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

Sagalpreet Singh
Indian Institute of Technology Ropar
Rupnagar, India
2019csb1113@iitrpr.ac.in

Shweta Jain
Indian Institute of Technology Ropar
Rupnagar, India
shwetajain@iitrpr.ac.in

Shashi Shekhar Jha
Indian Institute of Technology Ropar
Rupnagar, India
shashi@iitrpr.ac.in

ABSTRACT

There are several classification tasks where neither the human nor the model is perfectly accurate. Some recent works, therefore, focus on the Human-AI team model, where the AI model's probabilistic output is combined with the human-predicted class label. The combined decision is shown to consistently outperform the model's or human's accuracy alone. All the previous works, however, restrict to the setting where they consider a single human to combine with the AI model. Motivated by the crowdsourcing literature, which combines labels from multiple humans, we show that combining multiple human labels with the model's probabilistic output can lead to significant improvement in accuracy. This paper further shows that while combining multiple humans helps, a naive combination of humans with AI model can lead to poor accuracy. Hence, there is a strong need for an intelligent strategy to select a subset of humans and combine their labels. To this end, we present an approach to merge the predicted labels from multiple humans with the model's probabilistic output. We then provide an efficient algorithm to find the optimal subset of humans whose combined labels offer the most accurate output. Finally, we empirically demonstrate that the combined model outperforms the AI model or any human alone in terms of accuracy. Besides this, our subset selection algorithm and combination method outperforms the single human model and other naïve combination techniques.

CCS CONCEPTS

• Human-centered computing; **• Computing methodologies → Supervised learning**;

KEYWORDS

Human-AI Team; Subset Selection; Confusion Matrix; Combining Predictions; K-way classification

ACM Reference Format:
Sagalpreet Singh, Shweta Jain, and Shashi Shekhar Jha. 2023. On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

The advancements in Artificial Intelligence (AI) have shown exemplary improvements in the performance of various tasks. AI-based systems are slowly becoming an intricate part of domains such as health care, finance, job recruitment, cyber-security, criminal justice, intrusion detection, and many more. With more data and better algorithms, most of the AI research is focused on developing highly accurate models so that these models can be used autonomously. Hence to harness the power of AI in high-stakes domains, the adoption of AI is moving towards the paradigm of *AI-assisted decision-making*. In this paradigm, humans and AI models work as a team, also referred to as *Human-AI team* to make accurate decisions on highly complex tasks efficiently while ensuring the overall team's productivity. The hybrid Human-AI approaches are being favoured for various reasons on several machine learning tasks [12, 14, 18, 26, 32, 33, 36]. In literature, the Human-AI team collaboration has been considered through two different perspectives. One is in the form of deferred outputs [16, 22, 24] where the objective is to maximize accuracy by learning a deferred model that predicts whether the incoming instance should be deferred to the human or not. Typically, humans are considered experts with high costs; however, they provide the correct answer with a very high probability. There are two issues with this approach, first, in many applications, humans may not be available with high accuracy, and second, the AI model is trained from scratch alongside human in order to learn the deferred model [21, 22, 24]. With high human costs, training an AI model alongside human is a very costly proposition. Instead, in this paper, we consider the second approach, wherein human works alongside a trained AI model to make the decision in a combined manner. Such an approach does not require high expertise of human, so an independently trained AI model can be readily used without much fine-tuning to achieve better accuracy of the combined outputs.

Research has shown that humans and AI models make different types of errors [6, 27, 29], which necessitates to developing an approach for combining the predictions from humans and the AI model. This has been explored in [15] for combining class-level prediction from a human with probabilistic output from an AI model. Their work shows that the combined model achieves much better accuracy, but they limit themselves to combining a single human's predictions with the AI model's output. Limiting to one human-predicted label may pose a significant restriction on the accuracy of the combined approach as the combination may get biased with the accuracy of an individual. In order to achieve high combined accuracy, in this paper, we explore the combinations of multiple humans with varying levels of expertise on the task.

Methods to combine predictions from multiple models are an active area of research [4, 17, 19, 28]. Extensive work combining predictions from multiple humans show that diverse combinations outperform any individual alone [10, 11, 20]. Therefore, it is natural to ask the following questions:

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

- How can the labels from multiple humans be combined together with that of the 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?

This paper addresses all these questions.

We propose a non-trivial method ComHAI to combine multiple human labels with the AI model's probabilistic outputs. We show that using ComHAI, the resulting accuracy of the combined model is significantly better than the case when a single human prediction is used, or other naive methods for combining predictions are used, such as taking the mode of the human predictions as a single human label. We consider a $\mathcal{K}$-way classification problem, with each human making predictions of the class labels given the input instance and the classification AI model providing class-conditioned probabilities. We further show that the accuracy is non-monotone in terms of the number of humans considered, even if the accuracy of every human is more than 50%. Hence, it is important to find the subset that results in the maximum accuracy given the instance. We then propose our algorithm called as GreedySubsetSelection that provides the best subset of humans whose output labels, when combined with AI model's probabilistic output leads to maximum accuracy. In particular, our key contributions are three-fold:

- We propose an approach to combine the predicted class labels from multiple humans with the probabilistic output of an AI model.
- We empirically validate our approach on the CIFAR-10H image classification dataset and show that Human-AI combinations are more accurate than any individual human, the model alone, or the model with a single human.
- We provide an efficient algorithm to select a subset of humans whose class-level outputs combined with probabilistic output of the AI model leads to maximum accuracy.

In the rest of the paper, we first discuss the state-of-art in the Human-AI team modeling in Section 2. We provide a set of preliminaries in Section 3. Next, we discuss our proposed ComHAI in Section 4 for combining the outputs from multiple humans with the AI model and a strategy to select the subset of humans in order to improve the overall accuracy of the combined model. Section 5 presents the experiments and results using the CIFAR-10H dataset. Finally, Section 6 discusses the scope of further improvements and the conclusions.

2 EXISTING MODELS FOR HUMAN-AI TEAMS

As discussed earlier, humans and AI models make different kinds of errors on decision-making tasks. On one hand, the humans bring their experience to make decisions on a task, while the AI model brings the common collective knowledge of the population in terms of its training from a collected dataset. Further, the AI model uses a level of abstraction that might be different from those of humans. Hence, various researchers have looked into ways to combine AI and human decisions such that they complement each other. A standard method in this context is to pass on all the instances to humans wherever model has low confidence [9]. However, in [22], the authors show that even in instances with low model confidence, the human may not always produce correct outputs. Hence, they extend the rejection learning approach into a framework called learning to defer, where a defer model is learned to pass an instance either to the AI model or to human for decisions. This deferred approach for Human-AI teaming has been studied extensively in recent years from different perspectives [5, 24, 35] . Keswani et al. [16] discuss an approach to defer an instance where the model has low confidence to multiple humans by modeling their individual expertise and biases. In [21], the authors discuss the limitations of the deferred approach for Human-AI teams. One significant issue is that the AI model is specialized on the instances of high confidence. This limits the AI model to generalize for all instances and makes it difficult to update the AI model if the data distribution changes.

In another approach to Human-AI team modeling, instead of deferring, the AI model's outputs are combined with that of the human to generate the final outcome. Bansal et al. [1] use an AI-assisted setting wherein the human can choose to either accept the AI's recommendation or solve for the same. The authors propose to train the AI model by directly optimizing the Human-AI team's performance. The same authors in [2] discuss the role of understanding of the AI's error boundary by human's mental model for developing effective Human-AI team with complementary functions. In Träuble et al. [31], the authors define the problem of prediction updates in AI/ML models and present a probabilistic approach for backward-compatible prediction updates. The authors highlight the importance of such backward compatible prediction updates in AI assisted tasks. Martinez et al. [23] extend the AI assisted framework to multiple humans with personalized loss functions for specific users in order to increase the performance of Human-AI teams with the personalized compatibility. The focus in these approaches is to optimize the AI model's performance considering how humans would interact with the model by navigating the performance-compatibility trade-off.

The above-mentioned approaches necessitate the training of the AI model in the presence of human decision maker. However, few approaches [15, 16] focus on combining the human decision on a given instance with the output of an independently trained AI model in order to improve the overall accuracy. In particular, Kerrigan et al. [15] present how the probabilistic output of the AI model can be combined with the class-level output of a human to improve the overall team's accuracy. Our paper extends the state-of-art in this line of work by considering the class-level outputs of multiple humans of varying expertise to combine with the probabilistic output of an independently trained AI model. Combining decisions from multiple humans is specifically studied in the crowdsourcing domain [11, 20], to find the optimal subset of workers for a task in order to optimize the performance of the overall decisions. Our work takes inspiration from the crowdsourcing domain to select a subset of humans in order to improve the accuracy of a combined decision model of the Human-AI team. Though inspired by crowdsourcing literature, we emphasize that combining the probabilistic output of an AI model with multiple humans brings in many non-trivial challenges. Combining outputs only from multiple humans (not AI model) leads to simple error functions coming from majority voting or weighted majority voting [11]. Whereas the combined accuracy function from multiple humans and AI model's probabilistic output does not turn out to be sub-modular and monotone, making the design of subset selection algorithm non-trivial.

3 PRELIMINARIES

We consider a k-way classification problem to predict label $y \in \mathcal{Y} = \{1, \ldots, k\}$ for a given feature vector $x \in \mathcal{X}$. For this task, we represent the AI model by $\mathcal{M}$ and let $m(x)$ denote the k-dimension normalized probability vector output of $\mathcal{M}$. We further assume that for each task, we have the labels from n humans. We denote i^{th} human as h_i, and $h_i(x)$ denotes the prediction made on an instance x by the i^{th} human. Note that humans provide the hard labels i.e. $h_i(x) \in \{1, 2, \ldots, k\}$. Also, let $h(x)$ denote the collection $\{h_1(x), h_2(x), \ldots, h_n(x)\}$. The prediction made by our combined model on an instance x is denoted by $c(x)$. Finally, the ground truth label of x is denoted by $y(x)$ which needs to be predicted. We posit that the combined model depends on two important measures, one, the accuracy of the model $\mathcal{M}$, and second, the confusion matrix of the humans. Let $\phi^{[i]}$ denote the estimated confusion matrix corresponding to the i^{th} human such that $\phi^{[i]}_{st} = p(h_i(x) = s | y(x) = t)$. In the next section, we explain how the confusion matrix of each human can be estimated given their labels.

3.1 Confusion Matrix of Humans

One possibility to estimate the confusion matrix is via the maximum likelihood estimate given as:

$$\phi^{[i]}_{st} = \frac{\sum_{x \in \mathcal{X}} \mathbb{I}(h_i(x) = s \wedge y(x) = t)}{\sum_{x \in \mathcal{X}} \mathbb{I}(y(x) = t)}$$

where $\mathbb{I}(.)$ is an indicator function. However, this method requires human labels on too many examples, which leads to poor accuracy when k is large. Previous works [3, 15, 34] have shown that instead of using the maximum likelihood estimate for $\phi^{[i]}_{st}$, using a Dirichlet prior over each column of the confusion matrix leads to more efficient estimation. This needs less number of samples to estimate the confusion matrix as we have incorporated additional information by taking the Dirichlet prior. Hence,

$$\phi^{[i]}_{st} = Dirichlet(\alpha_t)_s, \quad \forall t$$

The prior parameter $\alpha_t \in \mathbb{R}^{\mathcal{K}}$ is chosen such that

$$(\alpha_t)_k = \begin{cases} \beta, \ k \neq t \\ \gamma, \ k = t \end{cases}$$

where $\beta, \gamma \in \mathbb{R}_+$. The resultant prior matrix thus has γ along the diagonal and β on the off-diagonal. We choose a Dirichlet prior over each column such that all the off-diagonal prior values are equal. The posterior estimate is obtained by conjugacy.

3.2 Calibrating Model Probabilities

Neural networks tend to be overconfident in their predictions, particularly for the classification task where the output is a member of probability simplex. Thus, post-prediction calibration can be used to map $m(x)$ to $m^{\theta}(x)$ [7] where θ denotes the calibration parameters. Similar to [7, 15], we use the Bayesian version of temperature scaling for calibrating model probabilities. We assume a Gaussian prior on the log-temperature, $logT \sim \mathcal{N}(\mu, \sigma)$ with $\mu = 0.5$ and $\sigma = 0.5$ [15]. The maximum a posteriori is estimated by gradient-based optimization since the prior is non-conjugate. Existing works [15] have shown that this approach is effective in such a setting.

3.3 Single Human Prediction & Model Probabilities

We consider the work by Kerrigan et al. [15] as a baseline to compare our combination method. Their work establishes a framework for combining predictions from a single human with model probabilities. We briefly describe their combination method. The training comprises of learning temperature scaling parameter and estimating the confusion matrix for human as described above. Then the prediction from human h are combined with model probabilities m using the following relation:

$$p(y(x)|h(x), m(x)) \propto p(h(x)|y(x))p(y|m(x))$$

Our approach is more generalized allowing predictions from multiple humans to be combined with the model probabilities. Further, our approach is robust to missing data i.e. our approach is applicable even in case predictions are not available from some of the humans. Moreover, we re-model the combination method, abstracting out the subset selection task.

4 PROPOSED APPROACH

In this section, we describe our proposed ComHAI for combining multiple human predictions with the model probabilities. We further establish that combining multiple humans is non-monotone and discuss a subset selection strategy to select human labels for improving the accuracy of the combined model.

4.1 Estimating a Single Label from Multiple Humans

There are many possible ways to combine predictions from multiple humans with the model probabilities to improve the combined model's performance. Previous work on crowd sourcing focuses on combining predictions from multiple humans and considers that as a human-predicted label. Majority voting and weighted majority voting have been explored in such works [11]. Hence, we describe some naive combination techniques in order to derive a single label from multiple humans:

- **Best Human:** Consider only the label predicted by the best human in the lot. The best human is the one having the highest accuracy among other humans. The accuracy of a human can be approximated from the confusion matrix corresponding to that human.
- **Best Majority Human:** Compute the mode (majority prediction) from the labels predicted by all humans. The best majority human is the one having the highest accuracy among the humans with the majority prediction label. Combine predictions from that human with the model probabilities.
- **Best Weighted-Majority Human:** Consider that the predictions by each human are weighed by their accuracy. Compute the weighted majority based prediction from the labels predicted by all humans and choose the most accurate human with that prediction. Combine the predictions from that human with the model probabilities.

In all the above-mentioned combination methods, we are effectively combining model probabilities with only a single human label. This can be done using the pre-existing work [15] on combining predictions from a single human with model probabilities, as described in Section 3.3.

We present the results obtained by using such naïve combination methods in Section 5. Although these combination techniques result in better accuracy than any single human or model prediction alone, we present a more sophisticated approach to combine predictions that outperforms these naïve methods significantly in terms of accuracy.

4.2 ComHAI: Combining Multiple Humans Labels with AI Model Output

We now present ComHAI to combine multiple human class labels with the model's probabilistic output. Similar to existing works on crowdsourcing, which essentially combine multiple human labels without AI model [11] and ensemble methods which combine predictions from multiple AI models but not considering humans [28], we assume that human predictions are independent of each other. With this assumption, we can apply Bayes' rule to get the following set of inequalities:

$$
\begin{aligned}
p(y(x)|h(x), m(x)) &\propto p(y(x)|m(x))p(h(x)|y(x), m(x)) \\
&\propto p(y(x)|m(x)) \prod_{i\in[n]} p(h_i(x)|y(x))
\end{aligned}
$$

This naturally leads to the following combination method:

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

Further, as discussed in Section 3.2, we can map the model's output $m(x)$ to a calibrated model $m^{\theta}(x)$. In order to establish the improvement in the performance of the combination approach, we develop a lower bound on the accuracy when the predictions are combined together using Equation 1.

Lemma 4.1. *Given n human labels and a calibrated model's output probabilities $m^{\theta}(x)$, the lower bound on the accuracy of the combined model is given as:*

$$
\mathbb{E}[\mathbb{1}(c(x) = y(x))] \geq \mathbb{P}\left\{\prod_{i\in[n]} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1 - \phi^{[i]}_{h_i(x)y(x)}} > \frac{1 - m^{\theta}_{y(x)}(x)}{m^{\theta}_{y(x)}(x)}\right\} \quad (2)
$$

Proof. Let's consider the accuracy of the combined model as:

$$
\begin{aligned}
\mathbb{E}[\mathbb{1}(c(x) = y(x))] &= \mathbb{P}\left\{y(x) = \arg\max_{k} m^{\theta}_{k}(x) \prod_{i\in[n]} \phi^{[i]}_{h_i(x)k}\right\} \\
&= \mathbb{P}\left\{m^{\theta}_{y(x)}(x) \prod_{i\in[n]} \phi^{[i]}_{h_i(x)y(x)} > \max_{k\neq y(x)} m^{\theta}_{k}(x) \prod_{i\in[n]} \phi^{[i]}_{h_i(x)k}\right\} \\
&\geq \mathbb{P}\left\{m^{\theta}_{y(x)}(x) \prod_{i\in[n]} \phi^{[i]}_{h_i(x)y(x)} > \max_{k\neq y(x)} m^{\theta}_{k}(x) \prod_{i\in[n]} \max_{k\neq y(x)} \phi^{[i]}_{h_i(x)k}\right\} \\
&\geq \mathbb{P}\left\{m^{\theta}_{y(x)}(x) \prod_{i\in[n]} \phi^{[i]}_{h_i(x)y(x)} > \left(1 - m^{\theta}_{y(x)}(x)\right) \prod_{i\in[n]} \left(1 - \phi^{[i]}_{h_i(x)y(x)}\right)\right\} \\
&= \mathbb{P}\left\{\prod_{i\in[n]} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1 - \phi^{[i]}_{h_i(x)y(x)}} > \frac{1 - m^{\theta}_{y(x)}(x)}{m^{\theta}_{y(x)}(x)}\right\}
\end{aligned}
$$

□

In the above expression, note that $\frac{1-m^{\theta}_{y(x)}(x)}{m^{\theta}_{y(x)}(x)}$ is fixed and is dependent on the already learnt model. The only way to maximize the accuracy is via maximizing the term $\prod_{i\in[n]} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}}$. Let us denote $f_S(x) = \prod_{i\in S} \frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}}$, which essentially denotes the left hand expression inside the probability in Equation 2 when instead of all humans, only a subset of humans S is considered. We first make the following observation:

Lemma 4.2. *For any $x \in \mathcal{X}$, $f_S(x)$ is not monotone in S even when the accuracy of all the humans is above* 0.5.

Proof. We compare $f_S(x)$ and $f_{S\cup\{i\}}(x)$. It is easy to see that $f_{S\cup\{i\}}(x) > f_S(x)$ iff $\frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}} > 1$ which obviously is not true unless $h_i(x) \neq y$. □

The above proof also gives the idea that combining all the humans who provide the correct label will always increase accuracy. However, the true label $y(x)$ is not known. Below, we provide GreedySubsetSelection outlining the procedure for optimally selecting the subset of humans without using the knowledge of true label such that it maximizes the lower bound on accuracy.

4.3 GreedySubsetSelection: Efficient Subset Selection Algorithm for Combined Predictions

Apart from the above theoretical results, we also observe experimentally that combining predictions from all humans does not necessarily lead to a better prediction, and the accuracy may drop drastically. This drop is more prevalent in scenarios where the human accuracies are close to 0.5. We now form a theoretical framework to select a subset of human predictions for the combination instead of all the human predictions. Let the selected subset be denoted by S such that $i \in \mathcal{N}$ and $1 \leq i \leq \mathcal{K}, \forall i \in S$. Under such a setting, we can modify the lower bound in Equation 2 as:

$$\mathbb{E}[\mathbb{1}(c(x) = y(x))] \geq \mathbb{P}\left\{\prod_{i\in S}\frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}} > \frac{1-m^{\theta}_{y(x)}(x)}{m^{\theta}_{y(x)}(x)}\right\} \tag{3}$$

The goal now is to select a subset such that the lower bound is maximized. We approach this problem by maximizing the term $\left(\prod_{i\in S}\frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}}\right)$ in Equation 3. Hence, the optimal subset S^* for combination is as follows:

$$S^* = \arg\max_S\left(\prod_{i\in S}\frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}}\right) \tag{4}$$

To select an optimal subset, we need the knowledge of $y(x)$ i.e. the true label. Since, $y(x)$ is not available, a very natural choice would be to choose the label which maximizes the probability (in equation 3) as this would indicate the label which is maximizing the probability of output label being correct. Thus, we select the pseudo-optimal subset S^{**} as follows:

$$S^{**} = \arg\max_S \max_{1\leq j\leq\mathcal{K}}\left(\prod_{i\in S}\frac{\phi^{[i]}_{h_i(x)j}}{1-\phi^{[i]}_{h_i(x)j}}\right) \tag{5}$$

The idea is to basically consider that label as the true label for a subset S, which maximizes the expression $f_S(x)$. Since the true labels are not assumed to be known in this case, we call the selected subset the pseudo-optimal subset.

Our combination method considers a subset of human predictions for the combination using Equation 5. Algorithm 1 presents our subset selection algorithm for the selection of a set of humans given an instance x. Instead of iterating over all the subsets and computing the value of $f_S(x)$, we can greedily construct this subset. Note that $f_S(x)$ does not increase by considering any human i such that $\frac{\phi^{[i]}_{h_i(x)y(x)}}{1-\phi^{[i]}_{h_i(x)y(x)}} \leq 1$. This naturally leads to an obvious greedy algorithm for selecting the optimal subset as per Equation 4 where we select all and only those humans corresponding to which this term is greater than 1.

The greedy algorithm for the selection of pseudo optimal subset requires a little more work. Let the pseudo optimal value of $f_S(x)$ be arrived at by $S = S^{**}$ and $j = k^{**}$. Merely with the knowledge of $j = k^{**}$, we can easily figure out the pseudo-optimal subset using our observation that the subset will have all and only those humans for which $\frac{\phi^{[i]}_{h_i(x)k^{**}}}{1-\phi^{[i]}_{h_i(x)k^{**}}} > 1$.

So, the problem reduces to the following 3 steps. Compute $\phi^{[i]}_{h_i(x)j}$ for all pairs of humans and class labels [lines 1-4 of Algorithm 1]. Then for each class j, compute $f_S(x)$ for the candidate pseudo optimal subset, denoted by $C[j]$ [lines 6-13 of Algorithm 1]. Finally, compute the pseudo optimal subset corresponding to the class label with maximum $C[j]$ value [lines 14-16 of Algorithm 1].

Algorithm 1 GreedySubsetSelection: Pseudo Optimal Subset Selection

```
Require: $n, \mathcal{K} \in \mathbb{N}$
Require: $\mathcal{K} \times \mathcal{K}$ dimension matrices $\phi^{[i]}, 1 \leq i \leq n$
Require: $h_i(x), 1 \leq i \leq n$
 1: for $1 \leq i \leq n$ do
 2:     for $1 \leq j \leq \mathcal{K}$ do
 3:         $M[i][j] \leftarrow \frac{\phi^{[i]}_{h_i(x)j}}{1-\phi^{[i]}_{h_i(x)j}}$
 4:     end for
 5: end for
 6: for $1 \leq j \leq \mathcal{K}$ do
 7:     $C[j] \leftarrow 1$
 8:     for $1 \leq i \leq n$ do
 9:         if $M[i][j] > 1$ then
10:             $C[j] \leftarrow C[j] \times M[i][j]$
11:         end if
12:     end for
13: end for
14: $y^* \leftarrow \arg\max_{1\leq j\leq\mathcal{K}} C[j]$
15: $S^{**} \leftarrow \{1 \leq i \leq n \mid M[i][y^*] > 1\}$
16: return $S^{**}$
```

5 EXPERIMENTS AND RESULTS

We performed experiments to evaluate the performance of our proposed ComHAI against multiple baselines. The code is available at github.com/sagalpreet/Human-AI-Team[30]. We evaluate our proposed approach ComHAI against multiple baseline combination methods on CIFAR-10H [25], an image classification dataset with human annotations. CIFAR-10H contains 10-way human labels from multiple humans for 10,000 images from the standard CIFAR10 test dataset.

5.1 Human Labels and the Classification Model

We simulate the human predictions according to the annotations in the dataset. The human prediction made by a particular human is parameterized by the accuracy of that human. Let there be n human annotations available for an image x given by $h(x) = \{h_1, h_2, \ldots, h_n\}$ and the true label be $y(x)$. Let us denote the fraction of humans who predicted class k as $g_x(k)$. We denote the set of all class labels except the true label as $\mathcal{Y}'(x) = \{1, 2, \ldots, \mathcal{K}\} - \{y(x)\}$. We also define a probability distribution over all the classes as follows:

$$p(k|x) = \begin{cases} 0 & , \; k = y(x) \\ \frac{g_x(k)}{\sum_{i\in\mathcal{Y}'(x)} g_x(i)} & , \; k \neq y(x) \text{ and } g_x(y(x)) \neq 1 \\ \frac{1}{\mathcal{K}-1} & , \; k \neq y(x) \text{ and } g_x(y(x)) = 1 \end{cases}$$

We simulate the prediction $h^{[\Psi]}(x)$ for a human with accuracy Ψ as follows:

$$h^{[\Psi]}(x) = \begin{cases} y(x) & , \text{ with probability } \Psi \\ t \sim p(k|x) & , \text{ with probability } 1-\Psi \end{cases}$$

where t is sampled according to the probability distribution $p(k|x)$.

For the AI model's probabilistic prediction, we use a custom CNN model with enough room for improvements in the predictive

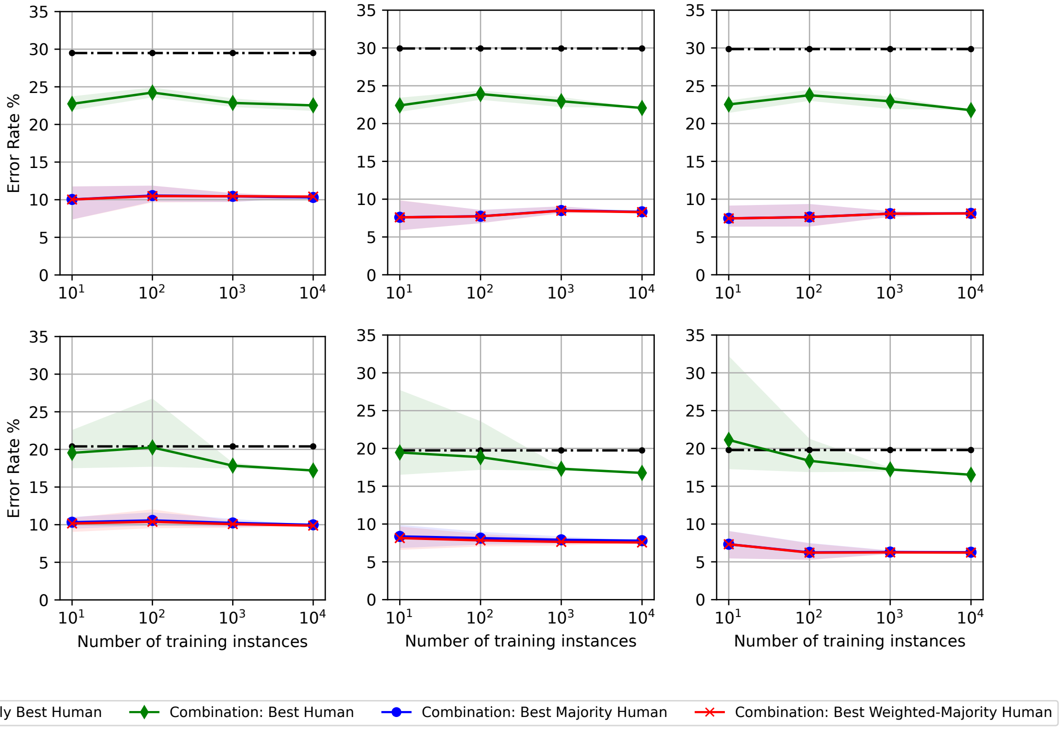

Figure 1: Learning curves on CIFAR-10H using naïve combination methods. Each plot corresponds to a different set of human labelers where every human is characterized by accuracy. The number of humans considered for plots in the first row is 5, 10, and 15, respectively, all having an accuracy of 70%. Plots in the second row correspond to 4, 7, and 13 humans, with accuracies ranging from 0.5 to 0.8. Custom CNN is being used as the AI model for probabilistic outputs that are to be combined with human labels.

performance. The CNN model used has 3 convolution layers with 32 3 × 3 convolution filters and ReLU activation followed by a 2 × 2 maxpooling layer. This was followed by 128 nodes dense layer with ReLU activation and a 10 node dense layer at the output with softmax. We have used categorical cross entropy loss and stochastic gradient descent for optimization. The model is trained on the CIFAR-10 training set consisting of 50,000 images.

The accuracy of our custom CNN model is 56.74% on CIFAR-10H test dataset consisting of 10,000 images. We have also evaluated our framework with probabilistic outputs from a complex network, Resnet-110[8]. Our framework works well in either case. For instance, ResNet-110, alone, is 93.89% accurate on CIFAR-10H dataset. Our combination method using ResNet-110 as AI model achieves 99.15% accuracy (averaged over 10 runs). This is when the number of humans is 7 with accuracies ranging from 0.5 to 0.8. On the other hand, in the same setting, our combination method with CNN as AI model achieves 96.65% accuracy (averaged over 10 runs).

5.2 Baseline Combination Methods

As an evaluation yardstick, we plot the error rate on the evaluation dataset as a function of dataset size. We first evaluate the naïve combination methods as described in Section 4 wherein we consider a single predicted class label to combine with model probabilities for a given instance. These naïve combination methods are compared against the most accurate single human ('Only Best Human') as baseline. We also evaluate the more sophisticated combination techniques alongside our proposed ComHAI method. More specifically, following are the sophisticated combination techniques that we consider:

- **All Humans Selected:** Use predictions from all humans to combine with model probabilities using the Equation 1 with calibrated model probabilities.
- **Random Subset Selection:** Select a subset of humans randomly (a human is selected or not with 0.5 probability) and use predictions from those humans to combine with model probabilities using the Equation 1 with calibrated model probabilities.
- **True LB Subset Selection:** Select the optimal subset of humans as per Equation 4 so as to maximize the lower bound on accuracy according to the Equation 2 and use predictions by these humans to combine with model probabilities. Note that this method is not practical since the subset selection requires knowledge of ground truth. We, however, evaluate this combination method as it maximizes the true lower bound.

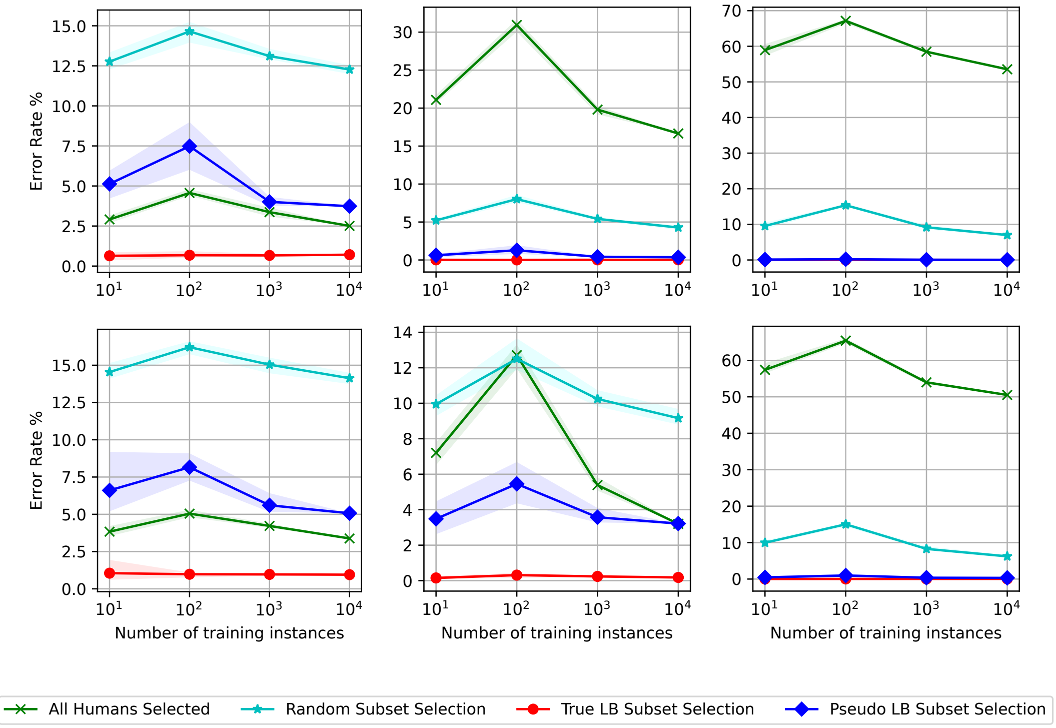

Figure 2: Learning curve on CIFAR-10H using subset selection based methods. Each plot corresponds to a different set of human labelers available, where every human is characterized by accuracy. The number of humans considered for plots in the first row is 5, 10, and 15, respectively, all having an accuracy 70%. Plots in the second row correspond to 4, 7, and 13 humans, with accuracies ranging from 0.5 to 0.8. Custom CNN is being used as the AI model for probabilistic outputs that are to be combined with human labels

- **Pseudo LB Subset Selection (GreedySubsetSelection):** Select a pseudo-optimal subset of humans according to the Equation 5 and use predictions by these humans to combine with model probabilities.

All the combination algorithms described are very efficient and can be easily run on the CPU. Reproducing results from the experiment above should take less than an hour of computation on any decent processor with good enough (≥8 GB) RAM. This estimation excludes the training time for CNN. The memory requirements are linear in the number of humans and quadratic in the number of classes in the classification task.

5.3 Inferences

The plots in Figure 1 depict the performance of naïve combination methods against the *Only Best Human* with different number of humans considered. The plots in the first row in Figure 1 are for 5, 10, and 15 humans, respectively, all humans having an accuracy of 70%. Plots in the second row correspond to 4, 7, and 13 humans, with their accuracies ranging from 0.5 to 0.8. It may be noted that the CNN model's accuracy is $\sim$ 57%. Among the naïve combination methods, *Only Best Human* and *Combination: Best Weighted Majority Human* methods are sensitive to our estimate of confusion matrix which improves with more number of evaluation instances. Further, as can be observed, the error rate falls significantly when the combination method is used with the *Best Majority Human* and *Best Weighted-Majority Human*. However, the improvement in the performances is stabilized even with the increasing number of instances.

The plots in Figure 2 depict the performance of sophisticated combination baselines with our proposed subset selection method using probabilistic predictions from our simple custom CNN. We observe that *Random Subset Selection* method tends to perform quite well in certain settings, although still inferior to *Pseudo LB Subset Selection*. It can be noted that for random selection, we include a human in the subset with 0.5 probability, so the expected size of the subset would be equal to half the number of human predictions available. If the size is closer to that of the optimal subset, then the combination is effective, especially in cases where the accuracy of humans is not very varied. Similar results are obtained when using probabilistic outputs from a complex model like ResNet-110 as is evident from Figure 3.

All Humans Selected method does perform well when fewer human labelers are available. However, the accuracy reduces drastically for when a larger set of human labelers are available. This empirically establishes the non-monotone property of combining multiple humans with the AI model's probabilistic output. On other

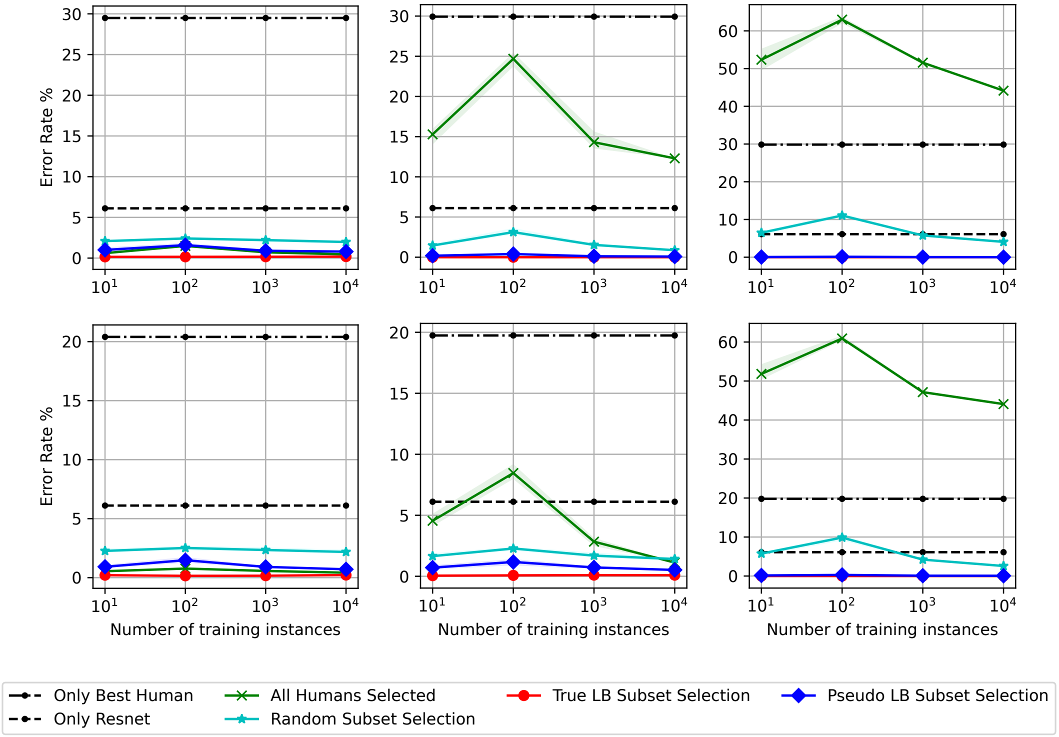

Figure 3: Learning curve on CIFAR-10H using subset selection based methods. Each plot corresponds to a different set of human labelers available, where every human is characterized by accuracy. The number of humans considered for plots in the first row is 5, 10, and 15, respectively, all having an accuracy 70%. Plots in the second row correspond to 4, 7, and 13 humans, with accuracies ranging from 0.5 to 0.8. Resnet is being used as the AI model for probabilistic outputs that are to be combined with human labels

hand *Pseudo LB Subset Selection* consistently performs very well in general.

We based *Pseudo LB Subset Selection* as an approximation to *True LB Subset Selection* method, which uses knowledge of true labels for subset selection. We have empirically verified that it is indeed a good approximation, especially when the number of human labelers is large. The method achieves more than 99.9% accuracy in some scenarios even when none of the individual humans is more than 80% accurate and the AI model is just 56.74% accurate.

6 DISCUSSIONS AND CONCLUSIONS

We have established a theoretical framework for combining human predictions with model probabilities to achieve significantly high accuracies and evaluated the results empirically on an image classification task. Our proposed framework is also robust to missing human predictions. As our combination method is based only on a subset of human labels, we can simply ignore the humans for whom the predictions are not available and choose a subset from those present. Present works [16] on extended learning to defer for multiple experts setting has this drawback of requiring human predictions from every human on every instance as was identified in the review by Leitão et al. [21].

Our combination method requires predictions from humans to select a subset of humans whose predictions are then actually used for the combination. There is a scope for making the combination method cost-effective if the subset selection could be simply based on the knowledge of input instance (images, in our experiments) without requiring the predictions from each human. Associating a cost with the prediction from each human and having a modified optimization problem with a penalty associated with the selection of each human can also help in achieving a similar goal. Another interesting aspect to the combination method is the analysis of fairness and human biases. For existing Human-AI team models such as the deferred approach, it has been shown that failing to consider biases may lead to aggravation of unfairness [13, 21].

Although we have evaluated our combination method only on an image classification task, we expect similar performances of our proposed framework in other settings. However, non-stationarity factors are known to render AI models useless, even Human-AI collaboration systems may additionally suffer from change in human behavior due to exogenous factors or adaption of human behavior as was noted in [21]. In our proposed approach, such factors may affect the estimated human confusion matrix that must be re-learned from time to time in order to ensure good performance of the combination approach.

ACKNOWLEDGMENTS

We would like to express our gratitude to Google Research for their support through the travel grant awarded to Sagalpreet. This grant allowed him to attend the conference and present the work, which would not have been possible otherwise.

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[3] Olivier Caelen. 2017. A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence* 81, 3 (2017), 429–450.

[4] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[5] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI Collaboration with Bandit Feedback. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1722–1728. Main Track.

[6] Robert Geirhos, Kristof Meding, and Felix A Wichmann. 2020. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems* 33 (2020), 13890–13902.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Hkg4TI9xl

[10] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[11] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y Narahari. 2018. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence* 254 (2018), 44–63.

[12] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[13] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2021. Selective Classification Can Magnify Disparities Across Groups. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=N0M_4BkQ05i

[14] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[15] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems* 34 (2021), 4421–4434.

[16] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 154–165.

[17] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence* 20, 3 (1998), 226–239.

[18] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[19] Ludmila I Kuncheva. 2014. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

[20] PJ Lamberson and Scott E Page. 2012. Optimal forecasting groups. *Management Science* 58, 4 (2012), 805–810.

[21] Diogo Leitão, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. 2022. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. *CoRR* abs/2206.13202 (2022). https://doi.org/10.48550/arXiv.2206.13202 arXiv:2206.13202

[22] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[23] Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi HS Lelis. 2021. Improving the Performance-Compatibility Tradeoff with Personalized Objective Functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5967–5974.

[24] Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*. PMLR, 7076–7087.

[25] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9617–9626.

[26] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.

[27] Amir Rosenfeld, Markus D Solbach, and John K Tsotsos. 2018. Totally looks like-how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1961–1964.

[28] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.

[29] Thomas Serre. 2019. Deep learning: the good, the bad, and the ugly. *Annual review of vision science* 5, 1 (2019), 399–426.

[30] Sagalpreet Singh, Shweta Jain, and Shashi Shekhar Jha. 2023. *ComHAI*. https://doi.org/10.5281/zenodo.7680856

[31] Frederik Träuble, Julius Von Kügelgen, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Peter Vincent Gehler. 2021. Backward-Compatible Prediction Updates: A Probabilistic Approach. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[32] Laura Trouille, Chris J Lintott, and Lucy F Fortson. 2019. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1902–1909.

[33] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.

[34] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*. 155–164.

[35] Rajeev Verma and Eric Nalisnick. 2022. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*. PMLR, 22184–22202.

[36] Zahra Zahedi and Subbarao Kambhampati. 2021. Human-AI Symbiosis: A Survey of Current Approaches. *CoRR* abs/2103.09990 (2021). arXiv:2103.09990 https://arxiv.org/abs/2103.09990