2018 SIIM Conference on
Machine Intelligence in Medical Imaging
September 9-10, 2018 | Hilton SF Financial District | San Francisco, CA

Radiology SWARM: Novel Crowdsourcing Tool for CheXNet Algorithm Validation

Safwan Halabi, MD, Stanford University; Matthew Lungren, MD, MPH; Louis Rosenberg, PhD; David Baltaxe; Bhavik Patel, MD; Jayne Seekins, MD; Francis Blakenberg, MD; David Mong, MD; Timothy Amrhein, MD; Pranav Raipurkar, MS; David Larson, MD, MBA; Jeremy Irvin; Robyn Ball; Curtis P. Langlotz, MD, PhD, FSIIM; Gregg Willcox

Introduction

Researchers at Stanford University School of Medicine and Unanimous AI conducted a study in which a "swarm" of radiologists (i.e. a group connected by Swarm AI algorithms) reviewed a set of 50 chest x-rays and for each predicted the likelihood that the patient has pneumonia. The predictive accuracy of the Swarm AI system was then compared to that of the machine learning program CheXNet, which has been shown in prior studies to significantly outperform individual human radiologists in pneumonia screening tasks. Thus, while previous research shows that a software-only solution like CheXNet can outperform individual radiologists, the current study explores if small groups of radiologists, when networked together as a real-time collaborative system moderated by AI algorithms, can amplify their collective accuracy to levels that rival or exceed the current state-of-the-art in purely algorithmic diagnosis.

Methods

A group of eight radiologists, at separate locations and connected by Swarm AI algorithms, were tasked with diagnosing a set of 50 chest x-rays as a unified system (i.e. as an "intelligent swarm"). For each of the 50 trials, a chest x-ray was presented simultaneously to all eight radiologists. After a few seconds of individual assessment, the group worked together as a swarm, converging on a probabilistic diagnosis to reflect the likelihood that the patient has pneumonia. In this way, a set of 50 probabilities were generated for the 50 test cases. Separately, the same set of 50 chest x-rays were run through the CheXNet software algorithm, a state-of-the-art 121-layer convolutional neural network, to generate probabilities as to whether each patient has pneumonia. These two sets of probabilities were then scored against ground truth and compared using a variety of statistical techniques.

Results

We compared the performance of the Swarm AI system, which uses a small group of human radiologists connected by swarm intelligence algorithms, against the software-only CheXNet system. We assess the two methods across three different performance metrics – (i) binary classification accuracy, (ii) Mean Absolute Error, and (iii) ROC analysis.

i. **Binary Classification**: Using fifty-percent probability as the cutoff for classifying a positive diagnosis, the CheXNet system achieved 60% diagnostic accuracy against Ground Truth across the 50 test cases, while the Swarm AI system achieved 82% accuracy across the same 50 cases. To assess statistical significance, bootstrap analysis was performed on 10,000 samples, as shown in Figure 1a. The swarm was found to be significantly more accurate in binary classification than the ML system ($p<0.01$, $\mu_{difference} = 21.9\%$).

ii. **Mean Absolute Error**: MAE is calculated as the absolute value of the Ground Truth minus the Predicted Probability. A bootstrap analysis of MAE revealed that the swarm of radiologists had significantly higher probabilistic accuracy than the ML system ($p<0.001$, $\mu_{difference} = 21.6\%$), as shown in Figure 1b. To address the possibility that Ground Truth could be error prone, we also looked at "Agreed Truth", defined as only those cases where the Swarm AI system and the CheXNet system 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 1c.

iii. **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 Swarm AI system scores significantly higher than the pure ML system ($p<0.01$, $\mu_{difference} = 0.198$), as shown in Figure 1d.

Discussion

We compared the state-of-the-art in ML diagnosis of chest x-rays with a hybrid system comprised of eight radiologists connected by Swarm AI algorithms and found that the swarm significantly outperformed the 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 Swarm AI system to outperform. Additional research is warranted using more definitive Ground Truth and a wider range of cases. It is likely that the Swarm AI system excels in certain types of cases, while the ML system excels in others. We believe future research should identify these differences, so each method can be applied to those cases which are most appropriate.

Conclusion

We find that a Swarm AI system that combines real-time human input with intelligence algorithms is significantly more accurate in diagnosing pneumonia than a state-of-the-art software-only ML system. This suggests that Swarm AI may be a powerful tool for establishing Ground Truth for use in training and for validating machine learning systems.

Keywords

artificial intelligence; radiology; machine learning; pneumonia; chest radiography; crowdsourcing

Figures: Figure 1a (top left): Bootstrapped Percent Correct, Figure 1b (top right): Bootstrapped Mean Absolute Error; Figure 1c (bottom left) Bootstrapped Mean Absolute Error for Agreed Truth; Figure 1d (bottom right): Bootstrapped AUROC Analysis.

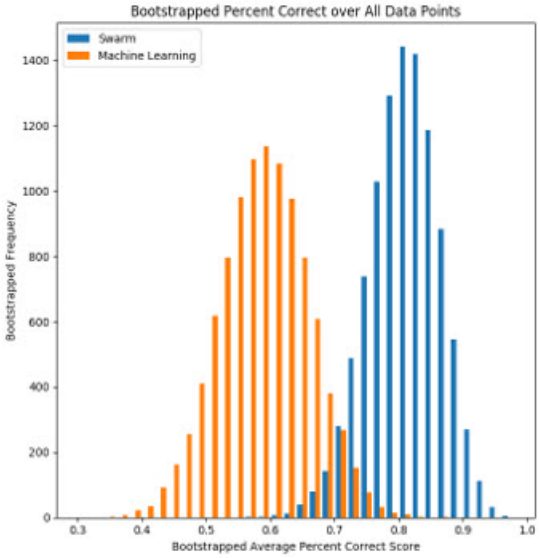

Figure 1a

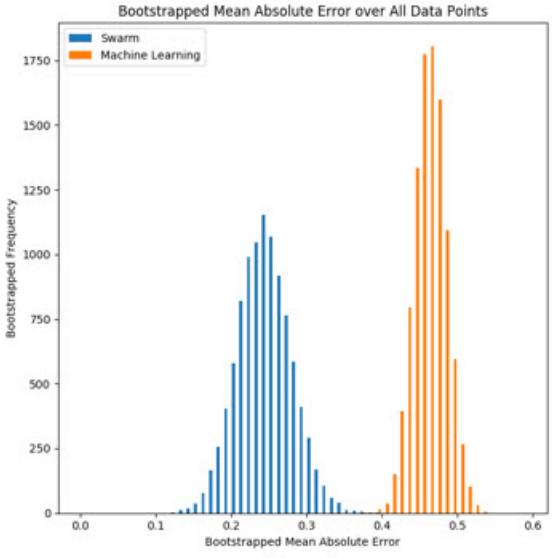

Figure 1b

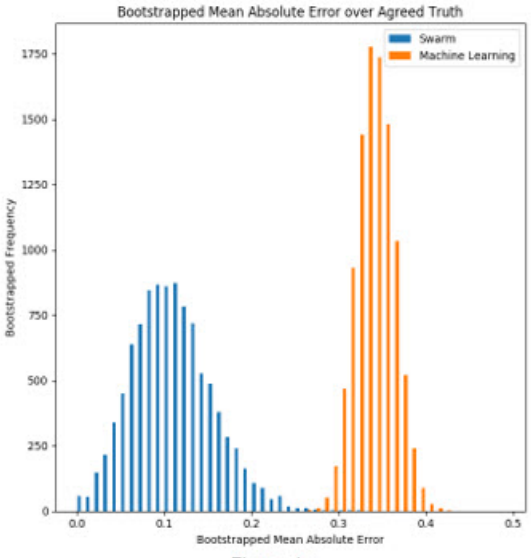

Figure 1c

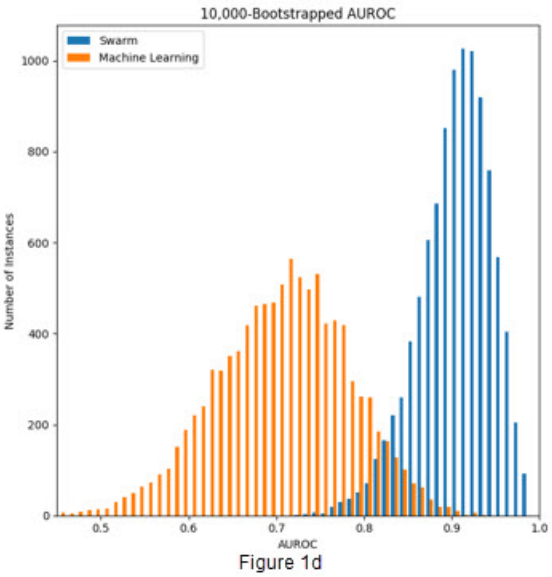

Figure 1d