Landmark-Based Screening: Femoral Head Coverage and Graf Classification in Infant Developmental Dysplasia of the Hip

Allison Clement[1,2][0000−0003−0339−3674], Abhinav Singh[1,2][0000−0002−7329−6792], 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

Abstract. Infant Development Dysplasia of the Hip (DDH) is a disorder where the hip joint does not form properly. The Graf method and The Femoral Head Coverage (FHC) method are ultrasound-based screening techniques which use anatomical landmarks to classify disease severity and guide treatment.
Deep learning has been used for either Graf Classification or for FHC, yet there has been only minimal investigation into combining the two. No work to-date has detected FHC using landmarks alone.
This paper develops a model which predicts both Graf Classification and FHC from landmarks only. In this method, Recall (Precision) improved when combining methods compared to FHC and Graf methods alone. Two external datasets were used to evaluate model performance under domain shift. Improvements are needed to generalise the model to new datasets. Since the model encompasses both techniques, it gains a clinical understanding of automated methods for DDH screening and improves clinical use.

Keywords: Landmark detection · Ultrasound · DDH · Screening

1 Introduction

Automated landmark detection is important for several tasks in computer vision such as face recognition or pose estimation. More specifically in the medical domain, the identification of anatomical landmarks is emerging as a method for measuring geometric variables. Angle measurements derived from landmarks help with screening for orthopaedic diseases. Most such methods use a single measurement method in the decision-making [19]. This work explores the combined use of more than one geometric classification method for a clinical decision.

1.1 Clinical Application

Infant Development Dysplasia of the Hip (DDH) is a condition where the femoral head is not aligned with the acetabular socket. When diagnosed early it can be

Table 1: Graf Classification Groupings by Clinical Experts. Showing the Graf Class and the associated α angle range. These class and angle ranges are each associated with a specific class description and recommended treatment.

Graf Class	α **Range**	**Description**
G1	$\geq 60^{\circ}$	Normal: Discharge Patient
G2	≥ 43 and $<60^{\circ}$	Borderline: Clinical Review ± Brace
G3&4	$<43^{\circ}$	Abnormal: Brace

treated effectively with a brace. Late diagnosis requires complex surgical procedures and prolonged time in the hospital. There is considerable debate on the optimal method for quantifying the severity of DDH requiring treatment [2, 24]. Ultrasound (US) imaging is generally used for screening due to its accessibility and non-ionising radiation [2]. However, the US modality is subjective and interpreting scans for diagnosis can be difficult for borderline cases. Clinical classifications exist which can aid the interpretation of DDH US images [22].

The Graf method provides an angle-based anatomical assessment of the condition's severity. The Graf angles (called alpha α and beta β) are derived from manually placed landmark annotations. A recent UK surgeons' consensus [1] has decided that treatment decisions should use only the α angle (Figure 1). The α angle corresponds with recommended treatments reported in Table 1 (See US image presentation of classes in Figure 2). These groupings have been determined through discussion and review with clinical experts.

Another common technique used for DDH screening from US imaging is the Femoral Head Coverage (FHC) method. In this method, a line is first drawn parallel to the upper edge of the ilium. The FHC percentage is then determined by the femoral head proportion below this line (Figure 1). The clinical decision boundary is abnormal for FHC percentages $\leq 50\%$ [21], and normal otherwise.

The Graf α angle and the FHC percentage have been correlated [6, 8, 27]. Combining the Graf and FHC methods by simply thresholding the clinically defined boundaries has proven to improve Accuracy [8].

1.2 Automated Detection

Deep Convolutional Neural Networks (CNNs) have automated Graf classification without predicting landmarks or calculating intermediate angles [25]. However, predicting the Graf classification from an US image without these intermediate metrics and landmarks does not replicate the clinical process. To promote clinical adoption, it is essential to replicate the process by predicting both landmarks and angles.

Both U-Nets and Transformers have predicted landmarks and segmentation of critical structures to make Graf angle predictions. This better replicates clinical workflow which allows for a better model-decision understanding [7, 9, 11, 16]. Recent work has shown superior performance in automating Graf calculations by incorporating heatmaps into predictive models [4, 17]. Chen et al. use a VGG

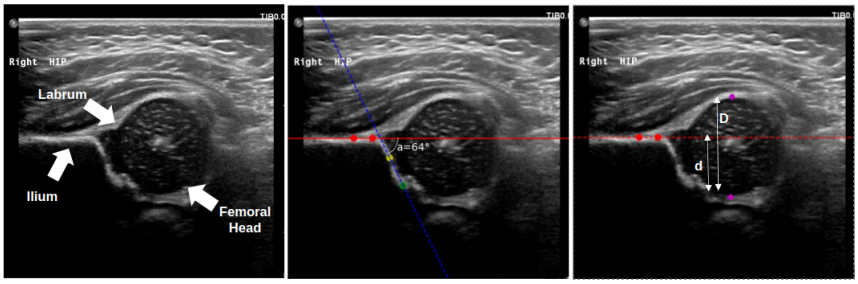

Fig. 1: The left shows the relevant DDH anatomy (the Labrum, Ilium and Femoral Head). The centre shows landmarks used to calculate the Graf α angle. Specific landmarks: the ilium points (red), the turning point (yellow), and the lower limb point (green). The lines connect the landmarks to create the baseline (red) and the cartilage roof line (blue). The right shows landmarks relevant to the FHC method. The two ilium points (red) are used as the iliac line. Femoral head landmarks are purple. The FHC percentage is given by the Euclidean distance of the bony acetabular depth (d) over the Euclidean distance calculated for the entire cartilaginous femoral head (D).

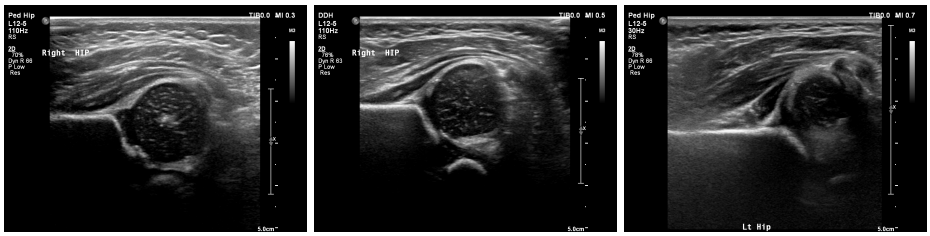

Fig. 2: Example US images for all Graf Classes. Highlighting the difficulty in determining borderline cases. The left shows a normal Graf 1 hip. The right shows an abnormal Hip, Graf 3/4. The centre image shows a borderline case, Graf 2, which is more difficult to classify.

encoder-decoder network to predict heatmaps and Graf angle outputs but do not evaluate FHC [4]. Hu et al. developed a Multi-task Cross Collaboration network which combines heatmaps, segmentations and Graf lines [17]. This work however requires many segmentation maps, which are time-consuming by nature.

To automatically calculate the FHC percentages, segmentation-based methods have been developed. These methods report an accuracy of 89.9% when compared to clinicians [26]. Generating ground truth segmentation is much more time-consuming than landmark identification. Thus, a landmark-based FHC method would be less cumbersome to develop. To our knowledge, nothing has been published on automating landmark annotations to calculate FHC alone.

Limited work has investigated the combination of automating both screening methods. One method used a U-Net for predicting FHC and Graf simultaneously [13]. This was a pilot study with limited data. Recent work has shown promising results for automatically calculating both using the segmentation of

Table 2: Clinical Ground Truth (CGT) Classification Breakdown. The number of scans in each Graf class for all datasets. The number of scans is reported with the percentage relative to the entire set.

	Graf 1	**Graf 2**	**Graf 3&4**
Primary Dataset (D0)	667 (60.3%)	216 (19.5%)	224(19.3%)
External Dataset (D1)	13 (26.0%)	30 (60.0%)	7 (14.0%)
External Dataset (D2)	21 (42.0%)	25 (50.0%)	4 (8.0%)

anatomical structures as intermediates to measurement [10]. This work, however, has not reported combined classification metrics.

To our knowledge, no work has created a landmark-based automated method including both the Graf and the FHC methods to make the final diagnostic class prediction.

1.3 Contributions

The proposed method includes both Graf and FHC to allow for improved clinical adoption. This work improves current methods for automating DDH screening by:

1. *Developing an automated landmark-based Graf-FHC screening method,*
2. *Creating a combined method for Graf-FHC predictions,*
3. *Evaluating generalisation of methods using external datasets.*

2 Methods

2.1 Dataset

1107 US scans were acquired from Alder Hey Children's Hospital (D0). The pixel size of the dataset was reported to be 0.07mm by 0.07mm. Two additional datasets (50 scans each) were collected from external hospitals to simulate domain shift. External Dataset (D1) was obtained from the Royal National Orthopaedic Hospital, Stanmore. The second External Dataset (D2) was obtained from Milton Keynes University Hospital. The pixel sizes of the External Dataset images were not available. The image size for D0 was (1024x768 pixels). The image size for D1 and D2 varied ([947-960]x[583-587] and [271-569]x[278-571] pixels, respectively).

Landmarks were placed manually on five key points for the Graf method. Four of these were used for the Graf method α calculation as pictured in Figure 1 and the fifth landmark was used for clinical observation. Two additional landmarks were placed on the top and bottom extremities of the femoral head (as pictured in Figure 1) to calculate the FHC. All landmarks were confirmed by consulting two additional expert clinicians. These annotations are considered the Clinical Ground Truth (CGT).

Network Architecture. The network input was US images. During training the output landmarks were applied as Gaussian heatmaps (σ=5), with one landmark per channel. This created a per-pixel probability of localisation. During data loading, all images were padded and resized to 512x352 using `imgaug` [14].

A UNet++ [12] with a ResNet34 encoder was pre-trained on ImageNet data using `PyTorch` [23]. The decoder had five layers (256, 256, 256, 128, and 64) each later batch normalized, followed by a rectilinear (ReLU) activation function. Attention in the decoder applied Spatial and Channel 'Squeeze and Excitation Blocks'. Implementation was employed using `Segmentation Models Pytorch` [12].

A spatial-softmax function was applied to each channel for a probability-like distribution. Image augmentation used `imgaug` [14]. The augmentation included rotation (factor$\leq$5), intensity shift (factor$\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 with L2 regularisation to optimise the model.

Ethical Considerations. The primary dataset (D0) was collected retrospectively from routine screening images at Alder Hey Children's Hospital. A Philips EP1Q5G (L12-5 linear probe) ultrasound machine was used for all scans.

Data was anonymised before being used for research. All data was stored securely in the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS). Models were run with the advanced computational facilities (Advanced Research Computing, ARC) available at the University of Oxford. The study was approved by the National Health Services (NHS) Health Research Authority (HRA) and the University of Oxford.

The External Datasets (D1 and D2) both followed the same registration, anonymisation, and data storage methods as the primary dataset.

3 Experiments

The main test set (D0) and the external datasets (D1 and D2) were evaluated. D0 was split into training, validation and testing (70%:15%:15%, respectively). The class balance of the entire D0 dataset was replicated in each subset: training, validation and test. This was done by ensuring iterating through potential subsplits as defined by the previous literature [20]. External datasets, D1 and D2, were used exclusively as test sets.

It is important to compare all metrics in the landmark-based DDH clinical pipeline to understand the best-performing automated method [5]. This work will evaluate landmark, angle, and classification metrics.

Landmark Metrics. The *hottest point* (highest value) in each channel was selected as the final predicted landmark position. All landmarks (n=7) were included in the landmark metrics.

To evaluate the performance of landmark localisation the Average Mean Radial Error (aMRE) and the Successful Detection Rate (SDR) were calculated.

The aMRE is the Euclidean distance between a predicted landmark and its equivalent in the CGT (the 'a' in aMRE is the average distance). The SDR was calculated as the percentage of landmarks within defined aMRE thresholds (7, 14 and 28 pixels). These thresholds were chosen to be able to compare with the current published literature [4,17].

Angle Metrics. The α angle for each image was calculated from the predicted landmarks (see centre image in Figure 1). The test sets were compared to CGT angles by calculating the Mean Absolute Average Distance (MAAD). The percentage of the test set where MAAD fell within $\alpha \leq 1^\circ$ and 2° of the CGT were reported.

Classification Metrics. The Graf classes were grouped for screening into abnormal and normal groups. One grouping considered Graf 2 as abnormal (Graf 1 vs. 2-3-4) and the second grouping considered Graf 2 as normal (Graf 1-2 vs. 3-4). This was due to the clinical relevance of the separation of these groups. Some clinicians choose to focus on identifying and discharging patients with normal scans (Graf 1), whereas other clinicians wish to prioritise the abnormal cases (Graf 3-4). Accuracy, Precision and Recall were reported for both groupings.

FHC (normal vs. abnormal) was determined from the landmarks and lines, as per Section 1. The Accuracy, Precision and Recall for FHC were reported.

A combined classification was calculated by taking the more conservative prediction. This means that, if the FHC or Graf screening method predicts the patient is abnormal, the patient will be considered abnormal. This, more conservative approach, will hopefully ensure no patients needing treatment are missed. This was done by implementing Equation 1.

$$Combined\ Decision = \begin{cases} Normal, & if\ FHC = Normal, \\ & \quad Graf = Normal \\ Abnormal, & otherwise \end{cases} \tag{1}$$

From Equation 1, 'otherwise' represents cases where either FHC or Graf resulted in an abnormal screening. This combined decision was calculated separately for each grouping of Graf screening (1 vs. 2-3-4 and 1-2 vs. 3-4).

4 Results

Landmark Results. The landmark values for aMRE and SDR were found to be best in D0 test set compared to D1 and D2 (Table 3). SDR was within similar ranges within 14 and 28 pixels away, however at a threshold of 7 pixels, D0 outperformed the external sets. The aMRE could not be reported in this table as we did not have access to pixel sizes for D1/D2.

Table 3: Landmark Metrics. The Average Mean Radial Error (aMRE) and Successful Detection Rate (SDR) (averaged across all seven landmarks). The literature row compares the most recently published competing results [4, 17]. Some values and metrics could not be reported (NR) due to missing pixel sizes.

	aMRE		**SDR**		
	pix	mm	7 pix	14 pix	28 pix
D0	**3.9**	**0.27±0.04**	**89.8%**	**99.0%**	**99.8%**
D1	9.0	NR	72.0%	90.6%	93.4%
D2	12.0	NR	54.3%	80.3%	92.8%
Literature	NR	0.59-1.22	37.0%	79.6%	96.7%

Angle Results. The MAAD was lowest for the D0 test set, indicating the best performance in this set. The percentage of the test set falling within 1 ° and 2 ° was highest for D0 (See Table 4).

Classification Results. Classification for the Graf method and FHC were reported alone. The grouping, considering Graf 2 as abnormal (1 vs. 2-3-4) had greater accuracy than when considering Graf 2 as normal (1-2 vs. 3-4) (See Table 5). Precision and Recall however were superior when Graf 2 was in the normal group (1-2 vs. 3-4).

The Graf method (1 vs. 2-3-4) outperformed the FHC method for accuracy, where Precision was greater for all datasets than in the FHC method (Table 5 and 6). FHC performed better for Accuracy and Precision than considering Graf 2 as normal (1-2 vs. 3-4). Recall of Graf (1-2 vs. 3-4) was better in all test sets.

The combined FHC and Graf method is reported in Table 7. The combined method had a greater Accuracy when combined with FHC than with Graf alone and considering Graf 2 as normal (1-2 vs. 3-4). When combining using our conservative method (Equation 1) to create a combined Graf (1 vs. 2-3-4) and FHC decision, the accuracy for D0 decreased compared to the Graf method alone and was similar to the FHC method alone. The Precision increased for D0 when combining both the Graf (1 v 2-3-4) and FHC compared to Graf (1 vs. 2-3-4) alone, but was less than FHC alone. The combined FHC-Graf method improved Recall in D0 compared to both the Graf (1 vs. 2-3-4) method alone (91.6%) and the FHC method alone.

Figure 3 shows visual differences in both external test sets. Figure 4 illustrates poor-performing results for a borderline image from D2. In this image landmarks with large errors are highlighted and associated heatmaps for these landmarks show a widespread.

5 Discussion

This work demonstrates the ability of a landmark-based automated method to produce Graf and FHC predictions simultaneously, thereby refining and streamlining the clinical screening process.

Table 4: Angle Prediction Metrics. Mean Absolute Average Distance (MAAD) between Clinical Ground Truth (CGT) angle and the predicted angle and the α values within 1° and 2° were reported. The literature row reports recent published results [4, 17]. Some values were not reported (NR).

	Angle Metrics		
	MAAD	$\alpha \leq 1°$	$\alpha \leq 2°$
D0	3.8°	**62.0**%	**69.3**%
D1	5.9°	42.0%	44.0%
D2	8.6°	34.0%	38.0%
Literature	**2.2**-7.0°	NR	NR

Table 5: On the left, classification metrics for Normal compared to Abnormal of Graf 1 vs. 2-3-4. On the right Normal compared to Abnormal when grouping Graf Classes 1-2 vs. 3-4. The literature row reports recent published results [4, 17].

	Graf Method (1 vs. 2-3-4)			**Graf Method (1-2 vs. 3-4)**		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
D0	98.3	81.7	**91.6**	82.8	92.3	95.8
D1	**100**	62.5	70.0	62.5	83.3	92.0
D2	93.7	48.4	66.0	57.1	57.1	88.0
Literature	84.4-86.0	**85.4**	88.2	**93.0-94.4**	**97.0**	**97.5**

Landmarks. Table 3 reports aMRE and SDR for direct comparison with the literature. The aMRE for state-of-the-art (SOTA) DDH methods is 0.59-1.22mm [4, 17]. Our work improves the aMRE to 0.27mm for this task.

Since the mm-per-pixel size of the datasets D1 and D2 is not available, we report the aMRE in pixels. The aMRE for D1 and D2 was more than double D0. We report averages for 7 landmarks (not just 5 as in other DDH literature), which could have also affected overall results.

Qualitative analysis of landmark errors in the external datasets indicates that changes in intensity and image size affect the model's ability to place landmarks accurately (see Figure 4). These relatively larger landmark localisation errors cause, in turn, a large discrepancy in aMRE.

Denoising and other image-prepossessing steps may help with generalising the model [15]. In the future, our heatmaps can be leveraged to flag erroneous landmarks seen in the external sets [18].

When converting the SDR from pixels to mm (7 pixels (0.5mm), 14 pixels (1mm) and 28 pixels (2mm)), our work achieves 89.9-99.8% SDR in D0, and 54.3-93.4% for D1/D2. This compares favourably to existing work ranges of 37.0-96.7% [4](refer back to Table 3)

Table 6: FHC Classification Results. Accuracy, Precision and Recall metrics for FHC compare normal and abnormal. Current results published in the literature are shown [26]. Some values were not reported (NR).

	FHC Method		
	Accuracy (%)	Precision (%)	Recall (%)
D0	**91.6**	93.8	**89.4**
D1	88.0	**100.0**	76.9
D2	80.0	96.4	75.0
Literature	89.9	NR	NR

Table 7: Combined Method Classification Results. This shows the combined classification using both FHC and Graf, as per Equation 1. Combining Graf groupings for screening of 1 vs. 2-3-4 on the left and 1-2 vs. 3-4 to the right.

	Combined Method (1 vs. 2-3-4)			**Combined Method (1-2 vs. 3-4)**		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
D0	**92.8**	**90.0**	94.7	**91.6**	93.8	**89.4**
D1	86.0	58.8	**100.0**	88.0	**100.0**	76.9
D2	74.0	62.5	78.9	74.0	85.7	72.7

Angles. The MAAD in our work was 3.8° for D0 and 5.9-8.6° for D1/D2. This does not compete well with the SOTA α angles reported (2.2-7° [3, 4, 17, 25].

Table 4 illustrates that, for D0, 62% of the measured angles were within 1° of the CGT. Unsurprisingly, this frequency was lower for D1/D2 because the landmark detection was less precise. This indicates the model did not generalise well under domain shift for angle calculations.

Our results are, nevertheless, within the clinically reported intra– and inter-rater angle measurements for 1-2° (61.5-68.9% our method, 18.2-33.2% clinically) [5].

Classification Performance. In previous work [3, 9, 16], the evaluation of model performance for the Graf Method has focused on Accuracy which gives equal priority to false positives and false negatives. It is critical to report and prioritise Recall as it allows clinicians to understand the ability of the model to identify abnormal hips.

Recent work has reported Accuracy (84.4-90.2%), Precision (85.4-98.2%) and Recall (73.9-88.2%) [4, 5, 17]. Evaluating Graf screening groupings (1 vs. 2-3-4 and 1-2 vs. 3-4) showed all metrics were comparable or exceeded the ranges reported in the literature (See Table 5). These classification metrics listed were all significantly reduced in D1/D2.

Surprisingly, the Graf methods 1-2 vs. 3-4 reported a lower Accuracy and a much higher Precision and Recall than Graf methods 1 vs. 2-3-4. Since Graf 2

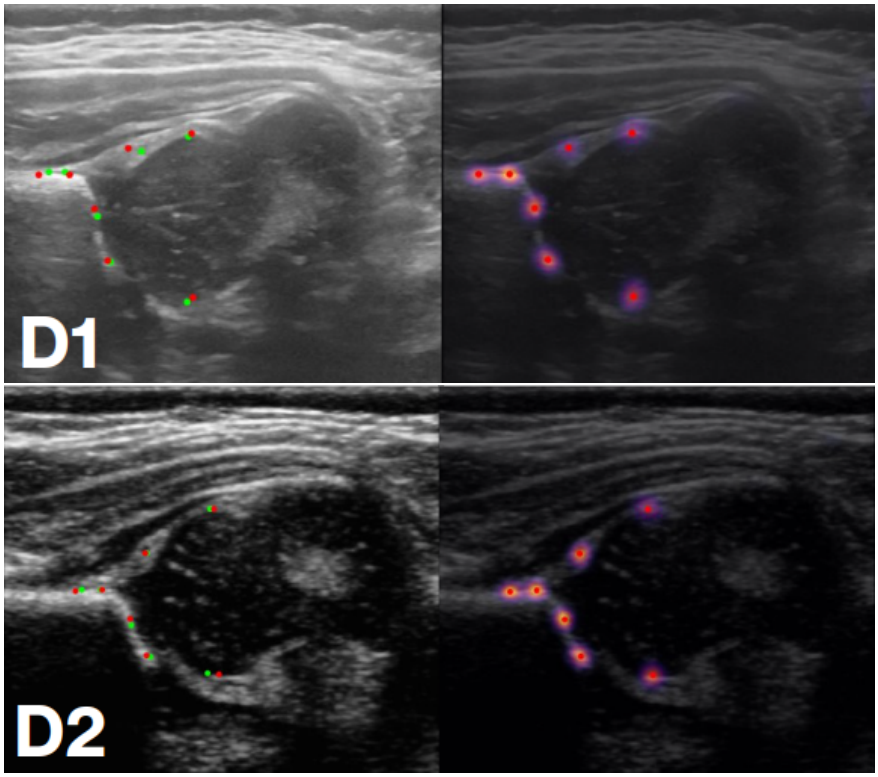

Fig. 3: External Dataset Results. Sample 'normal patient' output on D1 (top) and D2 (bottom). The left-hand side shows US raw image with CGT landmarks (green) and predicted model outputs (red). The right-hand side landmark heatmap output, with 'hottest point' in red.

(borderline) cases are more difficult to visually separate from Graf 1 classes, we would have expected accuracy to go up in the 1-2 vs. 3-4 method. This could potentially be an effect of class imbalances present in our data.

The classification metrics are all similar to the reported clinically. Where clinical inter– and intra– reliability has been reported as 74.6% for Accuracy, 98.2% for Precision, and 69.3% for Recall [5].

The accuracy of a line-based FHC method, trained on anatomical segmentations was 89.9% [26]. Our landmark-based method shows accuracy for D0 has improved to 91.6%, and is reasonable for D1 and D2, at 80-88%. Future work will explore the clinical variability between reviewers of a landmark-based FHC percentage calculation.

Combined Method. When combining the Graf and FHC methods using simple threshold boundaries of clinical methods, the values have improved compared

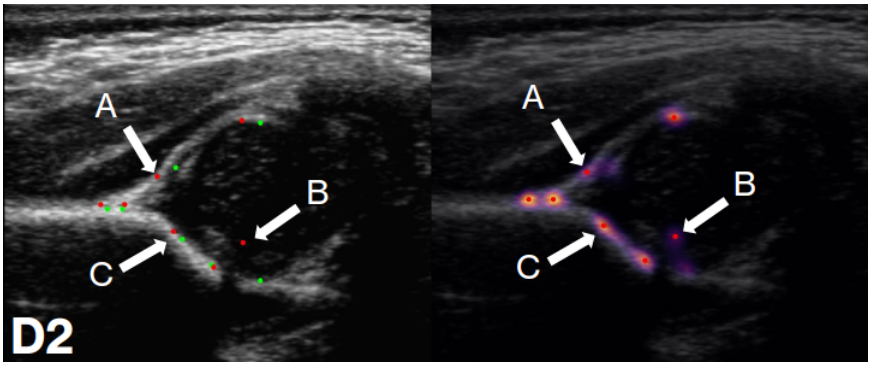

Fig. 4: External Dataset Results with Poorly Identified Landmarks. Sample 'borderline patient' (Graf 2) output on D2. The left side shows US raw image with CGT landmarks (green) and predicted model outputs (red). The right-hand side landmark heatmap output, with 'hottest point' in red. A: Showing the landmarks and heat map area predicted incorrectly for the labrum, B: showing an incorrectly identified point and the heatmap identified C: showing the incorrectly identified bony rim point.

to either method alone (sensitivity 79-82% and specificity 100%) [8]. Recall (sensitivity) was similar in our work (See Table 7).

Recent work has shown promising results for automatically calculating FHC and Graf simultaneously. This work however requires segmentation of anatomical structures which is more time-consuming annotations than landmark placement [10]. This work showed angle differences of 2.2-2.3°, but only evaluated the Graf methods classification metrics.

When combining using our conservative method (Equation 1) values for Accuracy and Precision decreased (1 vs. 2-3-4). The Precision increased when compared to Graf alone, which was unexpected (see Table 7). As expected, the decrease in Precision of the combined method when compared to FHC alone as there would be an increase in the false positive identification by adding Graf 2 to the abnormal group.

The Recall in the combined method increased for datasets. This was expected as the model was trying to follow a conservative approach. In this approach, we hoped to force the model to go with the more severe case, increasing the amount of 'false negatives' (wrongly identified hips as normal when they are in fact abnormal). Increasing the false negatives and in turn recall will help create a method which decreases the likelihood of a patient being 'missed'. This shows the model will help to avoid patients being sent home when they require treatment, reducing the high-risk consequences for the child.

The training distribution for the model replicated the D0 class distribution (see Table 2 and Section 2). However, we can note a large percentage of abnormal scans in the D1/D2. This can be attributed to the fact D1 hospital was a referral hospital where DDH may already be suspected. D2 has a class distribu-

tion closer to that of D0. Future work should investigate the effect of that these class imbalances may have on model generalisation.

Overall model generalisation requires improvements. This work showed the utility and importance of out-of-distribution datasets in assessing model performance.

6 Conclusion

To the best of our knowledge, no prior methods use landmarks for DDH screening using both the Graf method and the FHC percentages. This work fills that gap, providing a landmark-based Graf-FHC prediction comparable to clinician performance. Although within an acceptable range for most published metrics, further improvements are needed for the model to generalise to new datasets.

Improving understanding and trust in automated decisions is critical for healthcare adoption. Future work will evaluate the use of confidence measures from derived heatmaps to help clinicians infer a level of certainty in model-based decisions. Developing robust techniques for automating the Graf-FHC method for DDH could make screening accessible to all newborns.

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. Al-Essa, R.S., Aljahdali, F.H., Alkhilaiwi, R.M., Philip, W., Jawadi, A.H., Khoshhal, K.I.: Diagnosis and treatment of developmental dysplasia of the hip: A current practice of paediatric orthopaedic surgeons. Journal of Orthopaedic Surgery **25**(2), 2309499017717197 (2017)
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. 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 (2024)
6. Fan, W., Li, X.j., Gao, H., Yi, X., Liu, Q.j.: Exploration of femoral head coverage in screening developmental dysplasia of the hip in infants. Journal of Medical Ultrasonics **46**, 129–135 (2019)
7. 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)

8. Gunay, C., Atalar, H., Dogruel, H., Yavuz, O., Uras, I., Saylı, U.: Correlation of femoral head coverage and graf α angle in infants being screened for developmental dysplasia of the hip. International orthopaedics **33**, 761–764 (2009)
9. 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)
10. 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)
11. Huang, T., Shi, J., Li, J., Wang, J., Du, J., Shi, J.: Involution transformer based u-net for landmark detection in ultrasound images for diagnosis of infantile ddh. IEEE Journal of Biomedical and Health Informatics (2024)
12. Iakubovskii, P.: Segmentation models pytorch (2019), `https://github.com/qubvel/segmentation_models.pytorch`
13. Jaremko, J., Hareendranathan, A., Bolouri, S., Frey, R., Dulai, S., Bailey, A.: Ai aided workflow for hip dysplasia screening using ultrasound in primary care clinics. Scientific Reports **13** (06 2023). `https://doi.org/10.1038/s41598-023-35603-9`
14. Jung, A.B.: imgaug. `https://github.com/aleju/imgaug` (2018), [Online; accessed 30-Oct-2018]
15. Kang, M., Kang, M., Jung, M.: Total generalized variation based denoising models for ultrasound images. Journal of Scientific Computing **72**, 172–197 (2017)
16. 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`
17. Liangni, H.: Ccmt: Cross collaboration mult-task network for neonatal hip bone intelligent diagnosis. In: 2024 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE (2024)
18. 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)
19. 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)
20. 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)
21. Morin, C., Harcke, H., MacEwen, G.: The infant hip: real-time us assessment of acetabular development. Radiology **157**(3), 673–677 (1985)
22. Õmeroğlu, H.: Use of ultrasonography in developmental dysplasia of the hip. Journal of children's orthopaedics **8**(2), 105–113 (2014)
23. 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`

24. 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)
25. 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)
26. Stamper, A., Singh, A., McCouat, J., Voiculescu, I.: Infant hip screening using multi-class ultrasound scan segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2023)
27. Striano, B., Schaeffer, E.K., Matheney, T.H., Upasani, V.V., Price, C.T., Mulpuri, K., Sankar, W.N., et al.: Ultrasound characteristics of clinically dislocated but reducible hips with ddh. Journal of Pediatric Orthopaedics **39**(9), 453–457 (2019)