Automated Measurement of Migration Percentage in Hip Surveillance Radiographs: Development and Testing of a Deep-Learning "Artificial Intelligence" Algorithm

Chun-Hsiao Yeh, MS [a]; Justin Krogue, MD [b]; Patrick Donahue, BS [c]; Marie Villalba, BS [d]; Sangryul Jeon, PhD [a]; Stella X. Yu, PhD [a]; Vedant A. Kulkarni, MD[d]

[a] University of California, Berkeley, and International Computer Science Institute

[b] Department of Orthopaedic Surgery, University of California, San Francisco

[c] Washington University School of Medicine

[d] Department of Orthopaedic Surgery, Shiners Hospitals for Children – Northern California

Background & Objectives: Hip surveillance in children with cerebral palsy (CP) requires accurate assessment of the migration percentage (MP) on radiographs. Poor access to radiologists trained in the measurement of MP and inconsistent reporting of MP are barriers to hip surveillance implementation. This study aims to develop and test a deep-learning algorithm that automatically measures MP on hip surveillance radiographs.

Study Participants & Setting: In this diagnostic study, we included anterior-posterior (AP) pelvis radiographs of children with CP between the ages of 2 – 18y undergoing hip surveillance at a referral children's surgical hospital

Materials & Methods: Radiographs were de-identified and an online image annotation tool was used to label relevant pelvic and femoral landmarks in two steps for the calculation of "ground truth" MP [Figure 1]. Subjects were randomly allocated into three image sets: training (70%), validation (15%), and test (15%). Two convolutional neural network (CNN) deep-learning models were trained to calculate the MP. The first CNN model based on ResNet 18 architecture was trained to calculate the degrees of rotation needed to level the pelvis, while the second model based on Cascaded Pyramid Network architecture was trained to detect the key landmarks for calculation of MP on the leveled image. The measurement error and the reliability of the deep learning algorithm on the test image set were calculated referenced against expert-labeled "ground truth" MP.

Results: A total of 1275 radiographic images from 588 subjects with CP were identified for the study (53%M, 47%F; mean age at x-ray 6.8y; 19% GMFCS I, 18% GMFCS II, 13% GMFCS III, 35% GMFCS IV, and 32% GMFCS V). The training (894 images), validation (186 images), and test (195 images) sets were comparable groups with no significant differences in age at x-ray and MP ($p=0.22$ and $p=0.69$). The deep-learning algorithm had a mean error of 8% $\pm$ 10% [Figure 2]. Increased error was weakly but significantly correlated with increasing MP (R-squared = 0.1, $p < 0.0001$), but was not significantly associated with the presence of hip or spine implants ($p=0.21$). When a ground truth MP > 30% is considered a "positive" case, the deep learning algorithm had a sensitivity of 70%, specificity of 94%, positive predictive value of 85%, and negative predictive value of 87%. The deep-learning algorithm performed with an AUC of 0.923, indicating a model with excellent discriminatory characteristics. The deep-learning model functions most optimally when an automated measurement of MP $\geq$ 27% constitutes a "positive case," with sensitivity and specificity reaching 85% [Figure 3].

Conclusions/Significance: A deep-learning algorithm can automatically determine MP with high accuracy. As a screening tool, the algorithm has excellent discriminatory characteristics, highlighting its role in a hip surveillance program. Further refinement of the algorithm using larger data sets will improve its generalizability for use in hip surveillance programs worldwide.

Figure 1

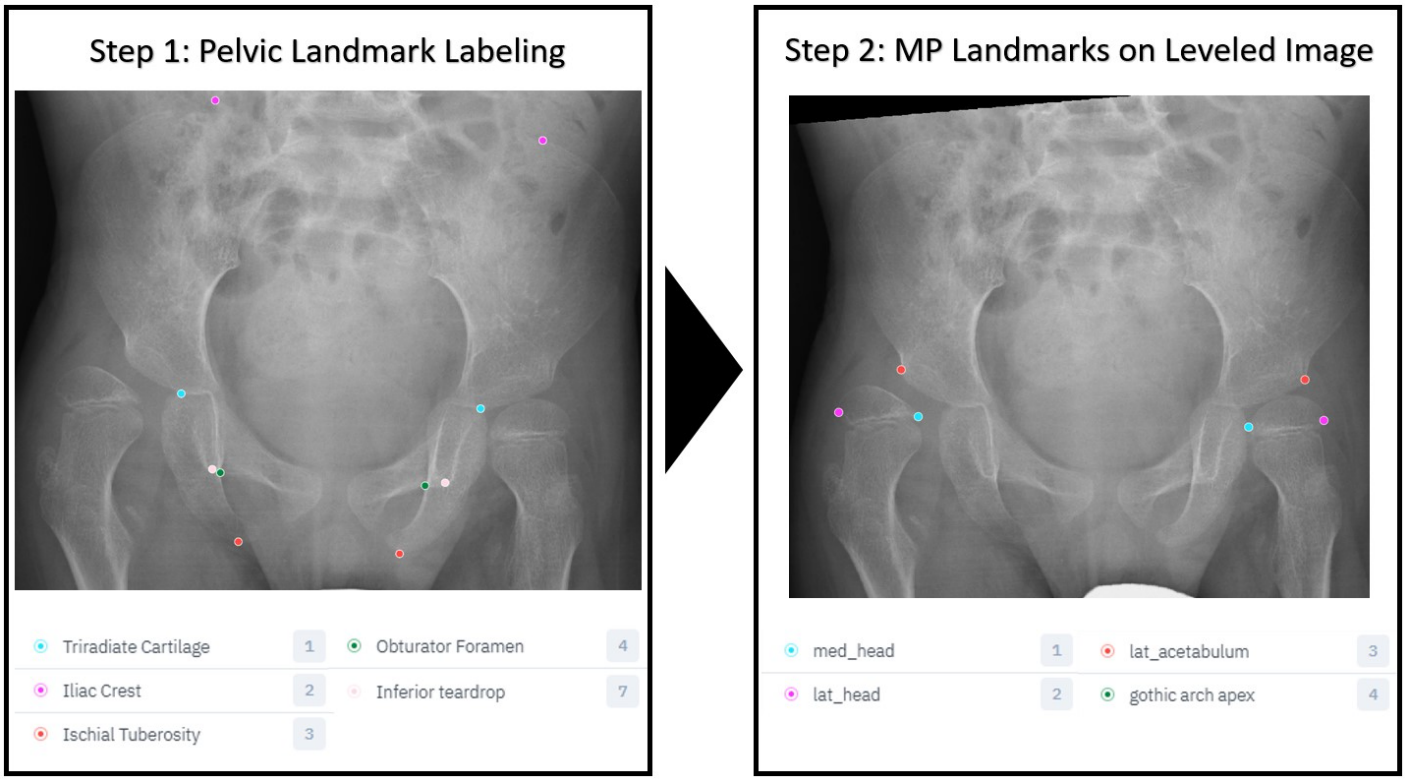
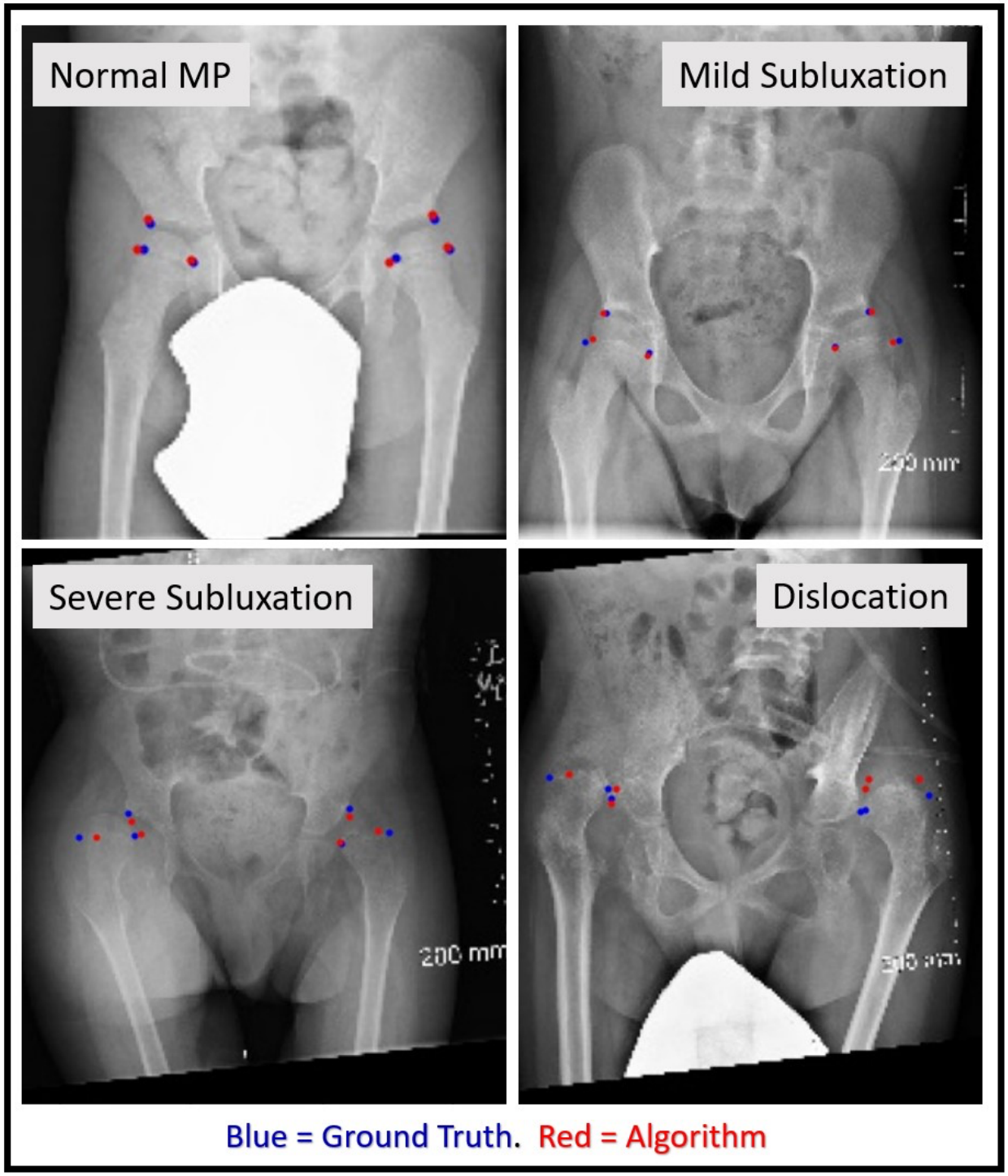
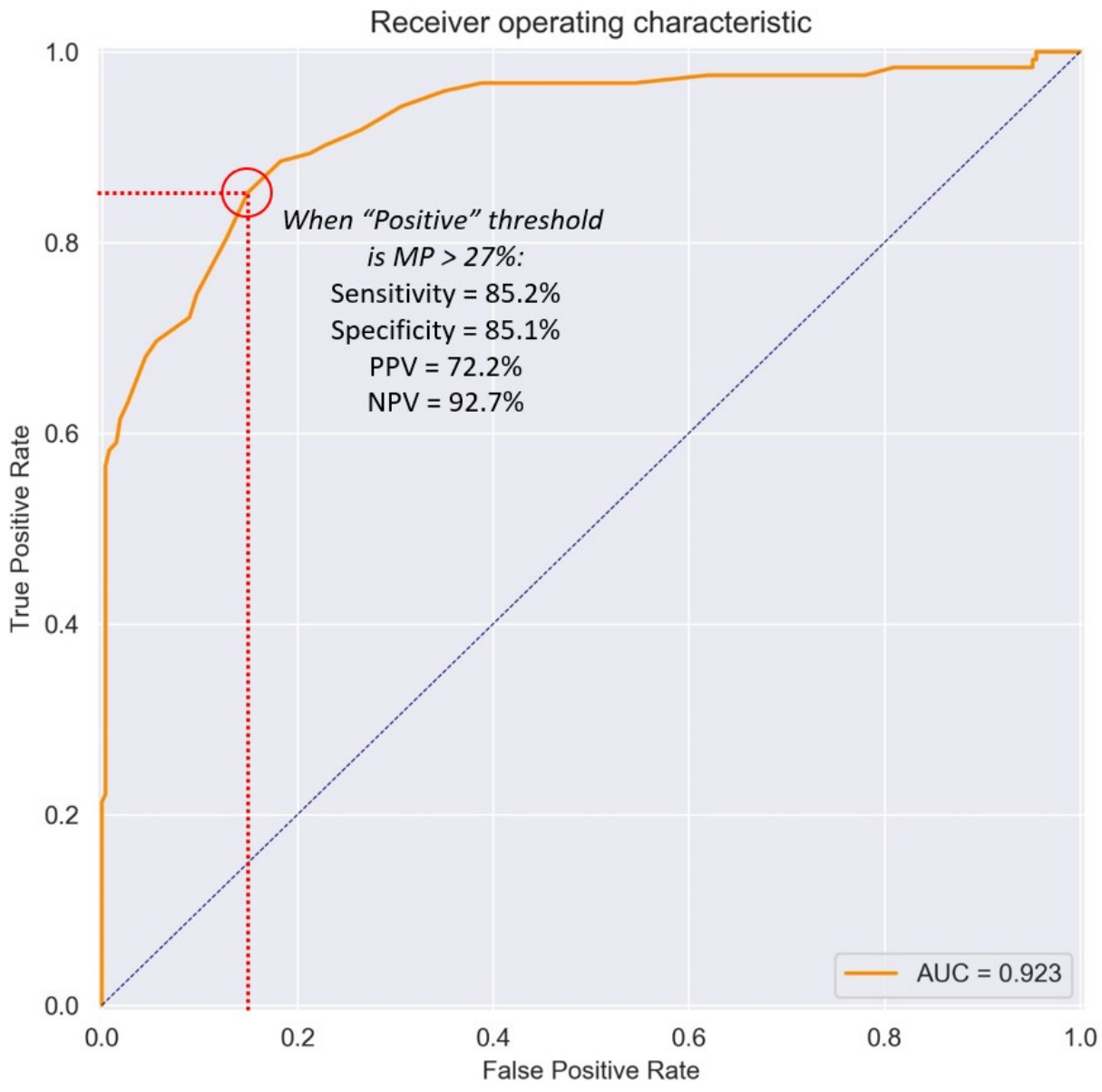

Step 1: Pelvic Landmark Labeling
Triradiate Cartilage 1
Iliac Crest 2
Ischial Tuberosity 3
Obturator Foramen 4
Inferior teardrop 7
Step 2: MP Landmarks on Leveled Image
med_head 1
lat_head 2
lat_acetabulum 3
gothic arch apex 4

Figure 2:


Normal MP
Mild Subluxation
Severe Subluxation
Dislocation
Blue = Ground Truth. Red = Algorithm

‘

Figure 3:


Receiver operating characteristic
When "Positive" threshold
is MP > 27%:
Sensitivity = 85.2%
Specificity = 85.1%
PPV = 72.2%
NPV = 92.7%
AUC = 0.923
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0