Development and Validation of a Deep Learning Model to Identify Proximal Humerus Fracture Features on Radiographs

Maria Cordero Romero[1], D. Xiaofeng Yang[2], Hudson Smith[2], Sarah B. Floyd[1]
[1]Department of Public Health Sciences, Clemson University, Clemson, SC
[2]Department of Mathematical and Statistical Sciences, Clemson University, Clemson, SC

Disclosures: The authors report no disclosures.

INTRODUCTION: Proximal Humerus Fractures (PHF) are among the most common fractures in the elderly, yet clear guidance on how these fractures should be treated is lacking. PHF treatment is contingent on a myriad of factors including patient factors like age, function level, and preferences for care. Additionally, the complexity or features present in the fracture may dictate initial management decisions. However, identification of fracture features has historically been inconsistent across physicians. Advances in computer vision modeling provides an opportunity to develop novel tools to assist orthopaedic surgeons in their identification of important fracture features and assist in treatment decision-making. The objective of this work was to develop and assess the performance of a deep learning model in identifying the presence of important PHF features on X-ray.

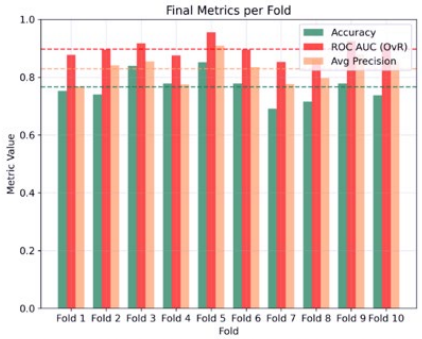
METHODS: A Delphi consensus method was previously conducted with orthopaedic surgeons to generate a comprehensive list of important PHF features for the model to identify. The Delphi process resulted in 8 features for modeling (topographical parts, displacement of parts, dislocation, head shaft angulation, head shaft translation, head split fracture, head impaction, and bone quality). The sample included patients with index PHF treated at a Level 1 Trauma Center in the Southeast United States. All X-ray images taken from the index visit through two weeks following the index visit were used to develop X-ray composites for each patient. X-ray composites were then labeled by orthopaedic surgeons in Labelbox, an online AI labeling platform. A total of 1,623 X-ray composite images were labeled by five orthopaedic surgeons to create the gold standard labeled dataset for model training. A Multiple Instance Learning (MIL) approach was implemented using a modified ResNet-50d backbone architecture with tanh attention pooling mechanism. The model processed variable-sized patient bags containing multiple X-ray views per patient. The tanh attention mechanism extends beyond simple dot-product attention by incorporating non-linear transformations using learnable matrices (V) and hyperbolic tangent activations, enabling the capture of more complex feature relationships between image instances. The feature extractor was initialized with pre-trained weights and frozen during training to prevent overfitting. Entropy regularization (weight=0.01) was applied to encourage peaked attention distributions, promoting focus on the most diagnostically relevant images within each patient bag. Training employed cross-entropy loss combined with entropy regularization, with the model handling variable bag sizes through custom batching procedures. The data was split into 10 partitions for training and testing, and performance metrics were computed based on the held-out partition. The performance across all splits was aggregated to estimate the average generalization error and the uncertainty in this error. This work was reviewed and approved by the IRB at the institution where the work was conducted.

RESULTS SECTION: The MIL model with tanh attention was evaluated on 1,623 composite image sets from patients with PHF. Modeling to date has focused on two of the 8 PHF features, head shaft angulation (HSA) and head shaft translation (HST). For HSA classification (Figure 1), 10-fold cross-validation achieved a patient-level accuracy of 76.63% ± 4.78% and ROC AUC of 89.75% ± 3.02%, with balanced accuracy of 73.88% ± 5.69% and average precision of 82.89% ± 4.56%. For HST detection (Figure 2), the model achieved superior performance with patient-level accuracy of 76.97% ± 3.67% and exceptional ROC AUC of 93.83% ± 1.18%. The model converged efficiently with average best epochs of 14.1 ± 7.2 for HSA and 9.1 ± 9.4 for HST.

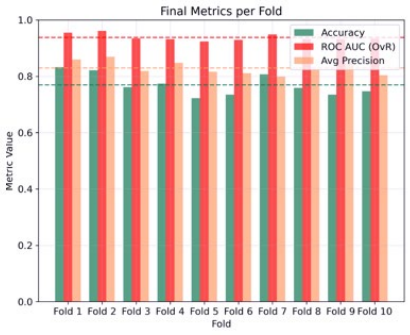
DISCUSSION: The expected outcome of this project is a validated deep learning model that can automate and standardize the identification and reporting of clinically relevant fracture features that can be widely disseminated across orthopaedic practice settings. The results of this work will facilitate the next step in enhancing our ability to diagnose a difficult shoulder injury, develop evidence-based treatment recommendations, and improve the quality of fracture care. Additionally, the availability of computer vision models to assist physicians with reading and classifying X-ray images has the potential to improve efficiencies in care delivery. Ongoing work includes the expansion of models to the remaining 6 fracture features.

SIGNIFICANCE/CLINICAL RELEVANCE: We present a model capable of identifying HSA and HST on PHFs. These findings represent a first step towards the development of a deep learning model to augment the diagnosis process of current surgeons in identifying PHF features. Consistent identification of fracture features can be paired with outcome data to develop treatment guidelines for patients with PHF.

IMAGES AND TABLES: Figure 1. Head-Shaft Angulation Performance Results; Figure 2. Head-Shaft Translation Performance Results




```
==================================================
PATIENT-LEVEL RESULTS SUMMARY
==================================================
Patient-Level Accuracy: 0.7663 ± 0.0478
Patient-Level Balanced Accuracy: 0.7388 ± 0.0569
Patient-Level AUC (OvR): 0.8975 ± 0.0302
Patient-Level Avg Precision: 0.8289 ± 0.0456
Patient-Level Test Loss: 0.6099 ± 0.0775
Average Best Epoch: 14.1 ± 7.2
```




```
==================================================
PATIENT-LEVEL RESULTS SUMMARY
==================================================
Patient-Level Accuracy: 0.7697 ± 0.0367
Patient-Level Balanced Accuracy: 0.7169 ± 0.0419
Patient-Level AUC (OvR): 0.9383 ± 0.0118
Patient-Level Avg Precision: 0.8299 ± 0.0233
Patient-Level Test Loss: 0.5585 ± 0.0623
Average Best Epoch: 9.1 ± 9.4
```