A Comparison of Convolutional Neural Networks versus Orthopaedic Physicians in Diagnosing Knee Osteoarthritis Severity

Sarah Lu, Michael Kevin Fei[1], Joseph Galal Elsissy, Brian Andrew Schneiderman[2]
[1]Creighton University School of Medicine, [2]Loma Linda University Medical Center

INTRODUCTION:

Convolutional neural networks are deep machine learning models used to analyze images. Potential applications of this emerging technology are vast, particularly in orthopaedics where the utilization of radiography and advanced imaging is foundational to the field. One such use may be in the diagnosis of degenerative knee osteoarthritis, as accurate staging may influence treatment. A convolution neural network regression model was recently created, which scored a knee X-ray with Kellgren and Lawrence (KL) Grading Scaling on a continuous spectrum. While there has been a recent focus on the introduction of new convolutional neural networks, very few have been evaluated for performance in comparison to practicing physicians. The primary aim of this study was to determine whether this regression neural network model can grade knee osteoarthritis radiographs as effectively as practicing orthopaedic surgeons.

METHODS:

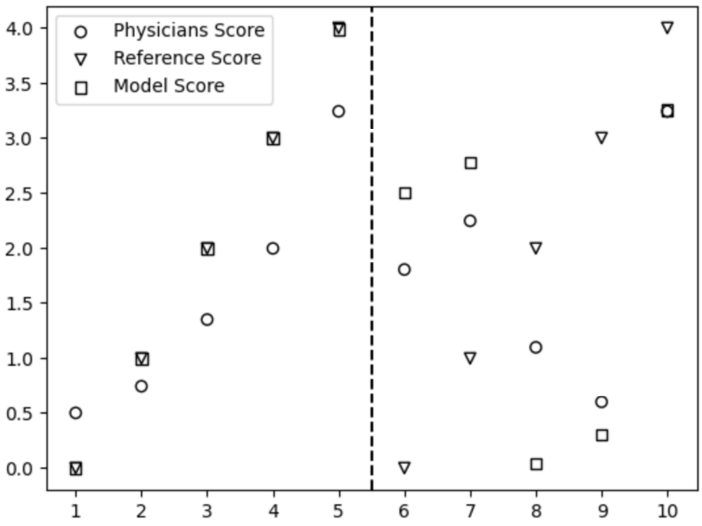
Pre-labeled radiographs from The Osteoarthritis Initiative were used as reference values to train the EfficietNet CNN architecture according to the Kellgren-Lawrence (KL) grading scale, using a 6604/826/830 train/validation/test split and achieving a 0.83 AUC score. The model produced continuous KL scores. Next, 10 sample images with 2 sample images from each KL grade were selected for orthopaedic physicians to evaluate. Of the images, 5 “good performing” images, images where the model and reference agreed, and 5 “poor performing,” images where the model predicted a value different than the reference, were selected for the survey. T-tests and mean absolute errors were calculated from the samples: reference, model, and physician values.

RESULTS:

Twenty orthopaedic_surgeons participated in this study, including 14 attending surgeons and 6 fellows subspecialty trained in adult reconstruction, sports medicine, or traumatology. No significant difference was noted in grading performance between physicians and the model ($p=0.351$). The mean absolute error for all 10 images was 1.025 between reference and physician values and 0.574 between model and physician values. For the poor-performing images, the mean absolute error between reference and physician values were 1.42 and 0.519. The difference in performance between the model and physician were compared to the difference in performance between the reference and physician. A statistically significant difference was found ($p=0.032$).

DISCUSSION AND CONCLUSION:

This study is one of the first to evaluate the performance of a convolutional neural network in comparison to practicing orthopaedic surgeons. It found that model and physician performance in determining radiographic KL grade were not significantly different. The low mean absolute error between model and physician values indicate that the model performed similarly to practicing clinicians. Considering the model used a regression machine learning model on a continuous spectrum rather than the discrete values of the KL scale, the low mean absolute error also suggests that the machine learning algorithm was predictive of the average aggregate physician score. For images that the machine learning model performed poorly, physician performance conformed more with the model rather than the reference. Overall, this novel study employed a convolutional neural network machine learning model to determine radiographic KL grade and found its performance to be comparable to practicing orthopaedic surgeons.


Physicians Score
Reference Score
Model Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10