# On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy

Sagalpreet Singh    Shweta Jain    Shashi Shekhar Jha

Indian Institute of Technology Ropar

AAMAS 2023

# Introduction

# Motivation

- Highly accurate and fully autonomous AI models
  - Huge amount of data
  - Better algorithms
- Hybrid approaches are preferred in some situations
  - Improving trust between humans and machines
  - Humans and machines make different types of errors

# Motivation

- Empirical evidence that humans and AI classifiers do not make the same types of errors in image classification
  - Geirhos et al.[1]; Rosenfeld et al.[2]; Serre[3]
- Hybrid human-machine teams have been effective in real world applications
  - Speech Transcription: Gaur et al.[4]
  - Face Identification: Philips et al.[5]
  - Clinical Radiology: Rajpurkar et al.[6]

# Human-AI Team

- AI assisted decision-making
  - Humans and AI models work as a team
  - Ensure overall model's productivity
- Ways to collaborate
  - Deferred model
  - Combined model

# Combining Predictions

- How can the labels from multiple humans be combined with probabilistic output of an AI model?
- Can the combined model with multiple humans lead to better accuracy as compared to a single human and AI model?
- How to intelligently select humans so as to improve the accuracy of the combined model?

# Methodology

# Problem Setting

- $k$-way classification problem
- Label for each instance by every human
- Class conditioned probabilities from AI model

# Single Human Prediction & AI Model

true label of an instance $x$

also denoted by $\phi_{h(x),y(x)}$

also denoted by $m_{y(x)}(x)$

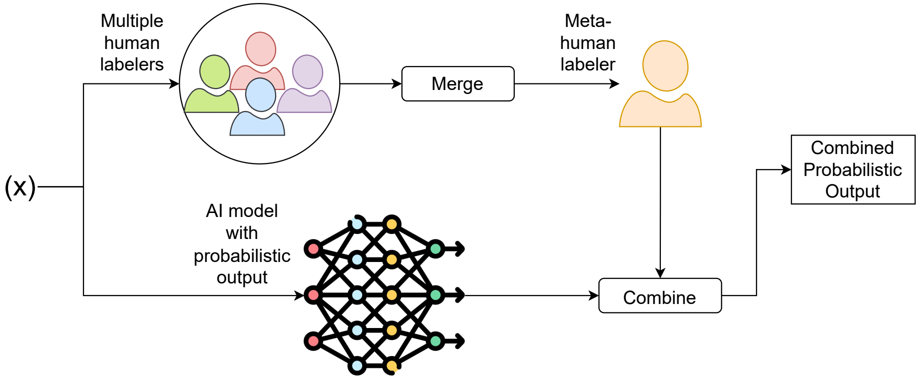
$$p(\,y(x)\mid h(x)\,,\,m(x)\,) \propto p(h(x)|y(x))\;\; p(y(x)|m(x))$$ [1]

prediction by human on instance $x$

model probability on instance $x$

[1] Kerrigan et al., Combining human predictions with model probabilities via confusion matrices and calibration. NeurIPS 2021

# Naïve Combination Methods



# Merge Heuristics

- **Best Human**: The one with the highest accuracy
- **Best Majority Human**: The one with the highest accuracy among those whose prediction is same as the mode of all the predictions
- **Best Weighted-Majority Human**: The one with the highest accuracy among those whose prediction is same as the accuracy-weighted mode of all the predictions

# Merge Heuristics

- **Best Human**: The one with the highest accuracy
- **Best Majority Human**: The one with the highest accuracy among those whose prediction is same as the mode of all the predictions
- **Best Weighted-Majority Human**: The one with the highest accuracy among those whose prediction is same as the accuracy-weighted mode of all the predictions

All these methods are sub-optimal in terms of the accuracy

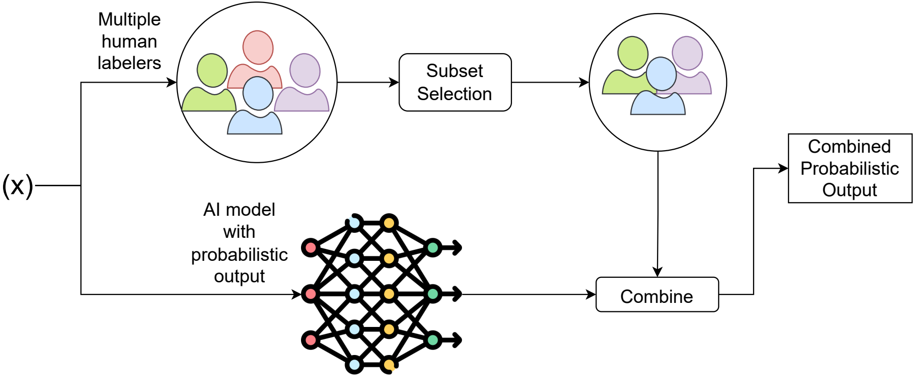
# ComHAI

- Assuming that all human predictions are independent of each other:

$$p(y(x) = j|\boldsymbol{h}(x) = \{l_1, l_2, \ldots, l_n\}, m(x)) = \frac{m_j(x)\prod_{i\in[n]}\phi^{[i]}_{l_i j}}{\sum_{k=1}^{\mathcal{K}} m_k(x)\prod_{i\in[n]}\phi^{[i]}_{l_i k}}$$

- Should labels from all humans be considered?
  - Answer is NO!
  - Accuracy is not monotone in terms of the number of humans considered
  - Remodel the problem as subset selection

# Subset Selection of Humans



# Lower Bound on Accuracy

$$\mathbb{E}[\mathbb{1}(c(x) = y(x))] \geq \mathbb{P}\left\{\prod_{i\in[n]} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}} > \frac{1-m_{y(x)}(x)}{m_{y(x)}(x)}\right\}$$

- To maximize the lower bound, maximize the following term:

$$\prod_{i\in[n]} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}}$$

# Subset Selection of Humans

- **True LB Subset Selection**: Select optimal subset of humans that maximize the true lower bound on accuracy, where the optimal subset, $S^*$ is given as:

$$S^* = \arg\max_{S} \left( \prod_{i \in S} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1 - \phi^{[i]}_{h_i(x)y(x)}} \right)$$

# Subset Selection of Humans

- **True LB Subset Selection**: Select optimal subset of humans that maximize the true lower bound on accuracy, where the optimal subset, $S^*$ is given as:

$$S^* = \arg\max_{S} \left( \prod_{i \in S} \frac{\phi^{[i]}_{h_i(x) y(x)}}{1 - \phi^{[i]}_{h_i(x) y(x)}} \right)$$

$\boxed{y(x) \text{ is not known}}$

# Subset Selection of Humans

- **True LB Subset Selection**: Select optimal subset of humans that maximize the true lower bound on accuracy, where the optimal subset, $S^*$ is given as:

$$S^* = \arg\max_{S} \left( \prod_{i \in S} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1 - \phi^{[i]}_{h_i(x)y(x)}} \right)$$

$\boxed{y(x) \text{ is not known}}$

- **Pseudo LB Subset Selection**: Select pseudo-optimal subset of humans that maximize the pseudo-lower bound on accuracy, where the pseudo optimal subset, $S^{**}$ is given as:

$$S^{**} = \arg\max_{S} \max_{1 \leq j \leq \mathcal{K}} \left( \prod_{i \in S} \frac{\phi^{[i]}_{h_i(x)j}}{1 - \phi^{[i]}_{h_i(x)j}} \right)$$

# Subset Selection Algorithm

- Subset selection is exponential in the number of humans
- Greedy choice exists: For any $j \in [\mathcal{K}]$, only those human satisfying the following constraint should be selected:

$$\frac{\phi^{[i]}_{h_i(x)j}}{1 - \phi^{[i]}_{h_i(x)j}} > 1$$

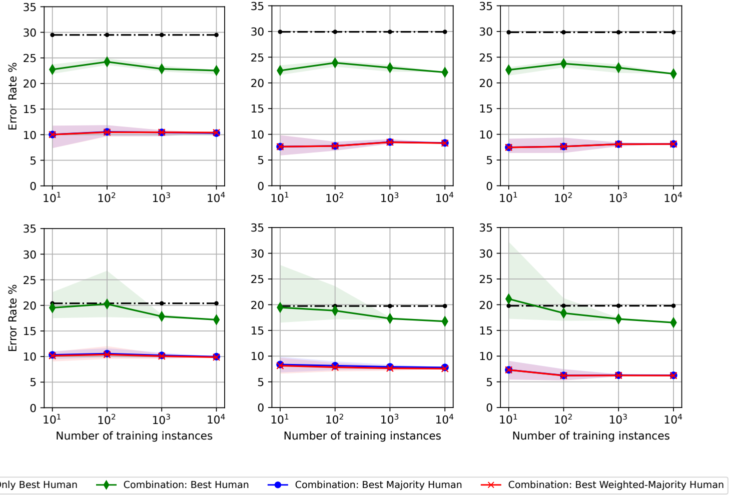
- This leads to a polynomial time algorithm

# Experiments and Results

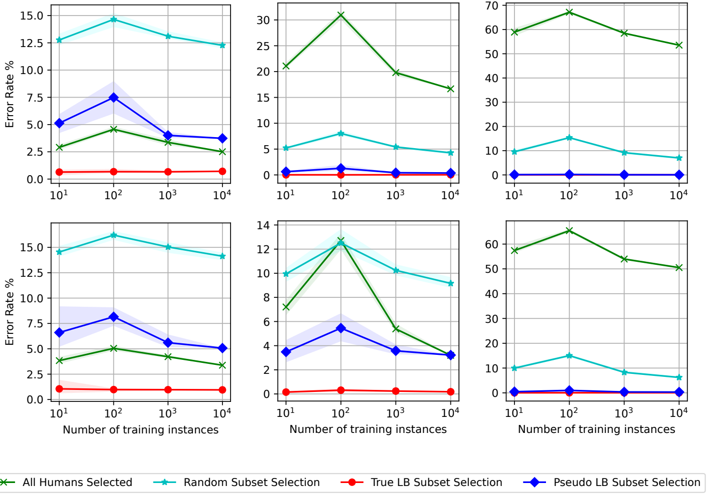
# Experimental Setup

- CIFAR-10H dataset
  - Simulate human labels by sampling from underlying distribution
- AI models with class conditioned probabilistic output
  - Custom CNN: 56.74% accuracy
  - ResNet-110: 93.89% accuracy

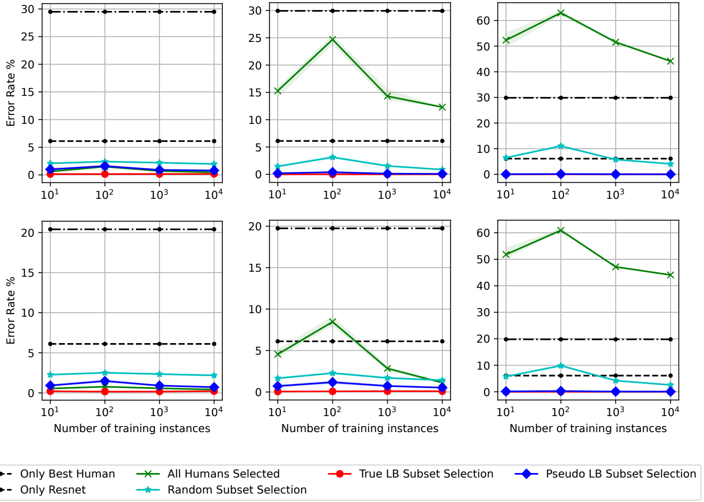
# Naïve Combination and Custom CNN



# ComHAI and Custom CNN



# ComHAI and ResNet-110



# References I


[1] R. Geirhos, K. Meding, and F. A. Wichmann, "Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 890–13 902, 2020.

[2] A. Rosenfeld, M. D. Solbach, and J. K. Tsotsos, "Totally looks like-how humans compare, compared to machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1961–1964.

[3] T. Serre, "Deep learning: The good, the bad, and the ugly," *Annual review of vision science*, vol. 5, no. 1, pp. 399–426, 2019.

[4] Y. Gaur, F. Metze, Y. Miao, and J. P. Bigham, "Using keyword spotting to help humans correct captioning faster," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

# References II


[5] P. J. Phillips, A. N. Yates, Y. Hu, *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018.

[6] P. Rajpurkar, C. O'Connell, A. Schechter, *et al.*, "Chexaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv," *NPJ digital medicine*, vol. 3, no. 1, p. 115, 2020.

# Thank you!


Reach us

sagalpreet60@gmail.com shwetajain@iitrpr.ac.in shashi@iitrpr.ac.in