

Improving clinician trust in automated hip screening: Utilising heatmaps to express confidence in anatomical landmarks

Singh Abhinav, Clement Allison, Perry Daniel, Hemmadi Sandeep, Voiculescu Irina

Background: The UK screening programme has not improved early detection of developmental dysplasia of the hip (DDH) since 1986. The Graf method has intra/inter-observer differences ranging from 4 to 10.1 degrees. Artificial intelligence (AI) assisted diagnostics can improve accuracy and objectivity. Improving clinician trust remains paramount in facilitating the adoption of these techniques.

Methods: This study utilised an anonymised dataset obtained by experienced radiologists during routine NHS practice. The dataset contained 190 2D hip ultrasound images from 100 multi-ethnic babies (aged 4-12 weeks). Two clinical experts provided the five anatomical landmarks representing the base of ilium (1st and 2nd points), turning point (3rd), lower limb (4th) and labrum (5th). An alpha angle >60 degrees was normal (Graf 1) and <60 degrees was abnormal (Graf 2a,2c,D,3,4). We adapted a convolutional neural network (modified U-Net++, 70:15:15 training:validation:testing data split) to generate a confidence heatmap around each landmark prediction. The Graf angles were calculated from the 'hottest' (most confident) point to provide a screening diagnosis.

Results: Within the test dataset- ilium, femoral head and labrum heatmaps were identified in all images. In 93% of cases (n=27) the true and predicted alpha angles were within the same correct diagnostic threshold. An addition of 1-pixel radii outward from the centre of the ground truth led to a difference of 5.41 ± 0.82 degrees. The largest heatmap radius corresponded with the turning point and labrum which reflects identification difficulty in clinical practice.

Conclusions: Our AI heatmap successfully indicates its 'level' of confidence in identifying key landmarks within scans. It automatically calculates Graf angles and provides a diagnosis which agrees with expert clinicians in 93% of cases, exceeding the existing state-of-the-art performance.

Implications: This method will improve clinician confidence when selecting diagnostic landmarks. It will allow non-experts to participate in universal screening.

Disclosure: None.