Balancing Humans and Machines: A Study on Integration Scale and Its Impact on Collaborative Performance

Rui Zou[1,2], Sannyuya Liu[1,2], Yawei Luo[3], Yaqi Liu[4], Jintian Feng[1,2], Mengqi Wei[1,2], Jianwen Sun*[1,2]

[1] National Engineering Research Center of Educational Big Data, Central China Normal University, Wuhan 430079, China
[2] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[3] School of Software Technology, Zhejiang University, Hangzhou 310027, China
[4] School of Information and Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China
zouruixyz@mails.ccnu.edu.cn, liusy027@ccnu.edu.cn, yaweiluo@zju.edu.cn, liuyaqi@zuel.edu.cn,
{fjt2018,weimengqi}@mails.ccnu.edu.cn, sunjw@ccnu.edu.cn

Abstract

In the evolving artificial intelligence domain, hybrid human-machine systems have emerged as a transformative research area. While many studies have concentrated on individual human-machine interactions, there is a lack of focus on multi-human and multi-machine dynamics. Our study introduces a statistical method for assessing ensembles of any size and investigates the optimal human-machine ratio alongside the impact of ensemble size on performance in human-machine collaboration. This paper delves into these nuances by introducing a novel statistical framework that discerns integration accuracy in terms of precision and diversity. Empirical studies reveal that performance increases consistently with scale, either in human or machine settings. However, hybrid systems present complexities. Their performance is intricately tied to the human-to-machine ratio. Interestingly, as the scale expands, integration performance growth isn't limitless. It reaches a threshold influenced by model diversity. This introduces a pivotal 'knee point', signifying the optimal balance between performance and scale. This knowledge is vital for resource allocation. Grounded in rigorous evaluations using public datasets, our findings emphasize the framework's robustness in refining integrated systems.

Introduction

In the last decade, AI has made significant strides, particularly in deep learning (LeCun, Bengio, and Hinton 2015; Goodfellow, Bengio, and Courville 2016). This progress has advanced fields like computer vision (Redmon et al. 2016), speech recognition (Graves, Mohamed, and Hinton 2013), and NLP (Vaswani et al. 2017). The interplay between machine cognition and human intelligence has been pivotal in this surge (Wu et al. 2022). For example, tools such as ChatGPT empower individuals with limited writing proficiency to match expert-level work (Noy and Zhang 2023). Despite AI's progress, it cannot mimic the complexity of the human brain entirely. AI models can have unforeseen vulnerabilities (Papernot et al. 2016; Serre 2019; Zhang et al. 2020). For instance, advanced image and text classifiers occasionally commit unexpected errors (Recht et al. 2019; Hendrycks et al. 2021; Ribeiro et al. 2020). The human brain excels in abstract reasoning and learning from limited data, and this distinction highlights the complementary nature between humans and AI. Hence, blending human judgment with AI has become crucial in addressing these challenges (Wilder, Horvitz, and Kamar 2021; Bansal et al. 2021; De et al. 2020) and is a focal point in human-machine interaction studies (Riedl 2019; Bansal et al. 2019; Zhang et al. 2021; Zahedi and Kambhampati 2021).

An agent is an entity designed for specific tasks. An agent can be a machine capable of independently performing a specific task (like an image recognition model), or a human. Homogeneous agent collaboration refers to purely human-to-human or machine-to-machine interactions, while mixed-type collaboration refers to human-machine interactions. However, many current studies focuses on homogeneous agent collaborations, such as ensemble techniques in machine learning (Breiman 1996; Freund and Schapire 1996; Friedman 2001) or studies on human societies (Aggarwal et al. 2019; Lamberson and Page 2012), whereas mixed-type collaborations, especially large-scale human-machine partnerships, have been relatively unexplored (Steyvers et al. 2022). Present studies mostly delve into one human interacting with one machine (Tejeda et al. 2022; Steyvers et al. 2022). Though individual agents often have limited problem-solving abilities, collaboration amplifies their potential (Wu and Wu 2019). Yet, large-scale human-machine collaboration research is notably sparse.

In the field of ensemble learning, the mainstream opinion is that the success of ensemble learning relies on the trade-off between the accuracy and diversity of individual learners (Zhou 2020; Breiman 1996; Kuncheva and Whitaker 2003; Ho 1998). Inspired by this, we propose a statistical framework that breaks down ensemble performance into individual accuracy and diversity. Subsequently, using this framework, we investigate the relationship among ensemble performance, ensemble size, and diversity, particularly in settings involving collaboration between multiple humans and machines. Firstly, Our method reveals the relationship between ensemble performance and size, theoretically finding that both homogeneous and heterogeneous agent ensembles have a performance ceiling proportional to diver-

*Corresponding author.

sity. Secondly, we observed that the performance of homogeneous agent ensembles always improves with increasing size. However, mixed-type agents exhibit a more complex pattern: adding more of one type of agent, while keeping the other constant, does not consistently lead to better outcomes. Instead, a balance emerges due to the optimal ratio. Moreover, this optimal ratio changes as the ensemble size increases. We refer to the path formed by the changing optimal ratio as the ensemble size varies as the 'optimal trajectory'.Our proposed statistical framework can predict this trajectory, thereby enabling more efficient ensemble strategies. For instance, for a specific task with an ensemble size of five, the optimal ratio is two humans to three machines. Finally, we discovered that the curve between ensemble performance and size is increasing, and the growth rate is faster and then slower, so there exists a 'knee point'. Both theory and experiments indicate that beyond this knee point, the rate of improvement in ensemble performance significantly decreases. Our method can effectively estimates this knee point, aiding in balancing ensemble performance and computational costs.

In image classification tasks, humans and machines exhibit distinct behaviors (Geirhos, Meding, and Wichmann 2020; Geirhos et al. 2018). Leveraging this, we examined collaborative human-machine performances on large-scale datasets: CIFAR10H (Peterson et al. 2019) and ImageNet-16H (Steyvers et al. 2022). The latter integrates human predictions with representative CNN outputs, while CIFAR10H is abundant with human insights. Our results align with our theory, confirming that our statistical framework adeptly predicts ensemble performance trajectories. Compared to homogeneous methods, our approach showcases superior precision.

Our key contributions are: **(1)** We introduce a statistical framework that breaks down ensemble performance into individual accuracy and diversity, emphasizing their multiplicative relationship on ensemble effectiveness. **(2)** We propose a statistical framework for multi-human and multi-machine analyses. It explores ensemble performance versus size, uncovering a performance ceiling governed by inter-model diversity. Notably, its utility extends beyond image classification tasks. **(3)** We ascertain that homogeneous agent ensemble performance scales with ensemble size. However, mixed-agent ensemble performance escalates along an optimal trajectory. **(4)** Our tool adeptly predicts this optimal trajectory, guiding ensemble strategy refinement. It also pinpoints the knee point in performance growth versus ensemble size, optimizing computational resources.

Related Works

Human-machine collaboration is a budding field, with limited yet significant research. Present work largely focuses on human group decisions and aggregated machine learning outcomes. In the realm of **human group decision-making**, there's a consensus that group decisions, enriched by participant diversity, often outperform individual ones. The concept of a collective intelligence factor has surfaced as a key predictor of diverse team performance (Woolley et al. 2010). (Aggarwal et al. 2019) pinpointed the beneficial role of collective intelligence on team learning rates, underscoring its bridge between cognitive style diversity and team learning. Through models and studies, (Page 2008) spotlighted diversity's boost to team performance, highlighting the value of varied backgrounds. (Davis-Stober et al. 2015) discerned an balance between diversity and prediction accuracy in large teams, proposing methods for optimal team formation. Lastly, in the context of image searches, (Juni and Eckstein 2017) found that both weighted average and mean methods outclass majority-vote strategies in accuracy, underscoring the collective strength of diverse human visual systems. Research into ensemble machine learning reveals a positive link between classifier diversity and performance. The importance of diversity in machine ensembles has been well-documented, with **ensemble error decomposition techniques** being particularly salient. The Error-ambiguity decomposition by (Krogh and Vedelsby 1995) categorizes ensemble error into individual errors and an ambiguity tied to diversity. This was further developed by (Ueda and Nakano 1996) into the Bias-variance-covariance decomposition. (Brown et al. 2005) affirmed the connection between these methods. However, while diversity's significance is clear, existing decomposition methods largely serve to regression, leaving a void in classification that our framework seeks to fill. Certain **ensemble methods emphasize increasing diversity**. For instance, (Tumer and Ghosh 1996) showcased that diversity can be enhanced by minimizing inter-classifier correlations through strategies like cross-validation. The Bagging approach by (Breiman 1996) amplifies diversity via subset resampling, while Boosting variants such as AdaBoost (Freund and Schapire 1996) and GBM (Friedman 2001) drive diversity by advocating learner complementarity. Regarding **Human-AI Collaboration**, there's a growing trend of humans leaning on AI for decision-making. (Tejeda et al. 2022) introduced a cognitive model detailing human reliance on AI. Concurrently, (Steyvers et al. 2022) proposed a Bayesian method combining human and AI predictions and established an accuracy metric for ensemble sizes $M \in \{1, 2\}$. For $M > 2$, since deriving an exact expression becomes complex, it is not applicable in this case. Yet, neither study comprehensively addresses the challenges of multi-human-multi-machine collaborations.

Statistical Framework

In this section, we introduce a statistical framework designed for accuracy prediction in multi-class ensemble models. Given N samples, the ensemble's correct prediction condition for the i^{th} sample is defined by:

$$\sum_{m=1}^{M} \lambda_{k,m}^{(k)[i]} > \sum_{m=1}^{M} \lambda_{k,m}^{(j)[i]} \quad (\forall j \neq k). \tag{1}$$

Here, $\lambda_{k,m}^{(j)[i]}$ denotes the logit prediction of the m^{th} model for class j on the i^{th} sample, with its true label being class k. **Broadening to all samples labeled as class** k, the ensemble accuracy (ACC) is probabilistically depicted as:

$$A_{k,M} = p\left(\sum_{m=1}^{M} \lambda_{k,m}^{(k)} > \sum_{m=1}^{M} \lambda_{k,m}^{(j)} \quad \forall j \neq k\right). \tag{2}$$

For a deeper ensemble analysis, our framework adopts two statistical assumptions: Gaussian and Conditional Independence (CI). **Gaussian Assumption**: For a truth label of class k, correct predictions are Gaussian distributed as:

$$\lambda_{k,m}^{(k)} \sim \mathcal{N}(a_m, \sigma_{am}^2). \tag{3}$$

Similarly, incorrect predictions follow:

$$\lambda_{k,m}^{(j)} \sim \mathcal{N}(b_m, \sigma_{bm}^2) \quad (\forall j \neq k). \tag{4}$$

Leveraging Gaussian distribution for data approximation is widely accepted (Bishop and Nasrabadi 2006; Murphy 2012), substantiated by the Central Limit Theorem (CLT). The CLT implies that large samples from independent and identically distributed random variables tend to converge to a Gaussian distribution. **Conditional Independence Assumption (CI)**: With the true label as class k, predictions across classes are considered mutually independent. This assumption frequently surfaces in ensemble prediction literature, either implicitly or explicitly (Kuncheva 2014; Sagi and Rokach 2018; Steyvers et al. 2022; Kerrigan, Smyth, and Steyvers 2021). Moreover, (Kuncheva 2006) postulates that even with a CI assumption breach, a CI model can remain an optimal discriminant. Given the Gaussian and CI Assumptions, we can simplify X from Eq 2: $X = \sum_{m=1}^{M}[\lambda_{k,m}^{(k)} - \lambda_{k,m}^{(j)}]$ as a Gaussian-distributed variable: $X \sim \mathcal{N}(\mathbb{E}[X], \mathrm{Var}[X])$. Let's calculate its mean and variance: $\mathbb{E}[X] = \sum_{m=1}^{M}(a_m - b_m)$; $\mathrm{Var}[X] = \sum_{p=1}^{M}\sum_{q=1}^{M}(\sigma_{a,p}\sigma_{a,q}\rho_{a,pq} + \sigma_{b,p}\sigma_{b,q}\rho_{b,pq})$. Here, $\rho_{a,pq}$ and $\rho_{b,pq}$ denote correlations between the correct and incorrect predictions of models p and q, respectively. Incorporating X into Eq 2, we get:

$$A_{k,M} = p\{X > 0 \quad (\forall j \neq k)\} \tag{5}$$

$$= \Phi^{L-1}\left(\frac{\sum_{m=1}^{M}(a_m - b_m)}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\sigma_{a,p}\sigma_{a,q}\rho_{a,pq} + \sigma_{b,p}\sigma_{b,q}\rho_{b,pq})}}\right). \tag{6}$$

With L denoting the total class count, Φ representing the Cumulative Distribution Function (CDF), Eq 6 captures the ensemble prediction accuracy for M models with respect to the k^{th} class. Extending our analysis to **all samples** and assuming a total of N samples with N_k samples in class k, the ensemble accuracy for M models becomes:

$$A_M = \sum_{k=1}^{L}\frac{N_k}{N}A_{k,M}$$

$$= \sum_{k=1}^{L}\frac{N_k}{N}\Phi^{L-1}\left(\frac{\sum_{m=1}^{M}(a_m - b_m)}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\sigma_{a,p}\sigma_{a,q}\rho_{a,pq} + \sigma_{b,p}\sigma_{b,q}\rho_{b,pq})}}\right). \tag{7}$$

For datasets like ImageNet-16H and CIFAR10H used in our study, each class has an equal sample size ($N_k = N_j$ for all k, j). Leveraging this, Eq 7 simplifies to:

$$A_M = \Phi^{L-1}\left(\frac{\sum_{m=1}^{M}(a_m - b_m)}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\sigma_{a,p}\sigma_{a,q}\rho_{a,pq} + \sigma_{b,p}\sigma_{b,q}\rho_{b,pq})}}\right). \tag{8}$$

Decomposition of Ensemble Accuracy

We delve into the methodology for decomposing ensemble accuracy into individual model accuracy and diversity components. When considering a single model (i.e., $M = 1$), its performance can be represented as:

$$A_{M=1} = \Phi^{L-1}\left(\frac{\Delta}{\sigma\sqrt{2}}\right). \tag{9}$$

Here, $\Delta = a - b$, encapsulating the performance of this singular model. From Eq 8, we separate parameters into two categories: those denoting the model's inherent performance (a, b, σ) and those outlining the inter-model relationships (ρ). Our objective is to partition the ensemble accuracy into individual performance and inter-model interactions. Assumptions aiding this decomposition include:

$$\begin{aligned} &\Delta = a_m - b_m, \quad \forall m = 1, \ldots, M, \\ &\sigma = \sigma_{a,m} = \sigma_{b,m}, \quad \forall m = 1, \ldots, M. \end{aligned} \tag{10}$$

In the context of ensemble accuracy, view Δ as the bias and σ as variance. Integrating Eq 10 into Eq 8, we split the ensemble performance as:

$$A_M = \Phi^{L-1}\left(\frac{\Delta}{\sigma}\frac{M}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\rho_{a,pq} + \rho_{b,pq})}}\right)$$

$$= \Phi^{L-1}\left(\frac{\Delta}{\sigma}V\right). \tag{11}$$

Here, $\frac{\Delta}{\sigma}$ denotes individual accuracy, and V captures the model diversity. Hence, ensemble accuracy can be seen as a product of individual accuracy and diversity, emphasizing the balance between the two. Specifically, the ensemble's efficacy is shaped by: **(1)** Individual Accuracy, $\frac{\Delta}{\sigma}$: Enhanced performance in standalone models directly boosts the ensemble's capability. **(2)** Diversity, represented by V: A diminished correlation among individual models increases the ensemble's diversity, enriching its performance.

It's pivotal to recognize that in Eq 11, we presume biases and variances across models to be fairly consistent. This is particularly apt in human-AI collaborations where agent performance variance is typically marginal, supporting our assumption. Nonetheless, in occasional situations with significant variations in the $\frac{\Delta}{\sigma}$ term, our estimation may be off. In these cases, we resort to the average individual accuracy, $\frac{\Delta}{\sigma} = \mathbb{E}[\frac{\Delta_m}{\sigma_m}]$, to give a comprehensive estimate. Although Eq 11 may not be as precise as Eq 7, its theoretical delineation of the interplay between individual accuracy and diversity is insightful.

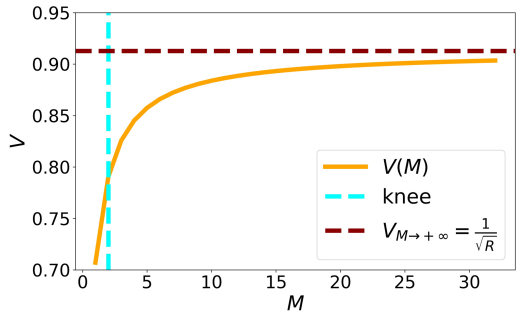

Figure 1: Plotting diversity V against ensemble size M. The red curve signifies the ensemble's maximum threshold; the blue curve showcases the 'knee'.

Study on Diversity and Ensemble Size

Our focus shifts to how ensemble performance fluctuates with increasing ensemble size. Assuming the individual accuracy term $\frac{\Delta}{\sigma}$ remains consistent, ensemble accuracy, as shown in Eq 11, chiefly hinges on diversity. To elucidate this, we express diversity via Eq 12:

$$V = \frac{M}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\rho_{a,pq}+\rho_{b,pq})}}. \tag{12}$$

Subsequently, we explore two scenarios: incorporation of uniform agents versus a mix of diverse agents.

Analysis of Homogeneous Agents Considering homogeneous agents, we classify ensembles as purely machine-based (MM...M) or entirely human-driven (HH...H). We postulate that pairwise correlations, be it correct or incorrect, are fairly uniform:

$$\forall p \neq q: \quad \rho_a = \rho_{a,pq}; \quad \rho_b = \rho_{b,pq}. \tag{13}$$

This derivation yields:

$$\begin{aligned} V(M) &= \frac{M}{\sqrt{2M + M(M-1)(\rho_a+\rho_b)}} \\ &= \sqrt{\frac{M}{2+(M-1)(\rho_a+\rho_b)}} \qquad (14) \\ &= \sqrt{\frac{M}{2+(M-1)R}}, \qquad (15) \end{aligned}$$

with $R = \rho_a + \rho_b \in [0, 2]$. From Eq 14, diversity impacts ensemble diversity equally during correct and incorrect decisions. Additionally, Eq 15 denotes: **(1)** $V(M)$ increases, yet has an upper bound. **(2)** As ensemble size grows, the growth rate of $V(M)$ tapers. This reveals a 'knee', indicating an optimal trade-off between ensemble size (resource usage) and accuracy. Figure 1 illustrates the $V(M)$ curve, its upper limit, and the knee. The ensemble performance for homogeneous agents is directly related to ensemble scale. Considering $R \in [0, 2]$, the upper bound for ensemble diversity $V_{M\to\infty}$ spans $[\frac{1}{\sqrt{2}}, +\infty]$. Theoretically, with significant ensemble diversity, agent ensemble performance can continually improve. The knee's calculation can adopt various strategies; for discrete datasets, we adopt the method from (Satopaa et al. 2011).

Analysis of Heterogeneous Agent Heterogeneous agents involve a mix of humans and machines (H...HM...M). Assuming n_H humans and n_M machines, the combinations of HM, HH, and MM are:

$$n_{HM} = n_H n_M; \quad n_{HH} = \binom{n_H}{2}; \quad n_{MM} = \binom{n_M}{2}. \tag{16}$$

Incorporating the above into Eq 12, the diversity for heterogeneous agents $V(n_H, n_M)$ becomes:

$$\begin{aligned} V(n_H, n_M) &= \frac{M}{\sqrt{2M + 2\sum_{q=p+1}^{M}\sum_{p=1}^{M-1}(\rho_{a,pq}+\rho_{b,pq})}} \\ &= \frac{M}{\sqrt{2(M + n_{HM}R_{HM} + n_{HH}R_{HH} + n_{MM}R_{MM})}}. \end{aligned} \tag{17}$$

Define $R_{HM} = \rho_{a,HM} + \rho_{b,HM} \in [0, 2]$. Likewise, R_{HH} and R_{MM} are defined. Based on the human-machine collaboration characteristic: $R_{HM} < R_{HH} < R_{MM}$, we depict the relationship between diversity and ensemble scale as the surface $V(n_H, n_M)$ in Figure 2. The optimal trajectory $V(n_H^*, n_M^*)$ is extracted by identifying the surface's extremal values. Given ensemble scale M, the optimal human and machine count are:

$$\begin{aligned} n_H^* &= \frac{-MR_{HM} + MR_{MM} + \frac{R_{HH}}{2} - \frac{R_{MM}}{2}}{R_{HH} - 2R_{HM} + R_{MM}}, \\ n_M^* &= \frac{MR_{HH} - MR_{HM} - \frac{R_{HH}}{2} + \frac{R_{MM}}{2}}{R_{HH} - 2R_{HM} + R_{MM}}. \end{aligned}$$

Incorporating these into Eq 17 yields the optimal trajectory $V(n_H^*, n_M^*)$ for $M^* = n_H^* + n_M^*$. While homogeneous agent ensembles consistently improve with size, as illustrated in Figure 1, mixed-type agents exhibit a nuanced behavior.

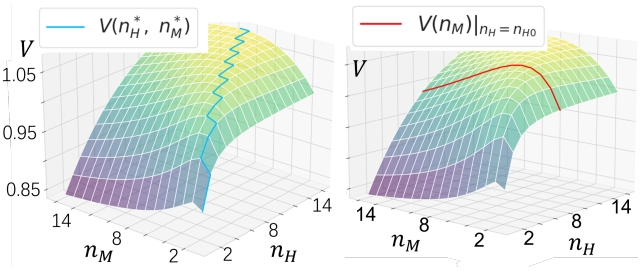

Figure 2: The left and right surfaces both illustrate $V(n_H, n_M)$. The *left* blue trajectory marks the optimal human-machine combination (n_H^*, n_M^*) for maximal diversity at a given ensemble scale M, termed the optimal trajectory $V(n_H^*, n_M^*)$. On the *right*, the red curve depicts diversity changes with machine count when $n_H = n_{H0}$.

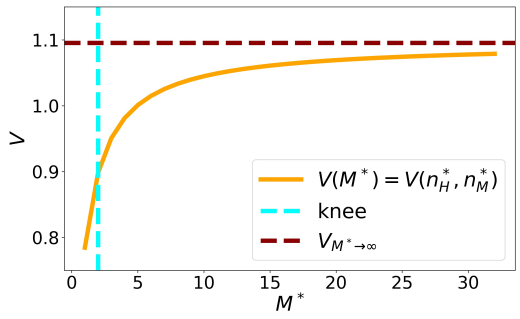

Figure 3: The relationship between mixed-type agent diversity along the optimal trajectory and ensemble size.

Specifically, they have an optimal agent ratio where enlarging the ensemble might sometimes hinder performance. As demonstrated by the red curve in Figure 2, with a constant human count $n_H = n_{H0}$, diversity $V(n_M)|_{n_H=n_{H0}}$ initially grows with machine number n_M but then wanes. Furthermore, the optimal trajectory $V(M^*)$ enables knee estimation. Figure 3 reveals that mixed-type ensembles, akin to their homogeneous counterparts, possess both an asymptote and a knee.

Summary of Accuracy Estimation Methods

Assumptions can influence both the precision and scope of estimations. Herein, we outline two techniques employed in our study to gauge ensemble accuracy. The ACC_1 accuracy metric is defined by Eq 7:

$$ACC_1 = \sum_{k=1}^{L} \frac{N_k}{N} \Phi^{L-1}\left(\frac{\sum_{m=1}^{M}(a_m - b_m)}{\sqrt{\sum_{p=1}^{M}\sum_{q=1}^{M}(\sigma_{a,p}\sigma_{a,q}\rho_{a,pq} + \sigma_{b,p}\sigma_{b,q}\rho_{b,pq})}}\right). \tag{18}$$

The ACC_2 metric, derived from Eq 11, is given by:

$$ACC_2 = \begin{cases} \Phi^{L-1}\left[\frac{\Delta}{\sigma}V(M)\right] & \text{for } N_{type} = 1, \\ \Phi^{L-1}\left[\frac{\Delta}{\sigma}V(n_H, n_M)\right] & \text{for } N_{type} = 2. \end{cases} \tag{19}$$

Here, $V(M)$ and $V(n_H, n_M)$ depict diversity expressions for homogeneous and mixed-type agents respectively, based on Eq 15 and Eq 17.

Theoretically, when it comes to measuring true accuracy, ACC_1 edges out ACC_2 because of its fewer built-in assumptions. Nonetheless, empirical results, especially on datasets like ImageNet-16H and CIFAR10H, suggest that the accuracies of the two methods are remarkably comparable. Delving deeper, ACC_2 grants a more clear understanding of the interplay between ensemble performance and individual accuracy and diversity, underscoring the significance of diversity. This method further clarifies the trade-offs between ensemble size and performance. In real-world applications, ACC_2 steers ensemble strategies, allowing for initial projections of the optimal trajectory and pivotal points.

Experiments

Experimental Setup

Human and machine image recognition capabilities differ, prompting us to utilize challenging image datasets: CIFAR10H and ImageNet-16H for experiments. **CIFAR10H**, an offshoot of CIFAR10 (Krizhevsky, Hinton et al. 2009), consists of 10,000 testing images. CIFAR10H provides an expansive 511,400 human annotations, averaging 50 per CIFAR10-test image. Given that the predictions lack confidence scores, we used knowledge distillation for human predictions, employing 50 LeNets (LeCun et al. 1998). This yielded the LeNet-H models, trained on CIFAR10H for 16 epochs. Similarly, 64 LeNet models (LeNet-M) were trained on the CIFAR10 training set for the same duration. Both model variants exhibited similar CIFAR10-test accuracies. **ImageNet-16H** fuses machine and human predictions from the ILSRVR dataset (Russakovsky et al. 2015). With 207 categories mapped to 16 classes in ImageNet-16H, each class has 75 images across four noise levels $\omega \in \{80, 95, 110, 125\}$. This aggregates to 4,800 evaluation images. Each image receives six unique human predictions from 145 participants, along with "low, medium, high" confidence scores, which, after normalization, resemble a Gaussian distribution. The dataset also contains machine predictions, including those from AlexNet (Krizhevsky, Sutskever, and Hinton 2012), DenseNet161 (Huang et al. 2017), GoogleNet (Szegedy et al. 2015), ResNet152 (He et al. 2016), and Vgg19 (Simonyan and Zisserman 2015). Each machine underwent fine-tuning for an epoch. Our research was divided: **(1)** Assessing our framework's precision and comparing real ensemble accuracy with various projections. **(2)** Employing the framework for multi-agent ensembles, exploring relationships between ensemble accuracy, scale, the optimal trajectory, and the knee point.

Estimated vs. Actual Ensemble Accuracy

We evaluated ensemble performance for varied ensemble sizes M. We compared the actual ensemble accuracy ACC with our proposed estimations ACC_1 and ACC_2[1] , and another statistical framework's estimate ACC_3 (Steyvers et al. 2022). Experimental results are available in Table 1. For clarity, the table presents averaged data from multiple scenarios. Taking ImageNet-16H with $M = 2$ as an example: the average is obtained from all combinations of four noise levels ($\omega \in [80, 95, 110, 125]$), three ensemble types (HH, MM, HM), and the mean results from the first four models in each type. This amounts to an aggregate of 256 combinations. Our experiments underscore that ACC_2, rooted in minimal assumptions, aligns closely with the actual ACC. ACC_1 offers commendable accuracy, though not quite as high, while ACC_3 trails in precision among the trio. The assumptions made by ACC_3 – such as equal variances in correct and incorrect predictions and consistent correlations – distinguish it from our framework. The results in Table 1 provide averages over varied noise levels. These details fur-

[1]Our work is available at https://github.com/Ticus0228/HM-decision-making.

		ACC	ACC_1		ACC_2		ACC_3	
	M	$\mu \pm \sigma$	$\mu \pm \sigma$	Δ	$\mu \pm \sigma$	Δ	$\mu \pm \sigma$	Δ
ImageNet-16H ($\omega \in [80, 95, 110, 125]$)	1	0.770±0.109	0.655±0.228	-0.115	0.763±0.169	-0.006	0.097±0.088	-0.673
	2	0.815±0.089	0.753±0.177	-0.062	0.845±0.124	0.030	0.240±0.110	-0.575
	4	0.864±0.074	0.841±0.133	-0.023	0.908±0.086	0.045		
	8	0.910±0.052	0.901±0.088	-0.009	0.957±0.048	0.046		
$\overline{\lvert\Delta\rvert}$				0.052		<u>0.032</u>		0.624
CIFAR10-H	1	0.625±0.011	0.434±0.019	-0.191	0.434±0.019	-0.191	0.371±0.025	-0.253
	2	0.676±0.010	0.499±0.015	-0.177	0.501±0.015	-0.176	0.411±0.009	-0.265
	4	0.708±0.011	0.538±0.013	-0.170	0.540±0.014	-0.167		
	8	0.723±0.010	0.554±0.011	-0.169	0.558±0.011	-0.165		
	16	0.737±0.011	0.567±0.010	-0.170	0.571±0.010	-0.166		
	32	0.744 ±0.010	0.572 ±0.009	-0.172	0.576 ±0.009	-0.168		
$\overline{\lvert\Delta\rvert}$				0.175		<u>0.172</u>		0.259

Table 1: Comparison of actual and estimated ensemble accuracies. ACC_1 and ACC_2 are outcomes of our framework, while ACC_3 follows (Steyvers et al. 2022). Note, ACC_3 is inapplicable when $M > 2$. Accuracy is presented as mean ± standard deviation ($\mu \pm \sigma$). Δ represents the difference between estimated and actual accuracies, while $\overline{\lvert\Delta\rvert}$ signifies the average absolute estimation error. The smallest average error is highlighted with an underline.

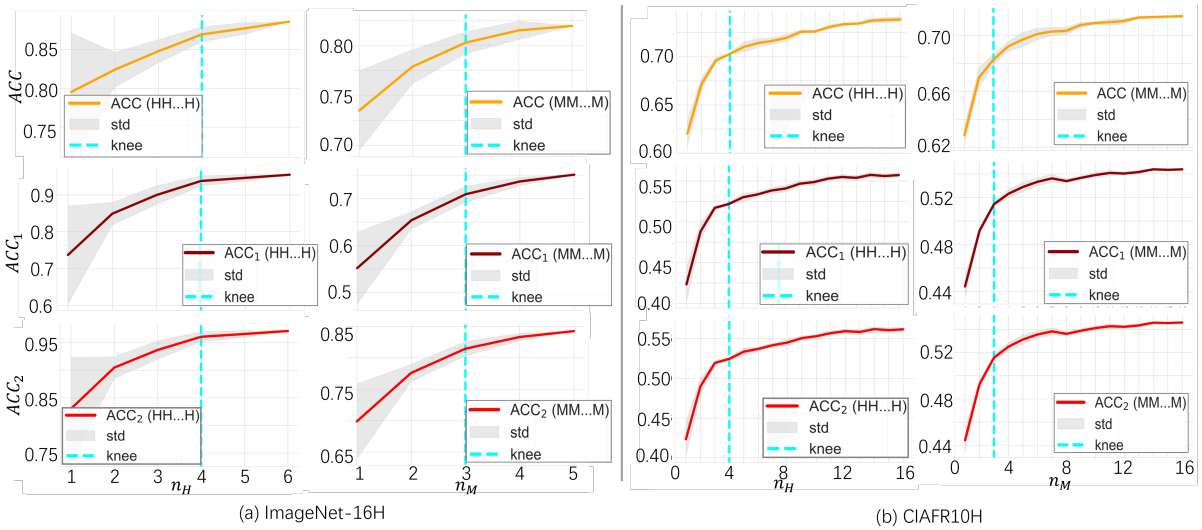

Figure 4: Homogeneous ensemble performance trends with ensemble size M on ImageNet-16H and CIFAR10H datasets. The three curves, from top to bottom, represent ACC, ACC_1 and ACC_2. Knees in the trends are delineated with vertical lines.

ther prove the strength and advantage of our framework compared to others

Homogeneous Agent Ensemble Experiments

Homogeneous ensembles, comprising either solely humans or machines, show distinct performance patterns. Figure 4 illustrates how ensemble performance varies with ensemble size M on ImageNet-16H (at $\omega = 110$) and CIFAR10H datasets. The averaging technique adheres to earlier descriptions, selecting either four random combinations or the entirety if fewer than four, and determining their mean. As seen, ACC generally amplifies with a growing M, but the rate decelerates, suggesting a performance ceiling – consistent with our theoretical findings in Figure 1. When ACC_1 and ACC_2 are employed to estimate the true ACC, the precision of their estimates is notably comparable. The identified 'knees' in the trends, marked by vertical lines, affirm that estimated knees from both ACC_1 and ACC_2 are closely aligned with actual ones. The concurrent trend trajectory across the trio further reinforces our confidence in statistical approaches for pinpointing knees

Heterogeneous Agent Ensemble Experiments

We use mixed-type intelligent agent ensemble to denote collaborations involving both humans and machines. Figure 5 shows the actual performance (ACC) alongside the predicted ensemble performance (ACC_2) for these agents on the ImageNet-16H ($\omega = 110$) and CIFAR10H datasets, using a consistent averaging approach. The heatmap reveals a basic congruence between the actual and estimated trends of ensemble performance. Predominantly, ACC ascends with a swelling ensemble size ($M = n_H + n_M$), resonating with

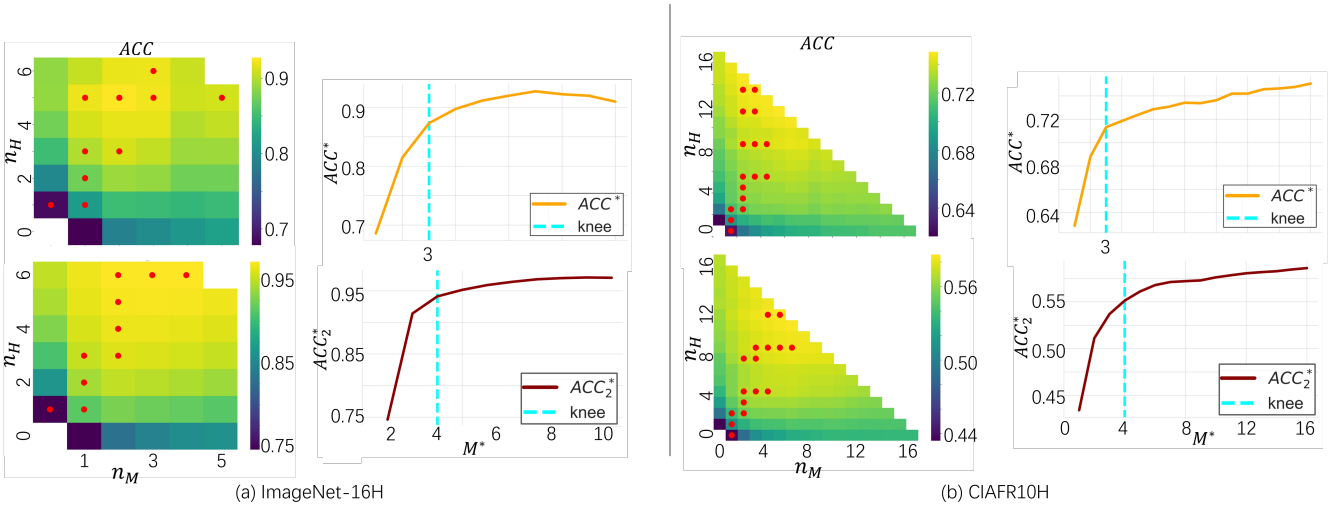

Figure 5: Performance dynamics for ensembles comprised of both humans and machines are illustrated across different sizes on the ImageNet-16H and CIFAR10H datasets. The *heatmap* displays both actual and estimated accuracies, varying with the count of humans (n_H) and machines (n_M). The optimal paths are marked by red dots, with actual paths shown above and estimated paths below. The *curves* compare ACC (top) with ACC_2 (bottom), each corresponding to the optimal ensemble size ($M^* = n_H^* + n_M^*$). Knee points in the curves are highlighted with vertical markers.

our ACC_2 anticipations. Red markers pinpoint the optimal trajectory, consistently mapped across datasets. From this heatmap, we distill and graph the ACC^* trend.

Upon analysis, the optimal path ACC^* on both datasets consistently increases with M^*, suggesting a potential saturation point. However, it's important to note that on the ImageNet-16H dataset, a drop in ACC^* is observed when $M^* \in \{7, 8, 9\}$. While this might seem contradictory to our initial findings, it can actually be attributed to the limited numbers of both humans and machines involved. For instance, at $M = 10$, given only 6 humans and 5 machines, optimal configurations such as $(n_M, n_H) = (3, 7)$ might be overlooked. The vertical markers on the curves denote the knee points. On ImageNet-16H, the observed and estimated knees align, whereas on CIFAR10H, there's a slight deviation with the actual knee at $M^* = 3$ and the estimated one at $M^* = 4$. Notwithstanding these small variations, the estimated optimal path predominantly aligns with the empirical data.

Experimentally, the estimation we proposed, ACC_1 and ACC_2, showcase superior precision in approximating the true ACC when compared to the precision of ACC_3. Additionally, ACC_2's trend aligns closely with the genuine ACC. On the theoretical front, we've validated numerous propositions intrinsic to ACC_2, yielding deep insights into aspects such as the upper limits of intelligent agent ensembles and the nexus between ensemble performance and size, among others. From a practical standpoint, we've illustrated how the equations tied to ACC_2 can be harnessed to pre-emptively predict optimal trajectories, knee points, and the like. These revelations are pivotal in devising ensemble strategies, striking an optimal balance between ensemble performance and resource utilization.

Conclusion

This paper presented a statistical approach to evaluate homogeneous and heterogeneous agent ensembles. Our research indicates that while ensemble accuracy for homogeneous agents usually correlates with size, mixed-agent ensembles follow a distinct trajectory of improvement along the optimal path. Our analysis identified a pivotal knee in the performance-size curve, providing insights into optimizing performance while economizing resources. The ongoing debate in ensemble learning revolves around balancing individual accuracy with diversity. Our framework aligns with prevailing theories, suggesting ensemble performance can be distilled into these two components. More innovatively, we clarified their interrelationship through a simple multiplicative model, offering clarity on their combined influence on ensemble outcomes. While our insights are profound, our model's assumptions carry inherent constraints. Future efforts aim to develop a more detailed decomposition framework, further revealing the intricate relationship between accuracy and diversity in ensemble learning.

Acknowledgments

This work was financially supported by the National Key R&D Program of China (2022ZD0117103), National Natural Science Foundation of China (62293554, 62077021), Hubei Provincial Natural Science Foundation of China (2023AFA020, 2022CFB414), and Fundamental Research Funds for the Central Universities (CCNU22LJ005).

References

Aggarwal, I.; Woolley, A. W.; Chabris, C. F.; and Malone, T. W. 2019. The impact of cognitive style diversity on implicit learning in teams. *Frontiers in Psychology*, 10: 1–11.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021. Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11405–11414.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2–11.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*. Springer.

Breiman, L. 1996. Bagging predictors. *Machine Learning*, 24(2): 123–140.

Brown, G.; Wyatt, J. L.; Tino, P.; and Bengio, Y. 2005. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6(9): 1621–1650.

Davis-Stober, C. P.; Budescu, D. V.; Broomell, S. B.; and Dana, J. 2015. The composition of optimally wise crowds. *Decision Analysis*, 12(3): 130–143.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2611–2620.

Freund, Y.; and Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*, 148–156.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232.

Geirhos, R.; Meding, K.; and Wichmann, F. A. 2020. Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33: 13890–13902.

Geirhos, R.; Temme, C. R.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31: 7538–7550.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT Press.

Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Ho, T. K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8): 832–844.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Juni, M. Z.; and Eckstein, M. P. 2017. The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, 114(21): E4306–E4315.

Kerrigan, G.; Smyth, P.; and Steyvers, M. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34: 4421–4434.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Master's Thesis, Department of Computer Science, University of Toronto*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25: 1097–1105.

Krogh, A.; and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, 7: 231–238.

Kuncheva, L. I. 2006. On the optimality of Naïve Bayes with dependent binary features. *Pattern Recognition Letters*, 27(7): 830–837.

Kuncheva, L. I. 2014. *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons.

Kuncheva, L. I.; and Whitaker, C. J. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2): 181–207.

Lamberson, P.; and Page, S. E. 2012. Optimal forecasting groups. *Management Science*, 58(4): 805–810.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Murphy, K. P. 2012. *Machine learning: A probabilistic perspective*. MIT Press.

Noy, S.; and Zhang, W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654): 187–192.

Page, S. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, 372–387.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, 9617–9626.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning*, 5389–5400.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond accuracy: Behavioral testing of NLP models with checkList. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 4902–4912.

Riedl, M. O. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1): 33–36.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Sagi, O.; and Rokach, L. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4): e1249.

Satopaa, V.; Albrecht, J.; Irwin, D.; and Raghavan, B. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems Workshops*, 166–171.

Serre, T. 2019. Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, 5: 399–426.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119–e2111547119.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Tejeda, H.; Kumar, A.; Smyth, P.; and Steyvers, M. 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 5(4): 491–508.

Tumer, K.; and Ghosh, J. 1996. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4): 385–404.

Ueda, N.; and Nakano, R. 1996. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks*, 90–95.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wilder, B.; Horvitz, E.; and Kamar, E. 2021. Learning to complement humans. In *Proceedings of International Conference on International Joint Conferences on Artificial Intelligence*, 1526–1533.

Woolley, A. W.; Chabris, C. F.; Pentland, A.; Hashmi, N.; and Malone, T. W. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004): 686–688.

Wu, M.; and Wu, X. 2019. On big wisdom. *Knowledge and Information Systems*, 58(1): 1–8.

Wu, X.; Wang, X.; Jin, B.; Yu, Z.; and Wu, M. 2022. *Human-machine synergy*. China Science Publishing & Media.

Zahedi, Z.; and Kambhampati, S. 2021. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*.

Zhang, R.; McNeese, N. J.; Freeman, G.; and Musick, G. 2021. "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–25.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3): 1–41.

Zhou, Z.-H. 2020. *Ensemble methods: Foundations and algorithms*. Publishing House of Electronics Industry.