may be fairly accurate in degree difference, a one-degree difference may result in a change of class. This can be reflected in our angle agreement metric.

Within 1° and 2°, all of our proposed methods are superior (73−94%) to the ground truth (18−33%). No work to date has presented this ground truth variability between clinicians. Additionally, we report the 'agreement' between reviewers. This shows how difficult it is for clinicians to agree on the exact landmark location, and how small differences in pixel placement propagate to large differences in angle measurements.

Several authors have stated the angle difference between the ground truth and predicted angles (1-3.4°) [3, 11, 6] however, no corresponding frequency has been reported. Our best-performing model predicted within 1-2°, 91.7-93.7% of the time (Table5).

Hu et. al, in addition to reporting the angle error, add the percent of alpha values within 5°(93%) [6]. As seen in Table5, our model can detect within 5°with a percentage of 95.8%.

The NLL with a weighted γ 0.6 showed the most angle predictions within 1° and 2°. We would have expected models to have a higher agreement for the class loss functions, however, the NLL image-alone was found to have slightly better performance. Further experimentation is needed to understand why this may be occurring.

All angle agreement metrics for the External Dataset are much higher than the Clinical Ground Truth (Table 5). The success of the External Dataset indicates the ability of this model to generalise however, the relatively small size of the dataset should be taken into consideration.

Classification Metrics. For G1 vs G2/G3&4, all automated models had better Accuracy (90.2%, Table 6) than clinicians (74.6%). This is comparable to the values reported in the literature for the same task (92.3-97.7%) [16, 3, 4]. When weighting the class using NLL-wMSECL Precision improved (100% compared to 94.4%, see Table 6 and Equation 3). However, as expected due to the mathematical relation between Precision and Recall, the Recall is reduced. We prioritised maximising Recall because clinicians prefer to be overcautious in reviewing a patient rather than incorrectly classifying them as normal when further treatment is needed.

This work also reports differences in class G1/G2 compared to G3&4. Overall Accuracy, Precision, and Recall are higher as visually it is easier to tell the difference between these groups than between G1 and G2 (see Figure 3. This is the first work which compares the clinician performance to the model performance for DDH. For most metrics, γ of 0.6 is the best, with other gammas following within close range. As expected, this indicates the importance of adding class input to the loss to improve output classification.

External Dataset The generalisability of the model is critical for cross-hospital usage. The model performed better when testing on an External Dataset. For the Angle metrics, the External Dataset had the best performance. This may be because the model that performed best included the class in the training, aiding the model in learning the relation between points rather than relying on image

features alone. Future work will collect data from another hospital to test this theory and confirm overall model generalisation.

Even in metrics such as, MRE and classification, where the External Dataset did not outperform other methods, the metrics are within a comparable range. To the best of our knowledge, this is the first work to test methods on an External Dataset for DDH classification using the Graf method.

Limitations Due to time constraints, we tested the γ parameter in intervals of 0.2. Ideally, we would optimize this parameter by creating a more frequent interval or have it as a learnable parameter for the model.

The dataset was limited in amount however, due to the difficulty and time-consuming nature of annotations it was all that was available. More annotated data would allow for comparisons of these models trained with splits to inherently mimic the natural class imbalance in the DDH population. Additionally, image de-noising on External Datasets will be done to further improve generalisability [16].

Finally, not using the confidence measures in class predictions is a limitation. Our method creates a baseline which can incorporate confidence measures to improve output metrics. This method provides a baseline with visual feedback, based on which clinicians can infer a level of certainty in the model's prediction.

Future Directions Future work will explore improving the model by leveraging the uncertainty in the predicted heatmaps to calculate the probabilities of angles. Developed from previous work in our group, Expected Radial Error (ERE) will be computed to flag erroneous predictions ([12]) We also propose a method which takes into account non-Gaussian probability distributions to evaluate the effect this may have on angle range predictions.

We did not compare NLL to MSE for the addition of the class. We intend to propose calculating the probability an angle belongs to a class which can then be used to evaluate the use of NLL as a part of the class loss function.

As explained in Section 2.2 (Model Architecture) the NLL loss function was chosen due to its success in previous work [12]. We plan to investigate adding the Graf class to the loss function after calculating the class probabilities. Finally, an MSE loss function for image-alone methods can also be applied.

We understand that one method of weighting or image-alone has not been superior in all metrics. We do however prove that all methods above are comparable to the current Gold Standards or values reported in the literature. In future, we hope to use the metrics and techniques in this work to optimize all metrics with one model.

7 Conclusion

This work summarises geometric evaluations relevant to the clinical workflow which should be considered to determine the best-performing AI models for

angle-based clinical decisions. The landmarks affect the final treatments for infants with DDH therefore, clinicians carefully consider how they place landmarks and may consciously or unconsciously adjust them to influence treatment. To the best of our knowledge, there has been no prior evaluation of the effect of adding an angle-based clinical decision to help the model during training. This work provides an automated method that improves the performance of most metrics by including a final class prediction into the training loss functions.

Adopting this method in a clinical setting, it is important to have all aspects of the workflow replicate clinical procedures. Although some studies show the output angles and landmarks along with the image classification, these studies have lacked evaluation of all geometric criteria within the clinical pipeline.

An improved understanding of performance metrics and how they compare to clinical variability can help clinical adoption by improving understanding and trust in automated methods. Future work can look at the incorporation of confidence measures to help clinicians infer a level of certainty in automated decisions. Automated methods can help reduce healthcare costs by making imaging and diagnosis more accessible to a wider range of patients and healthcare workers.

References

1. Aarvold, A., Perry, D., Mavrotas, J., Theologis, T., Katchburian, M.: The management of developmental dysplasia of the hip in children aged under three months. Bone & Joint Journal **105**(2), 209–214 (2023)
2. 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)
3. Chen, T., Zhang, Y., Wang, B., Wang, J., Cui, L., He, J., Cong, L.: Development of a fully automated Graf standard plane and angle evaluation method for infant hip ultrasound scans. Diagnostics (Basel) **12**(6), 1423 (2022)
4. 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)
5. Golan, D., Donner, Y., Mansi, C., Jaremko, J., Ramachandran, M., CUDL: Fully automating Graf's method for DDH diagnosis using deep convolutional neural networks. In: Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1. pp. 130–141. Springer (2016)
6. Hu, X., Wang, L., Yang, X., Zhou, X., Xue, W., Cao, Y., Liu, S., Huang, Y., Guo, S., Shang, N., et al.: Joint landmark and structure learning for automatic evaluation of developmental dysplasia of the hip. IEEE Journal of Biomedical and Health Informatics **26**(1), 345–358 (2021)
7. Iakubovskii, P.: Segmentation models pytorch (2019), `https://github.com/qubvel/segmentation_models.pytorch`
8. Janssen, S.J., Teunis, T., Ring, D., Parisien, R.C.: Cognitive biases in orthopaedic surgery. JAAOS-Journal of the American Academy of Orthopaedic Surgeons **29**(14), 624–633 (2021)

9. 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)
10. Jung, A.B.: imgaug. `https://github.com/aleju/imgaug` (2018), [Online; accessed 30-Oct-2018]
11. Lee, S.W., Ye, H.U., Lee, K.J., Jang, W.Y., Lee, J.H., Hwang, S.M., Heo, Y.R.: Accuracy of new deep learning model-based segmentation and key-point multi-detection method for ultrasonographic developmental dysplasia of the hip (DDH) screening. Diagnostics **11**(7) (2021). https://doi.org/10.3390/diagnostics11071174, `https://www.mdpi.com/2075-4418/11/7/1174`
12. McCouat, J., Voiculescu, I.: Contour-hugging heatmaps for landmark detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20597–20605 (2022)
13. 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)
14. 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`
15. Roposch, A., Liu, L.Q., Hefti, F., Clarke, N.M., Wedge, J.H.: Standardized diagnostic criteria for developmental dysplasia of the hip in early infancy. Clinical Orthopaedics and Related Research® **469**, 3451–3461 (2011)
16. Sezer, A., Sezer, H.B.: Deep convolutional neural network-based automatic classification of neonatal hip ultrasound images: A novel data augmentation approach with speckle noise reduction. Ultrasound in Medicine & Biology **46**(3), 735–749 (2020)
17. Ulziibat, M., Munkhuu, B., Schmid, R., Baumann, T., Essig, S.: Implementation of a nationwide universal ultrasound screening programme for developmental dysplasia of the neonatal hip in Mongolia. Journal of children's orthopaedics **14**(4), 273–280 (2020)