POLITECNICO DI TORINO
Repository ISTITUZIONALE

AI-Assisted Sequential Graf Assessment of Infant Hip Ultrasound: A Retrospective Single-Centre Evaluation in a Referral Cohort

Original
AI-Assisted Sequential Graf Assessment of Infant Hip Ultrasound: A Retrospective Single-Centre Evaluation in a Referral Cohort / Audisio, A., Aprato, A., Hu, J., Marullo, G., Sabbatini, M., Ulrich, L., Moos, S., Petraz, M., Vezzetti, E., Massè, A.. - In: JOURNAL OF IMAGING. - ISSN 2313-433X. - 12:11(2026). [10.3390/jimaging12110554]

Availability:
This version is available at: 11583/3017596 since: 2026-11-05T17:18:00Z

Publisher:
MDPI

Published
DOI:10.3390/jimaging12110554

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

15 November 2026

Journal of *Imaging*

MDPI

Article

AI-Assisted Sequential Graf Assessment of Infant Hip Ultrasound: A Retrospective Single-Centre Evaluation in a Referral Cohort

Andrea Audisio [1,2,*], Alessandro Aprato [1,2], Jessica Hu [3], Giorgia Marullo [4], Matteo Sabbatini [5], Luca Ulrich [4], Sandro Moos [4], Marco Petraz [6], Enrico Vezzetti [4] and Alessandro Massè [2]

1 Pediatric Orthopaedics and Trauma, Regina Margherita Children's Hospital, 10126 Turin, Italy
2 Department of Surgical Sciences, University of Turin, 10126 Turin, Italy
3 School of Medicine, University of Turin, 10126 Turin, Italy
4 Department of Management and Production Engineering, Politecnico di Torino, 10129 Turin, Italy; luca.ulrich@polito.it (L.U.)
5 Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
6 Pediatric Radiology, Regina Margherita Children's Hospital, 10126 Turin, Italy
* Correspondence: andrea.audisio@unito.it

Abstract

Graf-method ultrasound supports early assessment of developmental dysplasia of the hip (DDH). This study assessed a sequential artificial-intelligence-assisted workflow and clinicians' Checklist I decisions with and without computer-aided diagnosis (CAD). This retrospective single-centre study of referred infants used 1803 development images from 679 examinations and separate cohorts of 1024 workflow and 100 reader images; Graf IV hips were excluded. ResTransUNet segmented eight Checklist I structures to assess reportability. Fifteen clinicians assessed the reader images unaided and four weeks later with CAD. Stand-alone Checklist I accuracy was 93.0% (93/100; 95% CI 86.3–96.6). Of 1024 workflow images, 846 received final outputs; Checklist II non-compliance precision was 48.1%. Among accepted images, the non-type-I endpoint had 81.0% accuracy, 82.7% sensitivity and 80.4% specificity; exact Graf-type agreement was 40.1%. Across entering non-type-I images, 167/266 (62.8%) received a correct positive output. Mean absolute errors were 3.59° for the α angle and 11.91° for the β angle, with a β-angle bias of +11.07°. Clinician accuracy was 72.3% unaided and 90.2% assisted—an observed 17.9 percentage-point difference (95% CI 10.0–26.1); 26 correct decisions became incorrect on AI-error cases. The workflow supported Checklist I assessment, but fixed order precludes causal CAD inference; gate failures and angle errors require refinement and prospective evaluation before clinical implementation.

Keywords: developmental dysplasia of the hip; hip ultrasonography; Graf method; artificial intelligence; image quality; reader study; internal evaluation

Check for updates

Academic Editor: Luminița Moraru

Received: 19 August 2026
Revised: 9 October 2026
Accepted: 2 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Developmental dysplasia of the hip (DDH) comprises a spectrum of abnormal acetabular development and femoral-head instability, ranging from mild dysplasia to fixed dislocation. Its reported frequency varies according to population, diagnostic criteria, patient age and screening policy. Clinical examination and assessment of risk factors remain essential, while ultrasonography provides direct, radiation-free visualization of the cartilaginous hip during early infancy [1–3], including in complex skeletal anatomy [4].

The clinical value of ultrasound depends on acquisition and interpretation of diagnostically adequate images.

The Graf method provides a standardized morphological and quantitative framework [5,6]; international interdisciplinary consensus also emphasizes consistent evaluation of infant hips [7]. Graf assessment follows a prescribed sequence rather than relying on angle measurement alone. Checklist I identifies the chondro-osseous border, femoral head, synovial fold, joint capsule, labrum, cartilaginous roof, bony roof and turning point. Checklist II then assesses the standard plane, after which the α and β angles can be measured and a Graf type assigned. Acquisition and interpretation remain operator-dependent, and interobserver variation persists, particularly for angle placement and final classification [8,9]. An automated system can reproduce the sequence operationally without necessarily reproducing the clinical validity of every step.

Artificial intelligence (AI) has been applied to scan-adequacy assessment, standard-plane recognition, anatomical segmentation, landmark detection, angle measurement and Graf-type classification [10–13]. Scan quality affects automated interpretation [14], and AI-supported ultrasound has also been studied in primary-care settings [15]. QualityDDH already implements explicit Checklist I and II assessment using segmentation and structural priors and includes an external reader evaluation [16]. Reviews describe promising results but heterogeneous methods and limited external or clinically relevant validation [17–19]. We aimed to characterize a sequential research prototype in a referral cohort, quantify its performance among accepted images and its correct-output yield across all images entering the workflow, and compare Checklist I decisions between unaided and CAD-assisted sessions in a questionnaire-based reader study.

2. Materials and Methods

2.1. Study Design and Dataset Allocation

This retrospective single-centre study used static hip ultrasound images from infants referred to Regina Margherita Children's Hospital, Turin, Italy, for reassessment of suspected or previously identified DDH. Images were acquired on a Philips InnoSight system with an L12-4 linear transducer (4–12 MHz) by two orthopaedic surgeons experienced in Graf ultrasonography. The intended use of this version was research assessment of stored frames, not autonomous diagnosis or real-time acquisition guidance. Ethics approval was obtained from Città della Salute e della Scienza di Torino (protocol DB2021, approval 1489/2021, 14 November 2021) and covered the image and reader studies. Patient consent was waived for retrospective analysis of anonymized images; participating clinicians provided written consent.

The initial development cohort, collected from 1 March 2021 to 31 December 2025, comprised 1846 images from 537 infants, 682 examinations and 1074 distinct hips. Forty-three Graf IV images from 31 infants, 31 examinations and 34 hips were excluded. Twenty-eight of those infants also contributed retained images; only three infants were removed entirely. The included development cohort therefore comprised 1803 images from 534 infants, 679 examinations and 1068 distinct hips, with 1332 Checklist I-compliant and 471 non-compliant images. The exclusion reflected an investigator-defined limitation of an initial single-frame anatomical-presence approach in severely dislocated hips. The system was evaluated from Graf type Ia to III. Infant-level allocation produced training, tuning and internal test sets of 320, 107 and 107 infants, respectively, contributing 1082, 360 and 361 images (Table 1).

Table 1. Cohort accounting by infant, examination, distinct hip and image.

Dataset	Infants	Examinations	Hips	Images
Initial development	537	682	1074	1846
Contributors to excluded Graf IV images	31	31	34	43
Included development	534	679	1068	1803
Training	320	406	640	1082
Tuning	107	136	214	360
Internal test	107	137	214	361
Additional workflow cohort	468	528	904	1024
Additional reader cohort	92	92	100	100
Interobserver analysis within workflow	–	–	–	1017

Exclusion counts refer to contributors to the 43 removed images, not a disjoint patient cohort. Excluded and retained images share 28 infants, 28 examinations and 28 hips; three infants, three examinations and six hips were removed entirely. Dataset allocation was performed at the infant level, with both hips and repeated examinations from each infant kept within the same primary dataset. Because registry identifiers were removed from the archived prediction exports during anonymization, this separation cannot now be independently re-audited at the prediction-record level. The interobserver analysis is within workflow and is not added to the primary cohorts; its separate infant, examination and hip counts are unavailable without a validated image-to-registry link. Distinct hips are counted once across repeat examinations.

The workflow cohort contained 1024 additional images from 468 infants, 528 examinations and 904 distinct hips, with clinical angle measurements and Graf labels. Random selection was followed by adjustment to retain representation of different Graf types. This was a selected, Graf-type-enriched evaluation sample. A further 100 images acquired after January 2026 were randomly selected for the reader study from 92 infants, 92 examinations and 100 distinct hips. The observed 70 Checklist I-compliant and 30 non-compliant images resulted from random selection. Their reference Graf types were Ia ($n = 20$), Ib ($n = 34$), IIa ($n = 25$), IIb ($n = 7$), IIc ($n = 8$), D ($n = 3$) and IIIa ($n = 3$).

Dataset allocation was performed at the infant level, with images of both hips and any repeated examinations from the same infant assigned to the same primary dataset. The training, tuning, internal test, workflow-evaluation, and reader-study datasets were treated as separate datasets, and the workflow and reader cohorts were not used for training or tuning. The workflow cohort was additional to the development cohort. During anonymization, the linkage between archived prediction records and registry infant identifiers was removed to protect patient privacy. Consequently, the infant-level separation of the archived prediction records cannot now be independently re-audited from the retained analysis files. The expanded interobserver analysis used 1017 images from within the workflow cohort and replaced the earlier 100-image reliability analysis; it did not constitute an additional cohort.

Cohort counts and demographic summaries were calculated from the study registry supplied during revision. Sex was recorded once per infant, age was summarized at the first examination contributing to each specified cohort, and repeat attendance meant more than one contributing examination. A distinct hip was counted once across repeated views and examinations. The included development cohort comprised 346 female infants, 186 male infants and two with unrecorded sex; 129 infants had repeat examinations. Median age at the first included examination was 44 days (interquartile range 29–69; range 7–190), with age available for 516/534 infants. Corresponding median ages were 45 days (31–67; range 8–198; 459/468 infants) in the workflow cohort and 43 days (27.25–65.25; range 9–133; 90/92 infants) in the reader cohort. Recorded ages across all contributing examinations, including repeat visits, extended to 327 days in development and 255 days in the workflow cohort; the data are therefore not restricted to examinations performed within the first six months. Partition-specific demographics, missingness, laterality and sampling multiplicity

are reported in Tables S8 and S9. Missing ages and sex were not imputed. The study datasets and sequential workflow are summarized in Figure 1.

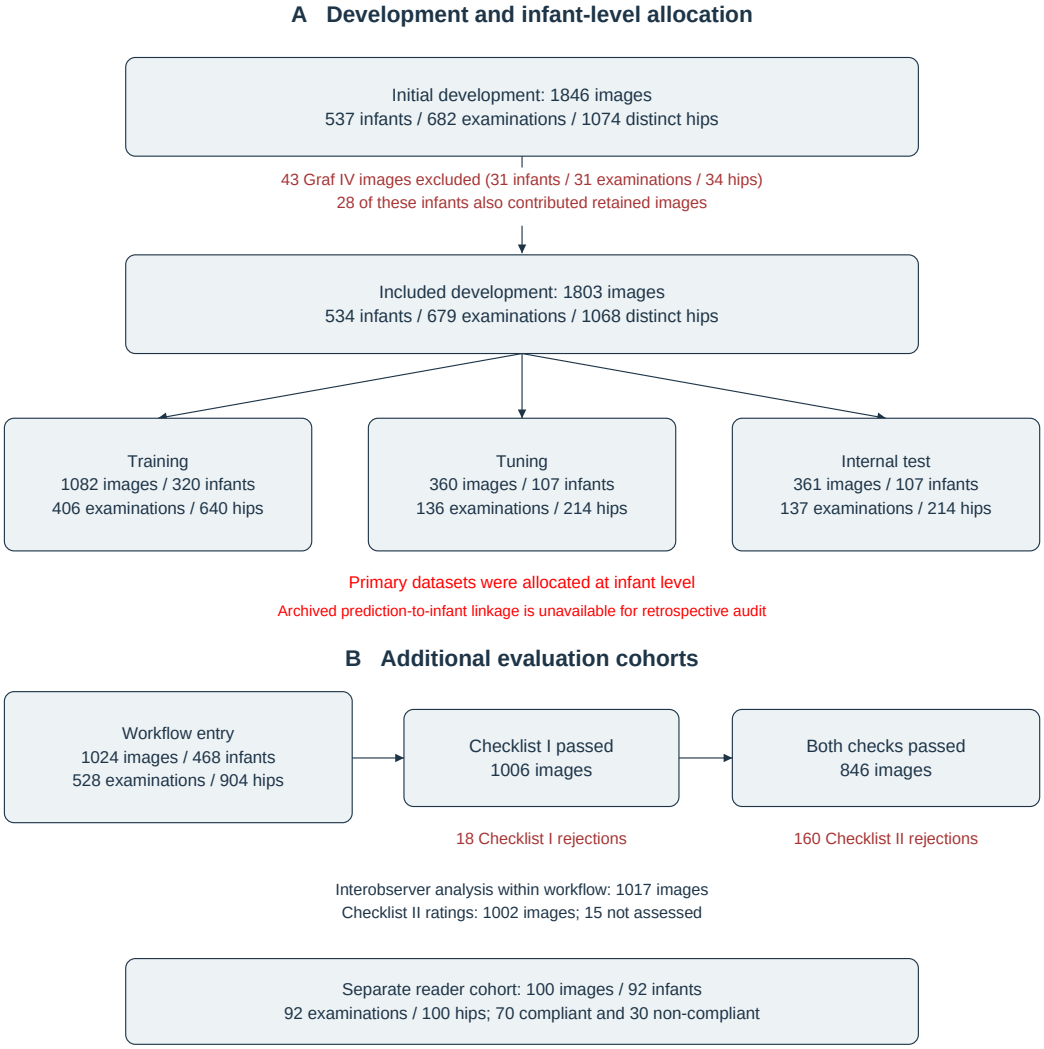


Figure 1. Study datasets and AI workflow: (**A**) development cohort and infant-level allocation; (**B**) additional evaluation cohorts and sequential workflow. Dataset allocation was performed at the infant level, and the workflow and reader datasets were separate from development and were not used for training or tuning. The linkage between archived prediction records and registry infant identifiers was removed during anonymization to protect patient privacy. We excluded 43 Graf IV images from 31 infants; 28 of those infants had other images retained in the study. The interobserver analysis used 1017 images from the workflow dataset. Numbers at each workflow step count images. Rejection at either checklist stopped automated angle measurement and Graf classification.

2.2. Reference Standard and Annotation

A pediatric radiologist and two orthopaedic surgeons independently assessed checklist compliance while blinded to one another's ratings. After the independent assessments, images without full agreement were reviewed jointly to establish consensus. The three experts also reviewed angle measurements and Graf labels to establish the final reference; a numerical angle-disagreement threshold was not documented. None of these experts participated as a reader. JH drew the anatomical masks in CVAT [20] under the training and supervision of AAp and AAu. Mask creation was therefore supervised. No second annotator independently redrew the anatomical masks. Therefore, interobserver reproducibility of the masks was not assessed, and systematic annotation error cannot be excluded. Measurements extracted from image overlays were manually checked. The expanded pre-consensus

export contained 1017 image rows with Checklist I ratings and angle measurements after seven non-image summary or identifier rows were excluded. These 1017 image rows represented a subset of the 1024-image workflow cohort; consequently, seven workflow images lacked recoverable individual pre-consensus ratings, although their final consensus labels remained in the workflow table. The seven excluded non-image rows and the seven workflow images without recoverable pre-consensus ratings were different records. Fifteen images were unanimously judged non-compliant at Checklist I and were not assessed for Checklist II, leaving 1002 images with three Checklist II ratings. The final consensus workflow table supplied the reference labels used for performance analysis, whereas the expanded pre-consensus export was used only for agreement analysis. Reference Graf labels were retained for all 1024 workflow records, final Checklist II labels for all 1006 records that reached automated Checklist II, and reference angles for all 846 records entering downstream evaluation. Checklist II interobserver agreement therefore used 1002 images because 15 of the 1017 independently rated images were unanimously judged non-compliant at Checklist I, whereas automated Checklist II performance used the 1006 images that passed automated Checklist I. For the 46 images judged non-compliant at reference Checklist II but accepted by the automated gate, the retained workflow records contained reference angles and Graf labels, but no linkage to a separately adequate view was preserved.

2.3. Sequential AI Workflow and Training

After cropping borders and acquisition overlays, images underwent noise reduction, contrast enhancement, normalization and resizing to 256 × 256 pixels. ResTransUNet used a residual encoder, multiscale dilated convolutions, bottleneck self-attention, attention-gated skip connections and a residual decoder. Its nine output channels comprise eight anatomical labels and background (Supplementary Figure S1). Class-specific connected-component filtering, gap filling and removal of small protrusions preceded area-based Checklist I assessment. The image passed only if every anatomical mask exceeded its respective minimum area (Table 2).

Table 2. Reported minimum anatomical-mask areas for Checklist I.

Structure	Minimum Area (Pixels)
Chondro-osseous border	150
Femoral head	150
Synovial fold	20
Joint capsule	30
Labrum	20
Cartilaginous roof	30
Bony roof	20
Turning point	100

Thresholds apply after post-processing on the 256 × 256 grid. All eight masks must exceed their thresholds. These are pixel areas, without documented depth/field-of-view normalization. The technical record describes tuning-set derivation; dated threshold-freeze records and threshold-sweep outputs were unavailable.

Training used focal Tversky loss (FT; weights $\alpha = 0.7$, $\beta = 0.3$, exponent $\gamma = 0.75$, smoothing 1), top-K cross-entropy on the 15% highest-loss pixels, and boundary loss weighted by 0.3. Background received weight 0.5; the chondro-osseous border, synovial fold and capsule received 1.5; other structures received 1. AdamW used learning rate and weight decay set to 10^{-4}, Kaiming-normal initialization, a 10-epoch warm-up and cosine annealing. Training used at most 200 epochs, early stopping after 20 epochs without tuning-Dice improvement, and a physical batch of eight with four-step accumulation. Augmentation comprised small rotations, scale changes, and brightness/contrast changes,

with paired geometric transformations of images and masks; exact ranges were not retained. The reported environment was PyTorch 2.8.0 with CUDA 12.8 on Windows and an NVIDIA RTX 4050 graphics processing unit (GPU; NVIDIA Corporation, Santa Clara, CA, USA). A conventional U-Net trained with Dice loss on the same partitions was an exploratory comparator.

Checklist II was evaluated only after Checklist I acceptance. Its implementation checked the lower iliac limb, standard-plane geometry and labrum, including a fitted iliac baseline with an inclination tolerance of $\pm 5°$.

Representative Checklist II decisions are shown in Figure 2.

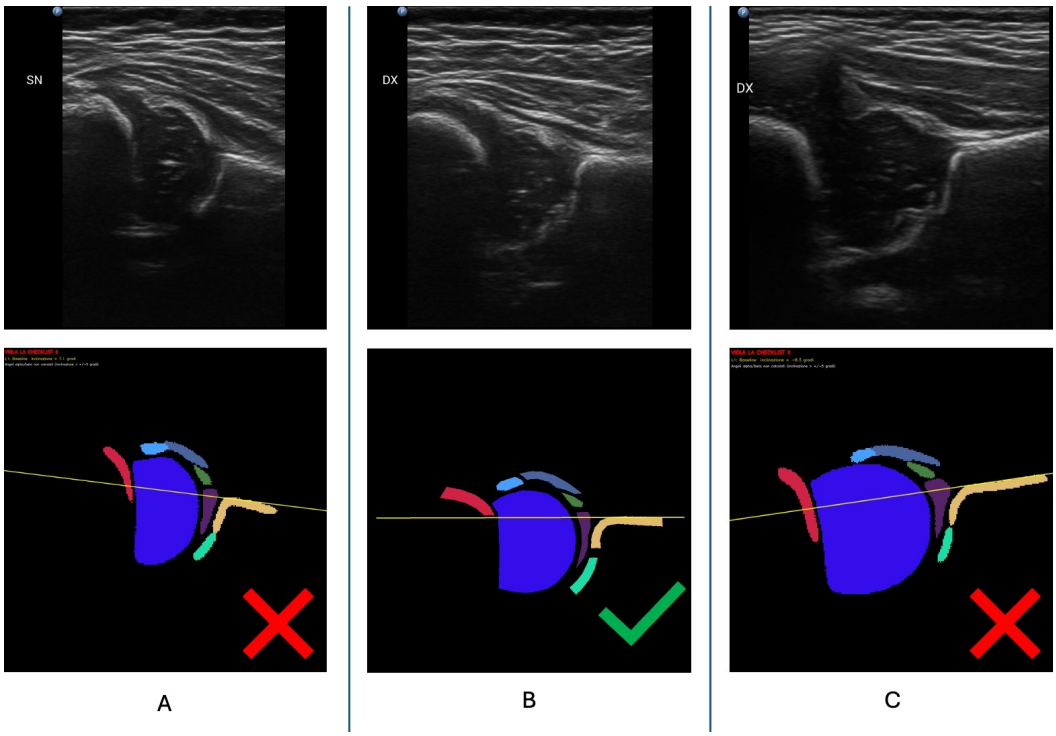


Figure 2. Representative Checklist II inclination assessment. The fitted iliac baseline (yellow) is outside the prespecified tolerance in (**A**,**C**) and within tolerance in (**B**). SN, left; DX, right. Colored regions indicate anatomical segmentation; red crosses and green check marks indicate rejection and acceptance, respectively.

For images passing both checks, the documented bony-roof construction used the turning point and segmented bony roof; the cartilaginous-roof line joined the turning point to the labral centroid. For a labral mask M_L, its geometric centroid is $c_L = |M_L|^{-1} \sum_{p \in M_L} p$, and the described cartilaginous direction is $v_c = c_L - t$, where t denotes the turning point. The angle between two direction vectors can be written as

$$\theta(u, v) = \frac{180}{\pi} \arccos\left(\frac{u \cdot v}{\|u\| \|v\|}\right). \quad (1)$$

Here, α compares the iliac baseline and bony-roof directions, and β compares the baseline and cartilaginous-roof directions. Figure 3 retains the submitted clinical and automated measurement example. This notation explains the reported construction; the precise contour fitting, direction orientation and supplementary-angle selection could not be verified without the inference implementation. A whole-mask centroid is not guaranteed to coincide with the clinician-selected labral landmark, and a line through a segmented region is not necessarily a tangent to the bony contour. The original model weights, executable inference code, complete contour-fitting and angle-orientation rules, per-image

mask areas, continuous Checklist II outputs, and dated threshold configuration were not retained. Consequently, the original inference pipeline cannot presently be independently reproduced, and alternative angle geometry, area-threshold sensitivity, and Checklist II tolerance sensitivity have not been tested.

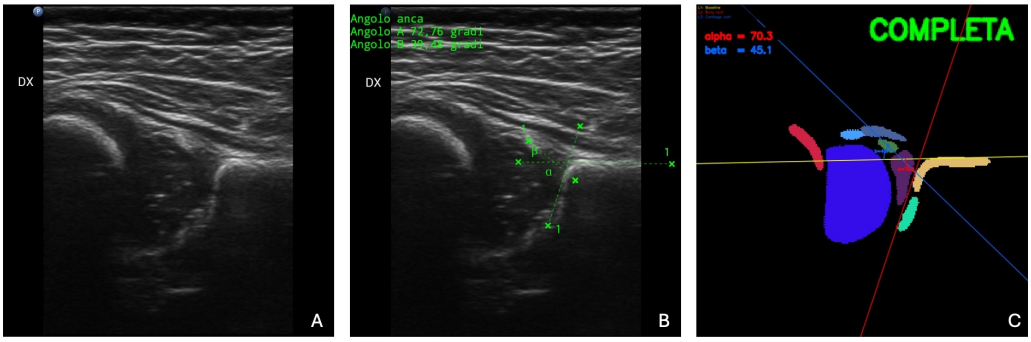


Figure 3. Graf measurement example: (**A**) original image; (**B**) clinician-generated angles; and (**C**) automated baseline, bony-roof line, cartilaginous-roof line, and α/β measurements. DX, right.

2.4. Endpoints and Reader Study

The binary imaging endpoint was Graf type I (Ia/Ib) versus non-type-I (IIa/IIb, IIc, D or III), replacing the original label "DDH versus no DDH". Although examination ages were supplied in the study registry during revision, its image codes cannot be matched reliably to the separately anonymized workflow prediction records. Workflow IIa and IIb outputs therefore could not be separated using age. The endpoint includes potentially physiological immaturity and is not an age-adjusted clinical DDH diagnosis. The simplified angle-based category rules are stated in Supplementary Methods S3. A post hoc restricted analysis excluded all reference IIa/IIb images and evaluated reference IIc/D/III versus I, with a positive prediction restricted to IIc/D/III.

Fifteen clinicians assessed the same 100 static images without CAD and four weeks later with CAD. The authors reported nine consultants/specialists and six residents, comprising 12 orthopaedic and three pediatric-radiology readers. The image order was randomized for presentation in each session, whereas the exported question sequence identified the same cases across sessions. Microsoft Forms displayed the original image and, in the assisted session, the segmentation and compliance suggestion. Readers could zoom but could not modify image intensity or window settings. Written instructions reported Checklist I elements and included a labelled anatomical example. The supplied assisted-session instructions are reproduced in Supplementary Methods S4. Forms start and finish times were retained as exploratory elapsed-session measures; pauses were not recorded or removed.

2.5. Statistical Analysis

Archived segmentation results were summarized using Dice and IoU. For both checklists, we report accuracy, sensitivity, specificity, precision, negative predictive value, F1-score and Matthews correlation coefficient (MCC). Graf classification was assessed using exact type agreement and the same binary measures for type I versus non-type I. We report classification performance among the 846 images accepted by both checklists and correct-output yield among all 1024 images entering the workflow. Angle agreement was summarized using mean absolute error, automated-minus-reference bias, Bland–Altman limits and the single-measure absolute-agreement intraclass correlation coefficient (ICC(A,1)). Agreement among the three experts was summarized using three-rater and pairwise kappa for the checklists and two-way random-effects ICC(A,1) for angles [21,22].

Proportions have 95% Wilson confidence intervals. F1-score, MCC, automated angle outcomes and automated-versus-reference ICC used 10,000 image-bootstrap replicates; checklist kappa intervals used 100,000. Expert-angle ICC intervals were calculated by Stata 18 (StataCorp LLC, College Station, TX , USA). The difference in reader accuracy between unaided and assisted sessions was estimated with a paired two-way bootstrap of readers and images (10,000 replicates) [23]. The anonymized reader-study cases could not be linked back to the 92 infant identifiers; the case dimension was therefore resampled at image level rather than infant level. No a priori sample-size calculation was performed. Reanalysis used Python and NumPy (versions and code in the analysis package), with random seed 20260917. Additional definitions and sensitivity analyses are in Supplementary Methods S4.

3. Results

3.1. Checklist I Segmentation and Reportability

Development results reported a mean Dice coefficient of 0.759 and mean intersection over union (IoU) of 0.641 for ResTransUNet. For Checklist I classification, the reported accuracy was 95.61%, precision 96.65%, recall 97.74%, and F1-score 97.19% (Table 3). The U-Net baseline yielded slightly higher mean Dice (0.766) and IoU (0.654), but lower reported Checklist I accuracy (91.67%), precision (96.47%), recall (92.66%), and F1-score (94.52%). At the class level, the femoral head had the highest overlap (Dice, 0.902; IoU, 0.845), whereas the synovial fold had the lowest (Dice, 0.654; IoU, 0.527). The underlying per-image development predictions and the numerator, denominator, and averaging convention used to calculate these percentages were not retained. Consequently, the reported percentages cannot be reconstructed as simple proportions from the 361-image internal test set. These values are presented only as unreconstructable historical summaries and are not treated as verified performance estimates. In the separate reader cohort, the stand-alone CAD system correctly classified 93 of 100 images in the Checklist I assessment.

Table 3. Archived segmentation and Checklist I reportability summaries.

(A). ResTransUNet segmentation performance		
Anatomical Class	**Dice**	**IoU**
Chondro-osseous border	0.795	0.681
Femoral head	0.902	0.845
Synovial fold	0.654	0.527
Joint capsule	0.720	0.586
Labrum	0.743	0.617
Hyaline cartilage roof	0.750	0.618
Bony roof	0.732	0.605
Turning point	0.776	0.646
Mean	0.759	0.641

(B). Archived model-level and Checklist I classification results						
Model	**Mean Dice**	**Mean IoU**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResTransUNet	0.759	0.641	95.61	96.65	97.74	97.19
U-Net baseline	0.766	0.654	91.67	96.47	92.66	94.52

The underlying per-image development predictions, numerator, denominator, and averaging convention were not retained. These values are unreconstructable historical summaries and are not treated as verified case-level performance estimates.

Three experts independently rated Checklist I within a 1017-image subset of the 1024-image workflow cohort before reaching consensus. Their ratings were unanimous for

1005/1017 images (98.8%), with a three-rater kappa of 0.786 (95% image-bootstrap CI, 0.626–0.893).

Figures 4 and 5 illustrate two different Checklist I errors: false acceptance when the bony roof was insufficiently visible and false rejection when the turning point was present.

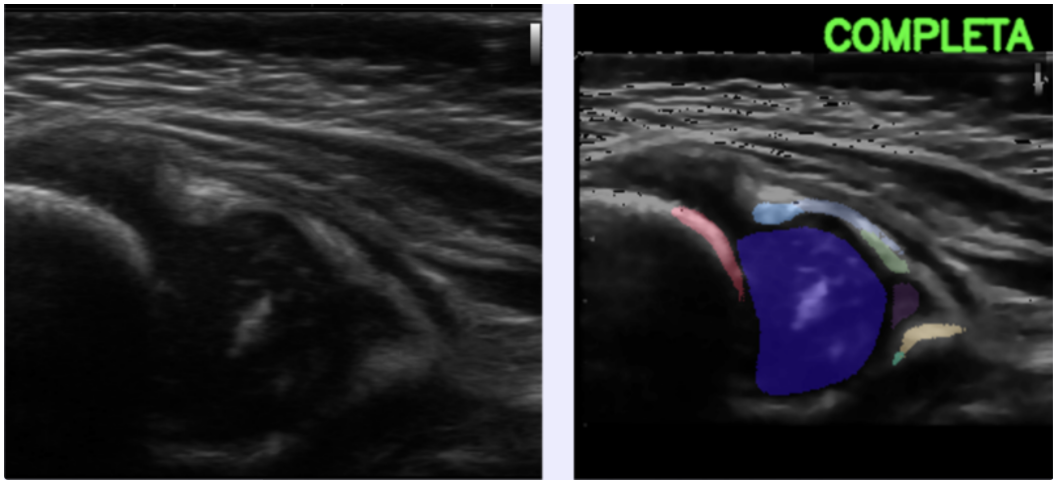


Figure 4. Illustrative false acceptance by the automated Checklist I assessment. (**Left**): ultrasound image without the segmentation overlay. (**Right**): automated segmentation and the output "COMPLETA" ("complete"), indicating Checklist I compliance. On clinical review, the bony roof (Checklist I item 7) was judged insufficiently delineated, despite the positive automated compliance decision.

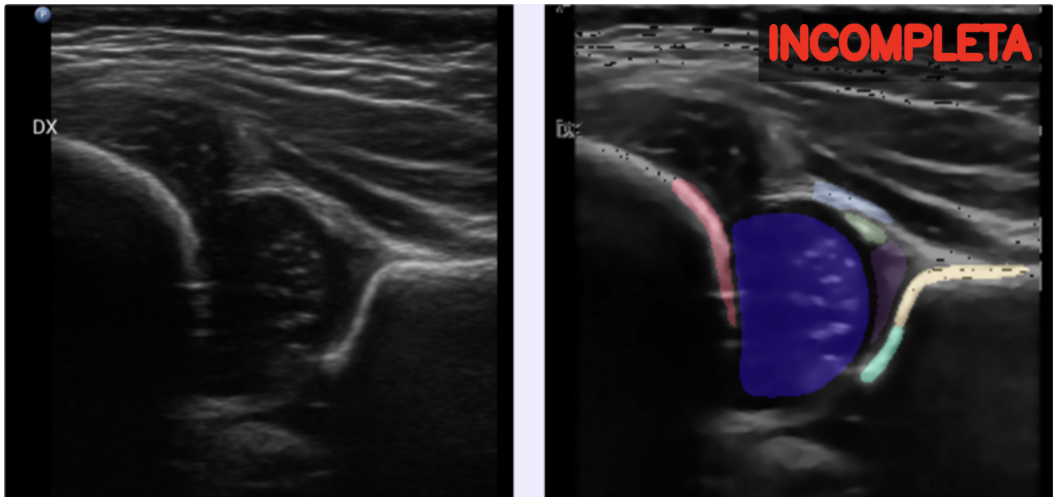


Figure 5. Illustrative false rejection by the automated Checklist I assessment. (**Left**): ultrasound image without the segmentation overlay. (**Right**): automated segmentation and the output "INCOMPLETA" ("incomplete"), indicating Checklist I refusal. On clinical review, the turning point (Checklist I item 8) was present, despite the negative automated compliance decision.

3.2. Sequential Workflow Performance

Of the 1024 images entering the sequential workflow, 18 were rejected at Checklist I and 160 of the remaining 1006 at Checklist II; 846 images therefore proceeded to angle measurement and classification (Table 4). Checklist II classification achieved 87.18% accuracy. Checklist II falsely rejected 83 reference-compliant images, including 22 reference non-type-I images, and allowed 46 reference-non-compliant images to proceed. Among the 846 images completing the workflow, the α-angle error was smaller than the β-angle error, with mean absolute errors of 3.59° and 11.91°, respectively. The reference Graf-type distribution at entry and at each automated rejection stage is shown in Table 5. Of the 266 reference non-type-I images entering the workflow, 64 were rejected before classification.

Table 4. Sequential workflow, quality-control classification, and angle-measurement performance.

(A). Image flow through the sequential workflow							
Stage		**Evaluated, *n***		**Not Forwarded, *n***		**Forwarded, *n* (% of Initial Set)**	
Checklist I		1024		18		1006 (98.24)	
Checklist II		1006		160		846 (82.62)	
Angle measurement and classification		846		–		846 (82.62)	
(B). Checklist II and binary Graf classification							
Endpoint	***n***	**Accuracy (%)**	**Precision (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**	**MCC**
Checklist II non-compliance	1006	87.18	48.13	62.60	90.60	54.42	0.477
Non-type-I Graf endpoint	846	80.97	57.00	82.67	80.43	67.47	0.565
(C). Automated versus clinical consensus angle measurements							
Angle	***n***	**MAE (°)**	**Bias (°)**	**SD (°)**	**95% Limits of Agreement (°)**	**ICC(A,1) (95% CI)**	
α	846	3.59	−1.66	4.20	−9.90 to 6.58	0.771 (0.739–0.798)	
β	846	11.91	11.07	8.36	−5.32 to 27.45	0.317 (0.277–0.356)	

MAE, mean absolute error; bias, automated minus clinical consensus measurement; MCC, Matthews correlation coefficient; ICC(A,1), single-measure absolute-agreement intraclass correlation coefficient; CI, confidence interval. For Checklist II, non-compliance was the positive class (TP = 77, FP = 83, TN = 800, FN = 46). For the binary Graf endpoint, non-type I was the positive class (TP = 167, FP = 126, TN = 518, FN = 35).

Table 5. Reference Graf types at workflow entry and after each automated decision.

Reference Type	All Entering	Checklist I	Checklist II	Accepted
Ia	473	4	61	408
Ib	285	7	42	236
IIa/IIb	218	6	37	175
IIc	21	1	5	15
D	11	0	5	6
III	16	0	10	6
Total	1024	18	160	846

Checklist I and II columns contain rejections at that stage. The latter is evaluated only among the 1006 images passing Checklist I. Counts refer to images, not independent infants; IIa and IIb could not be separated.

The corresponding 95% confidence intervals for reconstructable case-level performance measures are reported in Table 6.

Table 6. Reconstructable case-level model performance with 95% intervals.

Measure	Checklist I	Checklist II	Non-Type-I Endpoint
Accuracy	93.0 (86.3–96.6)	87.2 (85.0–89.1)	81.0 (78.2–83.5)
Sensitivity	97.1 (90.2–99.2)	62.6 (53.8–70.6)	82.7 (76.9–87.3)
Specificity	83.3 (66.4–92.7)	90.6 (88.5–92.4)	80.4 (77.2–83.3)
Precision	93.2 (84.9–97.0)	48.1 (40.5–55.8)	57.0 (51.3–62.5)
Negative predictive value	92.6 (76.6–97.9)	94.6 (92.8–95.9)	93.7 (91.3–95.4)
F1	95.1 (91.2–98.4)	54.4 (47.1–61.3)	67.5 (62.6–72.1)
MCC	0.831 (0.703–0.939)	0.477 (0.398–0.553)	0.565 (0.508–0.623)

All entries are percentages except MCC. Checklist I: reader cohort, $n = 100$, positive = compliant (TP/FP/TN/FN 68/5/25/2). Checklist II: $n = 1006$, positive = non-compliant (77/83/800/46). Non-type-I: accepted workflow images, $n = 846$, positive = IIa/IIb, IIc, D or III (167/126/518/35). Wilson intervals for proportions; image-bootstrap intervals for F1/MCC. All intervals are unadjusted for infant clustering.

Exact Graf-type agreement was 40.07% (339/846; 95% CI, 36.8–43.4; Table 7). Differentiation between types Ia and Ib contributed substantially to disagreement: 299 of 408 reference Ia images were classified as Ib. Errors also occurred in the other Graf categories. Binary type-I versus non-type-I classification achieved 80.97% accuracy, 82.67%

sensitivity for non-type I, and 80.43% specificity among images accepted by both checklists. Across all 266 reference non-type-I images entering the workflow, 167 received a correct positive output (62.8%). Thus, 82.67% represents sensitivity conditional on acceptance by both automated checklists, whereas 62.8% represents the correct-positive-output yield across the complete workflow. Across all 1024 entering images, 685 (66.9%) received a correct binary output and 339 (33.1%) received a correct exact Graf-type output; rejected images received no automated final classification.

Table 7. Graf-type classification and expert interobserver angle agreement.

(A). Confusion matrix for automated Graf classification

Reference Type	Automated Type						Total
	D	III	IIa/IIb	IIc	Ia	Ib	
D	1	1	4	0	0	0	6
III	2	2	0	2	0	0	6
IIa/IIb	16	2	116	6	5	30	175
IIc	4	2	5	4	0	0	15
Ia	0	0	47	0	62	299	408
Ib	1	0	78	0	3	154	236
Total	24	7	250	12	70	483	846

(B). Pairwise expert angle agreement in 1017 workflow images

Angle	Expert Pair	MAE (°)	Bias (°)	SD (°)	95% Limits of Agreement (°)
α	1–2	3.609	−0.445	4.625	−9.510 to 8.620
α	1–3	3.628	−0.356	4.601	−9.375 to 8.662
α	2–3	3.574	0.088	4.504	−8.739 to 8.915
β	1–2	7.044	−1.141	8.700	−18.194 to 15.912
β	1–3	6.867	−0.399	8.715	−17.480 to 16.683
β	2–3	6.809	0.743	8.604	−16.120 to 17.606

Angle	Expert ICC(A,1)	95% CI
α	0.886	0.875–0.896
β	0.726	0.701–0.749

Values on the diagonal of Panel A indicate exact agreement. In Panel B, bias was calculated as the first-listed minus the second-listed expert. The expert ICCs use a two-way random-effects, absolute-agreement, single-measure model.

As a separate sensitivity analysis, downstream evaluation was restricted to the 800 images that were accepted by the automated workflow and were also reference-compliant at Checklist II. Binary accuracy was 647/800 (80.9%; 95% CI, 78.0–83.5), sensitivity was 157/188 (83.5%; 95% CI, 77.5–88.1), and specificity was 490/612 (80.1%; 95% CI, 76.7–83.0). Exact Graf agreement was 319/800 (39.9%; 95% CI, 36.5–43.3). The α-angle MAE was 3.55°; the β-angle MAE and bias were 11.83° and +11.11°, respectively. Restriction to reference Checklist II-compliant images therefore did not materially improve downstream performance.

In a post hoc analysis excluding all 175 accepted images with reference type IIa/IIb, 671 accepted images remained: 644 type I and 27 type IIc/D/III. With predicted IIc/D/III counted as positive, sensitivity was 18/27 (66.7%; 95% CI, 47.8–81.4) and specificity was 643/644 (99.8%; 95% CI, 99.1–100.0). This restriction also excludes type IIb and changes the case spectrum.

After 15 images were unanimously judged non-compliant at Checklist I, the three experts assessed Checklist II on the remaining 1002 images. Checklist II ratings were unanimous for 970/1002 images (96.8%), with a three-rater kappa of 0.917 (95% image-bootstrap CI, 0.886–0.944). For the 1017 images with three angle ratings, interobserver agreement was higher for the α angle [ICC(A,1), 0.886; 95% CI, 0.875–0.896] than for the β

angle [ICC(A,1), 0.726; 95% CI, 0.701–0.749]. Pairwise mean absolute differences ranged from 3.574° to 3.628° for α and from 6.809° to 7.044° for β.

3.3. Paired Clinical Reader Study

Fifteen clinicians completed both sessions. Mean accuracy increased from 72.3% without CAD to 90.2% with CAD, an observed mean paired difference of 17.9 percentage points (95% paired reader-and-image bootstrap CI, 10.0–26.1; Table 8). All 15 clinicians had higher accuracy during the subsequent CAD-assisted session. Across the 1500 reader–image assessments, the assisted session contained 268 additional correct classifications. The stand-alone CAD model correctly classified 93 of the 100 reader-study images.

Table 8. Accuracy and recorded session duration in the paired clinical reader study.

(A). Reader accuracy				
Group	***n***	**Unaided Accuracy (%)**	**CAD-Assisted Accuracy (%)**	**Gain (Percentage Points)**
All clinicians	15	72.3 ± 16.2	90.2 ± 7.6	17.9 (95% CI, 10.0–26.1)
(B). Recorded session duration				
Outcome		**Unaided**	**CAD-Assisted**	
Time, min		13.12 (12.53–16.54)	8.95 (7.90–16.28)	

Accuracy values are mean ± SD across clinicians. The gain is the mean paired difference; its interval comes from a paired two-way bootstrap of readers and images. Time values are median (interquartile range) from questionnaire start and finish timestamps, not uninterrupted interpretation times.

Assisted reader decisions agreed with the CAD suggestion in 1344/1500 assessments (89.6%). On the seven images misclassified by CAD, readers were correct in 69/105 unaided assessments and 57/105 assisted assessments. Twenty-six initially correct decisions became incorrect with CAD displayed, whereas 14 initially incorrect decisions became correct. Across all 100 images, 334 reader decisions changed from incorrect to correct and 66 changed from correct to incorrect. These observations show that readers sometimes followed incorrect CAD suggestions; the fixed unaided-first session order does not establish that CAD caused any individual decision change.

Recorded median session duration was 13.12 min unaided and 8.95 min with CAD.

4. Discussion

This study evaluated an AI-assisted workflow designed to reproduce the sequential methodology described by Graf [5,6]. Required anatomical structures and the standard plane must be confirmed before angles are measured and a Graf type assigned. In the separate reader cohort, the stand-alone model classified 93 of 100 images correctly, while mean clinician accuracy was 72.3% unaided and 90.2% in the later CAD-assisted session. Performance declined downstream, particularly for β-angle estimation and exact Graf classification.

These findings should be interpreted within a heterogeneous literature. A review of 15 ultrasound studies reported promising performance but low certainty because of methodological differences and limited external validation [17]. Previous systems have addressed direct classification, standard-plane recognition, angle measurement, and Graf typing [11–13,15,24–26]. QualityDDH already implements explicit Checklist I and Checklist II assessment [16]. The contribution of the present study is the evaluation of how sequential quality decisions affect downstream outputs and how readers respond to AI suggestions.

Previous work on acquisition adequacy supports rejecting unsuitable images before measurement. Quader et al. reported an AUC of 0.985 for adequacy assessment, while Lee et al. found that only 320 of 921 images met both model and human validity criteria [10,27]. Other studies have evaluated standard-plane recognition or deliberately

included incorrectly positioned images [25,28,29]. Quality gates themselves, however, also require evaluation. In the present workflow, Checklist II rejected 83 reference-compliant images, including 22 non-type-I images, and allowed 46 reference-non-compliant images to proceed; its precision for non-compliance was 48.1%. Among 266 reference non-type-I images entering the workflow, 167 (62.8%) received a correct positive output, compared with 82.7% sensitivity among images accepted by both checklists.

Segmentation overlap and clinical adequacy are distinct. The femoral head had the greatest overlap and the synovial fold the lowest. Although U-Net had slightly higher aggregate Dice and IoU, the archived Checklist I classification metrics were numerically higher for ResTransUNet; these single-run results do not establish architectural superiority. A small omitted region may barely change overlap yet cause a required structure to fall below its presence threshold. Those thresholds were applied after resizing to 256×256 pixels, and their behavior across image scales and ultrasound systems was not established. This supports evaluating segmentation against its intended clinical endpoint [14]. Joint landmark–structure approaches may help preserve anatomical relationships [30].

Published measurement studies show both progress and persistent difficulty with β estimation [10,12,13,30,31]. In this study, β had a mean absolute error of 11.91°, positive bias of 11.07°, and absolute-agreement ICC of 0.317. Frequent Ia–Ib disagreement was accompanied by errors in IIa/IIb, IIc, D, and III. Exact Graf agreement was 40.07%, compared with 80.97% accuracy for the study-defined type-I versus non-type-I endpoint among accepted images. Use of the segmented labral centroid may contribute to β error. Differences in cohorts, reference standards, and metrics preclude direct ranking against other systems. Binary classification cannot replace exact Graf typing when category influences management.

The expanded interobserver analysis provides context for the clinical reference. Three-rater agreement was high for Checklist I ($\kappa = 0.786$) and Checklist II ($\kappa = 0.917$). Absolute-agreement ICCs were 0.886 for α and 0.726 for β, indicating lower reproducibility of the latter. These results support consistency within the expert group, but do not make its consensus an independent outcome-based diagnosis.

The paired reader study suggests potential for clinician-facing support but cannot establish that CAD caused the higher accuracy in the second session. All 15 clinicians had higher accuracy during the subsequent CAD-assisted session. Prior work supports clinician oversight, particularly for poor-quality images, and has shown benefits of AI assistance in other orthopaedic tasks [32,33]. Here, however, mean assisted accuracy (90.2%) remained below stand-alone model accuracy (93.0%). On seven model-error images, 26 previously correct reader decisions became incorrect in the assisted session, while 14 previously incorrect decisions became correct.

Three-dimensional and multisource approaches may reduce dependence on a manually selected frame, although acquisition quality remains consequential [34,35]. Future work should improve angle measurement and exact classification while verifying both checklists, then test the complete pathway prospectively across centres, devices, operators, and screening populations. Graf type IV hips and atypical anatomies should be included. A counterbalanced reader study using new cases would help separate CAD assistance from familiarity with previously assessed images.

Several limitations restrict generalizability. This retrospective, single-centre study used one ultrasound platform and a referred population; its workflow sample deliberately retained different Graf types and was not prevalence-representative. Type IV hips were excluded, and static frames did not test real-time acquisition or patient outcomes. The linkage between archived prediction records and registry infant identifiers was removed during anonymization to protect patient privacy. Consequently, infant-level separation of

the archived prediction records could not be independently re-audited, and image-level confidence intervals could not be adjusted for within-infant clustering and may be too narrow. Sequential gating also made angle and classification estimates conditional on automated acceptance. The original inference pipeline cannot presently be independently reproduced because the trained weights, executable inference code, and complete post-processing and geometric specifications were not retained. Anatomical masks were produced by one supervised annotator without independent mask reproduction; their interobserver reproducibility is therefore unknown, and systematic annotation error cannot be excluded. For the 46 reference Checklist II-non-compliant images that received downstream outputs, the retained records do not establish whether the reference measurements originated from the same non-standard frame or from another view in the examination. Finally, the type-I versus non-type-I endpoint is not an age-adjusted clinical DDH diagnosis, and the local consensus reference remains subject to uncertainty at poorly visualized boundaries [36].

The reader study involved 15 clinicians and 100 images, with a non-compliance frequency that may differ from screening practice. All readers completed the unaided session first and reassessed the same images four weeks later. Randomized presentation did not remove possible recall, learning, or period effects. The anonymized questionnaire cases could not be mapped back to the 92 infant identifiers, so infant-clustered resampling was not possible and the reported confidence interval may be too narrow. The paired reader-and-image bootstrap captured sampling variation across readers and image cases.

5. Conclusions

The sequential workflow shows potential as clinician-facing support for assessing stored hip-ultrasound images, but it is not ready for autonomous classification or clinical deployment. The β-angle mean absolute error of 11.91°, positive bias of 11.07°, and exact Graf-type agreement of 40.1% show that binary type-I versus non-type-I accuracy should not be interpreted as reliable exact Graf typing. Checklist II rejection also reduced the correct-positive-output yield across all entering non-type-I images to 62.8%. Prospective multicentre studies should address Checklist II, angle measurement, and exact Graf typing, including Graf IV and atypical hips.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jimaging12110554/s1. Figure S1. Operational workflow of the research prototype. The nine segmentation channels comprise eight anatomical structures and background. The schematic shows the documented stages, not every implementation parameter. Table S1. Class-specific minimum areas used for Checklist I reportability assessment. Table S2. Recorded rules used for automated Graf classification. Table S3. Archived development segmentation summaries. Table S4. Angle errors by reference Graf type and reference Checklist II status among accepted images. Table S5. Pairwise reference-expert angle differences in the 1017-image analysis. Table S6. Reader sensitivity analyses. Table S7. Illustrative prevalence-shift calculations for the non-type-I endpoint among accepted images. Table S8. Sex and age at the first contributing examination by cohort. Table S9. Repeat examinations, image laterality and within-cohort sampling. References [37–46] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, A.A. (Andrea Audisio); methodology, A.A. (Andrea Audisio), M.S., G.M., A.A. (Alessandro Aprato), L.U. and M.P.; software, M.S. and G.M.; validation, A.A. (Alessandro Aprato), M.P., A.M., E.V. and S.M.; formal analysis, A.A. (Andrea Audisio), M.S., G.M., A.A. (Alessandro Aprato) and M.P.; investigation, J.H.; resources, A.M., E.V. and S.M.; data curation, A.A. (Andrea Audisio) and J.H.; writing—original draft preparation, A.A. (Andrea Audisio); writing—review and editing, M.S., G.M., A.A. (Alessandro Aprato) and L.U.; supervision, A.A. (Alessandro Aprato) and L.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Città della Salute e della Scienza di Torino (Study Protocol: DB2021; Approval code: 1489/2021; Approval date: 14 November 2021).

Informed Consent Statement: Patient consent was waived for retrospective analysis of anonymized images. Clinicians participating in the reader study provided written informed consent.

Data Availability Statement: The retained de-identified study data are available on request from the corresponding author, provided that their release complies with institutional data-protection restrictions. Original ultrasound images and the clinical registry are subject to institutional data-protection restrictions. The linkage between archived prediction exports and registry infant identifiers was removed during anonymization to protect patient privacy. Retained materials include a de-identified 1024-row workflow output table, an anonymous 3000-row reader-response matrix, the pre-consensus reliability export, aggregate reliability results, and the scripts used for the revision reanalysis. The original trained weights, executable inference code, linked per-image development masks and predictions, per-image mask areas, continuous Checklist II inclination outputs, complete contour-fitting and angle-orientation rules, and dated threshold configuration were not retained and are unavailable. The retained materials reproduce the revision analyses but not model training or inference.

Acknowledgments: During the preparation of this manuscript, the authors used Codex (OpenAI; GPT-6.1 Sol) for language editing, preparation of analysis code, tables and figures, and drafting of responses to reviewers. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
CAD	Computer-aided diagnosis
CI	Confidence interval
CUDA	Compute Unified Device Architecture
CVAT	Computer Vision Annotation Tool
DDH	Developmental dysplasia of the hip
DX	Right (image laterality marker)
F1	Harmonic mean of precision and recall
FN	False negative
FP	False positive
FT	Focal Tversky loss
GPU	Graphics processing unit
ICC	Intraclass correlation coefficient
IoU	Intersection over union
MAE	Mean absolute error
MCC	Matthews correlation coefficient
SD	Standard deviation
SN	Left (image laterality marker)
TN	True negative
TP	True positive

References

1. De Pellegrin, M.; Tessari, L. Early Ultrasound Diagnosis of Developmental Dysplasia of the Hip. *Bull. Hosp. Jt. Dis.* **1996**, *54*, 222–225.
2. Agostiniani, R.; Atti, G.; Bonforte, S.; Casini, C.; Cirillo, M.; De Pellegrin, M.; Di Bello, D.; Esposito, F.; Galla, A.; Marrè Brunenghi, G.; et al. Recommendations for Early Diagnosis of Developmental Dysplasia of the Hip (DDH): Working Group Intersociety Consensus Document. *Ital. J. Pediatr.* **2020**, *46*, 150. https://doi.org/10.1186/s13052-020-00908-2.
3. Rosendahl, K.; Tomà, P. Ultrasound in the Diagnosis of Developmental Dysplasia of the Hip in Newborns: The European Approach—A Review of Methods, Accuracy and Clinical Validity. *Eur. Radiol.* **2007**, *17*, 1960–1967. https://doi.org/10.1007/s00330-006-0557-y.
4. De Pellegrin, M.; Filisetti, C.; Cossutta, M.; Colombo, M.; Consiglieri, G.; Tucci, F.; Aiuti, A.; Bernardo, M.E. Ultrasonographic Hip Morphology in Mucopolysaccharidosis Type I Hurler after Hematopoietic Stem Cell Gene Therapy. *J. Child. Orthop.* **2025**, *19*, 377–385. https://doi.org/10.1177/18632521251349438.
5. Graf, R. Hip Sonography: Background; Technique and Common Mistakes; Results; Debate and Politics; Challenges. *Hip Int.* **2017**, *27*, 215–219. https://doi.org/10.5301/hipint.5000514.
6. Graf, R.; Maizen, C.; Seidl, T. *Sonography of the Infant's Hip: Principles, Implementation and Therapeutic Consequences*; Springer: Cham, Switzerland, 2024. https://doi.org/10.1007/978-3-031-71949-3.
7. O'Beirne, J.G.; Chlapoutakis, K.; Alshryda, S.; Aydingoz, U.; Baumann, T.; Casini, C.; de Pellegrin, M.; Domos, G.; Dubs, B.; Hemmadi, S.; et al. International Interdisciplinary Consensus Meeting on the Evaluation of Developmental Dysplasia of the Hip. *Ultraschall Med.* **2019**, *40*, 454–464. https://doi.org/10.1055/a-0924-5491.
8. Peterlein, C.D.; Schüttler, K.F.; Lakemeier, S.; Timmesfeld, N.; Görg, C.; Fuchs-Winkelmann, S.; Schofer, M.D. Reproducibility of Different Screening Classifications in Ultrasonography of the Newborn Hip. *BMC Pediatr.* **2010**, *10*, 98. https://doi.org/10.1186/1471-2431-10-98.
9. Roposch, A.; Graf, R.; Wright, J.G. Determining the Reliability of the Graf Classification for Hip Dysplasia. *Clin. Orthop. Relat. Res.* **2006**, *447*, 119–124. https://doi.org/10.1097/01.blo.0000203475.73678.be.
10. Quader, N.; Hodgson, A.J.; Mulpuri, K.; Schaeffer, E.K.; Abugharbieh, R. Automatic Evaluation of Scan Adequacy and Dysplasia Metrics in 2-D Ultrasound Images of the Neonatal Hip. *Ultrasound Med. Biol.* **2017**, *43*, 1252–1262. https://doi.org/10.1016/j.ultrasmedbio.2017.01.012.
11. 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* **2022**, *12*, 1423. https://doi.org/10.3390/diagnostics12061423.
12. 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. *J. Ultrasound Med.* **2023**, *42*, 1235–1248. https://doi.org/10.1002/jum.16133.
13. Chen, Y.P.; Fan, T.Y.; Chu, C.C.J.; 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. *Biomed. J.* **2024**, *47*, 100614. https://doi.org/10.1016/j.bj.2023.100614.
14. Hareendranathan, A.R.; Chahal, B.; Ghasseminia, S.; Zonoobi, D.; Jaremko, J.L. Impact of Scan Quality on AI Assessment of Hip Dysplasia Ultrasound. *J. Ultrasound* **2022**, *25*, 145–153. https://doi.org/10.1007/s40477-021-00560-4.
15. Jaremko, J.L.; Hareendranathan, A.; Seyed Bolouri, S.E.; Frey, R.F.; Dulai, S.; Bailey, A.L. AI-Aided Workflow for Hip Dysplasia Screening Using Ultrasound in Primary Care Clinics. *Sci. Rep.* **2023**, *13*, 9224. https://doi.org/10.1038/s41598-023-35603-9.
16. Liu, R.; Zhang, Y.; Luo, X.; Zheng, Y.; Liu, Q.; Liu, M.; Jiang, L. QualityDDH: Visualized Standardization of Neonatal Hip Ultrasound via a Structural Prior Regression Framework. *Vis. Comput.* **2025**, *41*, 11589–11602. https://doi.org/10.1007/s00371-025-04121-2.
17. Bhavsar, S.; Gowda, B.B.; Bhavsar, M.; Patole, S.; Rao, S.; Rath, C. Artificial Intelligence to Detect Developmental Dysplasia of Hip: A Systematic Review. *J. Paediatr. Child Health* **2025**, *61*, 1712–1727. https://doi.org/10.1111/jpc.70172.
18. Azmi, A.; McArthur, A.; McDonald, S.; Hareendranathan, A.; Jaremko, J.L. Diagnostic Accuracy of Artificial Intelligence-Assisted Infant Hip Ultrasound Interpretation for Developmental Dysplasia of the Hip: Systematic Review and Meta-Analysis. *Pediatr. Radiol.* **2026**, *56*, 1165–1173. https://doi.org/10.1007/s00247-026-06601-6.
19. Carnazzo, S.M.; Balconara, D.; La Quatra, M.; Praticò, A.D. Artificial Intelligence in Neonatal Hip Ultrasound: A Scoping Review of Methods and Clinical Evidence. *J. Paediatr. Child Health* **2026**. https://doi.org/10.1111/jpc.70474.
20. CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT). *Zenodo* **2023**. https://doi.org/10.5281/zenodo.8070041.
21. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *1*, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
22. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012. Erratum in *J. Chiropr. Med.* **2017**, *16*, 346. https://doi.org/10.1016/j.jcm.2017.10.001.
23. Owen, A.B. The Pigeonhole Bootstrap. *Ann. Appl. Stat.* **2007**, *1*, 386–411. https://doi.org/10.1214/07-AOAS122.

24. Gong, B.; Shi, J.; Han, X.; Zhang, H.; Huang, Y.; Hu, L.; Wang, J.; Du, J.; Shi, J. Diagnosis of Infantile Hip Dysplasia With B-Mode Ultrasound via Two-Stage Meta-Learning Based Deep Exclusivity Regularized Machine. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 334–344. https://doi.org/10.1109/JBHI.2021.3093649.
25. Atalar, H.; Üreten, K.; Tokdemir, G.; Tolunay, T.; Çiçeklidağ, M.; Atik, O.Ş. The Diagnosis of Developmental Dysplasia of the Hip From Hip Ultrasonography Images With Deep Learning Methods. *J. Pediatr. Orthop.* **2023**, *43*, e132–e137. https://doi.org/10.1097/BPO.0000000000002294.
26. Kinugasa, M.; Inui, A.; Satsuma, S.; Kobayashi, D.; Sakata, R.; Morishita, M.; Komoto, I.; Kuroda, R. Diagnosis of Developmental Dysplasia of the Hip by Ultrasound Imaging Using Deep Learning. *J. Pediatr. Orthop.* **2023**, *43*, e538–e544. https://doi.org/10.1097/BPO.0000000000002428.
27. 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* **2021**, *11*, 1174. https://doi.org/10.3390/diagnostics11071174.
28. Li, X.; Zhang, R.; Wang, Z.; Wang, J. Semi-Supervised Learning in Diagnosis of Infant Hip Dysplasia Towards Multisource Ultrasound Images. *Quant. Imaging Med. Surg.* **2024**, *14*, 3707–3716. https://doi.org/10.21037/qims-23-1384.
29. Xu, Y.; Yu, H.; Lin, X.; Zhao, Y.; Huang, Z.; Xu, X.; Yang, X. Artificial Intelligence Technology in Ultrasound Screening of Infant Developmental Dysplasia of the Hip. *Chin. J. Med. Imaging Technol.* **2023**, *39*, 1229–1233. (In Chinese)
30. 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 J. Biomed. Health Inform.* **2022**, *26*, 345–358. https://doi.org/10.1109/JBHI.2021.3087494.
31. Oelen, D.; Kaiser, P.; Baumann, T.; Schmid, R.; Bühler, C.; Munkhuu, B.; Essig, S. Accuracy of Trained Physicians Is Inferior to Deep Learning-Based Algorithm for Determining Angles in Ultrasound of the Newborn Hip. *Ultraschall Med.* **2022**, *43*, e49–e55. https://doi.org/10.1055/a-1177-0480.
32. Ghasseminia, S.; Lim, A.K.S.; Concepcion, N.D.P.; Kirschner, D.; Teo, Y.M.; Dulai, S.; Mabee, M.; Kernick, S.; Brockley, C.; Muljadi, S.; et al. Interobserver Variability of Hip Dysplasia Indices on Sweep Ultrasound for Novices, Experts, and Artificial Intelligence. *J. Pediatr. Orthop.* **2022**, *42*, e315–e323. https://doi.org/10.1097/BPO.0000000000002065.
33. Tanzi, L.; Audisio, A.; Cirrincione, G.; Aprato, A.; Vezzetti, E. Vision Transformer for Femur Fracture Classification. *Injury* **2022**, *53*, 2625–2634. https://doi.org/10.1016/j.injury.2022.04.013.
34. He, J.; Cui, L.; Chen, T.; Lyu, X.; Yu, J.; Guo, W.; Wang, D.; Qin, X.; Zhao, Y.; Zhang, S. Study on Multiplanar Measurements of Infant Hips With Three-Dimensional Ultrasonography. *J. Clin. Ultrasound* **2022**, *50*, 639–645. https://doi.org/10.1002/jcu.23185.
35. Ghasseminia, S.; Seyed Bolouri, S.E.; Dulai, S.; Kernick, S.; Brockley, C.; Hareendranathan, A.R.; Zonoobi, D.; Rao, P.; Jaremko, J.L. Automated Diagnosis of Hip Dysplasia From 3D Ultrasound Using Artificial Intelligence: A Two-Center Multi-Year Study. *Inform. Med. Unlocked* **2022**, *33*, 101082. https://doi.org/10.1016/j.imu.2022.101082.
36. Audisio, A.; Zhu, T.; Joeris, A.; Sermon, A.; IJpma, F.F.A.; Giordano, V.; Desai, D.; Giannoudis, P.V.; Aprato, A. Pilot Validation Study for a Large Image Database of Proximal Femur Fracture Anteroposterior Radiographs: Searching for the Ground Truth. *Injury* **2026**, *57*, 113056. https://doi.org/10.1016/j.injury.2026.113056. Corrigendum in *Injury* **2026**, *57*, 113126. https://doi.org/10.1016/j.injury.2026.113126.
37. Smith, R. An Overview of the Tesseract OCR Engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition, 2007; Volume 2, pp. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991.
38. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer, 2015; Volume 9351, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
40. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122. https://arxiv.org/abs/1511.07122.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
42. Oktay, O.; Schlemper, J.; Le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999. https://arxiv.org/abs/1804.03999.
43. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging*; IEEE: Venice, Italy, 2019; pp. 683–687. https://doi.org/10.1109/ISBI.2019.8759329.
44. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LO, USA, 6–9 May 2019. https://doi.org/10.48550/arXiv.1711.05101.

45. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Santiago, Chile, 2015; pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.
46. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2019; Volume 32, pp. 8024–8035.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.