Radiology SWARM: Novel Crowdsourcing Tool for CheXNet Algorithm Validation

Safwan Halabi MD - @SafwanHalabi
Clinical Associate Professor of Radiology
Stanford University School of Medicine
Disclosures

- None
Introduction

- Researchers at Stanford University 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 radiographs and predicted the likelihood that pneumonia was present.

- Predictive accuracy of the Swarm AI system was then compared to that of the machine learning program CheXNet, which has been shown to outperform individual human radiologists in pneumonia screening tasks.

- This 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 algorithmic diagnosis.
Swarm Intelligence
(AI Algorithms)

Human Intelligence
- Knowledge
- Wisdom
- Insights
- Intuition

More Accurate
- Predictions
- Decisions
- Evaluations
- Forecasts
Swarm Intelligence

Nature has evolved methods for amplifying the Intelligence of groups

- Nature does not...
  - take polls, conduct surveys, or take votes
  - analyze historical data-sets (big data)

- Nature forms SYSTEMS
  - Real-Time Systems that think together as a group
  - Feedback loops converge on optimal solutions
  - Group → SMARTER TOGETHER than alone
Can Humans Swarm?

Natural Swarms

Artificial Swarms

#CMIMI18
Methods

- Eight board certified radiologists, at separate locations and connected by Swarm AI algorithms, were tasked with identifying the presence of pneumonia from a set of 50 chest radiographs as a unified system (i.e. as an “intelligent swarm”)

- For each of the 50 trials, a chest radiograph 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
Methods

T2. What is the probability that this patient has pneumonia?

https://bmjopen.bmj.com/content/8/1/e017964  #CMIMI18
Methods

- 50 probabilities were generated for the 50 test cases
- Separately, the same set of 50 chest radiographs were run through the CheXNet software algorithm, a state-of-the-art 121-layer convolutional neural network, to generate probabilities as to whether or not pneumonia was present on each image
- These two sets of probabilities were then scored against ground truth* and compared using a variety of statistical techniques

*Ground truth was determined by expert chest radiologist annotation of the set of 50 chest radiographs
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:
  - Binary classification accuracy
  - Mean Absolute Error (MAE)
  - ROC analysis
Results: Binary Classification

- Using 50% 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.

- The swarm was found to be significantly more accurate in binary classification than the ML system ($p<0.01$, $\mu$ difference = 21.9%).
Results: Binary Classification

Bootstrapped Percent Correct over All Data Points

+22% Accuracy
Results: Mean Absolute Error (MAE)

- 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%).
Results: Mean Absolute Error (MAE)

Bootstrapped Mean Absolute Error over All Data Points

- Swarm
- Machine Learning

#CMIMI18
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, μ difference = 21.3%)
Results: Mean Absolute Error (MAE)
Since 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.

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, μ difference = 0.198).
Results: ROC Analysis
Comparison of a state-of-the-art ML diagnosis of chest radiographs with a hybrid system comprised of eight radiologists connected by Swarm AI algorithms found that the swarm significantly outperformed the pure software system with respect to: (1) binary classification; (2) mean absolute error; and, (3) 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.

Future research should identify these differences so each method can be applied to those cases which are most appropriate.
Conclusion

- The 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.
Acknowledgements

Unanimous AI
- Louis Rosenberg PhD
- David Baltaxe
- Gregg Willcox
- Mimi Lyons

Stanford University / Radiologists
- Matthew Lungren MD PhD
- Bhavik Patel MD
- Jayne Seekins MD
- Francis Blankenberg MD
- David Mong MD
- Timothy Amrhein MD
- David Larson MD MBA
- Curt Langlotz MD PhD
- Jeremy Irvin
- Robin Ball