The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Human and AI Perceptual Differences in Image Classification Errors

Minghao Liu[1], **Jiaheng Wei**[2], **Yang Liu**[1], **James Davis**[1]

[1] Department of Computer Science and Engineering, University of California, Santa Cruz
[2] Data Science and Analytics Thrust, Hong Kong University of Science and Technology (Guangzhou)
mliu40@ucsc.edu, jiahengwei@hkust-gz.edu.cn, yangliu@ucsc.edu, davisje@ucsc.edu

Abstract

Artificial intelligence (AI) models for computer vision trained with supervised machine learning are assumed to solve classification tasks by imitating human behavior learned from training labels. Most efforts in recent vision research focus on measuring the model task performance using standardized benchmarks such as accuracy. However, limited work has sought to understand the perceptual difference between humans and machines. To fill this gap, this study first analyzes the statistical distributions of mistakes from the two sources and then explores how task difficulty level affects these distributions. We find that even when AI learns an excellent model from the training data, one that outperforms humans in overall accuracy, these AI models have significant and consistent differences from human perception. We demonstrate the importance of studying these differences with a simple human-AI teaming algorithm that outperforms humans alone, AI alone, or AI-AI teaming.

1 Introduction

One motivation of neural networks (NN) is creating artificial intelligence that can learn from human intelligence and mimic human behavior. In computer vision, researchers often build their work upon the assumption that neural networks learn a feature representation similar to visual cortex activity (Agrawal et al. 2014; Mur et al. 2013; Kuzovkin et al. 2018). It is believed that a well-trained machine network learns to represent input stimuli in a way that is similar to human visual perception (Battleday, Peterson, and Griffiths 2020). As a result, most current work in computer vision that aims to develop "better" computer models focuses on summary benchmark scores (e.g., prediction accuracy). As a result, the evaluation of human-to-machine similarity across the entire distribution of answers is ignored.

Human-centric studies do exist in some areas of AI research. Human-centric explainable AI conducts user studies to identify the "explanations" that are most meaningful to human users. Noisy label researchers utilize human disagreements as soft labels to train neural networks (Branson et al. 2010; Kovashka et al. 2016; Peterson et al. 2019; Wei et al. 2023). Human-in-the-loop work utilizes active learning to iteratively expand or update the training set by using human annotators (Han, Dong, and Demartini 2021; Joshi, Porikli, and Papanikolopoulos 2009; Yao et al. 2012; Taleb et al. 2021; Roels et al. 2019), or utilizing human feedback reinforcement learning methods to bridge the performance gap between humans and machines (Yang et al. 2024; Ouyang et al. 2022; Shnayder et al. 2016; Ramakrishnan et al. 2019). Human workers have also been requested to diagnose AI system failures given semantic explanations of the AI decision (Yang and Alonso 2024; Nushi, Kamar, and Horvitz 2018; Nushi et al. 2017). However, these works focus on developing training schemes to improve the model performance, and limited work explores the perceptual difference between human and machine classifiers. [1]

Our work quantifies the perceptual difference between humans and machines in the image classification task. We first analyzed the statistical distribution of mistakes made by human annotators and AI annotators (robust machine learning classifiers), where we found that *machines tend to make similar mistakes to other machines, while mistakes differ from those made by humans*. Next, we attempt to understand the differences using implicit information from machine and human classifiers, such as machine confidence scores, human annotation time, and agreement levels, to rank the difficulty of each classification task. Comparing the performance of humans and machines on various task difficulty levels, we observe that *despite being worse overall, humans outperform the machines in some conditions, such as when machines have low confidence or low agreement*. Finally, we demonstrate the potential of studying human-to-machine differences by using our findings to show that appropriate human-machine collaboration significantly outperforms collaboration between any two machine classifiers.

The contributions of this work are:

- A comparison of human-machine perceptual differences on an image classification task.
- Showing that understanding perceptual differences can improve overall system accuracy.

2 Preliminary

2.1 The Multi-Class Classification Task

In this work, we focus on the multi-class image classification task. Assume that labels $y \in [1, 2, ..., K]$ are generated by

[1] More related works are included in appendix (arXiv version).

Figure 1: **Stimuli for current study:** Representative images from the CIFAR-10 dataset, which includes ten categories of natural images. The perceptual difference between human and AI classifiers is studied using the distribution of mistakes made while predicting categorical labels.

a random variable Y. Ideally, a machine learner would have access to N training data points $D := \{(x_i, y_i)\}_{i \in [N]}$, where each image/instance x_i generates according to the random variable X. The goal of this classification task is to find the optimal classifier f by solving the optimization task:

$$f^* \leftarrow \arg\min_{f} \frac{1}{N} \sum_{i \in [N]} \ell(f(x_i), y_i),$$

where f is the classifier (e.g., the machine) and ℓ is the evaluation measure (i.e., loss function).

However, labels are often obtained through a crowdsourcing platform (e.g., Amazon Mechanical Turk (Turk 2012)) in practice, which can inevitably incur label/annotation errors to certain images (Xiao et al. 2015; Peterson et al. 2019; Wei et al. 2022c). In other words, the learner can only access a set of noisy labels $\tilde{y}_i$ generated by the random variable $\widetilde{Y}$, which may disagree with the clean label y_i.

2.2 Datasets and Model Training

The top-ranked machine vision models can achieve extremely high accuracy on CIFAR-10 (Krizhevsky, Hinton et al. 2009) image classification by training on clean labels. The models are so good with clean labels that this dataset is often considered a toy problem domain. However, label noise is prevalent in real-world classification data (Deng et al. 2009; Xiao et al. 2015; Wei et al. 2022c). Training models on datasets with label noise inevitably results in biased model prediction (Natarajan et al. 2013; Liu and Tao 2015; Patrini et al. 2017). Thus CIFAR-10 remains a widely used test dataset when studying image classification using noisy training data (Liu and Guo 2020; Cheng et al. 2020; Liu et al. 2020; Wei and Liu 2021; Liu et al. 2022; Wei et al. 2022a; Wang et al. 2022; Huang et al. 2023; Xia et al. 2023; Ortiz-Jimenez et al. 2023; Chen et al. 2023b,a; Park et al. 2024; Wang et al. 2024). The machine classifier must decide which of ten class labels (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) best describes each image. An example of images in this dataset is shown in Figure 1.

		Accuracy (%)
Humans	Human 1	82.8
	Human 2	81.9
	Human 3	82.4
Machines	CORES (Cheng et al. 2021)	84.5
	CE	85.0
	PLS (Lukasik et al. 2020)	85.9
	F-Div (Wei and Liu 2021)	86.0
	GCE (Zhang and Sabuncu 2018)	86.3
	FW (Patrini et al. 2017)	86.7
	PeerLoss (Liu and Guo 2020)	86.8
	BW (Patrini et al. 2017)	86.8
	NLS (Wei et al. 2022b)	88.4
	CAL (Zhu, Liu, and Liu 2021)	88.7
	Co-teaching+ (Yu et al. 2019)	89.1
	JoCoR (Wei et al. 2020)	89.5
	Co-teaching (Han et al. 2018)	90.0

Table 1: **Overall image classification prediction accuracies**: Robust machine learning models are trained on noisy human labels in CIFAR-N (40K training subset). The accuracies of humans and machines are calculated against the clean label in CIFAR-10 (10K test subset). Note that all machine models outperform all noisy human annotators.

In this paper, we adopt CIFAR-N (Wei et al. 2022c), a label-noise benchmark that provides three noisy human annotations for each image of the CIFAR-10 training dataset. We split the training set into a 40K training subset and a 10K test subset. We explore human perceptual differences using these noisy human annotations. For machine classifiers, there exists a family of robust designs for learning with noisy labels. We choose thirteen popular robust methods, allowing the machine to mitigate the impact of noisy labels.

A preliminary comparison of human and machine classifier accuracy is given in Table 1. Noisy human annotators from CIFAR-N achieve 81.9-82.8% accuracy. However, all the selected machine classifiers achieved higher accuracies than humans, ranging from 84.5% to 89.9%, when trained using data from CIFAR-N. A higher accuracy score is widely believed to imply that machines can replace human annotators in this classification task. Our study investigates: (1) Whether perceptual differences exist between humans and machines; and (2) Are these differences important to the overall machine performance?

3 Quantifying Perceptual Differences

In this section, we study the perceptual difference between humans and machines in image classification tasks. Our objective is to determine whether human capabilities are merely a subset of machine abilities or if they can complement each other. To represent their respective perceptual distributions, we use a set of machine predictions and noisy human annotations to conduct our analysis. To quantify and compare the distribution of predictions, Sec 3.1 makes use of confusion matrices to illustrate the pattern of their incorrect predictions, Sec 3.2-3.3 demonstrates their performance on tasks with various difficulty levels, and Sec 3.4 shows additional results when permuting training data and learning backbone.

3.1 The Prediction Confusion Matrix

To understand the overall perceptual distribution difference between human and machine classifiers, we visualize their incorrect predictions through confusion matrices.

Denote by f_{ML} a machine learning classifier, and f_{H} as a human annotator. For $f \in \{f_{\text{ML}}, f_{\text{H}}\}$, let $\mathbf{f}(x)$ be the model prediction probability vector given by the classifier f w.r.t. the sample $x \in X$, i.e., $\mathbf{f}(x) = [\mathbf{f}(x)_{[1]}, ..., \mathbf{f}(x)_{[K]}]$, where $\mathbf{f}(x)_{[i]} := \mathbb{P}(f(x) = i)$ denotes the probability of categorizing x as class i, in a K-class classification task. The model prediction could then be expressed as $f(x) := \arg\max_{i\in[K]} \mathbf{f}(x)_{[i]}$. Given a set of test data samples with clean labels $D^{\text{t}} := \{(x_i^{\text{t}}, y_i^{\text{t}})\}_{i\in[N]}$, for $f \in \{f_{\text{ML}}, f_{\text{H}}\}$, we aim to quantify the differences between f_{ML} and f_{H} through the following metric:

Confusion matrices of model predictions C: The elements of confusion matrix $\mathbf{C}$ are given by

$$\mathbf{C}_{p,q} := \mathbb{P}_{(X,Y)\sim D^{\text{t}}}(f(X) = q|Y = p).$$

To quantify the perceptual difference between humans and machines, one could replace the ground-truth (clean) label Y with the human annotation $f_{\text{H}}(X)$ and adopt $f = f_{\text{ML}}$.

Figure 2 shows the aggregated confusion matrices for samples with incorrect prediction between human-human, human-machine, and machine-machine. For instance, the "Humans v.s. Machines" figure compares three sets of human annotations against all thirteen designs of machine classifiers. By examining these confusion matrices, we can gain insight into the distribution of human and machine mistakes.

Comparisons between machines and machines: Denote by $\mathbf{C}^{\text{MM}}$ the confusion matrix of model predictions between a machine and a machine. Figure 2 (b) shows the aggregated confusion matrix $\mathbf{C}^{\text{MM}}$ between all pairs of machines. We observed a strong diagonal line in the plot, indicating that machines have a strong consensus and tend to make similar mistakes.

Comparisons between humans and humans: We define $\mathbf{C}^{\text{HH}}$ as the confusion matrix of model predictions between a human annotator and another human annotator. The human-human comparison $\mathbf{C}^{\text{HH}}$ is shown in Figure 2 (a). The diagonal line is weaker since human judgments sometimes diverge. By comparing $\mathbf{C}^{\text{HH}}$ and $\mathbf{C}^{\text{MM}}$ (Figure 2 a,b), we conclude that although humans in our test set have similar accuracies, they sometimes make different mistakes. In contrast, machines models, even those with different accuracies, tend to make similar mistakes.

Comparisons between machines and humans: Denote by $\mathbf{C}^{\text{HM}}$ the confusion matrix of model predictions between a human annotator and a machine. The human-machine comparison $\mathbf{C}^{\text{HM}}$ is shown in Figure 2 (c). By comparing the confusion matrices $\mathbf{C}^{\text{MM}}$ vs. $\mathbf{C}^{\text{HM}}$ (Figures 2 b,c), we see that humans and machines make significantly different mistakes.

Hypothesis testing is utilized to support our visual observations. With a significance level of $\alpha = 0.05$, the hypothesis testing results showed there exist significant differences between $\mathbf{C}^{\text{HH}}$ vs. $\mathbf{C}^{\text{MM}}$ and $\mathbf{C}^{\text{MM}}$ vs. $\mathbf{C}^{\text{HM}}$. Details are provided in Appendix B.1.

The perceptual difference between humans and machines suggests human capabilities aren't strictly a subset of machine abilities, implying the potential for complementary teaming.

3.2 Partitioning by Machine Difficulty

To better understand perceptual differences, we partition test images into subsets that may have performance variations. As a first attempt at partitioning, we group test cases based on the difficulty level for machine learning models f_{ML} to provide an accurate answer. The performance of humans and machines is evaluated for each subset. We try two difficulty measures: the machine's model confidence and agreement.

Machine's model confidence: Define model confidence as $\text{MC}(f = f_{\text{ML}}|x) := \mathbf{f}_{\text{ML}}(x)_{f_{\text{ML}}(x)}$ or $\arg\max_{j\in[K]} \mathbf{f}_{\text{ML}}(x)_j$, which indicates the model's prediction probability on sample x for its prediction $f_{\text{ML}}(x)$.

The machine confidence score is a commonly used metric to show the model prediction certainty level. A higher confidence score indicates a higher level of certainty and suggests that the task is easier for the machine to perform. In our experiments, we calculated the accuracy of humans and machines at different confidence levels using each machine classifier. Results are plotted in Figure 3 (a). The shaded band indicates the range of accuracies, and the solid line represents the average. The performance of the thirteen machine designs is shown in red, while the results of human classifiers are shown in blue.

Examining the plot, only the red band shows a strong upward trend when the confidence level increases, meaning that the machine performance strongly correlates to the machine confidence score. In contrast, human performance is only mildly correlated with machine confidence. In addition, the plot shows that humans are more accurate than machines when machines have low confidence. Even though machines have higher accuracy on the complete test set, humans and machines have different expertise on different subsets, neither can fully replace the other on this classification task.

Machine agreement: Given a list of machine learning classifiers, i.e., k classifiers with j-th one denoted by $f_{\text{ML},j}$, we could quantify the difficulty level for machines w.r.t. a sample x as $\text{MA}(f = f_{\text{ML}}|x) := \frac{1}{k}\sum_{j\in[k]} \mathbf{1}(f_{\text{ML},j}(x) = y)$. The model agreement term $\text{MA}(f = f_{\text{ML}}|x) \in [0, 1]$ indicates a high consensus rate for machine learning classifiers (low difficulty level) if it is of a high value.

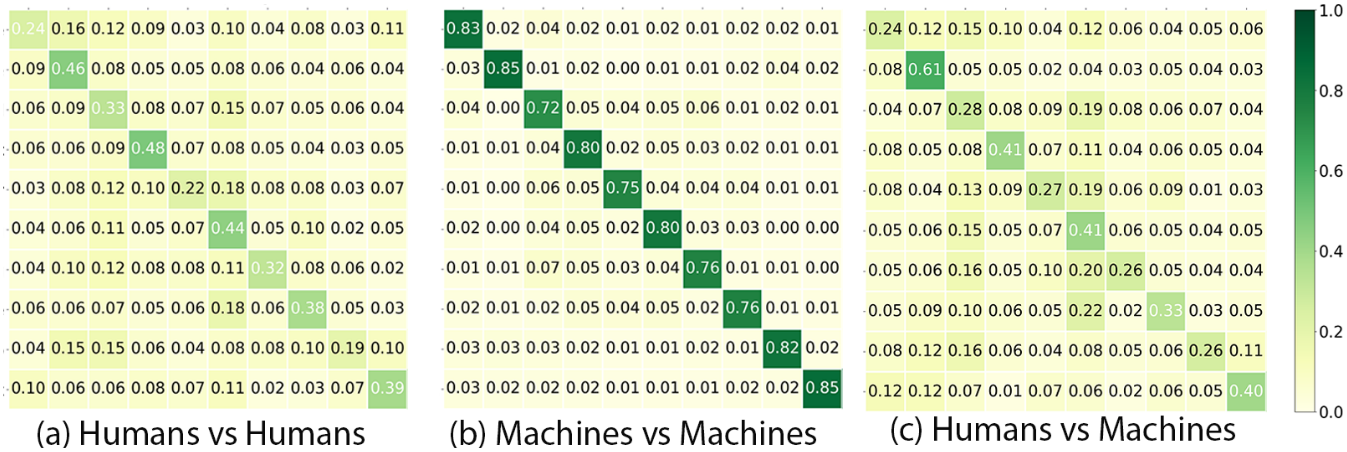


Figure 2: **Confusion matrices of incorrect answers:** The figure shows confusion matrices across permutations of machine classifiers or human annotators. The plot focuses on incorrect predictions from the test subjects to see if they make similar mistakes. For example, a darker cell in "Humans vs Machines" means a higher probability that the three human annotators make the same mistake as the thirteen designs of machine classifiers. (a) A mild diagonal line indicates that humans don't always make the same mistakes. (b) The strong diagonal line indicates that all the machine models tend to make similar mistakes. (c) The diagonal line is weak, indicating that the mistakes made by humans and machines diverge in this comparison.

Machine agreement uses a cluster of machine classifiers to rank the task difficulty. More machines making correct judgments on a task indicates the task is easier to solve, and vice versa. We repeat the experiment with a leave-one-out design, using 12 machine classifiers to determine the difficulty and one remaining classifier to test. We experimentally compare human and machine accuracies on tasks with different machine agreement levels, then visualize the difference in Figure 3 (b). The result again shows that machine performance strongly correlates with other machine classifiers, even though various machine designs have a wide range of performance. Human annotator performance, on the other hand, is only slightly correlated to machine agreement. Here, we draw a similar conclusion to using machine confidence levels: human and machine classifiers have different distributions of predictions.

Statistical hypothesis testing: We verify our visual observations using a linear regression fit, which shows the correlation between the accuracy and task difficulty levels. A larger slope indicates a stronger relation, while values closer to zero indicate a low correlation. We set the significance value to be $\alpha = 0.05$: the average of human classifiers has a slope of 0.11 w.r.t. the machine confidence and a slope of 0.08 with machine agreements, while machines have a slope of 1.00 in both cases. This illustrates that the difficulty level, as judged by machines, has a low correlation with human predictions. Details are provided in Appendix B.2, B.3.

3.3 Partitioning by Human Difficulty

We repeat the partitioning experiments via new criteria: this time partitioning by human difficulty rather than machine difficulty. To quantify the human difficulty levels, we use time spent labeling an image and an entropy-based measure of human agreement.

Time Spent: Assuming the i-th human annotator spent $t_i(x)$ time (in seconds) in annotating sample x, we adopt the average time spent on sample x as a measure to indicate the difficulty level of the given task x. Mathematically, $\bar{t}(x) := \frac{1}{k}\sum_{i\in[k]} t_i(x)$. A large $\bar{t}(x)$ means the sample x is relatively hard for human annotators since it requires humans to spend a long time on annotation. We calculated average time consumption using the CIFAR-H dataset (Peterson et al. 2019) since CIFAR-N lacks annotator time information.

Human agreement (entropy): Given k human annotators, the entropy of a sample x is given by:

$$\text{Entropy}(x) = -\sum_{i\in[K]} p_{\text{H},i}(x)\cdot\log\left(p_{\text{H},i}(x)\right),$$

where $p_{\text{H},i}(x) := \frac{1}{k}\sum_{j\in[k]} \mathbf{1}(f_{\text{ML},j}(x) = i)$.
Human agreement is a metric to evaluate human consensus on a task given multiple annotations. A higher agreement level implies less ambiguity and easier judgment. We calculate entropy using CIFAR-H because it provides at least 47 labels per image, enough to calculate agreement.

Analysis: We calculated human and machine performance under different human difficulty levels, then aggregated results as shown in Figure 3 (c,d). Both human and machine performance strongly correlate with annotation time and human agreement. Here, we conclude that machines will struggle with cases humans find difficult. While this result contrasts our findings using machine difficulty to partition tasks, it does not contradict those results. Machine and human difficulty metrics result in different partitions, and the statistics of these partitions are different.

3.4 Sensitivity of the Results

Comparing human and machine classifiers, we have found that machines are more similar to other machines and that

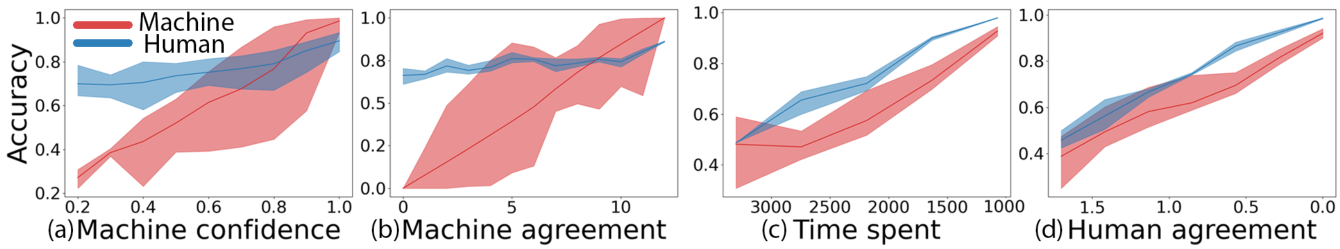


Figure 3: **Accuracy as a function of difficulty level:** The plots visualize the performance of humans and machines on tasks ranked by difficulty level. The shaded band indicates the range of accuracies for all classifiers, and the solid line represents the average performance. Task difficulty is measured by: **(a)** machine classifier confidence levels, **(b)** based on machine agreements, **(c)** based on human annotation time, **(d)** and human agreement levels. Plots (a,b) show machine classifier performance heavily correlates to machine difficulty levels, while human performance is significantly less correlated. Plots (c,d) indicate that both human and machine performance is correlated to human-derived difficulty levels.

humans have different statistics. This result may come from using the same training data for all machine classifiers or, alternately, from using the same vision backbone on all classifiers. To understand whether this is the cause, we conduct further experiments to provide variation to the machine classifiers during the training phase.

To ensure fairness in this test, we created a balanced test set to reduce the bias potentially caused by the majority of correct cases. A balanced set is a subset randomly selected from the test set, which includes half correct and half incorrect answers. To measure the similarity of two classifiers, we used *Matching percentage on a balanced set* as the numerical metric, defined as $\frac{P(A|B)+P(\overline{A}|\overline{B})}{2}$ for two sets A, B.

Training labels: We diversify the training labels to evaluate whether using the same training labels is the cause of different machine classifiers making similar mistakes. We train each machine on one of the three human annotations sets from the CIFAR-N dataset and compare their Matching percentage on a balanced test set in Figure 4. Higher matching percentages mean the two classifiers are more similar (darker colors). Notice that machines and humans tend to make less similar judgments, while machines tend to make similar judgments to other machines, even when trained with different sets of labels.

Vision backbones: We selected three commonly used vision backbones: ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2014), and Inception (Szegedy et al. 2015). The machine classifiers were retrained, each using a randomly selected backbone. The matching percentage was again used as a metric for similarity. For space reasons, the table of numbers from this experiment is in the supplementary materials (Appendix C.1). However, we draw the same conclusion. Machines and humans make different judgments, while two machine classifiers tend to make similar judgments, even when trained with different vision backbones.

4 Human-Machine Collaboration

We investigate the performance of joint systems using human and machine statistical distributions. Although machines have higher overall accuracies in our study, there exist cases in which humans perform better. The goal of this section is to demonstrate that the study of perceptual differences has practical value. We thus choose the simplest human-machine collaboration method possible, in which a joint classifier has access to both a machine guess and a human guess and returns one of these two answers. We investigate an upper bound on performance using an ideal collaboration and obtainable performance using a realistic collaboration based on simple thresholding.

4.1 Ideal Collaboration

Given predictions from a machine classifier and a human classifier, ideal collaboration selects the correct answer if either classifier generates it. Algorithms that use ground truth data to make perfect predictions are frequently called "oracles" in the computer vision literature, and we adopt this terminology, labeling ideal collaboration "oracle mode." If either classifier predicts the correct result, oracle mode will have the correct result. This is an upper bound for what is achievable in a real system.

Oracle mode performance is given in Figure 5(a), which shows the original machine classifier performance and the performance boost obtained from teaming. We show three options for teaming with each original machine predictor: teaming with a human classifier, teaming with an aggregated group of three human classifiers, and teaming with another machine classifier. The additional machine classifier is chosen by examining all options and reporting the pair with the greatest accuracy boost.

In all cases, teaming with a human has a greater upper bound on accuracy gain than teaming with another machine model. Using the machine cross-entropy (CE) classifier as an example, we see that it has an accuracy of (84.96%). When teaming with a human classifier (81.88%), the joint system can reach 95.82% (+10.86%) under this ideal collaboration condition. The boosted accuracy is significantly greater than each classifier alone. On the other hand, teaming the CE classifier with another machine classifier such as coteaching(Han et al. 2018) (89.95%) results in a teaming performance of 92.34% (+7.38%). Note that the lower accuracy human classifier results in better teaming performance than the higher accuracy co-teaching model. Indeed,

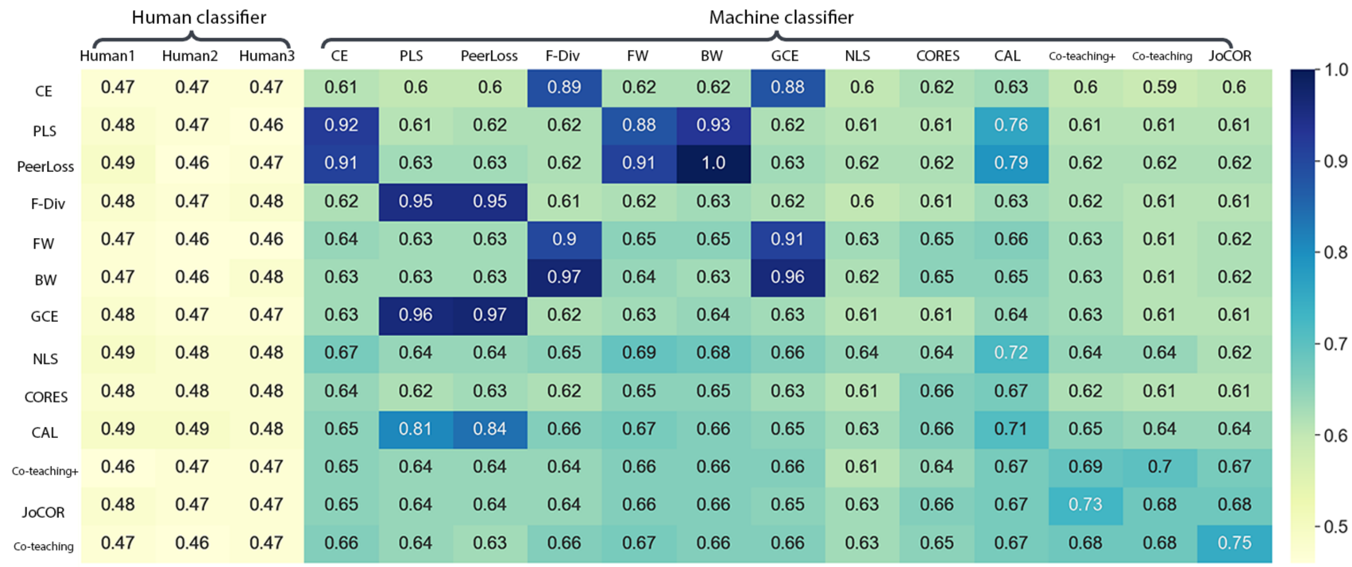


Figure 4: **Matching Percentage on balanced set:** The figure visualizes the matching percentage between each machine classifier and each other human/machine classifier on a balanced set. The machine classifiers are not all trained with the same training examples, yet the results show machines tend to make judgments that match other machines more than they match humans.

Pre-trained	$\text{Acc}_{\text{col, ml}}(0.4)$ v.s. $\text{Acc}^*_{\text{col, ml}}$		$\text{Acc}_{\text{col, ml}}(0.5)$ v.s. $\text{Acc}^*_{\text{col, ml}}$		$\text{Acc}_{\text{col, ml}}(0.6)$ v.s. $\text{Acc}^*_{\text{col, ml}}$	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
	-18.96	7.51e-6	-7.95	0.0005	-2.49	0.0555
Multi-nets	$\text{Acc}_{\text{col, ml}}(0.4)$ v.s. $\text{Acc}^*_{\text{col, ml}}$		$\text{Acc}_{\text{col, ml}}(0.5)$ v.s. $\text{Acc}^*_{\text{col, ml}}$		$\text{Acc}_{\text{col, ml}}(0.6)$ v.s. $\text{Acc}^*_{\text{col, ml}}$	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
	-66.25	0.0002	-37.57	0.0007	-34.12	0.0009

Table 2: Results of hypothesis test2 with paired student t-test: we adopt the significance value $\alpha = 0.05$.

the machine-machine teaming accuracy is only mildly better than co-teaching accuracy alone.

To investigate whether multiple human classifiers can further improve performance, we aggregate annotations from three human annotators using the majority vote to select the classification. This *aggre* classifier `(90.97%)`, has an accuracy much greater than a single human `(81.88%)`, but under the oracle teaming condition boosts accuracy only a little more to `97.13%(+12.17%)`.

The upper bound of teaming between machine and human classifiers outperforms any combination of machine-machine classifiers, revealing the benefit of leveraging human-machine complementary characteristics.

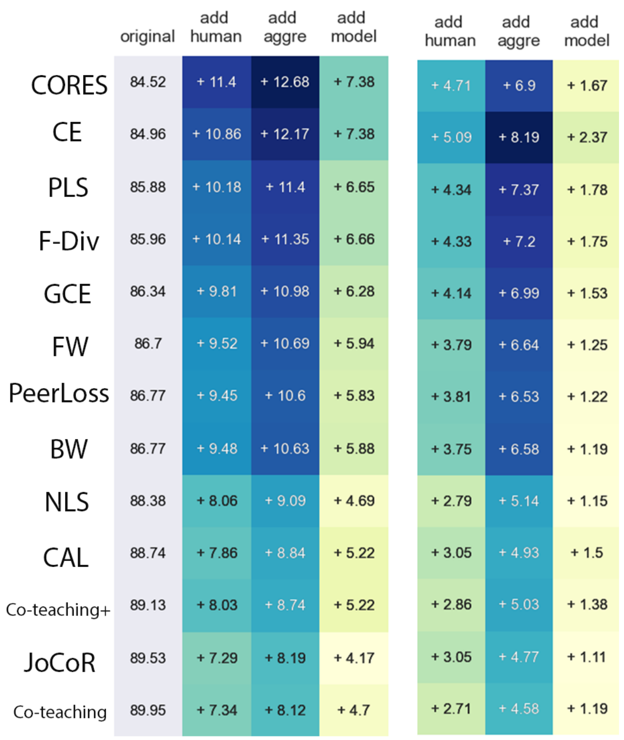
4.2 Realistic Collaboration

To understand performance in a realistic scenario without access to an oracle, we evaluate a simple algorithm that thresholds machine confidence to determine whether to use the machine or human guess.

The early memorization effect inspires thresholding. The learning-with-noisy-label literature (Liu et al. 2020; Wei et al. 2022c) observes that deep neural nets tend first to memorize clean patterns, then gradually fit on noisy patterns as training progresses. Samples that are of bad patterns usually have large losses, which degrades the generalization of the model. When evaluating the model performance on the test data, we would expect test samples that are similar to the learned training samples with clean labels to have relatively high confidence. Inspired by this effect, we replace low-confidence machine model predictions with those from a human classifier while keeping confident model predictions from the original model. This is not guaranteed to boost accuracy since the human classifier has lower overall accuracy, and sometimes, a correct classification will be replaced with an incorrect one. However, we saw previously that model confidence is strongly correlated with model accuracy, so we have reason to believe that the approach works.

We adopt a threshold value $\eta \in [0, 1]$ to determine when to swap the predictions of the machine predictors with the human classifier predictions. Selecting η to be 0 or 1 would use either all predictions from the machine predictors or all from the human classifier. To select the optimal threshold value, we use the paired student t-test to compare the algorithm's accuracy with different threshold values. The test shows that swapping model predictions with confidence $\eta \leq 0.6$ results in significant accuracy improvements, with a significance value less than $p = 0.05$. More details about the t-test and its results can be found in next section.

Fig 5(b) shows the performance boost under the same three teaming options used previously. Teaming with a human classifier produces a greater boost to accuracy than teaming with the best-choice machine model. All combi-



(a) Oracle mode (b) Realistic mode

Figure 5: **Post-hoc teaming:** The figure shows the original model performance and the boost from teaming options. *"Add human"* is teaming with a human classifier, *"Add aggre"* is teaming with a human classifier that aggregates answers from three humans, and *"Add model"* is teaming with another machine classifier. We compared all the permutations and visualized the best teaming combinations using a color map. A darker color indicates a greater boost. **(a)** Oracle mode is the upper bound from perfect teaming, **(b)** realistic mode is from a simplistic real algorithm. The results show the value of human-machine complementary teaming. Introducing a low-performance human to the teaming system causes more boost than introducing a higher-performance machine classifier.

nations of machine-machine teaming lead to only marginal boosts in accuracy. Using additional human classifiers in the *aggre* majority voting configuration leads to additional gains.

Since the upper bound performance is well above the simple realistic algorithm used, it seems likely that better real collaboration algorithms are possible.

Picking the threshold for human-machine collaboration

Given a confidence threshold η, we denote by $\text{Acc}_{\text{col, ml}}(\eta)$ as the test accuracy of replacing unconfident model predictions trained with a specific method by human annotations. We adopt paired t-test to see: for all robust methods, whether replacing unconfident model predictions (w.r.t. a fixed η^*) by human annotations could be consistently the best among such a human-machine collaboration (i.e., $\text{Acc}^*_{\text{col, ml}}$). Mathematically,

$\mathbf{H_0}$: There are no significant differences between $\text{Acc}_{\text{col, ml}}(\eta^*)$ and $\text{Acc}^*_{\text{col, ml}}$, for a list of robust methods;

$\mathbf{H_1}$: There are significant differences between $\text{Acc}_{\text{col, ml}}(\eta^*)$ and $\text{Acc}^*_{\text{col, ml}}$, for a list of robust methods.

The hypothesis testing results are summarized in Table 2. We are interested in two lines of methods: methods require a pre-trained model or multiple networks. For the pre-trained model, we observe that the p-value for $\eta = 0.6$ satisfies that $p \geq \alpha = 0.05$, indicating that we should accept the null hypothesis H_0, and conclude that there exists no significant differences between $\text{Acc}_{\text{col, ml}}(\eta^* = 0.6)$ and $\text{Acc}^*_{\text{col, ml}}$.

5 Limitations

We chose CIFAR-10 for our study because it's widely recognized and gives us a common ground for comparison. However it's possible that our findings here might not be generalizable to other datasets. In addition, our study necessarily looked at only some of the many machine classification methods. We hope that demonstrating that human machine perceptual differences exist in at least some cases is sufficient to motivate future research on additional data and methods.

6 Conclusions

In this work, we show that perceptual differences between human and machine classifiers exist on image classification tasks. Our analysis shows that while humans have lower overall accuracy than machines in our task, their perception is not a mere subset. All machine classifiers in our tests were strongly correlated with each other, while human classifiers produced a different distribution of answers. Given such differences in expertise, we show that human-machine teaming can lead to greater accuracy than machine-machine teaming.

This work has several practical implications. For scientists seeking to use AI as a digital twin for studying human decision-making, this work shows a weakness in current machine classification models. For application engineers with high-value image classification tasks, such as detecting cancer in medical images, this work demonstrates that human-machine collaboration may result in total system accuracy better than what is achievable by machines or humans alone. For computer vision researchers seeking to create fully automated machine classification systems, the findings in this study suggest that tuning algorithms solely for accuracy may not lead to the highest accuracy. If machine classifiers can be created with statistics more like humans, they will initially have lower total accuracy than the current state of the art. However, these new classifiers with differing statistics can then be combined with current methods to produce a fully automated joint system with higher performance than is currently possible.

Acknowledgements

Partial support for M.L. and J.D. was provided by WISEautomotive through the ATC+ Program award from MOTIE Korea 20014264

References

Agrawal, P.; Stansbury, D.; Malik, J.; and Gallant, J. L. 2014. Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.

Battleday, R. M.; Peterson, J. C.; and Griffiths, T. L. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1): 1–14.

Branson, S.; Wah, C.; Schroff, F.; Babenko, B.; Welinder, P.; Perona, P.; and Belongie, S. 2010. Visual recognition with humans in the loop. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 438–451. Springer.

Chen, H.; Shah, A.; Wang, J.; Tao, R.; Wang, Y.; Xie, X.; Sugiyama, M.; Singh, R.; and Raj, B. 2023a. Imprecise label learning: A unified framework for learning with various imprecise label configurations. *arXiv preprint arXiv:2305.12715*.

Chen, H.; Wang, J.; Shah, A.; Tao, R.; Wei, H.; Xie, X.; Sugiyama, M.; and Raj, B. 2023b. Understanding and mitigating the label noise in pre-training on downstream tasks. *arXiv preprint arXiv:2309.17002*.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2020. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *International Conference on Learning Representations*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Han, L.; Dong, X.; and Demartini, G. 2021. Iterative Human-in-the-Loop Discovery of Unknown Unknowns in Image Datasets. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 72–83.

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.

Huang, H.; Kang, H.; Liu, S.; Salvado, O.; Rakotoarivelo, T.; Wang, D.; and Liu, T. 2023. Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16719–16730.

Joshi, A. J.; Porikli, F.; and Papanikolopoulos, N. 2009. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, 2372–2379. IEEE.

Kovashka, A.; Russakovsky, O.; Fei-Fei, L.; Grauman, K.; et al. 2016. Crowdsourcing in computer vision. *Foundations and Trends® in computer graphics and Vision*, 10(3): 177–243.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Kuzovkin, I.; Vicente, R.; Petton, M.; Lachaux, J.-P.; Baciu, M.; Kahane, P.; Rheims, S.; Vidal, J. R.; and Aru, J. 2018. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1): 1–12.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, S.; Zhu, Z.; Qu, Q.; and You, C. 2022. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, 14153–14172. PMLR.

Liu, T.; and Tao, D. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3): 447–461.

Liu, Y.; and Guo, H. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, 6226–6236. PMLR.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Mur, M.; Meys, M.; Bodurka, J.; Goebel, R.; Bandettini, P. A.; and Kriegeskorte, N. 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in psychology*, 4: 128.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Nushi, B.; Kamar, E.; and Horvitz, E. 2018. Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, 126–135.

Nushi, B.; Kamar, E.; Horvitz, E.; and Kossmann, D. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Ortiz-Jimenez, G.; Collier, M.; Nawalgaria, A.; D'Amour, A. N.; Berent, J.; Jenatton, R.; and Kokiopoulou, E. 2023. When does privileged information explain away label noise? In *International Conference on Machine Learning*, 26646–26669. PMLR.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Park, D.; Choi, S.; Kim, D.; Song, H.; and Lee, J.-G. 2024. Robust data pruning under label noise via maximizing relabeling accuracy. *Advances in Neural Information Processing Systems*, 36.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1944–1952.

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

Ramakrishnan, R.; Kamar, E.; Nushi, B.; Dey, D.; Shah, J.; and Horvitz, E. 2019. Overcoming blind spots in the real world: Leveraging complementary abilities for joint execution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6137–6145.

Roels, J.; Vernaillen, F.; Kremer, A.; Gonçalves, A.; Aelterman, J.; Luong, H. Q.; Goossens, B.; Philips, W.; Lippens, S.; and Saeys, Y. 2019. A "human-in-the-loop" approach for semi-automated image restoration in electron microscopy. *bioRxiv*, 644146.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed Truthfulness in Multi-Task Peer Prediction. *ACM EC*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Taleb, A.; Lippert, C.; Klein, T.; and Nabi, M. 2021. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*, 661–673. Springer.

Turk, A. M. 2012. Amazon mechanical turk. *Retrieved August*, 17: 2012.

Wang, H.; Xiao, R.; Dong, Y.; Feng, L.; and Zhao, J. 2022. Promix: Combating label noise via maximizing clean sample utility. *arXiv preprint arXiv:2207.10276*.

Wang, W.; Feng, L.; Jiang, Y.; Niu, G.; Zhang, M.-L.; and Sugiyama, M. 2024. Binary classification with confidence difference. *Advances in Neural Information Processing Systems*, 36.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wei, H.; Tao, L.; Xie, R.; Feng, L.; and An, B. 2022a. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, 23615–23630. PMLR.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; Sugiyama, M.; and Liu, Y. 2022b. To Smooth or Not? When Label Smoothing Meets Noisy Labels. In *International Conference on Machine Learning*, 23589–23614. PMLR.

Wei, J.; and Liu, Y. 2021. When Optimizing f-Divergence is Robust with Label Noise. In *International Conference on Learning Representations*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2022c. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *International Conference on Learning Representations*.

Wei, J.; Zhu, Z.; Luo, T.; Amid, E.; Kumar, A.; and Liu, Y. 2023. To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2523–2535.

Xia, X.; Lu, P.; Gong, C.; Han, B.; Yu, J.; and Liu, T. 2023. Regularly truncated m-estimators for learning with noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Yang, D.; and Alonso, O. 2024. A bespoke question intent taxonomy for e-commerce. In *SIGIR 2024 eCommerce*.

Yang, D.; Chen, K.; Rao, J.; Guo, X.; Zhang, Y.; Yang, J.; and Zhang, Y. 2024. Tackling vision language tasks through learning inner monologues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19350–19358.

Yao, A.; Gall, J.; Leistner, C.; and Van Gool, L. 2012. Interactive object detection. In *2012 IEEE conference on computer vision and pattern recognition*, 3242–3249. IEEE.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 7164–7173. PMLR.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhu, Z.; Liu, T.; and Liu, Y. 2021. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10113–10123.