Development and External Validation of Automated Detection, Classification, and Localization of Ankle Fractures: Inside the Black Box of a Convolutional Neural Network

***Jasper Prijs, BS**; Zhibin Liao, PhD; Minh-Son To, MD; Johan Willem Verjans, MD; Paul C. Jutte, MD; Vincent Stirler, MD; Jakub Olczak, MD; Max Gordon, MD; Daniel Guss, MD; Christopher W. DiGiovanni, MD; Ruurd Jaarsma, FRACS; Frank Ijpma, MD; Job N. Doornberg, MD*
Flinders Medical Centre and University Medical Centre Groningen, Adelaide, AUSTRALIA

Purpose: Convolutional neural networks (CNNs) are increasingly being developed for automated fracture detection in orthopaedic trauma surgery. Studies to date, however, are limited to providing classification based on the entire image—and only produce heatmaps for approximate fracture localization instead of delineating exact fracture morphology. Therefore, we aimed to answer (1) what is the performance of a CNN that detects, classifies, localizes, and segments an ankle fracture; and (2) would this be externally valid?

Methods: The training set included 326 isolated fibula fractures and 423 non-fracture radiographs. The Detectron2 implementation of the Mask R-CNN was trained with labeled and annotated radiographs. The internal validation (or "test set") and external validation sets consisted of 300 and 334 radiographs, respectively. Consensus agreement between 3 experienced fellowship-trained trauma surgeons was defined as the ground truth label. Diagnostic accuracy and area under the receiver operating characteristic curve (AUC) were used to assess classification performance. The intersection over union (IoU) was used to quantify accuracy of the segmentation predictions by the CNN, where a value of 0.5 is generally considered an adequate segmentation.

Results: The final CNN was able to classify fibula fractures according to 4 classes (Danis-Weber A, B, C, and No Fracture) with AUC values ranging from 0.93 to 0.99. Diagnostic accuracy was 89% on the test set with average sensitivity of 89% and specificity of 96%. External validity was 89-90% accurate on a set of radiographs from a different hospital. Accuracies/AUCs observed were 100/0.99 for the "No Fracture" class, 92/0.99 for Weber B, 88/0.93 for Weber C, and 76/0.97 for Weber A. For the fracture bounding box prediction by the CNN, a mean IoU of 0.65 (standard deviation [SD] $\pm$ 0.16) was observed. The fracture segmentation predictions by the CNN resulted in a mean IoU of 0.47 (SD $\pm$ 0.17).

Conclusion: This study presents a look into the "black box" of CNNs (visually with many figures; however, these could not be attached for this abstract) and represents the first automated delineation (segmentation) of fracture lines on ankle radiographs. The AUC values presented in this paper indicate excellent discriminatory capability of the CNN and substantiate further study of CNNs in detecting and classifying ankle fractures.