Artificial Swarm Intelligence employed to Amplify Diagnostic Accuracy in Radiology

Louis Rosenberg
and Gregg Willcox,
Unanimous AI
San Luis Obispo, CA, USA

Safwan Halabi MD and
Matthew Lungren MD
Stanford University Medical School
Stanford, CA, USA

David Baltaxe and
Mimi Lyons,
Unanimous AI
San Francisco, CA, USA

***Abstract*— Swarm Intelligence (SI) is a biological phenomenon in which groups of organisms amplify their combined intelligence by forming real-time systems. It has been studied for decades in fish schools, bird flocks, and bee swarms. Recent advances in networking and AI technologies have enabled distributed human groups to form closed-loop systems modeled after natural swarms. The process is referred to as Artificial Swarm Intelligence (ASI) and has been shown to significantly amplify group intelligence. The present research applies ASI technology to the field of medicine, exploring if small groups of networked radiologists can improve their diagnostic accuracy when reviewing chest X-rays for the presence of pneumonia by "thinking together" as an ASI system. Data was collected for individual diagnoses as well as for diagnoses made by the group working as a real-time ASI system. Diagnoses were also collected using a state-of-the-art deep learning system developed by Stanford University School of Medicine. Results showed that a small group of networked radiologists, when working as a real-time closed-loop ASI system, was significantly more accurate than the individuals on their own, reducing errors by 33%, as well as significantly more accurate (22%) than a state-of-the-art software-only solution using deep learning.**

I. INTRODUCTION

Artificial Intelligence has made major advances in the field of Radiology in recent years, enabling automated diagnoses of medical images that rivals, and in some cases exceeds, the accuracy of human practitioners. For example, the CheXNet system developed at Stanford University School of Medicine was recently shown to diagnose the presence of pneumonia with significantly greater accuracy than expert radiologists [1]. In the field of dermatology, researchers recently found that a convolutional neural network outperformed a majority of human dermatologists tested in diagnosing melanoma [2,3]. And in the field of ophthalmology, a recent study by Google Deepmind has shown that algorithms trained by machine learning (ML) can be as good as human experts in detecting eye conditions [4].

Results like this have raised concerns in some medical fields about the future of the profession for human practitioners. This is particularly true in the field of radiology, where machine learning has made significant strides. This prompted, Geoffrey Hinton, a leading AI researcher to famously tell the New Yorker magazine last year that medical schools "should stop training radiologists now" [5]. These growing concerns, whether they prove justified or overblown, raise a significant question – what can be done to ensure that human judgement remains a valued and consequential factor in fields like radiology?

One approach is to use artificial intelligence to amplify the diagnostic abilities of human practitioners, rather than replace them. While there are numerous paths for exploring this notion, the present study looks at one promising technology known as Artificial Swarm Intelligence (ASI). Inspired by the natural principle of Swarm Intelligence (SI), this technology connects distributed groups of networked human participants into real-time systems modeled after natural swarms and moderated by AI algorithms. In layman's terms, this technology uses real-time networks and AI algorithms to build a "hive mind" of human practitioners, enabling the groups to converge on solutions together that are significantly more accurate than the individuals could achieve on their own [6-10]. In one recent study of ASI technology, researchers at Oxford and Unanimous AI tasked groups of financial traders with predicting four economic indicators: the S&P Index (SPX), the price of gold (GLD), the gold miners index (GDX) and the price of crude oil. Across three months of weekly forecasts, results showed a 26% increase in prediction accuracy ($p<0.001$) for the ASI-based predictions as compared to individual forecasts [11].

While prior studies have shown that ASI technology can amplify human accuracy in predictive tasks such as predicting sports and forecasting financial markets, no prior research has tested the use of distributed swarm-based technologies for medical diagnosis. The present study explores the use of ASI in the medical field, with a specific focus on diagnostic radiology. Specifically, we apply ASI technology to the diagnosis of chest x-rays for the presence of pneumonia. This diagnostic task was chosen because evaluating chest x-rays for pneumonia is the most commonly performed radiological procedure in the US and because machine learning systems like CheXNet have already shown that algorithms alone can outperform individual human practitioners. The question thus remains, can small groups of networked radiologists, working as a real-time "hive-mind," outperform the software only machine learning systems that currently exceeded individual human performance.

II. SWARMS AS INTELLIGENT SYSTEMS

When reaching decisions as an ASI system, distributed human groups "think together" as a real-time swarm in which participants act, react, and interact as a population, converging on optimized solutions in synchrony, moderated by intelligence algorithms. The swarming process is modeled on biological systems such as schools of fish, flocks of birds, and swarms of bees. The present study uses Swarm AI® technology from Unanimous AI, which is modeled primarily on the collective decision-making processes of honeybee swarms [6-10].

This framework was chosen because honeybee swarms have been shown to converge upon optimized solutions to complex problems that are far beyond the capabilities of their individual members [11]. The decision-making processes in honeybee swarms have been found to be surprisingly similar to the decision-making in neurological brains [12,13]. Both are distributed systems that employ large populations of simple excitable units (i.e., bees and neurons) that function in parallel to integrate noisy evidence, weigh competing alternatives, and converge on decisions in real-time synchrony [14-16].

III. ENABLING SWARMS

Unlike birds, bees and fish, humans have not evolved the natural ability to form closed-loop swarms, as we lack the subtle connections that other organisms use. Schooling fish detect vibrations in the water around them. Flocking birds detect motions propagating through the population. Swarming bees use complex body vibrations called a "waggle dance." To enable real-time swarming among networked human groups, unique interfaces, algorithms, and communication protocols are needed to close the loop around the full the set of members. To address this need, a software platform called *swarm.ai* was developed to enable distributed human groups, connected in real-time form anywhere in the world, to form closed-loop swarms over standard internet connections [10].

Modeled after the decision-making processes of honeybee swarms, swarm.ai enables networked groups to work in parallel to integrate noisy evidence, weigh competing alternatives, and converge on decisions in synchrony. As shown in Figure 1, the platform enables "human swarms" to answer questions by collaboratively manipulating a graphical puck to select from among a set of alternatives. Each participant provides input by moving a graphical magnet with a mouse, touchpad, or touchscreen. By positioning their magnet with respect to the puck, participants apply their will on the system. The input from each user is not a discrete vote, but a stream of vectors that varies freely over time. Because all members adjust their intent continuously, the swarm explores the decision-space, not based on the input of any single individual, but based on the dynamics of the system. This enables complex deliberations to emerge, empowering the group to converge on optimal solutions.

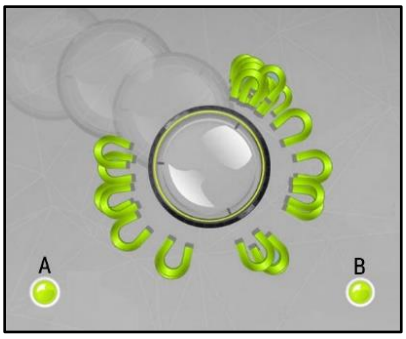

Fig. 1. ASI converging upon a solution as a real-time system

It is important to note that participants do not only vary the direction of their intent, but also modulate the magnitude of their intent by adjusting the distance between their magnet and the puck. Because the puck is in continuous motion across the decision-space, users need to continually move their magnet so that it stays close to the puck's outer rim. This is significant, for it requires participants to be engaged continuously throughout the decision process, evaluating and re-evaluating their intent. If they stop adjusting their magnet with respect to the changing position of the puck, the distance grows and their applied sentiment wanes.

Thus, like bees vibrating their bodies to express sentiment in a biological swarm, or neurons firing activation signals to express conviction levels within a biological neural-network, the participants in an artificial swarm must continuously update and express their changing preferences during the decision process, or lose their influence over the collective outcome. In addition, intelligence algorithms monitor the behaviors of all swarm members in real-time, inferring their implied conviction based upon their relative motions over time. This reveals a range of behavioral characteristics within the swarm population and weights their contributions accordingly, from entrenched participants to flexible participants to fickle participants.

IV. PNEUMONIA DIAGNOSIS STUDY

Researchers at Stanford University School of Medicine and Unanimous AI conducted a study in which a "hive mind" of eight radiologists connected by ASI swarming algorithms was tasked with diagnosing a set of 50 chest X-rays by working together as a real-time system. For each of the 50 trials, a chest X-ray was presented simultaneously to the radiologists. After a few seconds of individual assessment, the group worked together as an ASI swarm, converging on a probabilistic diagnosis as to the likelihood that the patient has pneumonia. All eight radiologists participated from their own unique locations, each connecting to the swarm.ai platform through a standard internet browser. For each of the 50 trials, the assessment was performed through a two-step process in which the swarm first converged on a coarse range of probabilities and then converged on a refined value for the probability. This is shown below.

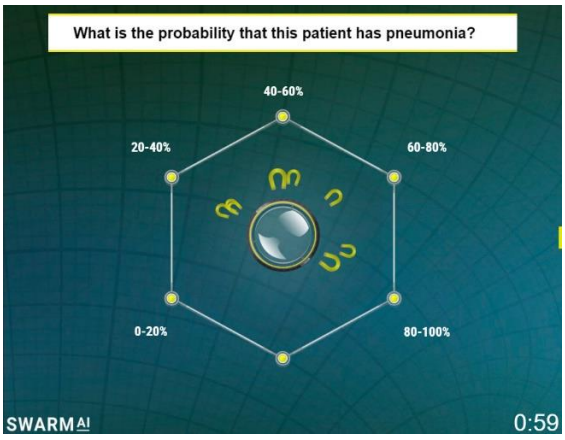

Fig. 2. ASI in the process of Diagnosing Pneumonia

Figure 2 shows a screenshot of the ASI system in the process of selecting the coarse range of probabilities for an X-ray image displayed to all participants at the same time. It's important to note that the screenshot above is a momentary snapshot of the system as the collaboratively controlled puck moves across the decision-space and converges upon an answer. This full process of AI moderated deliberation generally takes between 15 and 60

seconds. In the example shown above, the swarm converged on the range 40-60% within 18 seconds.

The swarm was then immediately tasked with selecting a specific value within the chosen coarse range. Figure 3 below shows a screenshot of the ASI system in the process of converging upon a probability that the patient has pneumonia. This generally takes an additional 15 to 30 seconds. In this way, each diagnosis was converged upon by the ASI system in under 90 seconds for each one of the 50 chest X-ray trials evaluated.

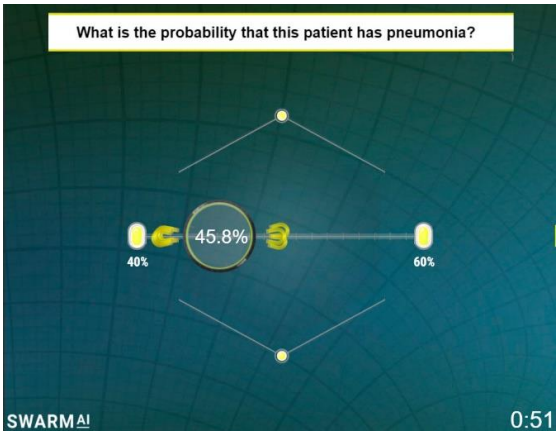

Fig. 3. ASI in process refining a diagnosis

Following this two-step process, the ASI system captured input from the distributed group of radiologists for each of the 50 trials. Because each participant is both a member of the system as well as a source of individual data, their initial input, before the swarm starts, was collected as a representation of their individual probabilistic forecast. This enables a performance comparison between the individuals, assessing on their own, and ASI system converging in synchrony.

Separately, the same set of 50 chest X-rays were run through a state-of-the-art ML system to generate algorithmic assessment for the presence of pneumonia. Specifically, the CheXNet deep learning software developed at Stanford University was used. This is a 121-layer convolutional neural network (CNN). It was employed to generate algorithmic probabilities as to whether each patient has pneumonia. In this way, three sets of diagnostic probabilities were generated, (a) individual diagnoses, (b) ASI diagnoses, and (c) software-only ML diagnoses. These three sets of probabilities were then scored against Ground Truth and compared using a variety of statistical techniques.

V. RESULTS

We compared the performance of the ASI system against both (a) individual human performance, and (b) the software-only CheXNet system. When comparing ASI to individual radiologist, we compute three metrics - (i) Binary Classification accuracy and (ii) Mean Absolute Error, and (iii) F1 scores, also known as harmonic mean. As shown in the figures below, the ASI system outperformed the individuals in all four metrics.

Binary Classification: Using fifty-percent probability as the cutoff for classifying a positive diagnosis, the individuals achieved 73% diagnostic accuracy (i.e. 27% error rate) against Ground Truth across the 50 test cases, while the ASI system achieved 82% diagnostic accuracy (i.e. 18% error rate) across the same 50 cases. This corresponds to a 33% reduction in errors when working as an ASI system as compared to direct individual performance. To assess significance, a bootstrap analysis was performed on 10,000 samples, as shown in Figure 4a. The swarm was found to be significantly more accurate than the individuals alone ($p<0.01$, $\mu_{difference} = 9.1\%$).

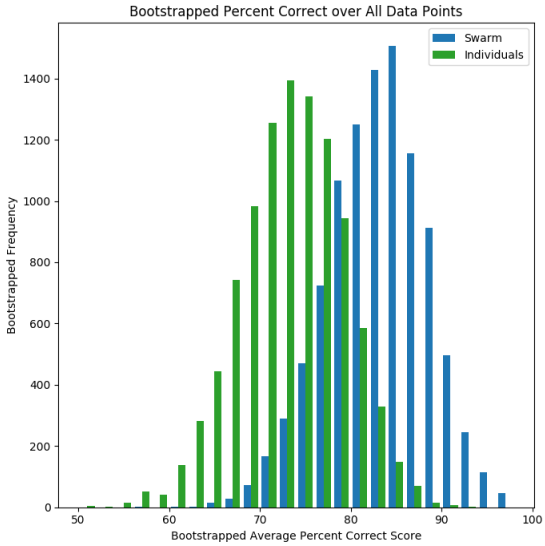

Figure 4a. Percent Correct (ASI vs Individuals)

Mean Absolute Error: MAE is calculated as the absolute value of the Ground Truth minus the Predicted Probability. A bootstrap analysis revealed that the swarm of radiologists had significantly higher probabilistic accuracy than the individuals ($p<0.002$, $\mu_{difference} = 8.6\%$), as shown in Figure 4b.

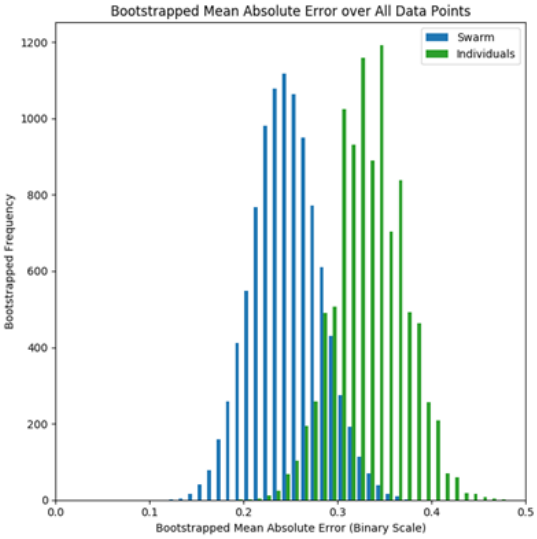

Figure 4b. Mean Absolute Error (ASI vs Individuals)

F1 Score: We compare the performance of the ASI system to individual radiologists on the F1 metric, which is defined as the harmonic average of the precision and recall achieved during binary classification. As shown in Figure 4c below, we find that the ASI system averages an F1 score of 0.75 while the individual radiologist achieve a lower average F1 score of 0.64. To assess statistical significance, we bootstrap across 10,000 samples. We

find that the F1 score of the Swarm was not sufficiently higher than the F1 for individuals for statistical significance ($p > 0.05$), suggesting that a set of 50 trials was not adequate to demonstrate statistical significance on F1 scores, which vary substantially in average magnitude based on the sample data set.

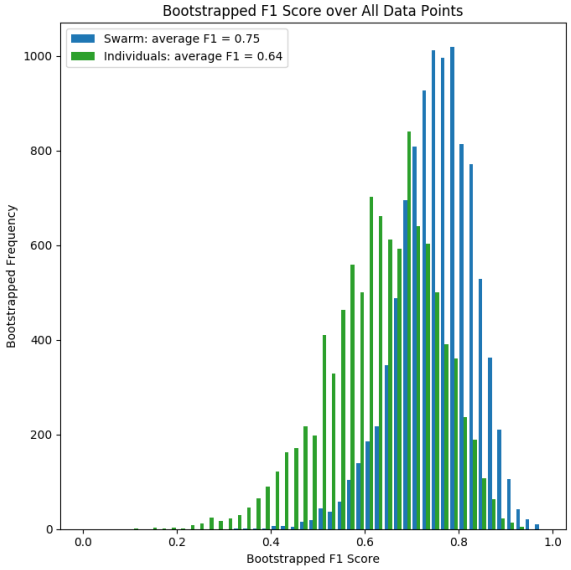

Figure 4c. Bootstrapped F1 Scores (ASI vs Individuals)

When comparing ASI to the ML system, we compute four metrics - (i) binary classification accuracy, (ii) Mean Absolute Error, (iii) ROC analysis, and (iv) F1 scores. As shown in the figures and text below, the ASI system outperformed the software-only deep learning system across all three metrics.

Binary Classification: Using fifty-percent probability as the cutoff for classifying a positive diagnosis, the ML system achieved 60% accuracy against Ground Truth across 50 trials, while the ASI system achieved 82% accuracy across the same 50 trials. To assess statistical significance, a bootstrap analysis was performed on 10,000 samples as shown in Figure 5a. The swarm was significantly more accurate in binary classification than the ML system ($p<0.01$, $\mu_{difference} = 21.9\%$).

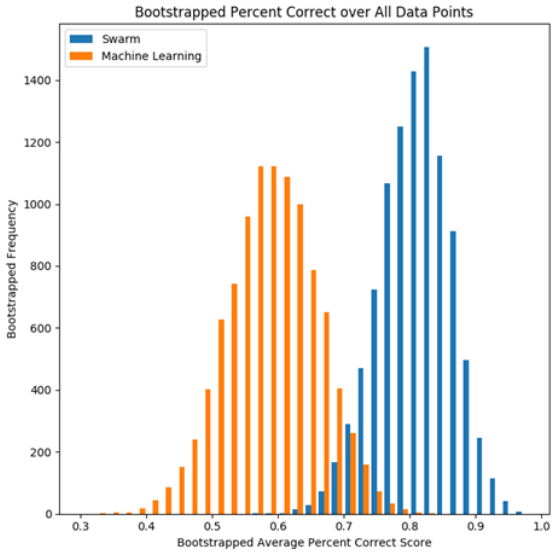

Figure 5a. Percent Correct (ASI vs ML)

Mean Absolute Error: MAE is calculated as the absolute value of the Ground Truth minus the Predicted Probability. A bootstrap analysis of MAE revealed the swarm had significantly higher probabilistic accuracy than the ML system ($p<0.001$, $\mu_{difference} = 21.6\%$), as shown in Figure 5b. To address the possibility that Ground Truth could be error prone, we also looked at "Agreed Truth", defined as only those cases where the ASI and ML systems agreed on the diagnosis. Even in this conservative case, the swarm significantly outperformed ML ($p<0.001$, $\mu_{difference} = 21.3\%$), as shown in Figure 5c.

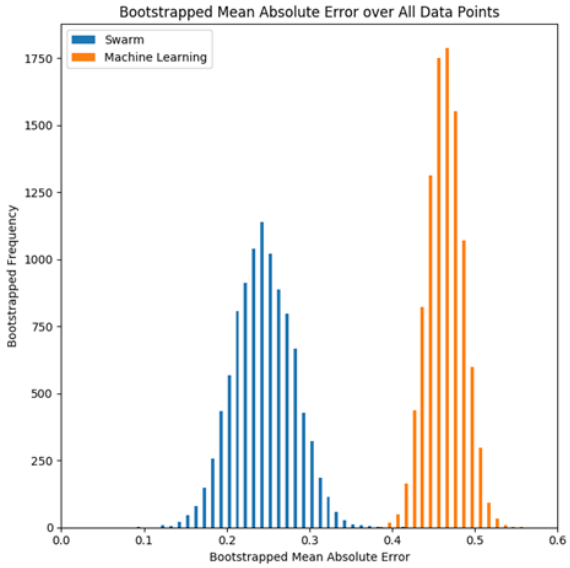

Figure 5b. Mean Absolute Error (ASI vs ML)

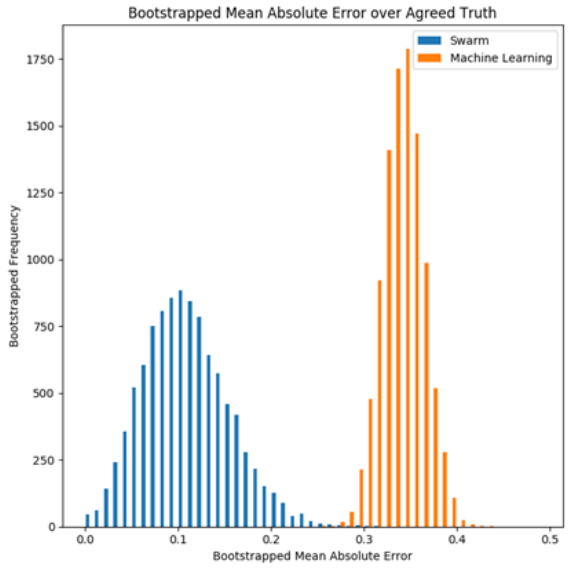

Figure 5c. Mean Absolute Error (ASI vs ML) using Agreed Truth

ROC Analysis: Because the Swarm AI system and the Machine Learning system have different approaches to probabilistic forecasting, a ROC analysis was performed to compare the true positive rate to the false positive rate across different cut-off points, the higher the ratio the better the

classification. We computed the Area Under the ROC Curve (AUROC) for both methods and found that the swarm of radiologists achieved an AUROC of 0.906, while the ML system achieved 0.708. Bootstrapping across 10,000 trials, we find that the ASI system scores significantly higher than the pure ML system ($p<0.01$, $\mu_{difference} = 0.197$), as shown in Figure 5d.

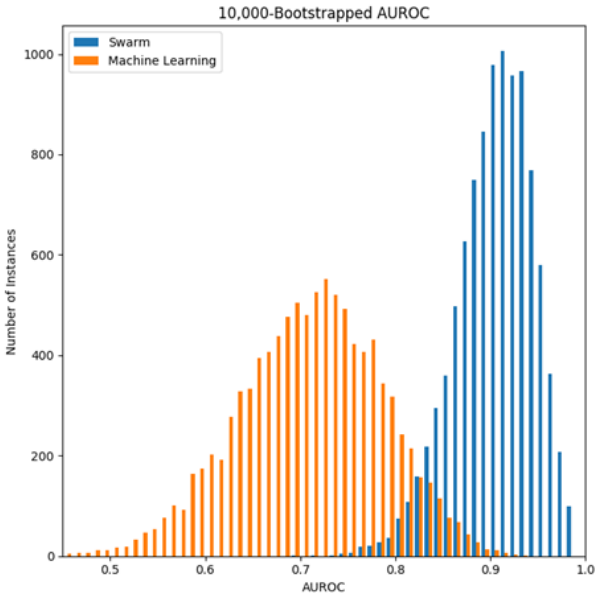

Figure 5d. AUROC Analysis (ASI vs ML)

We can also compare the ASI system to the ML system by plotting Receiver Operating Characteristic (ROC). As shown in Figure 5e below, the swarm outperforms the ML system across most discrimination levels, with higher true positive rates for each false positive. In fact, the swarm is able to find all instances of pneumonia in this dataset, while only mis-identifying 40% of the non-pneumonia cases. The AUROC of the ASI system is 0.91, while that of ML is 0.71 for this dataset.

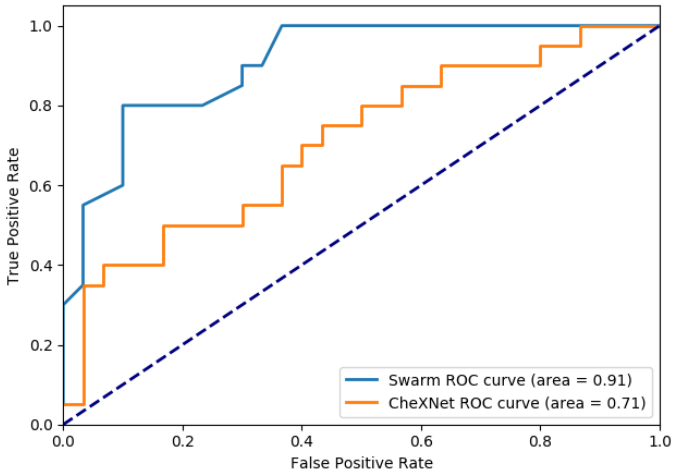

Figure 5e. Receiver Operating Characteristic (ASI vs ML)

F1 Score: We compare the performance of the ASI system to ML system on the F1 metric. As shown in Figure 5f below, we find that the ASI system achieves an average F1 score of 0.75 while the ML system achieves a lower average F1 score of 0.63. To assess significance, we bootstrap across 10,000 samples and find insufficient difference ($p > 0.05$), suggesting that a set of 50 trials was too small for significant F1 comparison.

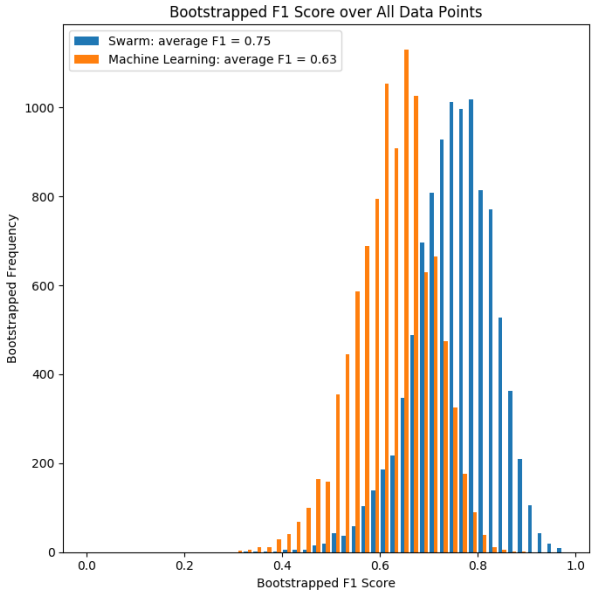

Figure 5f. Bootstrapped F1 Scores (ASI vs ML)

VI. CONCLUSIONS

We compared the ASI system to both individuals and to the state-of-the-art in ML diagnosis of chest X-rays for the presence of pneumonia and found that the hybrid ASI system that combined real-time human diagnosis and software optimization significantly outperformed both the individuals working on their own and a pure software system when compared with respect to (i) binary classification, (ii) mean absolute error, and (iii) ROC analysis. Because Ground Truth could be error prone, we also compared using "Agreed Truth" and still found the ASI system to outperform the ML system. Previous studies on the CheXNet system on a larger set of cases [1] achieved a higher AUROC (0.7680) as compared to 0.708 in this study, indicating that the 50 questions in this test set may be harder than average. Additional research is warranted using more definitive Ground Truth and a wider range of cases. In addition, the method for collecting individual responses in this study used only five levels of probability (0-20%, 20-40%, 40-60%, 60-80% and 80-100%). Future research should be performed that utilizes a higher resolution method for individual response mechanism.

Overall this study suggests that swarm-based technologies are quite promising for use in medical diagnosis, enabling small groups of medical professionals to combine their insights in real-time under software moderation and thereby achieve diagnostic accuracies that significantly exceed the accuracies of individual human practitioners as well as software-only solutions. It is likely that the ASI system excels in certain types of cases, while the software-only ML system excels in others. We believe future research should identify these differences so that each method can be applied to those cases which are most appropriate.

ACKNOWLEDGMENT

Thanks to Unanimous AI for use of the *Swarm.ai* platform and Stanford University School of Medicine for datasets used in diagnostic evaluations. Special thanks to additional medical contributors including Dr Bhavik Patel, Dr Jayne Seekins, Dr Francis Blankenberg, Dr David Mong, Dr Timothy Amrhein, Dr Pranav Rajpurkar, Dr David Larson, Dr Jeremy Irvin, Robyn Ball, Dr Curtis Langlotz, and Dr Evan Zucker.

REFERENCES

[1] P. Rajpurkar et al. (Dec. 2017). ‘‘CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning.’’

[2] V J Mar, H P Soyer. Artificial intelligence for melanoma diagnosis: How can we deliver on the promise? Annals of Oncology, 2018; DOI: 10.1093/annonc/mdy193

[3] H A Haenssle, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, L Uhlmann. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 2018;

[4] De Fauw et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease". Nature Medicine (2018),

[5] Siddhartha Mukherjee, A.I. Versus M.D., THE NEW YORKER, Apr. 3, 2017, http://www.newyorker.com/magazine/2017/04/03/ai-versus-md (last visited August 12, 2018).

[6] Rosenberg, L.B., “Human Swarms, a real-time method for collective intelligence.” Proceedings of the European Conference on Artificial Life 2015, pp. 658-659

[7] Rosenberg, Louis. “Artificial Swarm Intelligence vs Human Experts,” Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE.

[8] Rosenberg, Louis. Baltaxe, David and Pescetelli, Nicollo. "Crowds vs Swarms, a Comparison of Intelligence," IEEE 2016 Swarm/Human Blended Intelligence (SHBI), Cleveland, OH, 2016, pp. 1-4.

[9] Baltaxe, David, Rosenberg, Louis and N. Pescetelli, “Amplifying Prediction Accuracy using Human Swarms”, Collective Intelligence 2017. New York, NY; 2017.

[10] L. Rosenberg, N. Pescetelli and G Willcox, "Artificial Swarm Intelligence amplifies accuracy when predicting financial markets," 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), New York City, NY, 2017, pp. 58-62.

[11] Seeley T.D, Buhrman S.C 2001 “Nest-site selection in honey bees: how well do swarms implement the ‘best-of-N’ decision rule?” Behav. Ecol. Sociobiol. 49, 416–427

[12] Marshall, James. Bogacz, Rafal. Dornhaus, Anna. Planqué, Robert. Kovacs, Tim. Franks, Nigel. “On optimal decision-making in brains and social insect colonies.” Soc. Interface 2009.

[13] Seeley, Thomas D., et al. "Stop signals provide cross inhibition in collective decision-making by honeybee swarms." Science 335.6064 (2012): 108-111.

[14] Seeley, Thomas D. Honeybee Democracy. Princeton Univ. Press, 2010.

[15] Seeley, Thomas D., Visscher, P. Kirk. “Choosing a home: How the scouts in a honey bee swarm perceive the completion of their group decision making.” Behavioral Ecology and Sociobiology 54 (5) 511-520.

[16] Usher, M. McClelland J.L 2001 “The time course of perceptual choice: the leaky, competing accumulator model.” Psychol. Rev. 108, 550–592