Title

Utilizing a Digital Swarm Intelligence Platform to Improve Consensus Among Radiologists and Exploring Its Applications

Permalink
https://escholarship.org/uc/item/8r15x39v

Journal
Journal of Digital Imaging, 36(2)

ISSN
0897-1889

Authors
Shah, Rutwik
Astuto Arouche Nunes, Bruno
Gleason, Tyler
et al.

Publication Date
2023-04-01

DOI
10.1007/s10278-022-00662-3

Peer reviewed

Journal of Digital Imaging (2023) 36:401–413
https://doi.org/10.1007/s10278-022-00662-3

Utilizing a Digital Swarm Intelligence Platform to Improve Consensus Among Radiologists and Exploring Its Applications

Rutwik Shah[1,2] · Bruno Astuto Arouche Nunes[1,2] · Tyler Gleason[1] · Will Fletcher[1] · Justin Banaga[1] · Kevin Sweetwood[1] · Allen Ye[1] · Rina Patel[1] · Kevin McGill[1] · Thomas Link[1] · Jason Crane[1,2] · Valentina Pedoia[1,2] · Sharmila Majumdar[1,2]

Received: 21 October 2021 / Revised: 17 April 2022 / Accepted: 23 May 2022 / Published online: 22 November 2022

Abstract
Radiologists today play a central role in making diagnostic decisions and labeling images for training and benchmarking artificial intelligence (AI) algorithms. A key concern is low inter-reader reliability (IRR) seen between experts when interpreting challenging cases. While team-based decisions are known to outperform individual decisions, inter-personal biases often creep up in group interactions which limit nondominant participants from expressing true opinions. To overcome the dual problems of low consensus and interpersonal bias, we explored a solution modeled on bee swarms. Two separate cohorts, three board-certified radiologists, (cohort 1), and five radiology residents (cohort 2) collaborated on a digital swarm platform in real time and in a blinded fashion, grading meniscal lesions on knee MR exams. These consensus votes were benchmarked against clinical (arthroscopy) and radiological (senior-most radiologist) standards of reference using Cohen's kappa. The IRR of the consensus votes was then compared to the IRR of the majority and most confident votes of the two cohorts. IRR was also calculated for predictions from a meniscal lesion detecting AI algorithm. The attending cohort saw an improvement of 23% in IRR of swarm votes ($k = 0.34$) over majority vote ($k = 0.11$). Similar improvement of 23% in IRR ($k = 0.25$) in 3-resident swarm votes over majority vote ($k = 0.02$) was observed. The 5-resident swarm had an even higher improvement of 30% in IRR ($k = 0.37$) over majority vote ($k = 0.07$). The swarm consensus votes outperformed individual and majority vote decision in both the radiologists and resident cohorts. The attending and resident swarms also outperformed predictions from a state-of-the-art AI algorithm.

Keywords Swarm intelligence · Inter-rater reliability · Artificial intelligence · Consensus decisions · Workflow tools

Introduction

Consensus among radiologists is key for accurate disease diagnosis, patient care, and avoiding inadvertent medical errors [1]. Guidelines from the National Academy of Medicine recommend a team-based diagnosis, considered superior to individual interpretation [2]. Obtaining high inter-rater reliability among experts can be challenging when interpreting complex multifactorial diseases and grading lesions on multiclass scales. The phenomenon of variable inter-rater reliability has been widely documented across imaging subspecialities [3–7] and can result in both missed diagnoses and limit appropriate medical intervention at the right time [8] (Fig. 1).

Radiologists also perform an important role in training and benchmarking machine learning models. They classify and grade diseases, annotate lesions, and segment anatomical volumes on images [9, 10]. Opinion of the radiologists is often considered as "ground truth" for training models and against which its perfor/mance is measured.

Given that annotation tasks can be time-consuming, another approach is to have amateur labeling professionals (non-clinicians) annotate bulk of the images, with radiologists arbitrating discordant cases and performing a quality check of the dataset. However, the use of nonexperts is

Rutwik Shah and Bruno Astuto contributed equally and are co-first authors.

✉ Rutwik Shah
rutwik.shah@ucsf.edu

[1] Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA

[2] Center for Intelligent Imaging, University of California San Francisco, San Francisco, CA, USA

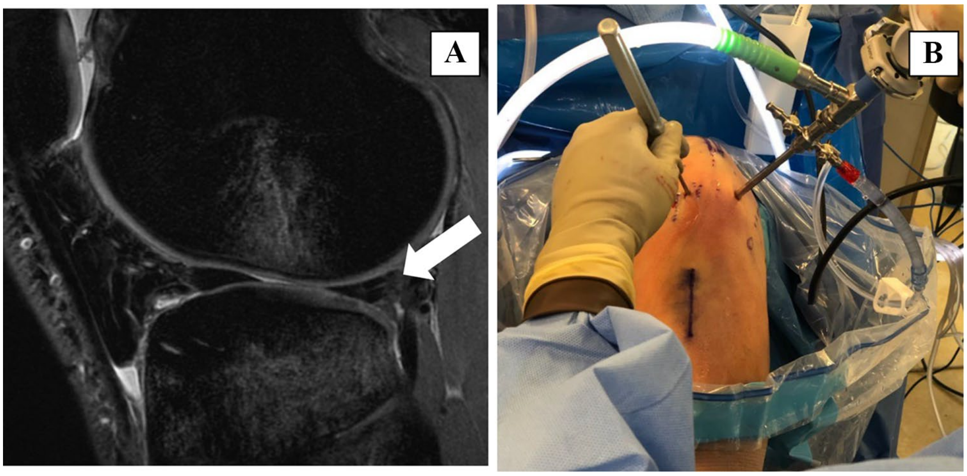

Fig. 1 **A** Sagittal sequence of a knee MR exam evaluated by multiple subspeciality trained musculoskeletal radiologists (arrow pointing to the ambiguous meniscal lesion) had discordant impressions of the presence and grade of lesions. **B** Swarm platform was used to derive consensus for the location of lesions, which matched with the arthroscopic findings considered as a standard of reference

fraught with risks and can create noisy labels [11, 12] or outright errors [13] which is consequential in high stakes artificial intelligence (AI) systems such as in medicine [14]. Numerous technical methods have been developed to mitigate the effects of label noise. These include techniques for label cleaning and denoising [15, 16], modifying loss functions [17, 18], or data re-weighting [19–21]. However, none of these methods fully mitigate the underlying cause of the noisy labels, which originate from interpersonal subjectivity at the time of label creation.

In both the approaches of expert and amateur data labeling, there is an assumption that the supervising radiologists being the experts provide true value but this fails to factor in the disagreement observed between multiple experts themselves.

Some common methods used to decide the consensus answer in medicine include the use of majority vote [22, 23], most confident vote [24], arbitration [25], and the Delphi technique [26, 27]. In this study, we investigate a novel technique called swarm intelligence, to improve consensus among expert participants. Inspired from observations made in birds and insects [28–30], swarm intelligence is a method to find the optimal answer in a group of multiple autonomous agents, who collaborate in real time. This concept has found applications in fields ranging from economic forecasting [31], robotics [31] to imaging AI [32].

Related Work and Key Concepts

Collective intelligence or wisdom of the crowds is defined as an emergent property of a quasi-independent multiagent system, where aggregated responses from the various agents outperforms individual responses [33]. This was perhaps best demonstrated by Galton's experiment demonstrating a crowd's average estimate of an ox's weight exceeding the best individual guess [34]. Multiple studies have demonstrated the phenomenon of collective intelligence and the various factors affecting it [35]. Individual conviction [36], level of expertise [37], cognitive diversity [38], personality traits [39], and social interaction [40] can all impact decision-making in groups. We describe key concepts of the team-based decision process in Table 1, relevant for understanding our study design.

Swarm intelligence (SI) is a specialized form of collective intelligence used to improve group decision-making in a wide range of biological species, from swarming bees and schooling fish to flocking birds. In recent years, a technology called artificial swarm intelligence (ASI) has been developed to enable similar benefits in networked human groups [41, 42]. A software platform called swarm was used in this study to enable networked human agents to make assessments by working together using the ASI technology. The software is designed to connect human agents with two distinguishing features; it requires *real-time participation* of all agents, and it has a *closed-loop feedback system* which updates and informs the agents of the combined group intent at each subsequent time step. It thus captures the dynamics of individual conviction, collaboration, negotiation, and opinion switching and is not simply a post hoc majority or an average vote analysis.

The primary aim of our study was to examine the effect of *synchronous*, blinded *nonsocial* interaction among clinical *experts* at different levels of expertise (radiologists, radiology residents), on a *specific task* (evaluation of meniscal lesion on knee MR) while answering a *fixed questionnaire*, and measure its effect on inter-rater reliability. Our secondary aim was to examine the effect of the number of participants (swarm size) in improving inter-rater reliability.

Methods

Radiographic and Clinical Dataset

The present study was conducted using previously acquired knee MRIs and corresponding clinical notes of 36 subjects

Table 1 Key concepts in team-based decision-making. Swarm intelligence requires real-time collaboration of all participants, with constant feedback of the group intent. Our study was designed to also be asocial to prevent any interpersonal bias

Key concepts	Options in team-based decision-making
Time of participation	Agents can participate in the prescribed activity asynchronously and then have results calculated post hoc, e.g., majority vote or average vote tabulation. Or agents can participate synchronously, where all participants answer questions at the same time, without exception. This is a key feature of the digital swarm platform
Expertise	Participating agents can all be domain experts (e.g., radiologists, radiology residents trained in specialized image interpretation) or nonexperts who may not possess specialized expertise relevant to the task at hand
Scope of task	The scope of the task for answering each question can be broad including multiple tasks (review images, clinical notes, and lab reports) or narrow and include a single task (image review) only
Questionnaire	The set of questions asked to the agents can be fixed and consistent for each item or can be adaptive based on previous responses, as seen in the Delphi technique
Communication	Communication can be either social or nonsocial. Social interaction allows agents to assess other's interests and preferences and also influence each other while performing the task at hand. This can lead to various interpersonal biases which can negatively impact overall results In contrast, nonsocial interaction allows agents to know group intent while being blinded to the identity, preferences, and level of expertise of other participants

enrolled for a longitudinal research study [43]. Subjects were recruited and scanned at one of three medical centers (UCSF, Mayo Clinic, Hospital for Special Surgery), to ensure patient data diversity. All subjects underwent arthroscopic evaluation and repair of the affected knee by an orthopedic surgeon, who recorded findings in the various compartments (meniscus, cartilage, bone, and ligaments) for lesions.

Distributions of patient demographics were $age = 42.79 \pm 14.75$ years, $BMI = 24.28 \pm 3.22$ kg/m2, and 64%/36% male/female. Study subjects were recruited with $age > 18$ years and exclusion criteria being concurrent use of any investigational drug, fracture or surgical intervention in the study knee, and any contraindications to MR. All subjects signed written informed consent approved by the Committee on Human Research of the home institution. The study was approved by the Institution Review Board.

Study Participants (Radiologists and Radiology Residents) and Task

Two cohorts of readers were recruited to evaluate the knee scans at multiple timepoints (Fig. 2). All readers examined only the sagittal CUBE sequence on the institutional Picture Archiving and Communication System (PACS). They were asked to answer the same question for each exam: "Select the regions of the meniscus where a lesion is observed," where a lesion was defined as Whole Organ Magnetic Resonance Imaging Score (WORMS) > 0 [44]. The six possible answer choices given were (1) none, (2–5) any one of the four meniscal horns (anterior and posterior; medial and lateral horns) compartments, or(6) more than one compartment (Table 2).

Cohort 1 included 3 board-certified musculoskeletal radiology attendings (averaging 19 of experience, range 4-–28) who read the MRI scans at two timepoints. First, at baseline, they independently graded the scan individually, also giving a self-reported confidence score for their reads (scale: 1 to 10). After a 15-day washout period, all 36 exams were reassessed by the attendings, while participating simultaneously in a swarm session (Unanimous AI, San Francisco), in real time.

Cohort 2 included 5 radiology residents (PG year 3–5). Similar to the attendings, they too first graded the scans independently at baseline with self-reported confidence scores. After a 15-day washout period, all 36 scans were reassessed by all 5 residents for a second time while participating simultaneously in a swarm session. After another 15-day washout period, 3 of the 5 residents (partial cohort 2) reassessed the 36 scans for a third time while participating in a second swarm session. This was done to measure the effect of swarm size on the inter-rater reliability.

Swarm Platform

To obtain the consensus answer of our participating radiologists and trainees, we utilized swarm platform (Unanimous AI, San Francisco), a platform which is modeled on the decision-making process of honeybees [45]. The platform allows multiple remotely located participants to collaborate in a blinded fashion over the Internet, in real time.

The platform consists of 2 key components: (1) a web-based application and (2) a cloud-based server that runs the proprietary swarm algorithm. Participants log into an online swarm session, using a standard web browser, and answer questions on the platform's hexagonal graphical user interface (GUI). The GUI captures real-time inputs from the full set of participants and provides immediate feedback based on the output generated from the swarm algorithm, essentially creating a closed-loop feedback system (Fig. 3).

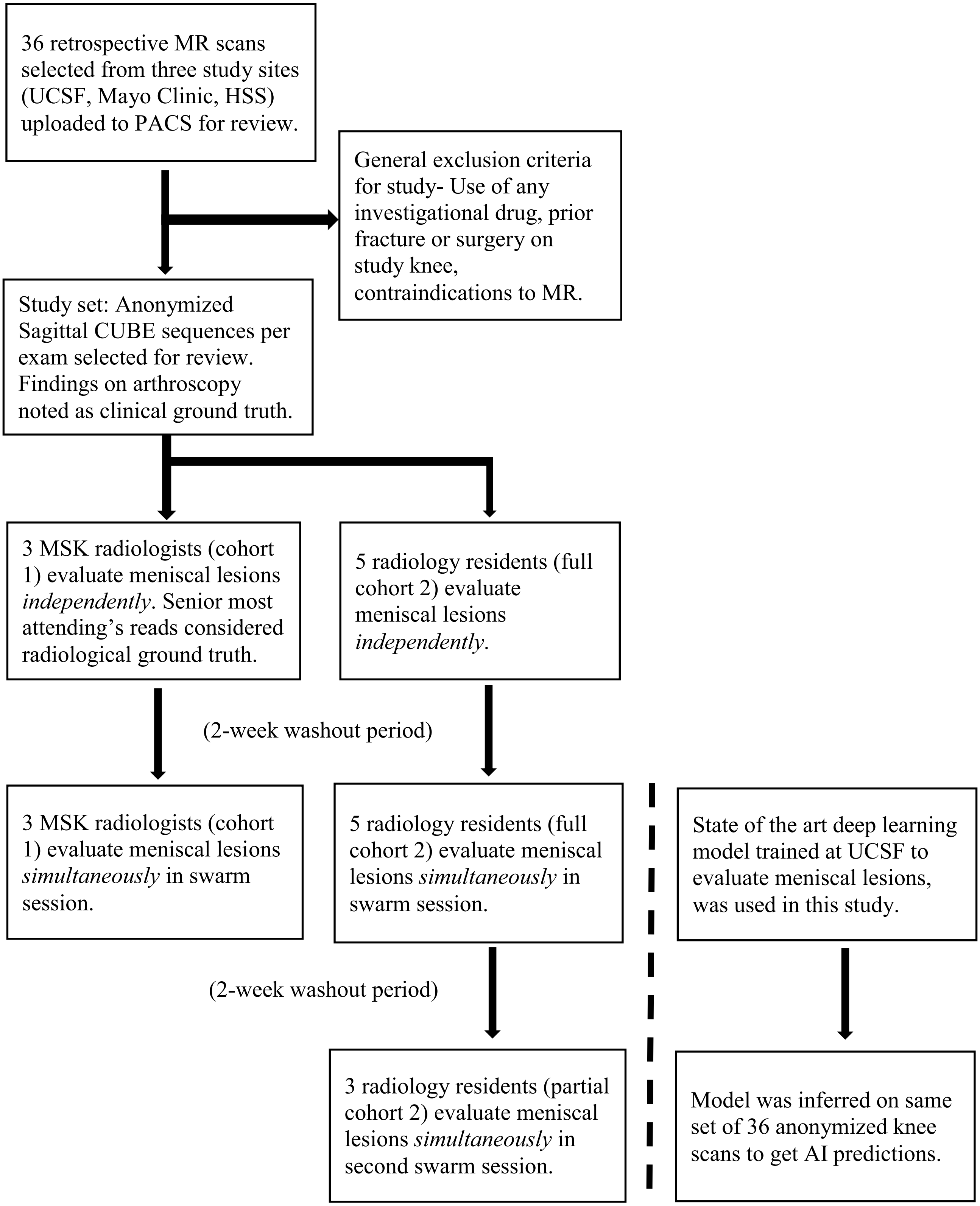
36 retrospective MR scans selected from three study sites (UCSF, Mayo Clinic, HSS) uploaded to PACS for review.
General exclusion criteria for study- Use of any investigational drug, prior fracture or surgery on study knee, contraindications to MR.
Study set: Anonymized Sagittal CUBE sequences per exam selected for review. Findings on arthroscopy noted as clinical ground truth.
3 MSK radiologists (cohort 1) evaluate meniscal lesions *independently*. Senior most attending's reads considered radiological ground truth.
5 radiology residents (full cohort 2) evaluate meniscal lesions *independently*.
(2-week washout period)
3 MSK radiologists (cohort 1) evaluate meniscal lesions *simultaneously* in swarm session.
5 radiology residents (full cohort 2) evaluate meniscal lesions *simultaneously* in swarm session.
State of the art deep learning model trained at UCSF to evaluate meniscal lesions, was used in this study.
(2-week washout period)
3 radiology residents (partial cohort 2) evaluate meniscal lesions *simultaneously* in second swarm session.
Model was inferred on same set of 36 anonymized knee scans to get AI predictions.

Fig. 2 Flowchart of various steps in the study. In total, 36 anonymized knee scans (sagittal CUBE sequences) were reviewed by a cohort of three MSK-trained radiologists and another cohort of five radiology residents, independently at first and then in swarm sessions. A deep learning model trained to evaluate meniscal lesions also inferred the same of 36 knee scans to obtain AI predictions for comparison

Table 2 Question and option choices to capture participant responses during swarm sessions

Question: Select the regions of the meniscus where a lesion is observed
Option 1: None
Option 2: Anterior horn of the medial meniscus
Option 3: Posterior horn of the medial meniscus
Option 4: Anterior horn of the lateral meniscus
Option 5: Posterior horn of the lateral meniscus
Option 6: More than one region

Using this system, both the cohorts answered questions in real time by collaboratively moving a graphical puck to select among a set of answer options. Each participant provided input by moving a graphical magnet to pull on the puck, thereby imparting their personal intent on the collective system. The preference is recorded as a continuous stream of inputs rather than just as a static vote. The conviction of each individual participant is indicated by distance between their magnets and the puck (strong versus weak pull). The net pull of all participants on the moves the puck in that direction, until a consensus is reached on one of the answer choices. The output of the collective answer is therefore also updated on the GUI in real time, as observed by the changing trajectory of the puck during an active swarm session. Because all users adjust their intent continuously in real time, the puck moves based on the complex interactions among all participants, empowering the group converge in synchrony.

Meanwhile, the swarm algorithm evaluates each user's intent at each instant by tracking the direction and strength of the pull of their magnets while comparing it with other participants. This is then used to (i) compute the consensus answer at each time step based on collective preferences and (ii) to provide instantaneous feedback to participants in the form of an updated puck trajectory, allowing them to stay with or switch their original answer choice, given the

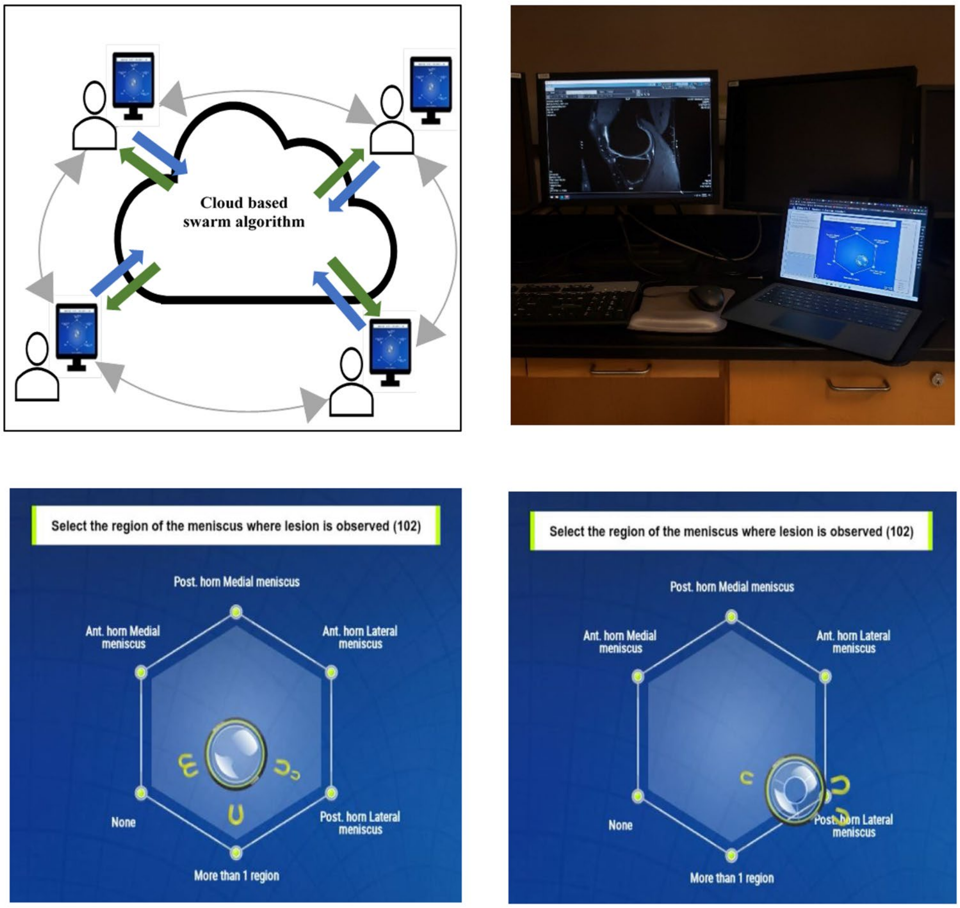
Cloud based
swarm algorithm
Select the region of the meniscus where lesion is observed (102)
Post. horn Medial meniscus
Ant. horn Medial meniscus
Ant. horn Lateral meniscus
None
Post. horn Lateral meniscus
More than 1 region
Select the region of the meniscus where lesion is observed (102)
Post. horn Medial meniscus
Ant. horn Medial meniscus
Ant. horn Lateral meniscus
None
Post. horn Lateral meniscus
More than 1 region

Fig. 3 **A** Schematic of the swarm platform. Multiple remote users are connected to each other in real time, via the web application. Inputs from users (blue arrows) are sent to the cloud server which runs the swarm algorithm, which then sends back continuous a stream of output (green arrows) to users in a closed-loop system. **B** Setup of the swarm session: Participants accessed the knee exams on a PACS workstation and logged into swarm sessions via a separate device. **C** Early time point in a session- multiple users pulling central puck in opposing directions using virtual magnets as seen in the graphical interface. **D** Late time point in the same session- users then converge onto a single answer choice after some negotiation and opinion switch

evolving group decision. The consensus decision computed by the swarm algorithm considers various factors such as (i) the number of swarm participants, (ii) the participants' initial preferences, (iii) participants' behavior (consistent versus changing in opinion), (iv) level of conviction, and (v) type of answer choices (ordinal versus categorical).

Swarm Sessions

Cohort 1 (3 MSK radiologists) participated in a single swarm session, after a washout period after the individual assessment of the knee scans. Cohort 2 (radiology residents) participated in two consecutive swarm sessions post a washout period after their individual assessment. The first resident swarm session had 5 residents. The second resident swarm session had 3 residents and was conducted to measure the effect of the swarm size.

To answer each question during our study, all participants in both cohorts were allowed 60 s to first review the knee scan and then another 60 s to actively participate in the swarm session, collaborate, and provide their consensus answer. In some instances of strong opposing opinions, a swarm may not be able to reach an answer within the time allotted to decide, in which case the platform records it as a "no consensus." All the participants in both the cohorts were blinded to each other and didn't communicate during the session to prevent any form of bias.

AI Model Inference

To benchmark a state-of-the-art AI model against swarm performance of the radiologists and residents, we ran the model over the same set of 36 knee MR scans (sagittal CUBE sequences only). An AI pipeline for localization and classification of meniscus lesions was trained and validated on a retrospective study conducted on 1435 knee MRIs (n = 294 patients; mean age, 43 ± 15 years; 153 women) [46]. The AI pipeline consisted of a V-Net convolutional deep learning architecture to generate segmentation masks for all four meniscus horns that were used to crop smaller sub-volumes containing these regions of interest (ROIs). Such sub-volumes were used as input to train and evaluate three-dimensional convolutional neural networks (3DCNNs) developed to classify meniscus abnormalities. Evaluation on the holdout set yielded sensitivity and specificity of 85% and 85% respectively on a binary assessment ("lesion" or "no lesion").

Statistical Analysis

All responses were binned into 3 classes (none, one compartment, more than 1 compartment) to enable comparisons between individual participant votes, swarm votes, and AI predictions. Confidence scores of the individual responses, among participants of the same cohort, were harmonized to evaluate for internal consistency using Cronbach's alpha. Sensitivity, specificity, and Youden index (measure of accuracy) were calculated for presence or absence of lesions.

The first time point responses were then used to calculate the majority vote and choose the most confident voter in each cohort. Cohen's kappa (k) values were tabulated with mean, standard deviation, and confidence intervals, bootstrapped 100 times resampling a full set of cases from the observations, to evaluate inter-rater reliability as described below.

Attending Inter-rater Reliability Compared with Clinical Standard of Reference (IRRc)

The first set of analyses was conducted comparing attending (cohort 1) responses to arthroscopic notes considered as clinical standard of reference (SOR). IRR of the individual attendings, their majority vote, and the most confident vote were calculated. The IRR of the attending swarm vote was also computed with respect to clinical SOR as well.

Resident Inter-rater Reliability Compared with Clinical Standard of Reference (IRRc)

The second set of analyses was conducted comparing residents (cohort 2) to the clinical SOR. Inter-rater reliability of the individual residents, their majority vote, and the most confident vote were calculated. The IRR of the swarm vote was also computed with respect to clinical SOR for both the 5-resident and 3-resident swarm votes.

Resident Inter-rater Reliability Compared with Radiological Standard of Reference (IRRr)

In many cases, clinical ground truth from surgical evaluation of lesions may not be available. Additionally, there may be low inter-rater reliability between radiologists and surgeons as well. In such instances, the interpretation of an experienced radiologist is often considered as standard of reference, especially when evaluating trainees.

To evaluate for swarm performance in such scenarios, we considered the responses of our senior-most participating attending as a radiological standard of reference. IRR of the individual residents, their majority vote, and the most confident vote was calculated. The IRR of the swarm vote was also compared with radiological SOR for both the 5-resident and 3-resident swarm votes.

Comparing AI Predictions with Clinical and Radiological Standards of Reference (IRRc and IRRr)

Similar to the resident and attending cohorts, the predictions of the model inference were compared with both the clinical and radiological SOR.

Results

The class balance as per clinical standard of reference was as follows: normal (15/36 exams), lesion in one compartment (13/36 exams), lesions in more than one compartment (8/36 exams). The class balance as per a radiological standard of reference was as follows: normal (8/36 exams), lesion in one compartment (8/36 exams), lesions in more than one compartment (20/36 exams).

Both the attending and resident cohorts show excellent internal consistency with Cronbach's alpha of 0.91 and 0.92, respectively. The sensitivity, specificity, and Youden index are described in Table 3. Both the cohorts had high sensitivity in detecting meniscal lesions, comparable between the majority votes, most confident votes, and the swarm votes. The swarm votes showed an improvement in specificity in all scenarios, and an increase in specificity was also observed with an increase in the resident swarm size. The attending swarm votes saw specificity improve by 40% (53.3%) over the attending majority vote (13.3%). The 3-resident swarm demonstrated an improvement in specificity of 20% over the majority vote and the most confident vote, for comparisons against the clinical SOR. The 3-resident swarm also showed an improvement in specificity of 37.5% over the majority vote and most confident votes, for comparisons based on radiological SOR. Similarly, the 5-resident swarm vote showed a specificity of 33% (based on clinical SOR) and 50% (based on radiological SOR), much higher than either the 5-resident majority and most confident vote. This has important clinical implications in preventing overdiagnosis of lesions.

Bootstrapped Cohen's kappa of the attending and resident cohorts' inter-rater reliability with the clinical and radiological standard of reference are mentioned in Table 4, with corresponding 95% confidence intervals. The swarm consensus votes consistently showed higher IRR than the individual voters, their majority vote, and the most confident voter. Superior IRR of swarm votes was observed for both the attending and resident cohorts. More importantly, an increase in swarm IRR was seen in both IRR_c and IRR_r. The swarm methodology thus improved agreement with either standard of reference, indicating its usefulness for assessment, even in scenarios when clinical and radiological observations may have discordance. An increase in IRR was also observed with an increase in resident swarm size. Interestingly, the 5-resident swarm IRR_c agreement was at a comparable level to the 3-attending swarm IRR_c. While the absolute kappa values reported in this study are in the slight to fair range, these should be viewed in light of the limited imaging exam (single sagittal MR sequence only) which was made available for the participants.

1. The IRRc for individual attendings ranged from $k = 0.08$ to 0.29. The 3 attending swarm vote IRRc was higher compared to the 3 attending majority vote and the 3 attending most confident vote (Fig. 4). Agreement on detecting normal cases increases significantly from 13% for majority vote (2/15) to 53% (8/15) for swarm vote. Since the senior-most radiologist was part of this cohort, no IRRr was calculated for the attendings.
2. IRRc for individual resident responses ranged from $k =$ 0.01 to 0.19 and was lower compared to the attendings.

Table 3 Sensitivity, specificity, and Youden's index for binary outputs for the attending and resident cohorts. Swarm consensus votes had higher specificity than the majority vote or most confident vote for both cohorts in all scenarios. The 5-resident swarm also shows higher specificity than that of the 3-resident swarm vote

	Clinical standard of reference			Radiological standard of reference		
	Sensitivity	Specificity	Youden index	Sensitivity	Specificity	Youden index
3 attending majority vote	100%	13.3%	0.13	N/A	N/A	N/A
3 attending most confident vote	95.2%	33.3%	0.28	N/A	N/A	N/A
3 attending swarm vote	90.4%	53.3%	0.43	N/A	N/A	N/A
3-resident majority vote	100%	0	0	100%	0	0
3-resident most confident vote	100%	0	0	100%	0	0
3-resident swarm vote	100%	20%	0.20	100%	37.5%	0.37
5-resident majority vote	100%	0	0	100%	0	0
5-resident most confident vote	95%	6.6%	0.01	96.2%	12.5%	0.08
5-resident swarm vote	95%	33%	0.28	92.5%	50%	0.42
AI prediction	100%	13.3%	0.13	100%	25%	0.25

Table 4 Cohen's kappa values in various comparisons of attendings, residents, and AI with clinical and radiological standards of reference (SOR). For both the attendings and residents, the swarm consensus vote has better agreement than either the majority vote or the most confident voter

	Mean (std) Kappa	95% CI	*p* value
3 attending majority vote versus clinical SOR	0.11 (0.06)	[0.02–0.24]	0.05
3 attending most confident vote versus clinical SOR	0.19 (0.09)	[0.02–0.35]	0.02
3 attendings swarm versus clinical SOR	0.34 (0.11)	[0.16–0.53]	0.18
3-resident majority voting versus clinical SOR	0.02 (0.04)	[−0.07–0.09]	0.16
3-resident most confident vote versus clinical SOR	0.08 (0.04)	[0.02–0.17]	0.85
3-resident swarm versus clinical SOR	0.25 (0.09)	[0.08–0.47]	0.85
5-resident majority vote versus clinical SOR	0.07 (0.06)	[−0.04–0.19]	0.79
5-resident most confident vote versus clinical SOR	0.12 (0.07)	[−0.02–0.26]	0.72
5-resident swarm versus clinical SOR	0.37 (0.10)	[0.16–0.61]	0.54
3-resident majority vote versus radiological SOR	0.27 (0.10)	[0.09–0.49]	0.24
3-resident most confident vote versus radiological SOR	0.15 (0.10)	[−0.03–0.37]	0.41
3-resident swarm versus radiological SOR	0.36 (0.14)	[0.08–0.63]	0.09
5-resident majority vote versus Radiological SOR	0.32 (0.09)	[0.16–0.52]	0.74
5-resident most confident vote versus radiological SOR	0.14 (0.14)	[−0.11–0.35]	0.18
5-resident swarm versus radiological SOR	0.39 (0.12)	[0.15–0.63]	0.03
AI versus clinical SOR	0.10 (0.09)	[−0.11–0.28]	0.001
AI versus radiological SOR	0.15 (0.14)	[−0.13–0.45]	0.015

The 3-resident swarm vote IRRc was higher compared to the 3-resident majority vote and 3-resident most confident vote (Fig. 5). The majority vote and most confident vote failed to identify any normal cases. Agreement on detecting normal cases is 20% (3/15) for swarm vote. The 5-resident swarm vote IRRc was again higher than the 5-resident majority vote and the 5-resident most confident vote. The 3-resident majority vote and most confident vote failed to identify any normal cases. The 5-resident majority vote failed to identify any normal cases. Agreement on detecting normal cases increases by 33% (5/15) for swarm vote.

3. IRRr for individual resident responses vs radiological observation ranged from $k = 0.09$ to 0.22. This was higher compared to the resident IRRc, indicating they had better agreement with their direct trainers i.e., the radiology attendings than with the orthopedic surgeons. In line with the earlier findings, both the 3 and 5 resident swarm vote IRRr was higher than their respective majority votes and most confident votes (Fig. 6). The majority vote and most confident vote failed to identify any normal cases. Agreement on detecting normal cases was 37.5% (3/8) for swarm vote. The majority vote and most confident vote failed to identify any normal cases. The

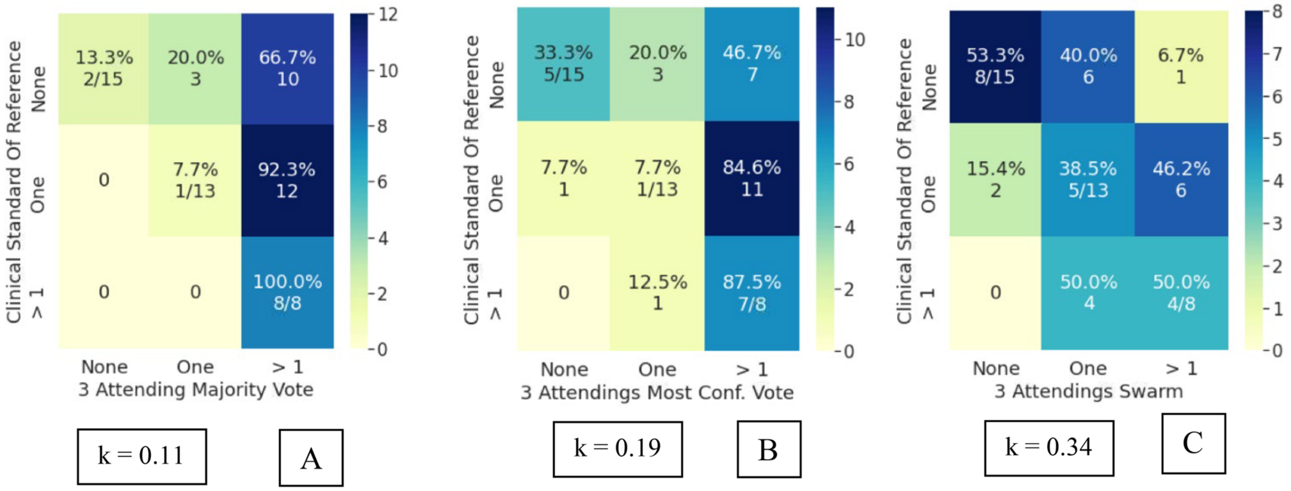

Fig. 4 Attendings grading compared to clinical standard of reference. **A** Confusion matrix (CM) for 3 attending majority vote (kappa: 0.11). **B** CM for 3 attending most confident vote (kappa: 0.19). **C** CM for 3 attending swarm vote (kappa: 0.34)

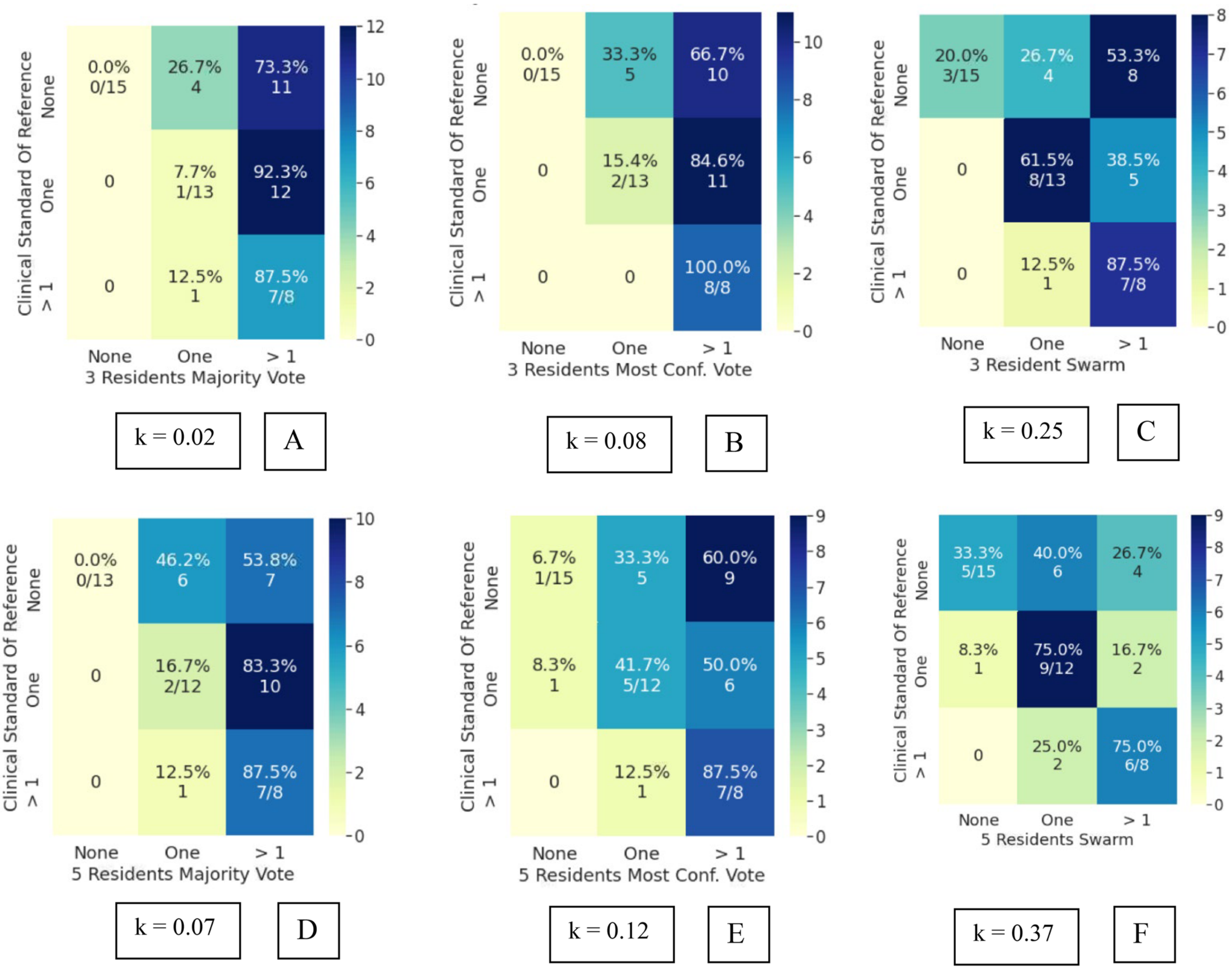

Fig. 5 Residents grading compared to clinical standard of reference. **A** Confusion matrix (CM) for 3-resident majority vote (kappa: 0.02). **B** CM for 3-resident most confident vote (0.08). **C** CM for 3-resident swarm vote (kappa: 0.25) D) CM for 5-resident majority vote (kappa: 0.07). **E** CM for 5-resident most confident vote (0.12). **F** CM for 5-resident swarm vote (kappa: 0.37). Note: The 5-resident swarm was unable to obtain a consensus in one exam. This exam was excluded during inter-rater reliability comparisons of 5-resident majority vote and 5-resident most confident vote for parity

5-resident majority vote failed to identify any normal cases. Agreement on detecting normal cases increased for the swarm vote in both size cohorts.

As opposed to the 3-resident and 3-attending swarms, the 5-resident swarm failed to reach a consensus in one exam, in the allotted time. This single occurrence was not enough to conclusively comment on relationship of swarm size and optimal time for decision and was subsequently excluded during comparisons with the majority and most confident votes.

4. AI predictions from the model inference had an IRRc of $k = 010$ and IRRr of $k = 0.15$. This is comparable to the range of individual resident inter-rater reliability (Fig. 7).

Discussion

Multiple studies have reported varying IRR among radiologists in interpreting meniscal lesions [47]. Differences in opinions can occur based on location, zone, tissue quality, and severity of lesion. Shah et al. reported prevalence and bias-adjusted kappa ranging from poor for medial meniscus zone 1($k = 0.22$) to excellent for lateral meniscus zone 3 ($k = 0.88$) [48]. Some imaging-related factors for the low agreement include limited image resolution, motion artifacts, and the limited time afforded to radiologists for image interpretation under an ever increasing workload [49].

Arthroscopic evaluation is often considered as the clinical standard of reference for evaluating radiological reads

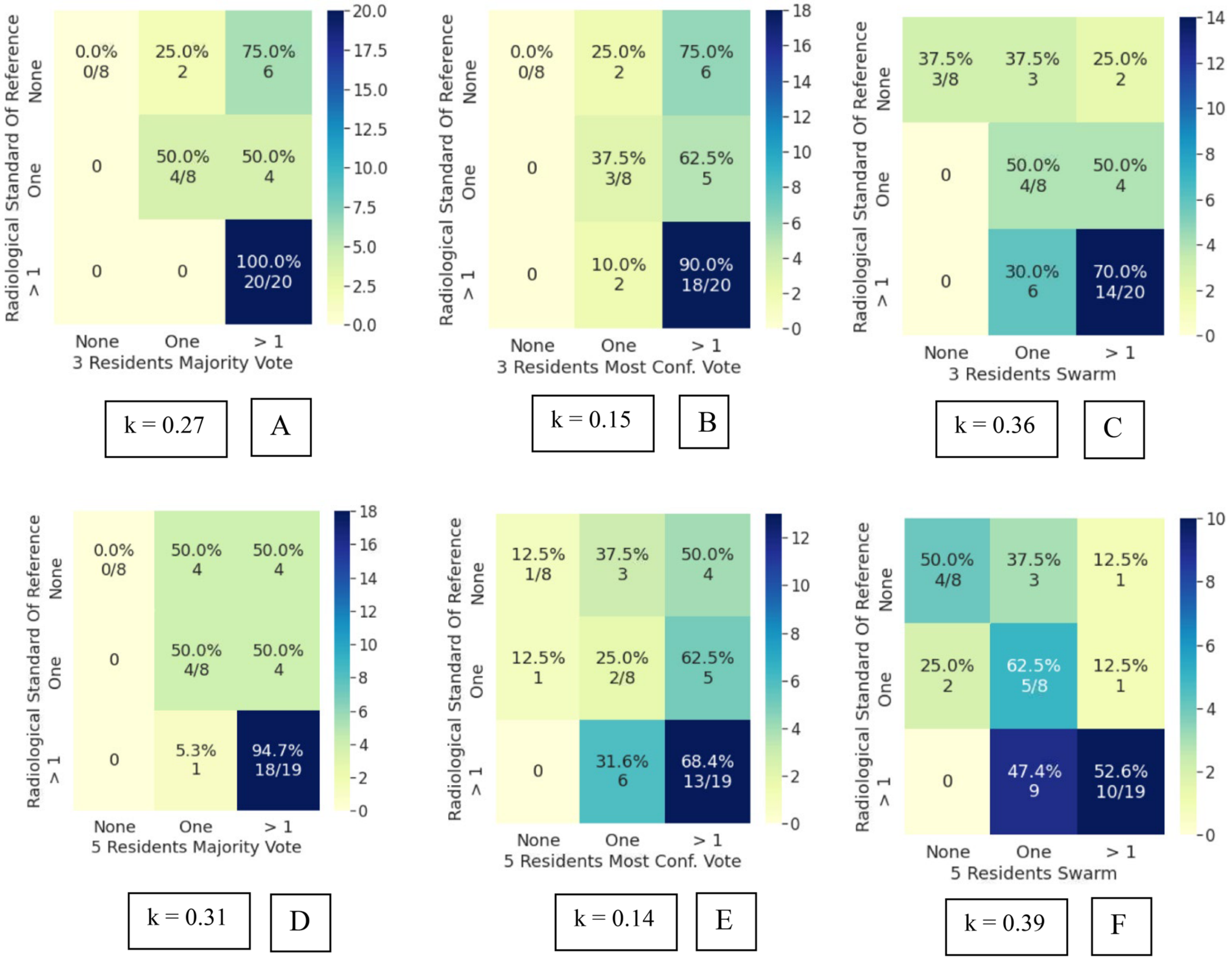

Fig. 6 Residents' responses compared to radiological standard of reference. **A** Confusion matrix (CM) for 3-resident majority vote (kappa: 0.27). **B** CM for 3-resident most confident vote (0.15). **C** CM for 3-resident swarm vote (kappa: 0.36). **D** CM for 5-resident majority vote (kappa: 0.31). **E** CM for 5-resident most confident vote (0.14). **F** CM for 5-resident swarm vote (kappa: 0.39). Note: The 5-resident swarm was unable to obtain a consensus in one exam. This exam was excluded during inter-rater reliability comparisons of 5-resident majority vote and 5-resident most confident vote for parity

[50]. However, surgeons have a narrower field of view during arthroscopy and lack the ability to view the region of interest in multiple orientations (sagittal, axial, coronal) simultaneously. These factors limit consideration of surgical observations as reliable clinical ground truth.

Additionally, there may be a lag time of days to weeks between imaging and arthroscopy allowing improvement or deterioration of lesion and which can further limit agreement with their radiology colleagues. Kim et al. reported inter-method reliability (radiology-arthroscopy) kappa values ranging from 0.40 to 0.71 depending on the laterality of lesion and presence of ACL tears [51]. Such differences in opinions are problematic for generating clinical consensus and defining ground truth labels for A.I. training. Given that the radiologist's report and arthroscopy evaluations can have some disagreement, we examined the use of swarm methodology against a radiological standard of reference (senior-most radiologist) as well.

Multiple investigators in the past have advocated the use of consensus voting to improve medical diagnoses [52] and demonstrated superior performance of majority or average vote [53]. However, no study till date had compared consensus votes from a real-time blinded collaboration to a post hoc majority vote. There have been varying opinions on what exactly improves the accuracy in a crowds-based answer, the effect of social interaction [54], or pure statistical aggregation. Social interaction can be further complicated by the interpersonal biases which can either improve or worsen crowd performance [55, 56]. Thus, it is pertinent to understand the exact influence of these factors especially when they are applied to make clinical decisions.

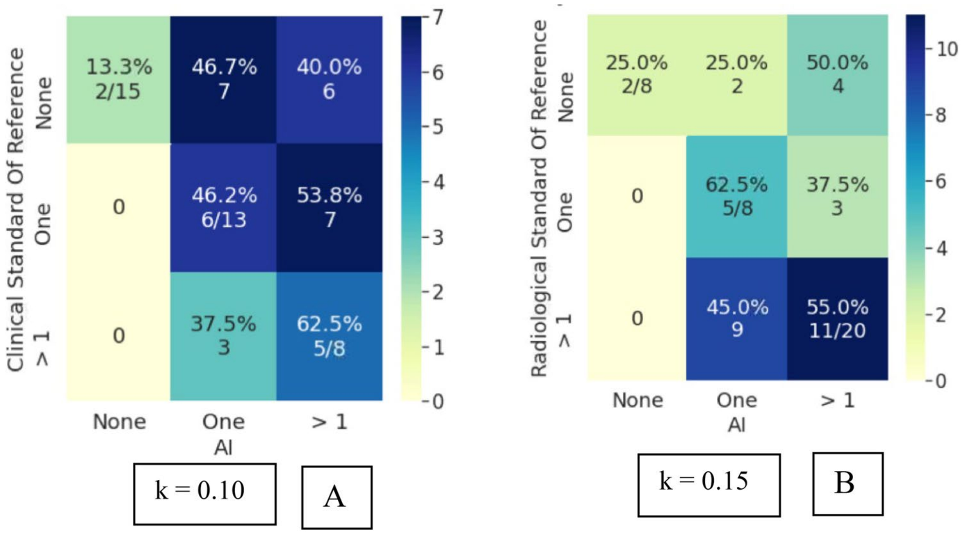

Fig. 7 AI prediction comparisons. **A** Confusion matrix for AI predictions compared to clinical standard of reference (kappa: 0.10). **B** Confusion matrix for AI predictions compared to radiological standard of reference (kappa: 0.15). Swarm votes of residents outperform AI in both sets of comparisons

Our current study explored these questions by first performing nonsocial interactions between blinded participants at equal levels of expertise (radiologists or residents' cohorts), in a bias-free environment. Next the resident cohort repeated a swarm session with fewer participants, to measure the effect of group size on the responses. Our results show both the group size and interaction influence performance, although conducting negotiations for the optimal answer under *anonymization* was key for resisting peer pressure.

A key aspect of our study was to evaluate the performance of an AI model on the same set of 36 knee exams. This model had been trained and tested on labels created by multiple radiologists and residents at our institution over time. The AI IRR_c was $k = 0.10$ and was comparable to the IRR_c of the 3-resident most confident vote. The AI IRR_r was $k = 0.15$, comparable to the IRR_r of the 3-resident most confident vote. In other words, the AI performance is already as good as its trainers. In both cases, however, the kappa was significantly lower than the kappa of either the resident or the attending swarms. A useful strategy to improve model performance beyond its current results would be to use swarm votes as labels in the training datasets. Leveraging swarm intelligence for AI training would provide higher quality labels which are more accurate, mitigate the problem of noisy labels, and reduce the need for large training datasets as currently needed for most deep learning models.

Swarm voting improved IRR by up to 32% in our study, which was based on a specific imaging modality (MR), and for a specific task of evaluating meniscal lesions. It would be important to investigate the increase in diagnostic yield by real-time consensus voting, in other diagnostic imaging scenarios across different modalities as well. The swarm platform would be a useful tool for expert radiologists to collaborate and evaluate complex or ambiguous cases. A potential first application would be for imaging workflows where multiple reads are already mandated, such as for double reads for breast mammograms, as practiced in Europe [57].

Our study had a few limitations. While we aimed to simulate the regular radiology workflow with the use of PACS, it did not capture the entire experience given the time constraints to run the swarm sessions. Normally, radiologists have access to multiple sequences and views in an MR exam, with prior exams and other relevant clinical notes for comparison. We speculate the inter-rater reliability in our study would have been higher and in line with other reported studies, with the availability of complete MRI exams.

Given scheduling challenges in the pandemic, we performed only necessary swarm sessions as required for this pilot study. While we observed improvements in swarm agreement with both the standards of reference, the overall dataset in this study was not large enough to power a statistically significant difference over individual or majority votes.

Though we were able to observe improved inter-rater reliability and specificity with an increase in swarm size (five- versus three-resident swarm), further investigation with additional participants is warranted to estimate optimal group size. Given the limited availability of expert radiologists, it will be important to understand if diagnostic gains made with larger groups peak at a certain participant number.

Conclusion

In conclusion, utilizing a digital swarm platform improved consensus among radiologists and allowed participants to express judgement-free intent. This novel approach outperforms traditional consensus methodologies such as a majority vote. Future direction of our work includes performing serial swarm sessions to generate more accurate labels for AI training. We also aim to explore the swarm platform for evaluating trainee performance. Residents at our center can self-assess their diagnostic opinions with peers, and the training program can assess group performance across cohorts over time, in an objective manner.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-022-00662-3.

Acknowledgements We would like to thank Unanimous AI, for providing us pro-bono access to the Swarm software platform for our study.

Funding R33 AR073552 grant from the NIH was utilized for conducting this study.

Declarations

Ethics Approval The following study was conducted in accordance with IRB process at the primary medical center. The study utilized retrospectively acquired data only, and the IRB committee determined that our study did not require an ethics approval.

Competing Interests The authors declare no competing interests.

References

1. Fink, A., Kosecoff, J., Chassin, M. & Brook, R. H. Consensus methods: characteristics and guidelines for use. *American journal of public health* **74**, 979-983 (1984).
2. Medicine, I. o., National Academies of Sciences, E. & Medicine. *Improving Diagnosis in Health Care*. (The National Academies Press, 2015).
3. Smith, C. P. *et al.* Intra- and interreader reproducibility of PI-RADSv2: a multireader study. *Journal of magnetic resonance imaging : JMRI* **49**, 1694-1703, https://doi.org/10.1002/jmri.26555 (2019).
4. van Tilburg, C. W. J., Groeneweg, J. G., Stronks, D. L. & Huygen, F. Inter-rater reliability of diagnostic criteria for sacroiliac joint-, disc- and facet joint pain. *Journal of back and musculoskeletal rehabilitation* **30**, 551-557, https://doi.org/10.3233/bmr-150495 (2017).
5. Melsaether, A. *et al.* Inter- and intrareader agreement for categorization of background parenchymal enhancement at baseline and after training. *American Journal of Roentgenology* **203**, 209-215, https://doi.org/10.2214/AJR.13.10952 (2014).
6. Tibrewala, R. *et al.* Computer-aided detection AI reduces interreader variability in grading hip abnormalities with MRI. *Journal of magnetic resonance imaging : JMRI*, https://doi.org/10.1002/jmri.27164 (2020).
7. Dunn, W. R. *et al.* Multirater agreement of arthroscopic meniscal lesions. *The American journal of sports medicine* **32**, 1937-1940, https://doi.org/10.1177/0363546504264586 (2004).
8. Bruno, M. A., Walker, E. A. & Abujudeh, H. H. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *RadioGraphics* **35**, 1668-1676, https://doi.org/10.1148/rg.2015150023 (2015).
9. Choy, G. *et al.* Current applications and future impact of machine learning in radiology. *Radiology* **288**, 318-328, https://doi.org/10.1148/radiol.2018171820 (2018).
10. Demirer, M. *et al.* A User interface for optimizing radiologist engagement in image data curation for artificial intelligence. *Radiology: Artificial Intelligence* **1**, e180095, https://doi.org/10.1148/ryai.2019180095 (2019).
11. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis* **65**, 101759, https://doi.org/10.1016/j.media.2020.101759 (2020).
12. Albarqouni, S. *et al.* Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging* **35**, 1313-1321 (2016).
13. Northcutt, C. G., Jiang, L. & Chuang, I. L. Confident learning: estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* (2021).
14. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15) (2021, May).
15. Northcutt, C. G., Wu, T. & Chuang, I. L. Learning with confident examples: rank pruning for robust classification with noisy labels. *arXiv preprint* http://arxiv.org/abs/1705.01936 *(2017).*
16. Lee, K.-H., He, X., Zhang, L. & Yang, L. Cleannet: Transfer learning for scalable image classifier training with label noise. In Proceedings of the IEEE conference on computer vision and pattern recognition 5447–5456 (2018).
17. Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G. & Mohd-Yusof, J. Combating label noise in deep learning using abstention. *arXiv preprint* http://arxiv.org/abs/1905.10964 *(2019).*
18. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C. & Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 11244-11253 (2019).
19. Veit, A., Nickel, M., Belongie, S., & van der Maaten, L. Separating self-expression and visual content in hashtag supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 5919-5927 (2018).
20. Shen, Y. & Sanghavi, S. Learning with Bad Training Data via Iterative Trimmed Loss Minimization. Proceedings of the 36th International Conference on Machine Learning, in Proceeding of *Machine Learning Research* **97,** 5739-5748 (2019).
21. Ren, M., Zeng, W., Yang, B. & Urtasun, R. Learning to Reweight Examples for Robust Deep Learning. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Researchin **80**,4334-4343 (2018).
22. Lehman, C. D. *et al.* Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* **290**, 52-58, https://doi.org/10.1148/radiol.2018180694 (2019).
23. Yan, Y. *et al.* in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* Vol. 9 (eds

Teh Yee Whye & Titterington Mike) 932--939 (PMLR, Proceedings of Machine Learning Research, 2010).

24. Kurvers, R. H. J. M. *et al.* Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences* **113**, 8777-8782, https://doi.org/10.1073/pnas.1601827113 (2016).
25. Posso, M. *et al.* Effectiveness and cost-effectiveness of double reading in digital mammography screening: a systematic review and meta-analysis. *European journal of radiology* **96**, 40-49 (2017).
26. Milholland, A. V., Wheeler, S. G. & Heieck, J. J. Medical assessment by a Delphi group opinion technic. *New England Journal of Medicine* **288**, 1272-1275 (1973).
27. Mamisch, N. *et al.* Radiologic criteria for the diagnosis of spinal stenosis: results of a Delphi survey. *Radiology* **264**, 174-179, https://doi.org/10.1148/radiol.12111930 (2012).
28. Seeley, T. D., Visscher, P. K. & Passino, K. M. Group decision making in honey bee swarms: when 10,000 bees go house hunting, how do they cooperatively choose their new nesting site? *American Scientist* **94**, 220-229 (2006).
29. Bonabeau, E. et al. *Swarm Intelligence: From Natural to Artificial Systems*. (OUP USA, 1999).
30. Krause, J., Ruxton, G. D. & Krause, S. Swarm intelligence in animals and humans. *Trends in Ecology & Evolution* **25**, 28-34, https://doi.org/10.1016/j.tree.2009.06.016 (2010).
31. Arrow, K. J. *et al.* The promise of prediction markets. *Science-new york then washington-* **320**, 877 (2008).
32. Rosenberg, L., Lungren, M., Halabi, S., Willcox, G., Baltaxe, D., & Lyons, M. Artificial Swarm Intelligence employed to Amplify Diagnostic Accuracy in Radiology. In S. Chakrabarti, & H. N. Saha (Eds.), 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018 (pp. 1186-1191). [8614883] (2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2018). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/IEMCON.2018.8614883 (2019).
33. Sulis, W. Fundamental concepts of collective intelligence. *Nonlinear Dynamics, Psychology, and Life Sciences* **1**, 35-53, https://doi.org/10.1023/A:1022371810032 (1997).
34. Galton, F. Vox Populi. Nature **75**, 450–451. https://doi.org/10.1038/075450a0 (1907).
35. Salminen, J. Collective intelligence in humans: a literature review. *arXiv preprint* http://arxiv.org/abs/1204.3401 *(2012).*
36. Bahrami, B. *et al.* Optimally interacting minds. *Science* **329**, 1081-1085 (2010).
37. Shanteau, J. How much information does an expert use? Is it relevant? *Acta psychologica* **81**, 75-86 (1992).
38. Kozhevnikov, M., Evans, C. & Kosslyn, S. M. Cognitive style as environmentally sensitive individual differences in cognition: a modern synthesis and applications in education, business, and management. *Psychological Science in the Public Interest* **15**, 3-33, https://doi.org/10.1177/1529100614525555 (2014).
39. McCrae, R. R. & Costa, P. T. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology* **52**, 81 (1987).
40. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The "Reading the mind in the eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry* **42**, 241-251 (2001).
41. Rosenberg, L. & Willcox, G. 1054-1070 (Springer International Publishing).
42. Rosenberg, L. in 2016 International Joint Conference on Neural Networks (IJCNN). 2547-2551.
43. Russell, C. *et al.* Baseline cartilage quality is associated with voxel-based T1ρ and T2 following ACL reconstruction: a multicenter pilot study. *Journal of Orthopaedic Research* **35**, 688-698, https://doi.org/10.1002/jor.23277 (2017).
44. Peterfy, C. G. *et al.* Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis. *Osteoarthritis Cartilage* **12**, 177-190, https://doi.org/10.1016/j.joca.2003.11.003 (2004).
45. Patel, B. N. *et al.* Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med* **2**, 111, https://doi.org/10.1038/s41746-019-0189-7 (2019).
46. Astuto, B. *et al.* Automatic deep learning assisted detection and grading of abnormalities in knee MRI studies. *Radiology: Artificial Intelligence* **0**, e200165, https://doi.org/10.1148/ryai.2021200165 (2021)
47. Phelan, N., Rowland, P., Galvin, R. & O'Byrne, J. M. A systematic review and meta-analysis of the diagnostic accuracy of MRI for suspected ACL and meniscal tears of the knee. *Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA* **24**, 1525-1539, https://doi.org/10.1007/s00167-015-3861-8 (2016).
48. Shah, J. *et al.* Correlation of meniscus tears on MRI and arthroscopy using the ISAKOS classification provides satisfactory intermethod and inter-rater reliability. *Journal of ISAKOS: Joint Disorders & Orthopaedic Sports Medicine* **5**, 201-207, https://doi.org/10.1136/jisakos-2019-000408 (2020).
49. Harolds, J. A., Parikh, J. R., Bluth, E. I., Dutton, S. C. & Recht, M. P. Burnout of radiologists: frequency, risk factors, and remedies: a report of the acr commission on human resources. *Journal of the American College of Radiology* **13**, 411-416, https://doi.org/10.1016/j.jacr.2015.11.003 (2016).
50. Fritz, B., Marbach, G., Civardi, F., Fucentese, S. F. & Pfirrmann, C. W. A. Deep convolutional neural network-based detection of meniscus tears: comparison with radiologists and surgery as standard of reference. *Skeletal radiology* **49**, 1207-1217, https://doi.org/10.1007/s00256-020-03410-2 (2020).
51. Kim, S. H., Lee, H. J., Jang, Y. H., Chun, K. J. & Park, Y. B. Diagnostic accuracy of magnetic resonance imaging in the detection of type and location of meniscus tears: comparison with arthroscopic findings. *Journal of clinical medicine* **10**, https://doi.org/10.3390/jcm10040606 (2021).
52. Kane, B. & Luz, S. Achieving diagnosis by consensus. *Computer Supported Cooperative Work (CSCW)* **18**, 357-392, https://doi.org/10.1007/s10606-009-9094-y (2009).
53. Kattan, M. W., O'Rourke, C., Yu, C. & Chagin, K. The wisdom of crowds of doctors: their average predictions outperform their individual ones. *Medical Decision Making* **36**, 536-540, https://doi.org/10.1177/0272989x15581615 (2016).
54. Brennan, A. A. & Enns, J. T. When two heads are better than one: Interactive versus independent benefits of collaborative cognition. *Psychonomic Bulletin & Review* **22**, 1076-1082, https://doi.org/10.3758/s13423-014-0765-4 (2015).
55. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* **108**, 9020-9025, https://doi.org/10.1073/pnas.1008636108 (2011).
56. Hertwig, R. Tapping into the wisdom of the crowd–with confidence. *Science* **336**, 303-304 (2012).
57. Perry, N. *et al.* European guidelines for quality assurance in breast cancer screening and diagnosis. -summary document. *Oncology in Clinical Practice* **4**, 74-86 (2008).