Poster Session C

Session: (L01–L14) Late-Breaking Posters

L04: Performance of an Artificial Intelligence Model Compared to Multiple Human Experts in Scoring Synovitis Severity and Osteophyte Severity on Joint Ultrasound Images

Monday, November 18, 2024 10:30 AM – 12:30 PM Eastern Time

Location: Hall C

NON-CME IP

Late-Breaking Poster Presenter(s)

SJ

Søren Andreas Just, MD, PhD

Section of Rheumatology, Department of Medicine, Svendborg Hospital, Odense University Hospital
Svendborg, Denmark

Disclosure information not submitted.

Anders Weber[1], Mads Ammitzbøll Danielsen[2], Bill Aplin Frederiksen[3], Hilde Berner Hammer[4], Benjamin Schultz Overgaard[3], Lene Terslev[2], Thiusius Rajeeth Savarimuthu[5] and **Soren Andreas Just**[3], [1]ROPCA, Odense, Denmark, [2]Center for Rheumatology and Spine Disease, Rigshospitalet, Glostrup, Denmark, [3]Section of Rheumatology, Department of Medicine, Svendborg Sygehus OUH, Svendborg, Denmark, [4]Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway, Faculty of Medicine, University of Oslo, Oslo, Norway, [5]Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, Denmark

Background/Purpose: To evaluate the agreement of an artificial intelligence (AI) model designed to assess greyscale and Doppler synovitis severity and osteophyte severity in hand joints compared to human expert raters, using a consensus score as the gold standard.

Methods: Ultrasound images of metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and interphalangeal (IP) joints were collected from patients with hand pain. Rheumatologists, all EULAR-certified ultrasound instructors, scored the images for synovial hypertrophy (SH) (5 raters), Doppler activity (3 raters), and osteophyte severity (4 raters) on a scale from 0 to 3 using the Global OMERACT-EULAR Synovitis Score (GLOESS) and the corresponding osteophyte scoring system. The AI model was trained, validated, and tested on 7314 images. The disease classifications of the AI model were tested against the raters on 1280 ultrasound images to assess SH, 840 ultrasound videos to assess Doppler activity and 351 ultrasound images to assess osteophytes. The agreement with the consensus was calculated as the AI's average agreement with all raters. Performance metrics, including Cohen's Kappa, Percent Exact Agreement (PEA), Percent Close Agreement (PCA), sensitivity, specificity, Positive

Predictive Value (PPV), and Negative Predictive Value (NPV), were calculated with 95% confidence intervals (CI).

Results: As illustrated in Figure 1, the AI and human raters achieved comparable results across all metrics.

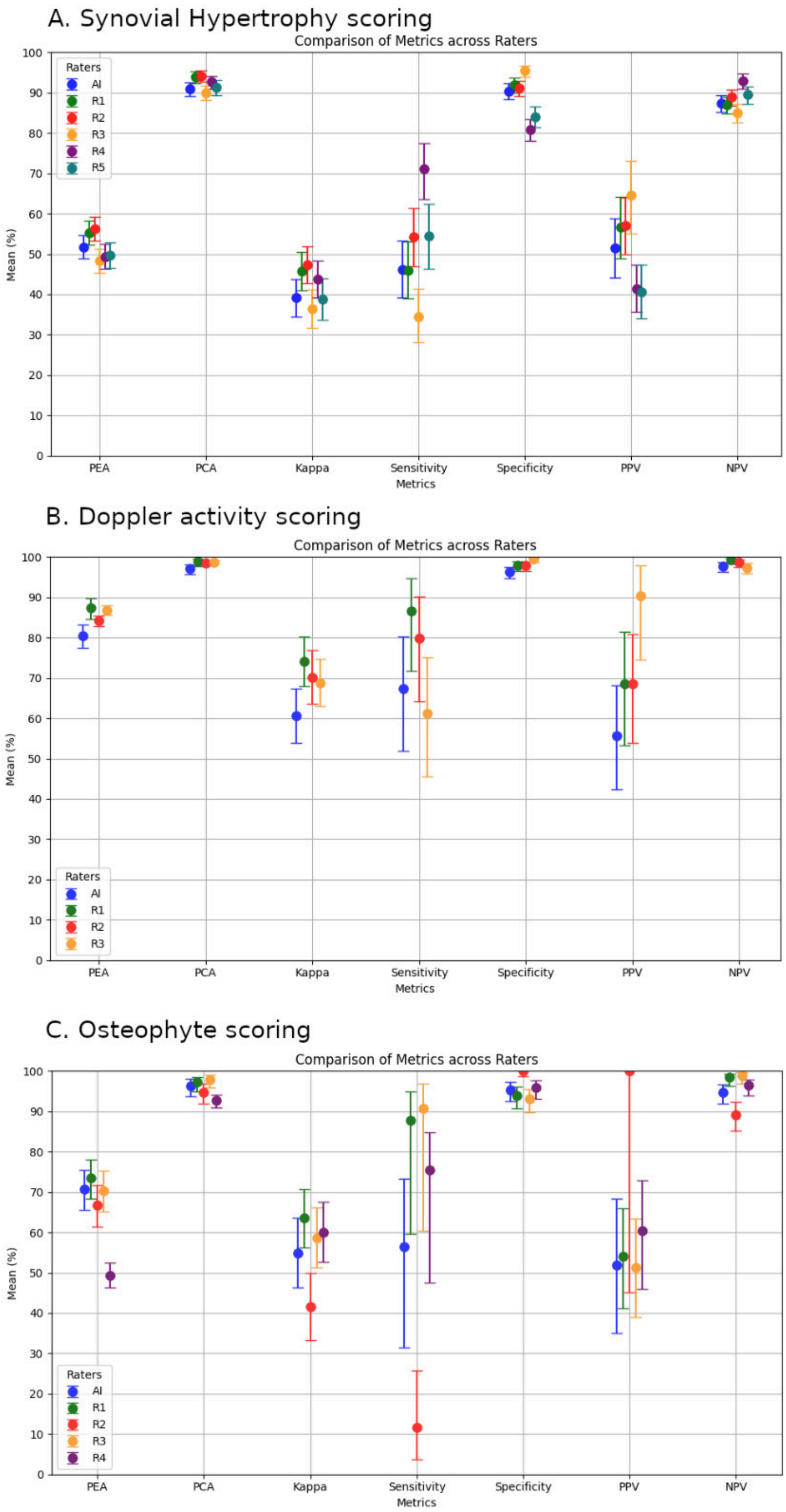
SH: The AI vs. consensus showed a Kappa of 0.39 (95% CI: 0.35–0.44), PEA of 51.77% (95% CI: 48.83–54.70%), PCA of 91.03% (95% CI: 89.21–92.63%), sensitivity of 46.19% (95% CI: 39.13–53.32%), and specificity of 90.43% (95% CI: 88.35–92.25%).

Doppler Activity: The AI vs. consensus had a Kappa of 0.61 (95% CI: 0.54–0.67), PEA of 80.49% (95% CI: 77.51–83.22%), PCA of 97.13% (95% CI: 95.69–98.18%), sensitivity of 67.31% (95% CI: 51.86–80.24%), and specificity of 96.29% (95% CI: 94.65–97.52%).

Osteophyte Grading: The AI vs. consensus showed a Kappa of 0.55 (95% CI: 0.46–0.63), PEA of 70.69% (95% CI: 65.57–75.45%), PCA of 96.28% (95% CI: 93.70–98.01%), sensitivity of 56.43% (95% CI: 31.56–73.36%), and specificity of 95.36% (95% CI: 92.44–97.36%).

These metrics, along with overlapping 95% confidence intervals depicted in Figure 1, indicate that the AI's performance is comparable to that of experienced human raters across all metrics.

Conclusion: The AI model performed at the level of expert human raters in assessing synovial hypertrophy, Doppler activity, and osteophyte severity in hand joints. This suggests that AI can be a reliable tool for evaluating joint ultrasound images, potentially aiding clinical decision-making by providing consistent and standardized assessments.



Performance metrics for synovial hypertrophy, Doppler activity, and osteophyte scoring for the AI and human raters, with 95% confidence intervals.

A. Weber: ROPCA ApS, 3; **M. Ammitzbøll Danielsen**: None; **B. Aplin Frederiksen**: None; **H. Berner Hammer**: None; **B. Schultz Overgaard**: None; **L. Terslev**: None; **T. Rajeeth Savarimuthu**: Ropca ApS, 8; **S. Just**: ROPCA, 2, 8.