Frederiksen *et al. BMC Rheumatology* (2026) 10:59
https://doi.org/10.1186/s41927-026-00654-8

BMC Rheumatology

RESEARCH

Open Access

Automated ultrasound with AI for osteophyte grading in hand osteoarthritis: comparison with expert rheumatologist assessment

Bill Aplin Frederiksen[1], Hilde Berner Hammer[2,3], Lene Terslev[4], Mads Ammitzbøll Danielsen[4], Thiusius Rajeeth Savarimuthu[5,6], Anders Bossel Holst Weber[6] and Søren Andreas Just[1,6*]

Abstract

Background The objective of this study was to characterise the agreement of the CE-certified automated robotic ultrasound system ARTHUR v.2.0, combined with the AI model DIANA v.2.0, for grading osteophytes in hand osteoarthritis (OA), using expert rheumatologist assessment as the reference standard.

Methods Thirty patients with hand OA underwent ultrasound of MCP, PIP, and DIP joints with ARTHUR v.2.0 and subsequently by an expert rheumatologist. Osteophytes were graded (0–3) using the OMERACT system. Agreement was assessed using weighted Cohen's kappa (κ), Percent Exact Agreement (PEA), Percent Close Agreement (PCA), sensitivity, and specificity. Comparisons were made against both the primary rheumatologist and an independent blinded external assessor (EA).

Results ARTHUR v.2.0 successfully scanned 703/840 joints (83.7%), with lower success in DIP joints. Agreement between ARTHUR+DIANA and the rheumatologist showed κ = 0.49, PEA 53.7%, and PCA 90.7%. Compared to the EA, the automated system showed κ = 0.46 and PCA 90.5%. Agreement between the rheumatologist and EA showed κ = 0.67 and PCA 97.5%. Binary agreement for disease presence (≥ Grade 2) for the automated system was 81.6% compared to the rheumatologist, although sensitivity was limited (36.5%).

Conclusions ARTHUR v.2.0 combined with DIANA v.2.0 achieved binary agreement comparable to expert–expert comparisons, although with limited sensitivity, and moderate agreement on the full 0–3 OMERACT scale. A decomposition analysis indicated that acquisition-related variability contributed substantially to the overall discrepancy. Refinement of AI threshold calibration and distal joint acquisition, and evaluation in larger and more diverse cohorts, is warranted to further improve sensitivity for osteophyte detection.

Keywords Osteoarthritis, Automated ultrasound, Artificial Intelligence (AI), Osteophyte grading, Robotic ultrasound

*Correspondence:
Søren Andreas Just
soeren.andreas.just@rsyd.dk
[1]Department of Medicine, Section of Rheumatology, Svendborg Hospital - Odense University Hospital, Baagøes Allé 15, Svendborg DK-5700, Denmark
[2]Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway
[3]Faculty of Medicine, University of Oslo, Oslo, Norway
[4]Center for Rheumatology and Spine Disease, Rigshospitalet, Glostrup, Denmark
[5]Maersk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, Denmark
[6]ROPCA ApS, Odense, Denmark

Background

Hand osteoarthritis (OA) is a highly prevalent rheumatic disease, affecting millions worldwide and contributing significantly to disability and healthcare costs [1]. It leads to chronic pain, stiffness, reduced hand function, and impaired quality of life, with substantial socioeconomic burdens due to healthcare utilisation and loss of productivity [1]. In the Netherlands, the annual societal cost of hand OA due to productivity loss has been estimated at €2452 per patient, while in the United States, the annual medical costs for patients with hand OA average $4274 per patient, highlighting the significant economic impact of the disease [2, 3].

Ultrasound can detect a range of pathological changes in hand OA, including osteophytes, synovial hypertrophy, cartilage thinning, and Doppler-detectable inflammation [4, 5]. Ultrasound has been shown to detect early structural changes before they are evident on radiographs and to diagnose severe OA with a sensitivity comparable to Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) [6–8].

Automated ultrasound systems as ARTHUR v2.0, combined with the AI-driven analysis tool DIANA v2.0, offer potential help in early hand OA diagnosis by providing standardized, reproducible imaging without requiring specialized training. We have previously published how DIANA was developed for assessing ultrasound images, including osteophyte scoring [9–11]. DIANA v.2.0 has recently been demonstrated to perform at the level of multiple expert rheumatologists in scoring both arthritis-related synovial changes and osteophytes [12]. Additionally, using ARTHUR is well tolerated by patients with rheumatic diseases [13]. In a recent study of patients with rheumatoid arthritis, we demonstrated that ARTHUR 2.0 + DIANA 2.0 achieves intra-robot agreement comparable to that of rheumatology experts and encouraging agreement when compared to a rheumatologist scanning the same patients [14]. The aim of this proof-of-concept feasibility study was to characterise the agreement of ARTHUR v.2.0 in combination with DIANA v.2.0 for grading osteophytes in hand OA compared to an expert rheumatologist and a blinded external assessor (EA).

Methods

Study design and participants

This study was conducted at the Section of Rheumatology, Svendborg Hospital, Odense University Hospital, as part of a protocol that also included a parallel cohort of patients with rheumatoid arthritis recently reported elsewhere [14]. Both studies were approved under the same ethical approval by the Danish Medical Research Ethics Committees (De Videnskabsetiske Medicinske Komitéer, VMK, Case number: 2400033), and all patients provided written informed consent prior to participation.

A total of 31 patients with established hand osteoarthritis were screened, of whom one declined participation. Patients were eligible if they fulfilled the American College of Rheumatology (ACR) classification criteria for hand OA and were ≥ 18 years of age [15]. All participants fulfilled the ACR classification criteria for hand osteoarthritis and had radiographic evidence of OA changes in the interphalangeal joints. Patients with CMC-1 involvement were not excluded, but none had isolated rhizarthrosis. Patients with severe joint deformities were excluded because such anatomical alterations interfere with ARTHUR's current ability to perform standardized, reproducible sweeps across joint surfaces. This was also done to ensure alignment with the OMERACT ultrasound scoring system, which has been validated primarily in non-deformed joints [4]. In total, 30 patients were enrolled, typically corresponding to mild-to-moderate structural disease without severe deformity.

Imaging procedures

Each participant underwent ultrasound of the metacarpophalangeal (MCP), proximal interphalangeal (PIP), and distal interphalangeal (DIP) joints bilaterally, first using the automated ARTHUR v.2.0 system (Fig. 1A) and subsequently by an experienced rheumatologist (Fig. 1B). For both automated and manual acquisitions, imaging was performed in the dorsal midline longitudinal plane according to OMERACT reference definitions. The rheumatologist followed the same slice orientation as ARTHUR to ensure methodological comparability and was not permitted to use paramedian or oblique planes.

ARTHUR v.2.0 is a CE-certified (MDR class IIa) robotic ultrasound platform capable of performing standardized sweeps across small joints. It incorporates real-time image analysis to select optimal acquisition frames, which are then transferred to DIANA v.2.0 for scoring. DIANA v.2.0 is a convolutional neural network-based AI model (also CE-certified, MDR class IIa) trained on more than 10,000 hand joint ultrasound images annotated by expert rheumatologists using OMERACT criteria [9–12]. The training dataset comprised images acquired in previous clinical trials in Denmark, primarily on General Electric Logiq 9 and 10 scanners. Importantly, no images from the present study were included in the development or validation of DIANA. DIANA v.2.0 not only provides a numeric OMERACT osteophyte score but also generates a segmentation overlay of joint structures (bone, synovium, tendon and osteophyte), allowing the assessor to understand which anatomical regions contributed to the score (Fig. 1D).

Both ARTHUR and the rheumatologist used the same ultrasound platform (GE Logiq E10, revision 3) with an ML6-15 linear probe.

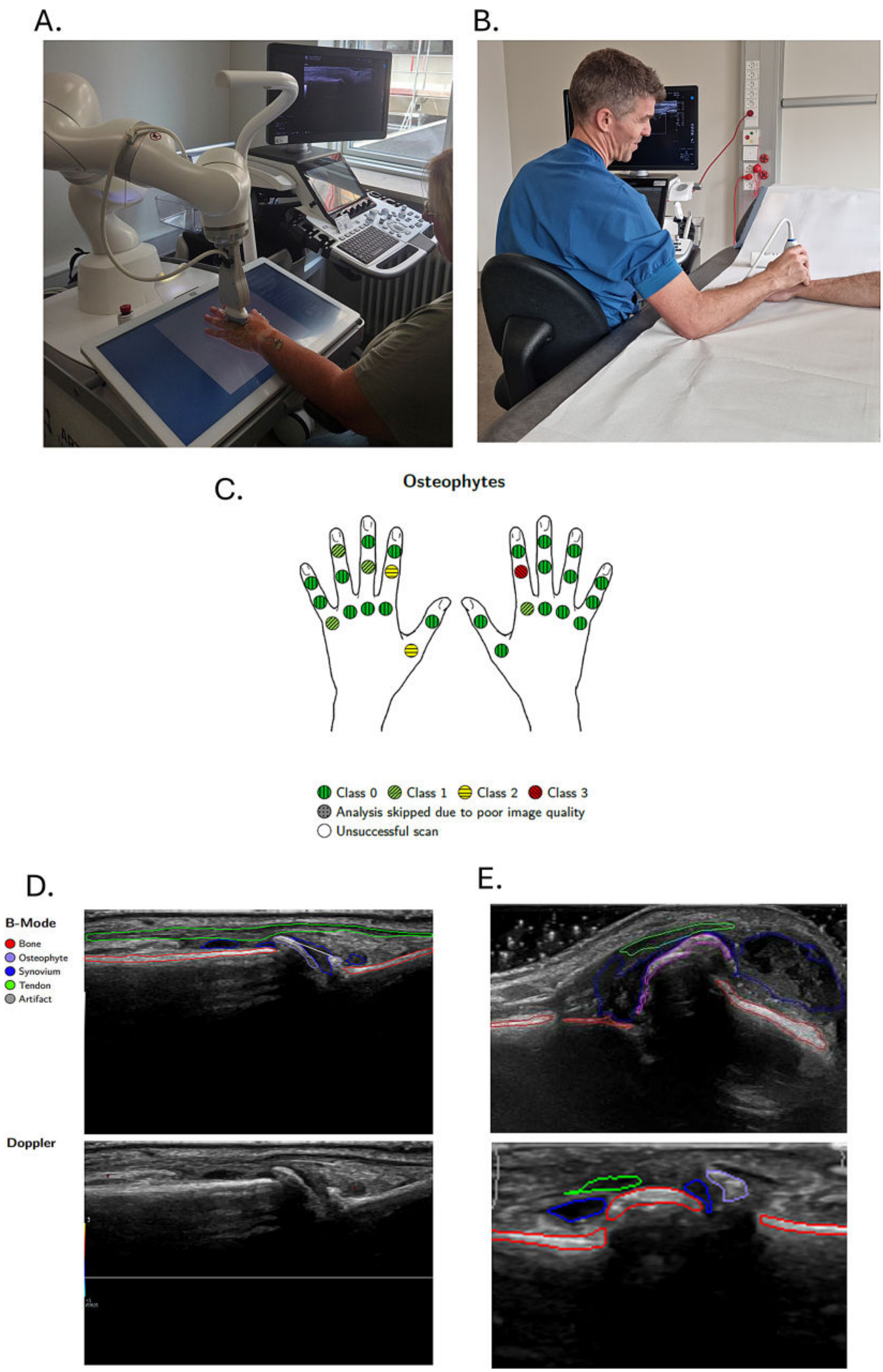

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Automated and expert ultrasound assessment of hand osteoarthritis. Automated and expert ultrasound assessment of hand osteoarthritis. **A**. Patient undergoing automated scanning with the robotic ultrasound system ARTHUR v.2.0. **B**. Ultrasound examination of a hand joint performed by the expert rheumatologist. **C**. Example of the PDF report overview generated by DIANA v.2.0, showing the results of automated osteophyte grading. **D**. Examples from the DIANA v.2.0 report with segmentation overlays highlighting anatomical structures and osteophyte regions. **E**. Representative concordant and discordant examples: the upper image (MCP joint) shows concordant grading between the rheumatologist and DIANA (both = grade 3), whereas the lower image (DIP joint) shows a discordant case where the rheumatologist and external assessor scored grade 3 while DIANA assigned grade 1

The rheumatologist (BF) had more than 10 years of experience in musculoskeletal ultrasound and serves as an ultrasound teacher and lecturer at national and EULAR levels. This rater's assessment of manually acquired images was considered the reference standard for the primary analysis. The rheumatologist graded the stored static images after acquisition, not during live scanning, to ensure methodological consistency with both the external assessor and DIANA, which evaluated the same still images (Fig. 2).

Independent external assessor

All stored images from both ARTHUR and the rheumatologist were independently reviewed by an external assessor (EA), a rheumatologist with more than 20 years of musculoskeletal ultrasound experience and active involvement in EULAR training programs (HBH) (Fig. 2). The EA acted as an independent comparator to assess reproducibility and inter-observer variability, whereas the rheumatologist served as the reference standard for the primary analysis.

The EA graded ARTHUR-acquired and rheumatologist-acquired images separately, ensuring that each comparison reflected the performance of a specific acquisition modality rather than a mixture of image sources (Fig. 2). This design allowed isolation of acquisition-related differences (robotic vs. manual) from interpretation-related differences (AI vs. human scoring). The EA graded osteophytes according to the OMERACT scoring system (0–3) [4]. Only joints deemed of sufficient quality by the EA were included in the analysis; images rejected due to inadequate quality were excluded. The EA's evaluation was used for supplementary comparisons to examine consistency across raters. The rheumatologist and the EA scored the ultrasound images using the CVAT (Computer Vision Annotation Tool) platform [16]. In this study, CVAT was used as a structured manual scoring interface, allowing side-by-side viewing of static images. All image grading was performed manually according to the OMERACT scoring system, and no automated or AI-assisted functions were used within the platform. All scoring was performed under blinded conditions. The external assessor (EA) was fully blinded to the image source, scoring all ARTHUR-acquired and rheumatologist-acquired images through the CVAT platform in randomized order without identifiers indicating the acquisition modality. The EA had no access to DIANA's scores or the rheumatologist's grades. Similarly, the primary rheumatologist (BF) graded only the manually acquired images without access to DIANA's output, as DIANA exclusively processes ARTHUR-acquired images.

DIANA v.2.0 evaluated ARTHUR's images and created a report showing why a given score was obtained (Fig. 1C and D).

Outcomes and definitions

Osteophytes were scored on the OMERACT 0–3 scale, where grade 0 indicates absence and grade 3 represents severe changes. For binary analysis, scores of 0 or 1 were classified as "normal," while scores of 2 or 3 were classified as "pathological" (3). For comparisons involving the EA, the reference score for each joint was defined as the maximum of the EA's grades on rheumatologist-acquired and ARTHUR-acquired images. This approach provides a conservative reference that captures pathology detected on either acquisition modality.

At the patient level, osteophyte presence was defined as at least one joint with a score ≥ 2, while absence was defined as all joints scoring ≤ 1.

Statistical analysis

The primary comparison was between ARTHUR+DIANA and the expert rheumatologist, with the latter's manual grading serving as the reference. Secondary comparisons were performed with the EA. Agreement was quantified at the joint level using Percent Exact Agreement (PEA), Percent Close Agreement (PCA; defined as a difference of ± 1 score), binary agreement, sensitivity, specificity, and quadratic weighted Cohen's kappa (Fleiss–Cohen weights). Sensitivity and specificity were calculated using the rheumatologist's grading as the reference standard, classifying scores of 0–1 as 'normal' and 2–3 as 'pathological' in accordance with the binary definition. Kappa values were interpreted according to Landis and Koch [17], where values < 0.20 represent poor agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, and ≥ 0.81 very good.

To explore whether automated–human discrepancies were related to inherent grading difficulty, a subgroup analysis was performed. Joints were stratified into a concordant subset, defined as joints where the rheumatologist and EA assigned the same osteophyte grade, and a discordant subset, where their grades differed. Agreement between DIANA and the rheumatologist (PEA,

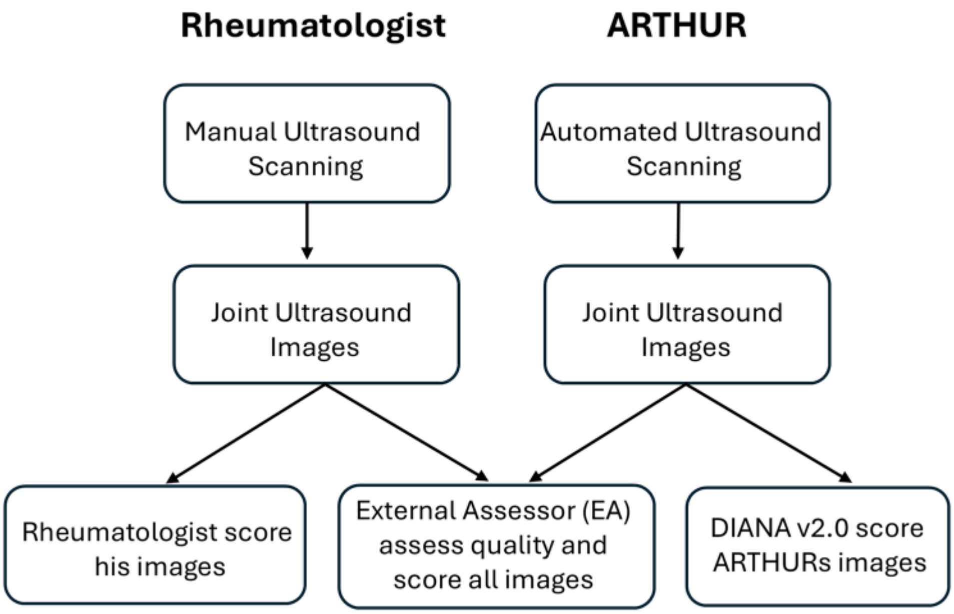

Fig. 2 Study workflow. Schematic representation of image acquisition by ARTHUR v.2.0 and manual acquisition, followed by scoring via DIANA v.2.0, the primary rheumatologist, and the external assessor (EA)

PCA, binary agreement, and quadratic weighted κ) was then calculated separately for each subset.

Differences in osteophyte grade distribution between joints with successful and failed ARTHUR scans were assessed using Fisher's exact test for the proportion of joints with moderate-to-severe osteophytes (≥ grade 2), the Mann–Whitney U test for mean osteophyte scores, and the chi-squared test for the full grade distribution. The same statistical tests were applied to compare osteophyte grade distribution between men and women.

To account for the effect of prevalence imbalance on kappa, the prevalence-adjusted bias-adjusted kappa (PABAK) was calculated for the binary comparison. To explore the relative contributions of acquisition-related and grading-related variability, a decomposition analysis was performed. Grading variability was assessed by comparing raters who evaluated the same image set: DIANA versus the EA on ARTHUR-acquired images, and the rheumatologist versus the EA on rheumatologist-acquired images. Acquisition variability was assessed by comparing the EA's scores across the two image sets (ARTHUR-acquired versus rheumatologist-acquired), as this comparison holds the rater constant and varies only the acquisition modality.

At the patient level, comparisons were made by calculating the percentage of patients classified as having at least one joint with osteophytes by each assessor. Each patient contributed multiple joints to the dataset, and analyses were conducted at the joint level. We acknowledge that this approach does not account for within-patient clustering, which may slightly underestimate uncertainty, but was chosen to maintain comparability with previous OMERACT ultrasound reliability studies. All analyses were performed in Stata version 18.0.

Table 1 Baseline characteristics of the study population

Characteristic	Value
Number of Patients	30
Age (years, mean ± SD)	72 ± 9
Gender (Male/Female)	16/14

Results

Study population

A total of 30 patients with established OA were included in the study. ARTHUR v2.0 was configured to scan 28 joints per patient. Table 1 presents the baseline characteristics of the study population.

Image quality assessment

All images were assessed by the EA, resulting in ARTHUR v2.0 successfully scanning 703 of 840 joints (83.7%). The highest scores for MCP 2–4 and PIP 2–4 (Table 2). The rheumatologist scanned 840 joints one time with a success rate of 100%. Table 2 provides an overview of ARTHUR's scanning success rate across different joint types.

Table 2 shows that ARTHUR achieved high scanning success rates for MCP and PIP joints (85–95%), whereas performance was lower for DIP joints (63%). Image rejection by EA was also highest for DIP joints (13%), suggesting greater technical challenges in imaging these joints. Supplementary Table S1 shows that joints with failed ARTHUR scans had a significantly higher proportion of moderate-to-severe osteophytes (≥ grade 2: 35.3% vs.

Table 2 Scanning success rate by joint type

Joint Type	MCP	PIP	DIP	All Joints
Left-hand Joint Images obtained by ARTHUR (%)	95	95	73	**94**
Right Hand Joint Images obtained by ARTHUR (%)	91	94	72	**92**
Images declined by EA (%)	7	10	13	**10**
Total Successful joint images obtained, after EA evaluation (%)	86	85	63	**84**

MCP, metacarpophalangeal joints; PIP, proximal interphalangeal joints; DIP, Distal interphalangeal joints

26.2%; $p = 0.034$), although the difference in mean OA scores did not reach statistical significance (1.2 vs. 1.1; $p = 0.064$). This indicates that joints excluded due to scan failure had higher disease severity in the binary domain.

Of note, ARTHUR acquired images for 782 joints in total, all of which were scored by DIANA. Of these, the EA accepted 703 (89.9%) as sufficient quality for human scoring. The primary comparison between ARTHUR+DIANA and the rheumatologist (Table 4) includes all 781 joints where both DIANA and the rheumatologist provided a score, as this comparison does not depend on EA image quality acceptance. Comparisons involving the EA (Supplementary Tables S3 and S4) are restricted to the 703 EA-accepted joints.

ARTHUR and DIANA versus the rheumatologist and the EA

In Table 3, the agreement on a joint level between ARTHUR+DIANA, the rheumatologist and the EA.

ARTHUR+DIANA achieved binary agreement comparable to the rheumatologist–EA agreement. On the OMERACT 0–3 scale, agreement was moderate with the rheumatologist (κ = 0.49) and good between the rheumatologist and EA (κ = 0.67). Indicating inherent interobserver variability against which the automated system's performance should be interpreted. Compared with the rheumatologist, ARTHUR+DIANA showed a sensitivity of 74.2% and specificity of 82.6%, while values versus the EA were 37.5% and 97.4%, respectively. Joint-specific metrics in Supplementary Table S2 show that concordance was highest for MCP and PIP joints and lowest for DIP joints. Table 4 presents the osteophyte OMERACT class distribution at the joint level for ARTHUR+DIANA and the rheumatologist.

Table 4 Agreement between ARTHUR+DIANA and rheumatologist osteophyte grades

ARTHUR+DIANA ↓ / Rheumatologist →	Grade 0	Grade 1	Grade 2	Grade 3	Total
Grade 0	313	41	7	1	362 (46%)
Grade 1	151	63	15	1	230 (29%)
Grade 2	50	56	36	20	162 (21%)
Grade 3	4	10	6	7	27 (4%)
Total	518 (66%)	170 (22%)	64 (8%)	29 (4%)	781 (100%)

Percentages represent the proportion of joints within each grading category relative to the total joints graded ($n = 781$).$N = 781$ joints where both DIANA and the rheumatologist provided a score. This includes joints where ARTHUR acquired an image that was scored by DIANA but subsequently not accepted by the EA for human scoring (see Table 2)

Table 4 presents the agreement matrix between ARTHUR+DIANA and the rheumatologist. Exact agreement was achieved in 419 joints across all grades. The automated system assigned grade 0 more frequently (66.3%) than the rheumatologist (46.4%). Discrepancies were primarily confined to the lower severity range, where ARTHUR+DIANA assigned a higher grade than the expert in 207 instances. Extreme disagreements (grade 0 vs. 3) occurred in five joints (< 1%). Supplementary Tables S3 and S4 provide agreement matrices for the external assessor (EA). The total joint counts for EA comparisons were 779 for ARTHUR images and 835 for rheumatologist images.

A subgroup analysis was performed to assess whether discrepancies between DIANA and the rheumatologist were related to inherent grading difficulty. Of 778 joints scored by all three assessors, the rheumatologist and EA agreed in 415 joints (53.3%). In this concordant subset, DIANA achieved a binary agreement of 86.0% (95% CI: 82.7–89.4) and a quadratic weighted κ of 0.51 (0.42–0.59). In the discordant subset (363 joints), binary agreement was 76.6% (72.2–80.9) and κ was 0.46 (0.38–0.54). Full results are presented in Supplementary Table S7. In the discordant subset, the EA scored higher than the rheumatologist in 77.1% of cases.

Binary 2 × 2 classification tables for all three comparisons are presented in Supplementary Table S8. Compared with the rheumatologist, DIANA achieved a sensitivity of 36.5% and a specificity of 95.9% for detecting pathological

Table 3 Agreement between ARTHUR v2.0 + DIANA v2.0, the rheumatologist and the external assessor (EA) for osteophyte scoring

Metric	ARTHUR+DIANA vs. Rheumatologist	ARTHUR+DIANA vs. EA	Rheumatologist vs. EA
Kappa (+/- 95% CI)	0.49 (0.42–0.54)	0.46 (0.40–0.51)	0.67 (0.63–0.70)
PEA (+/- 95% CI)	53.65 (50.08–57.19)	41.34 (37.85–44.88)	52.93 (49.48–56.36)
PCA (+/- 95% CI)	90.65 (88.39–92.60)	90.50 (88.22–92.47)	97.49 (96.18–98.44)
Binary (+/- 95% CI)	81.56 (78.66–84.22)	81.39 (78.47–84.06)	86.35 (83.83–88.60)

PEA: Percent Exact Agreement, PCA: Percent Close Agreement osteophytes (≥ grade 2). The rheumatologist achieved a sensitivity of 70.0% and specificity of 92.6% compared with the EA. Against the EA, DIANA showed a sensitivity of 37.5% and specificity of 97.4%. These figures indicate that while overall binary agreement was comparable across comparisons, DIANA's agreement is primarily driven by high specificity, with limited sensitivity.

Table 5 shows the osteophyte class distribution across the three assessors. The external assessor (EA) assigned grade 0 to 27.3% of joints and grade 3 to 9.9%. ARTHUR+DIANA classified 66.3% of joints as grade 0, compared to 46.4% by the rheumatologist. The EA assigned higher grades more frequently than both the rheumatologist and ARTHUR+DIANA. Table 5 also presents the osteophyte class distribution stratified by sex. There were no significant differences in the proportion of pathological osteophytes (≥ grade 2) between men and women (26.1% vs. 29.4%; $p = 0.314$), although the full grade distribution differed ($p = 0.030$), driven by a higher proportion of grade 3 osteophytes in women. DIANA's agreement with the rheumatologist was similar for both sexes (binary agreement: men 81.7%, women 81.4%).

The binary 2 × 2 classification table for ARTHUR+DIANA versus the rheumatologist is presented in Table 6. The standard binary kappa was 0.39, while the PABAK was 0.63, indicating that the unadjusted kappa was substantially deflated by the high prevalence of normal joints. Of the 120 joints classified as normal by DIANA but pathological by the rheumatologist, the majority ($n = 56$) were scored as grade 2 by the rheumatologist, reflecting DIANA's conservative threshold at the grade 1/2 boundary.

A decomposition analysis was performed to explore the relative contributions of acquisition and grading variability (Supplementary Table S9). On identical ARTHUR-acquired images, DIANA achieved a quadratic weighted κ of 0.57 against the EA ($N = 703$), compared with κ = 0.70 for the rheumatologist versus the EA on manually acquired images ($N = 815$). The EA's own agreement across the two image sets was κ = 0.74 ($N = 683$, PEA 61.3%, binary agreement 90.2%), quantifying the contribution of acquisition-related variability. The combined comparison between DIANA and the rheumatologist (κ = 0.49) was lower than either component alone.

At the patient level, defining disease presence as at least one joint with an osteophyte grade 2, ARTHUR+DIANA showed 86.7% (95% CI: 69.3 to 96.2) agreement with the EA. Both rheumatologist and the EA identified osteophytes in 96.7% of the cohort, while ARTHUR+DIANA identified presence in 90.0%. Detailed patient-level agreement and distributions are provided in Supplementary Tables S5 and S6.

Table 5 Osteophyte class distribution, overall and by sex

Grade	0 (%)	1 (%)	2 (%)	3 (%)	≥ 2 (%)
Rheumatologist (*n* = 839 joints)	46.4	29.5	20.7	3.5	24.2
Men (448 joints, 16 patients)	48.2	28.1	19.6	4.0	23.7
Women (391 joints, 14 patients)	44.8	29.2	23.5	2.6	26.1
ARTHUR+DIANA (*n* = 782 joints)	66.3	21.8	8.2	3.7	11.9
Men (420 joints, 16 patients)	67.1	22.1	7.6	3.1	10.7
Women (362 joints, 14 patients)	65.2	21.5	8.8	4.4	13.3
EA (*n* = 836 joints)	27.3	45.1	17.7	9.9	27.6
Men (445 joints, 16 patients)	26.7	47.2	18.9	7.2	26.1
Women (391 joints, 14 patients)	27.9	42.7	16.4	13.0	29.4

EA: External Assessor, n denotes the number of joints scored by each assessor. The number of joints differs between assessors because ARTHUR did not acquire images of all joints (see Table 2). The EA column represents the maximum osteophyte grade assigned by the EA across rheumatologist-acquired and ARTHUR-acquired images for each joint

Table 6 Binary 2 × 2 classification table (ARTHUR+DIANA vs. Rheumatologist):

	Rheumatologist Normal (0–1)	**Rheumatologist Pathological (2–3)**	**Total**
DIANA Normal (0–1)	568	120	688
DIANA Pathological (2–3)	24	69	93
Total	592	189	781

Discussion

This study evaluated the performance of the automated ultrasound system ARTHUR v.2.0 combined with the AI model DIANA v.2.0 for osteophyte grading in hand OA. The system achieved binary agreement comparable to that observed between two expert rheumatologists (81.6% vs. 86.3%), although with limited sensitivity for detecting pathological osteophytes (36.5% vs. 70.0%), driven primarily by high specificity (95.9%) and a conservative scoring threshold.

The agreement on the full 0–3 OMERACT scale was moderate. Notably, inter-observer agreement between the two human experts was good but not perfect (κ = 0.67), underscoring the inherent subjectivity of osteophyte grading and providing important context for evaluating automated performance.

This moderate level of agreement must be viewed in the context of inherent inter-observer variability; for instance, the external assessor (EA) assigned grade 1 to 45.1% of joints, whereas the primary rheumatologist only used this grade in 29.5% of cases. The EA's higher frequency of grade 1 scores likely reflects a lower threshold for interpreting minor cortical irregularities as borderline pathology, which the rheumatologist more frequently classified as either grade 0 (normal) or grade 2 (definite osteophyte). Such variability underscores the challenge of achieving perfect concordance between automated systems and human experts, particularly regarding low-grade structural changes.

The subgroup analysis showed that DIANA achieved higher agreement with the rheumatologist in joints where both human assessors agreed compared with joints where they disagreed (Supplementary Table S7). This suggests that a substantial proportion of automated–human discrepancies arise in cases with inherent grading ambiguity, where even experienced assessors disagree, rather than reflecting systematic model error. The decomposition analysis demonstrated that both acquisition-related and grading-related variability contribute to the overall discrepancy between the automated system and the rheumatologist (Supplementary Table S9). When evaluated on identical images, DIANA's grading agreement with the EA (κ = 0.57) was lower than the human–human grading agreement (κ = 0.70), indicating room for AI calibration improvement. However, the EA's own agreement across the two image sets (κ = 0.74) confirms that acquisition differences play a substantial role. These findings are consistent with a recent multi-expert comparison in which DIANA achieved osteophyte grading agreement (κ = 0.56, PEA 70.69%) within the range of interrater variability among four expert rheumatologists (κ range 0.42–0.64) when evaluated on manually acquired images [12]. This supports the interpretation that the lower agreement in the present study is largely attributable to acquisition-related variability. Having the rheumatologist independently grade ARTHUR-acquired images would have allowed a more direct within-study comparison and is recommended for future studies.

ARTHUR v.2.0 achieved high scanning success in MCP and PIP joints, but success was lower in DIP joints (63%), where technical challenges such as small joint size and curvature remain. This led to a modest underestimation of osteophyte prevalence in the distal joints. Importantly, joints with failed ARTHUR scans had a significantly higher proportion of pathological osteophytes than successfully scanned joints (35.3% vs. 26.2%; $p = 0.034$), indicating selection bias. The exclusion of more severely affected joints may lead to underestimation of osteophyte prevalence and could modestly inflate binary agreement metrics. This should be considered when interpreting the reported agreement figures, and future hardware improvements targeting distal joints with more advanced disease will be important to mitigate this bias.

A consistent observation was DIANA's conservative threshold, frequently assigning grade 0 in cases where experts identified low-grade pathology. The PABAK for the binary comparison (0.63) was substantially higher than the standard binary kappa (0.39), confirming that the high prevalence of normal joints deflates the unadjusted kappa statistic. This conservative scoring behaviour is primarily evident at the grade 0/1 boundary, where DIANA assigned grade 0 to 151 joints scored as grade 1 by the rheumatologist. Future calibration of DIANA's decision thresholds, potentially through adjustment of the training dataset composition or post-hoc score recalibration, may improve concordance with human experts for low-grade pathology without compromising specificity. Representative examples in Fig. 1E highlight that discrepancies often arise in technically challenging or lower-contrast regions, supporting the need for further refinement of the AI model for subtle structural changes.

From a diagnostic perspective, the EULAR evidence-based recommendations for the diagnosis of hand osteoarthritis define hand OA primarily by characteristic symptoms and radiographic findings [18]. As this study did not include healthy controls, it cannot address diagnostic accuracy. Nonetheless, automated systems may become valuable in research for developing reproducible datasets, early structural assessment, or as quantitative outcome measures in trials. Furthermore, automated scanning could improve accessibility in settings where expert sonographers are limited.

Several limitations must be addressed. Using a principal rheumatologist for both acquisition and scoring mirrors current clinical practice but may restrict generalizability. The study included only two human raters, precluding construction of a majority consensus reference standard. Future studies incorporating three or more independent assessors would allow majority-vote grading, potentially providing a more robust reference for evaluating automated systems. However, separate data comparing DIANA directly with multiple experts on manually acquired images demonstrated expert-level AI performance [12]. Suggesting that current discrepancies primarily reflect acquisition-related variation. The cohort contained a higher proportion of men (53%) than typically observed in epidemiological studies of hand OA. Although the full osteophyte grade distribution differed between sexes, with women showing a higher proportion of grade 3 osteophytes, there were no significant differences in the proportion of pathological osteophytes (≥ 2), and DIANA's agreement with the rheumatologist was similar for both sexes. Nevertheless, future studies should aim for sex-balanced cohorts to ensure generalisability. At the patient level, defining presence as at least one joint with a score 2 is clinically intuitive but may simplify within-patient variation. Future studies could explore cumulative or mean scores to provide a more nuanced representation of severity. Notably, ARTHUR v.2.0 does not currently scan the CMC-1 joint; thus, rhizarthrosis is not captured. Finally, the lack of clinical correlates, such as pain or functional measures, means imaging findings were not anchored to patient-reported outcomes. The study cohort represents mild-to-moderate structural hand OA without severe deformity, as patients with severe joint deformities were excluded because such anatomical alterations interfere with ARTHUR's current ability to perform standardised sweeps. Furthermore, the absence of healthy controls precludes evaluation of diagnostic accuracy. Together, these exclusions introduce spectrum bias that may affect both agreement metrics and the clinical applicability of the findings. The reported agreement figures should therefore be interpreted as representative of the mild-to-moderate OA population and may not generalise to the full spectrum of disease. Future hardware refinement is required to accommodate anatomical variations and joint deformities, and studies including healthy controls will be needed to establish diagnostic performance.

Although both ARTHUR and DIANA are CE-certified (MDR class IIa) and currently in clinical use, this study represents the first systematic characterisation of automated osteophyte grading agreement in hand OA, and the findings identify specific areas for AI calibration improvement. Current EULAR guidelines focus on symptoms and radiography rather than ultrasound, underscoring that the present system is primarily intended to support research and standardized acquisition [18]. Beyond hand OA, the osteophyte scoring capability of DIANA addresses a broader clinical need. In patients with established inflammatory joint diseases such as rheumatoid arthritis, concurrent degenerative changes may contribute to joint pain, and distinguishing inflammatory from degenerative pathology has implications for treatment decisions. DIANA's ability to provide standardised OMERACT-based scoring of both synovial changes and osteophytes within a single automated assessment is therefore relevant across both inflammatory and degenerative joint disease. The intended clinical population for ARTHUR + DIANA comprises patients with mild-to-moderate hand OA or inflammatory arthritis without severe deformity. In research, the system may facilitate standardised outcome assessment in clinical trials and longitudinal monitoring. In clinical practice, it could support structural assessment in settings where expert sonographers are not readily available, although it is not currently intended as a standalone diagnostic tool or as a replacement for expert assessment in advanced disease.

In conclusion, the CE-certified automated system ARTHUR v.2.0 combined with DIANA v.2.0 achieved binary agreement comparable to that between expert rheumatologists, although with limited sensitivity, and moderate agreement on the 0–3 scale. A decomposition analysis indicated that acquisition-related variability contributed substantially to the overall discrepancy, consistent with recent evidence that DIANA achieves expert-level osteophyte grading on manually acquired images. Refinement of AI threshold calibration and distal joint acquisition, and evaluation in larger and more diverse cohorts, is warranted to further improve sensitivity for osteophyte detection.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41927-026-00654-8.

Supplementary Material 1

Acknowledgements

Thank you to all patients participating in this study and all personnel at the Section of Rheumatology, Svendborg Hospital for helping.

Author contributions

SJ, LT, MAD and BAF designed the study. BAF and HBH contributed to data acquisition and clinical assessment. SJ, TS and ABHW developed AI models. ABHW performed the statistical analysis, and SJ drafted the manuscript. All authors read and approved the final manuscript.

Funding

Open access funding provided by University of Southern Denmark. The author(s) declare financial support was received for the research, authorship, and/or publication of the article. SJ is supported by a grant from the Region of Southern Denmark (21/17499).

Data availability

The data underlying this study cannot be shared publicly or deposited in a repository, as this was not part of the written consent given by the patients. According to Danish law and EU General Data Protection Regulation (GDPR), secondary sharing would require renewed patient consent and specific ethical approval. Enquiries regarding data access may be directed to the corresponding author (Søren Andreas Just, soeren.andreas.just@rsyd.dk).

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki at the Section of Rheumatology, Svendborg Hospital, Odense University Hospital, as part of a protocol that also included a parallel cohort of patients with rheumatoid arthritis recently reported elsewhere [14]. Both studies were approved under the same ethical approval by the Danish Medical Research Ethics Committees (De Videnskabsetiske Medicinske Komitéer, VMK, Case number: 2400033), and all patients provided written informed consent prior to participation.

Consent for publication

Not applicable.

Declaration of generative AI in scientific writing

During the preparation of this work, the author(s) used ChatGPT to improve readability and language in parts of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Competing interests

SJ and TS are cofounders of Ropca Aps, developing AI and producing the automated ultrasound scanning system called ARTHUR. ARTHUR's AI cannot currently assess osteophyte severity. ABHW is a full-time employee of Ropca Aps.

Received: 13 January 2026 / Accepted: 12 May 2026
Published online: 21 May 2026

References

1. Kloppenburg M, Kwok WY. Hand osteoarthritis–a heterogeneous disorder. Nat Rev Rheumatol. 2011;8(1):22–31.

2. Terpstra SES, Van De Stadt LA, Boonen A, Damman W, Rosendaal FR, Kloppenburg M. Hand osteoarthritis is associated with limitations in paid and unpaid work participation and related societal costs: the HOSTAS cohort. RMD Open. 2022;8(2):e002367.
3. Yelin E, Weinstein S, King T. The burden of musculoskeletal diseases in the United States. Semin Arthritis Rheum. 2016;46(3):259–60.
4. Hammer HB, Iagnocco A, Mathiessen A, Filippucci E, Gandjbakhch F, Kortekaas MC, Moller I, Naredo E, Wakefield RJ, Aegerter P, et al. Global ultrasound assessment of structural lesions in osteoarthritis: a reliability study by the OMERACT ultrasonography group on scoring cartilage and osteophytes in finger joints. Ann Rheum Dis. 2016;75(2):402–7.
5. Favero M, Belluzzi E, Ortolan A, Lorenzin M, Oliviero F, Doria A, Scanzello CR, Ramonda R. Erosive hand osteoarthritis: latest findings and outlook. Nat Rev Rheumatol. 2022;18(3):171–83.
6. Mathiessen A, Haugen IK, Slatkowsky-Christensen B, Boyesen P, Kvien TK, Hammer HB. Ultrasonographic assessment of osteophytes in 127 patients with hand osteoarthritis: exploring reliability and associations with MRI, radiographs and clinical joint findings. Ann Rheum Dis. 2013;72(1):51–6.
7. Mathiessen A, Cimmino MA, Hammer HB, Haugen IK, Iagnocco A, Conaghan PG. Imaging of osteoarthritis (OA): What is new? Best Pract Res Clin Rheumatol. 2016;30(4):653–69.
8. Husic R, Finzel S, Stradner MH, Dreu M, Hofmeister A, Beham-Schmid C, Graninger WB, Fessler J, Dejaco C. Ultrasound in osteoarthritis of the hand: a comparison to computed tomography and histology. Rheumatology (Oxford). 2022;61(Si):Si73–80.
9. Andersen JKH, Pedersen JS, Laursen MS, Holtz K, Grauslund J, Savarimuthu TR, Just SA. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. RMD Open. 2019;5(1):e000891.
10. Christensen ABH, Just SA, Andersen JKH, Savarimuthu TR. Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients. Ann Rheum Dis. 2020.
11. Overgaard BS, Christensen ABH, Terslev L, Savarimuthu TR, Just SA. Artificial intelligence model for segmentation and severity scoring of osteophytes in hand osteoarthritis on ultrasound images. Front Med (Lausanne). 2024;11:1297088.
12. Weber ABH, Terslev L, Ammitzbøll-Danielsen M, Frederiksen BA, Hammer HB, Overgaard BS, Savarimuthu TR, Just SA. Performance of an artificial intelligence model compared with multiple human experts in scoring synovitis and osteophyte severity on joint ultrasound images. EULAR Rheumatol Open. 2026;2(1):274–82.
13. Frederiksen BA, Schousboe M, Terslev L, Iversen N, Lindegaard H, Savarimuthu TR, Just SA. Ultrasound joint examination by an automated system versus by a rheumatologist: from a patient perspective. Adv Rheumatol. 2022;62(1).
14. Frederiksen BA, Hammer HB, Terslev L, Ammitzbøll-Danielsen M, Savarimuthu TR, Weber ABH, Just SA. Automated ultrasound system ARTHUR V.2.0 with AI analysis DIANA V.2.0 matches expert rheumatologist in hand joint assessment of rheumatoid arthritis patients. RMD Open. 2025;11(3).
15. Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K, Brown C, Cooke TD, Daniel W, Gray R, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hand. Arthr Rhuem. 1990;33(11):1601–10.
16. Computer Vision Annotation Tool. (CVAT). https://github.com/opencv/cvat.
17. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977;33(2):363–74.
18. Zhang W, Doherty M, Leeb BF, Alekseeva L, Arden NK, Bijlsma JW, Dincer F, Dziedzic K, Hauselmann HJ, Kaklamanis P, et al. EULAR evidence-based recommendations for the diagnosis of hand osteoarthritis: report of a task force of ESCISIT. Ann Rheum Dis. 2009;68(1):8–17.

Publisher's note