Development of an Artificial Intelligence Approach to Automatically Identify Reverse Shoulder Arthroplasty Implant Designs on Postoperative Radiographs

Linjun Yang, Austin Grove, Elizabeth Sun Kaji, Erick Marigi[1], Jonathan D Barlow[1], Joaquin Sanchez-Sotelo[1], John William Sperling[1]
[1]Mayo Clinic

INTRODUCTION: Since the introduction of the first commercially available reverse shoulder arthroplasty (RSA) implant, the number of different RSA implants produced by orthopedic companies has increased significantly. Radiographic identification of the surgical implant is particularly useful when considering revision surgery, as instruments and implants needed for the revision must be requested in advance. Although the operative report typically includes information about the components previously implanted, obtaining copies of the medical record is not always possible. Despite the availability of online libraries with radiographic examples of some prostheses, surgeons sometimes struggle to accurately identify implants on radiographs. Deep learning (DL) algorithms, commonly used in artificial intelligence computer vision, have demonstrated the capability to identify imaging features accurately and quickly from radiographic data. Therefore, the aim of this study was to develop and evaluate an automated DL tool to identify common RSA implants from anteroposterior or axillary radiographs automatically, saving time and providing surgeons with crucial information for revision surgery planning. If successful, this tool could also be used to process large numbers of images into registries or databases.

METHODS: The DICOM files of radiographs obtained after the implantation of 1542 RSA were retrieved and utilized in this study. We selected four manufacturers that are commonly used and represent design families of a classic Grammont style (Delta Xtend, DePuy, n=186), as well as more contemporary designs (Zimmer-Biomet Comprehensive, n=890; Stryker ReUnion n=291; Stryker Tornier Perform n=175). There were several radiographic projections obtained for each shoulder, for a total of 4,067 images that were labeled according to the 4 designs above. These images were randomly split into six equal-size folds at the patient level to avoid data leakage and also stratified by label to ensure that each fold had a similar number of radiographs for each design. Five-fold cross-validation was performed using the first five folds to train DL classification models to classify the correct manufacturer label. The five trained models were evaluated using the holdout fold (Figure 1) to find the best-performing model. The accuracy and F1 score, which is the harmonic mean of precision and sensitivity, were used to evaluate the classification performance.

RESULTS: The five DL classification models from the five-fold cross-validation achieved the average accuracy/F1 scores of 0.970/0.970, 0.974/0.973, 0.979/0.980, 0.973/0.974, and 0.971/0.971, respectively, when evaluated using the holdout test data fold. The best-performing model, which achieved the highest average accuracy/F1 score of 0.979/0.980, was further analyzed. The accuracy of this model on classifying Comprehensive, ReUnion, Delta Xtend, and Perform, were 0.990, 0.971, 0.990, and 0.964, respectively. The F1 score of this model on classifying Comprehensive, ReUnion, Delta Xtend, and Perform, were 0.985, 0.975, 0.990, and 0.978, respectively. Those metrics indicated excellent performance of the model to classify different manufacturers. It took the model 5.5 seconds to analyze 677 testing images.

DISCUSSION AND CONCLUSION: A fast, accurate, and automated DL tool was developed and validated to identify the manufacturer of the RSA implant from AP and axillary radiographs. This DL algorithm can facilitate surgical planning for revision surgery by identifying the manufacturer or design of the implant from the primary surgery. It can also be used to build registries and databases automatically.

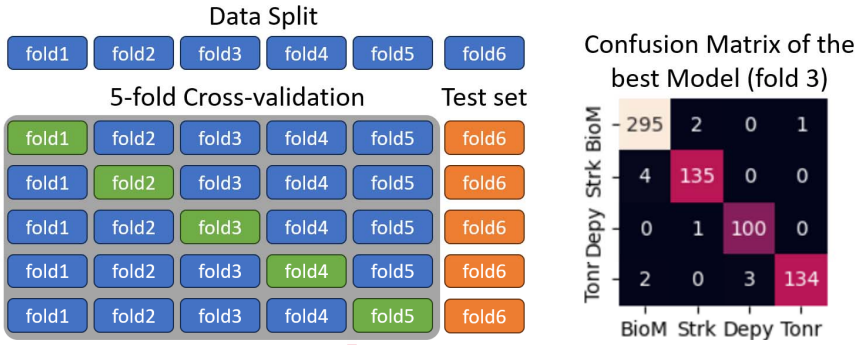

Figure 1. Left: Split of image data into 6 equal-size folds for deep learning (DL) model training and evaluation. The first five data folds were used for cross-validation, which resulted in five trained DL models. Each model was evaluated using radiographs from the fold 6, which helped determined the best-performing model (from fold 3); Right: confusion matrix that shows the classification performance of the best model on classifying the four manufacturers.