Image Resolution Affects The Performance Of Deep Learning Algorithms In Detecting Calcaneus Fractures On X-ray

Nicholas J. Yee[1,4,5,6], Atta Taseh[1], Samir Ghandour[1], Mansur Halai[3,4], Cari Whyne[4,5,6], Christopher W. DiGiovanni[1,2], John Y. Kwon[1,2], Soheil Ashkani-Esfahani[1,2]

[1]Foot & Ankle Research and Innovation Lab, Department of Orthopaedic Surgery, Mass General Brigham, Harvard Medical School, Boston, MA, [2]Foot and Ankle Division, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, [3]Division of Orthopaedic Surgery, St. Michael's Hospital, University of Toronto, Toronto, ON, [4]Division of Orthopaedic Surgery, Department of Surgery, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, [5]Institute of Biomedical Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, ON, [6]Orthopaedic Biomechanics Lab, Holland Bone and Joint Program, Sunnybrook Research Institute, University of Toronto, Toronto, ON
Nick.Yee@mail.utoronto.ca

Disclosures: N.J. Yee: None. A. Taseh: None. S. Ghandour: None. M. Halai: None. C. Whyne: None. C. DiGiovanni: None. J.Y. Kwon: None. S. Ashkani-Esfahani: None.

INTRODUCTION: Calcaneus fractures can be challenging orthopaedic injuries to diagnose on radiographs, especially in the urgent care setting or if minimally displaced. Deep learning (DL) models have shown promise in automating fracture detection, potentially improving diagnostic accuracy and consistency. Subtle fractures along the complex hindfoot anatomy are difficult to appreciate on low resolution images, but high-resolution images make model training and inferencing costly. This study aims to assess the impact of radiograph image resolution on convolutional neural networks (CNN) performance in automatically detecting calcaneus fractures on x-rays and to inform future DL model training strategies on identifying clinical features on foot radiographs.

METHODS: This retrospective study included foot radiographs from a single hospital between 2015 and 2022 for a total of 1,775 x-ray series (551 fractures; 1,224 without) and was split into training (70%), validation (15%), and testing (15%). ImageNet pre-trained ResNet models were fine-tuned on the dataset. Three training strategies were evaluated: 1) single size: trained exclusively on 128x128, 256x256, 512x512, 640x640, or 900x900 radiographs (5 model sets); 2) curriculum learning: trained exclusively on 128x128 radiographs then exclusively on 256x256, then 512x512, then 640x640, and finally on 900x900 (5 model sets); and 3) multi-scale augmentation: trained on x-ray images resized along continuous dimensions between 128x128 to 900x900 (1 model set). Inference time and training time were compared using a Kruskal-Wallis test with post-hoc Dunn's test.

RESULTS: Multi-scale augmentation trained models achieved the highest average area under the Receiver Operating Characteristic (AUROC) curve of 0.938 [95% CI: 0.936 - 0.939] for a single model across image resolutions compared to the other strategies without prolonging training or inference time (Figure 1). Using the optimal model sets, curriculum learning had the highest sensitivity on in-distribution low-resolution images (85.4% to 90.1%) and on out-of-distribution high-resolution images (78.2% to 89.2%). However, curriculum learning models took significantly longer to train compared to the other two training strategies (11.8 [IQR: 11.1–16.4] hours; single size: 8.0 [5.6–10.4] hours; multi-scale: 6.7 [4.9–8.9] hours; P<.001).

DISCUSSION: While 512x512 images worked well for fracture identification, curriculum learning and multi-scale augmentation training strategies algorithmically improved model robustness towards different image resolutions without requiring additional annotated data. The models trained with curriculum learning achieved the best sensitivity at each image resolution out of the 3 training strategies and the performance continued to improve up to 900x900 resolution. However, inference would require several model weights for the range of image resolutions in comparison to the single model weight from multi-scale training.

SIGNIFICANCE/CLINICAL RELEVANCE: This study demonstrated that DL could identify calcaneus fractures on radiograph of a variety of image resolutions making it a useful technique for automatic foot radiograph interpretation. Moreover, we demonstrated how different model training strategies impact fracture detection performance and its importance in future development of DL musculoskeletal radiograph interpretation pipelines.

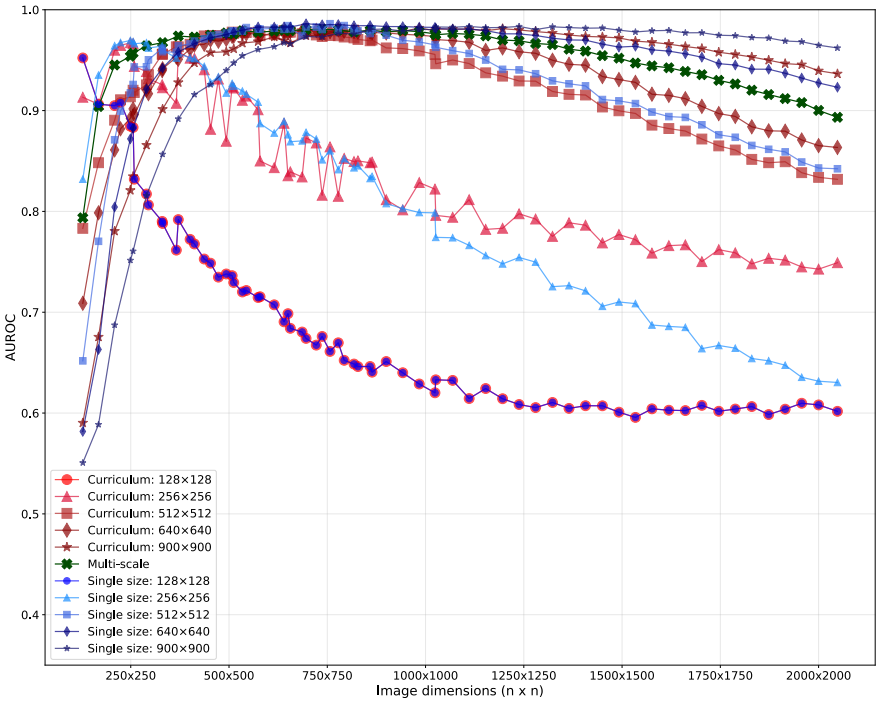

Figure 1: Average AUROC of the ResNet models trained under curriculum learning, multi-scale augmentation, and single size training strategies across image dimensions (128×128 to 2000x2000). Curriculum training started on 256x256 images and the weights initialized using the single size 128x128 models (represented by the blue dot with a red outline).