Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration

Gavin Kerrigan[1] **Padhraic Smyth**[1] **Mark Steyvers**[2]
[1]Department of Computer Science [2]Department of Cognitive Sciences
University of California, Irvine
gavin.k@uci.edu smyth@ics.uci.edu mark.steyvers@uci.edu

Abstract

An increasingly common use case for machine learning models is augmenting the abilities of human decision makers. For classification tasks where neither the human or model are perfectly accurate, a key step in obtaining high performance is combining their individual predictions in a manner that leverages their relative strengths. In this work, we develop a set of algorithms that combine the probabilistic output of a model with the class-level output of a human. We show theoretically that the accuracy of our combination model is driven not only by the individual human and model accuracies, but also by the model's confidence. Empirical results on image classification with CIFAR-10 and a subset of ImageNet demonstrate that such human-model combinations consistently have higher accuracies than the model or human alone, and that the parameters of the combination method can be estimated effectively with as few as ten labeled datapoints.

1 Introduction

One of the main goals of machine learning is to develop algorithms that can operate robustly in an autonomous fashion without human supervision. However, there are many applications where hybrid human-machine approaches are likely to be a preferred mode of operation, for a variety of different reasons, such as improving trust between humans and machines, and allowing for a human or a model to take over in situations where the one or the other lacks expertise [1, 2, 3, 4, 5, 6, 7].

The performance benefits of combining multiple predictors, rather than relying on a single predictor, have been clearly demonstrated in past work in a variety of fields. For example, in machine learning there is a rich vein of research over the past few decades on combining models using a variety of different estimation and algorithmic approaches [8, 9, 10, 11]. This existing line of work emphasizes that combinations of models that have diversity in how they make predictions can systematically outperform a single model. In parallel, in the behavioral science literature, there has been extensive prior work studying combinations of human opinions where, again, diverse combinations tend to outperform any single individual [12, 13].

This naturally leads to questions about hybrid combinations of human and machine predictions, rather than just combining one type or the other. For example, one motivation for hybrid combinations is empirical evidence that human and machine classifiers do not make the same types of errors for problems such as image classification [14, 15, 16], i.e., they are diverse in their predictions. These ideas have begun to have impact in real-world applications, where hybrid human-machine teams have been found to be effective in areas such as crowdsourcing [17], citizen science [18], speech transcription [19], face identification [20], and clinical radiology [21, 22, 23].

In this paper we focus on a specific, simple, and important instantiation of the general problem of hybrid combinations of human and machine predictions. In our problem we consider a K-way classification problem such as image classification, with a single human making hard classification

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

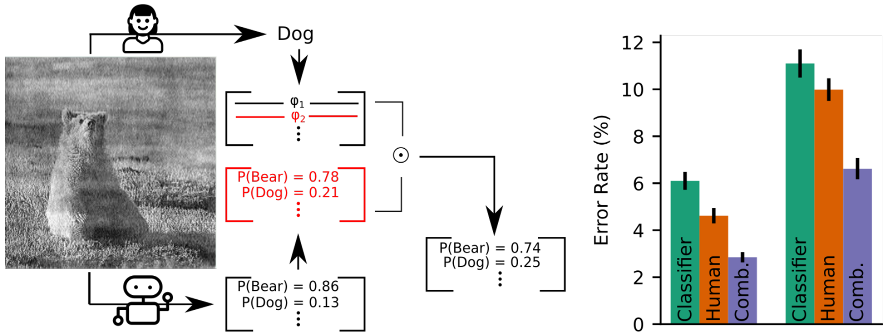

Figure 1: Left: Combining a human's label and a classifier's probabilities for an ImageNet-16H image (true label: bear). Right: Human-machine combinations (purple) achieve lower error rates on average than the human or classifier alone (ResNet-164 on CIFAR-10H, VGG-19 on ImageNet-16H).

decisions (no confidence estimates) and a single classification model providing class-conditional probability vectors. While humans can provide confidence estimates with their label predictions, calibrated self-assessment of confidence can be difficult [24, 25, 26] as well as time-consuming.

A key question in this context is whether the non-probabilistic information from the human predictor can be effectively combined with the probabilistic information from a machine learning model. We answer this question in the affirmative and show both theoretically and empirically how relatively simple probabilistic combination techniques can robustly outperform each of a human and machine on their own. In particular we show that a human can augment their predictions with those of a classification model, improving classification accuracy and producing calibrated predictions, even when the model is less accurate than the human. Similarly, from the model's perspective, accuracy can often be significantly improved by augmenting its' class probabilities with a human's labels, while improving calibration performance, even when the human is less accurate than the model.

Figure 1 shows an example using our proposed methodology for an image from the ImageNet dataset. The human incorrectly predicts the label `dog`. The classification model (a VGG-19 deep network) predicts the correct label `bear` with a probability of 0.86 (uncalibrated) and 0.78 (calibrated). The combined prediction is for `bear` with a confidence of 0.74, with `dog` having a higher confidence given the human's prediction. The histograms on the right show the overall average reduction in error rate on test images (ResNet-164 on CIFAR-10H [27], VGG-19 on ImageNet-16H [28]): even though the classifiers are on average less accurate (6.10%, 11.1% error) than the human (4.62%, 9.99% error), the resulting combinations have lower error rates than either (2.85%, 6.62% error).

The primary contributions of our work are as follows:

- We propose and investigate a general framework for combining predictions using instance-level confidence from a model and class-level information from a human. The methods we propose are straightforward to implement in practice and label-efficient.[1]
- We empirically validate our approach on the CIFAR-10H and ImageNet-16H image classification datasets and show that human-machine combinations in this context are systematically more accurate and better calibrated than either alone.
- We develop a theoretical understanding in this framework of the key tradeoffs related to calibration and accuracy for both the individual human and model, introducing the notion of model and human confidence ratios. We illustrate how these factors affect the combination, showing for example how two models with the same accuracy but with different calibration properties can have different performance when combined with human predictions.

The paper begins in Section 2 by introducing notation and background concepts. Section 3 discusses related work and in Section 4 we propose a number of different estimation methods. Section 5 describes our experimental results with two image classification datasets, with individual human labelers, individual models, and combinations. Complementing the experimental results in Section 6

[1] Code for our estimation methods and experiments is available at `https://github.com/GavinKerrigan/conf_matrix_and_calibration`.

we develop theoretical results that characterize combination performance. Section 7 concludes the paper including a discussion of potential societal impact and limitations.

2 Combining Human Labels and Model Probabilities

Notation. We consider a K-ary classification problem, where the goal is to predict a label $y \in \mathcal{Y} = \{1, \ldots, K\}$ from features $x \in \mathcal{X}$. The random variable (x, y) follows an unknown distribution with support $\mathcal{X} \times \mathcal{Y}$. We assume access to an individual human labeler represented by the function $h : \mathcal{X} \to \mathcal{Y}$, where $h(x) \in \mathcal{Y}$ is the label predicted by the human. In addition, we have access to a trained machine classifier $m : \mathcal{X} \to \mathbb{R}^K$, where $m(x)$ is the normalized probability vector output by the classifier. Both the human and classifier are assumed to be noisy labelers relative to the ground truth y. The true labels could be determined, for example, by expert labelers or additional information not contained in x.

Combining Predictions. Given an input $x \in \mathcal{X}$, our goal is to predict a true label conditioned on the predictions $h(x)$ and $m(x)$. The key challenge in combining human and classifier predictions is simultaneously leveraging both the class-level outputs from the human and the predictive distributions output by the classifier. Although we focus on the particular case where h is a human and m is a classifier, our setup could be applied more generally to combinations of a non-probabilistic labeler (whose output is categorical) and a probabilistic labeler (whose output is a distribution over classes).

There are a variety of functional forms that could be used to combine the predictions. We pursue a probabilistic approach, where the conditional distribution over labels that we seek can be factored via Bayes' rule as

$$p(y|h(x), m(x)) \propto p(h(x)|y, m(x))p(y|m(x)). \tag{1}$$

It is natural in this context to pursue a conditional independence (CI) approach, where the human labels $h(x)$ and the probabilistic predictions $m(x)$ are assumed to be conditionally independent given y. Under this assumption we can write

$$p(y|h(x), m(x)) \propto p(h(x)|y)p(y|m(x)) \tag{2}$$

where the right-hand side terms have a natural interpretation in terms of calibrated probabilities at the class level $\big(p(h(x)|y)\big)$ and at the instance level $\big(p(y|m(x))\big)$.

We parameterize the term $p(h(x)|y)$ by the confusion matrix for the labeler h, which we denote by φ with entries $\varphi_{ij} = p(h(x) = i|y = j)$. On the other hand, the probabilistic output of the classifier $m(x)$ may differ from $p(y|m(x))$. For example, modern neural networks tend to be overconfident in their predictions [29]. To remedy this, post-hoc calibration maps $m(x)$ to well-calibrated probabilities via a learned calibration map with parameters θ. In this work, we use $m^{\theta}(x)$ to denote the output of the classifier after applying such a calibration map. The second term in Equation (2) is then parameterized by the calibrated classifier probabilities $m^{\theta}(x)$. Altogether, our method expresses the predicted probability for class j as:

$$\boxed{p(y = j|h(x) = i, m(x)) = \frac{\varphi_{ij} m^{\theta}_j(x)}{\sum_{k=1}^{K} \varphi_{ik} m^{\theta}_k(x)}} \tag{3}$$

The CI assumption above is common (both implicitly and explicitly) in prior work on combining predictions, such as additive classifier ensembles [10, 11] and (log-)linear opinion pools [30, 31]. As our primary motivation is to develop a relatively simple and robust methodology for combining human and model predictions, the additional functional or parametric assumptions (and parameters) required to specify a joint model for $p(h(x), m(x)|y)$ are beyond our scope. In addition, although the CI assumption is unlikely to hold exactly, prior work [32] notes that a CI model can be an optimal discriminant even when the CI assumption is violated. As further motivation, for the two datasets we use in this paper, CIFAR-10H and ImageNet-16H, the conditional dependence of $h(x)$ and $m(x)$ appears to be relatively weak (see Appendix B for details).

3 Related Work

We summarize below relevant aspects of related literature. While there is a significant amount of prior work in machine learning and related fields on combining predictors, this work has in general not addressed the specific problem of combining hard label predictions from a human with probabilistic label predictions from a model.

Ensembles and Opinion Pools. There is a rich literature in machine learning on studying predictions based on ensembles of classification models. For non-probabilistic classifiers, the most common aggregation methods are variants of (weighted) majority voting [9, 11]. However, in our case of only two predictors, a weighted majority vote ensemble can never improve accuracy over its components. Beyond majority voting, naive Bayes aggregation [33, 10] fits a class-level confusion matrix to each predictor. Kim and Ghahramani [34] develop a fully Bayesian extension of this, which relaxes the independence assumption by explicitly modeling correlations between predictors. However, because these confusion-matrix aggregation methods are at the class level they are unable to take full advantage of the instance-level uncertainties produced by the probabilistic labeler.

In the context of aggregating predictions from multiple humans, there has been a considerable amount of prior work in the behavioral sciences and forecasting literature. Approaches include additive linear and log-linear opinion pools for subjective distributions [30, 31], techniques for weighting linear combinations of real-valued human predictions [13, 35], and voting methods for combining label predictions from more than two human predictors [36]. A key difference between these methods and our work is that they do not address the problem of how to combine probabilistic and non-probabilistic predictions in a human-machine context.

Leveraging Human and Model Predictions. Combining human predictions with model predictions to solve classification problems has been a topic of recent interest in a number of different areas. For example, in [37] simple averaging is used to combine the labels of multiple human annotators with the output of a classifier for astronomical image classification, achieving better performance than with either the humans or the classifier. In crowd-sourcing, classification models have been used to automatically filter examples to improve human annotation efficiency [17, 38, 2, 5]. A similar line of research focuses on algorithmic deferral techniques where a model defers to human predictions based on the model's confidence [39, 40], as well as work on adapting prediction models to the human decision maker [41, 42, 43, 44]. The results in [41] in particular describe experiments with the same CIFAR-10H dataset that we use in this paper. However, in addition to being different to our work in terms of its focus on deferral (rather than combining) we also note that in [41] the improvements in performance are demonstrated using relatively large numbers of human labels. In contrast, as we demonstrate in Section 5 on the CIFAR-10H and ImageNet-16H datasets, our methods require only a small number of human labels to yield combined predictions that are more accurate than either human or model alone. In general, existing work on filtering and deferral strategies complements the combining methods that we develop in this paper. All of these approaches are useful in a broad range of human-AI applications, but in different contexts.

4 Estimation Methods and Algorithms

Combining human and machine predictions via Equation (3) requires learning two sets of parameters: confusion matrix parameters for the human and calibration parameters for the classifier. The choice of procedure used to infer these parameters impacts the label efficiency and quality of the resulting combination. In this section, we detail several inference procedures and empirically evaluate them in the context of human-machine combinations.

To estimate our combination model, we assume access to a combination dataset $\mathcal{D}_C = \{(h(x_\ell), m(x_\ell), y_\ell)\}_{\ell=1}^{n}$ with human labels, machine probabilities, and ground truth labels. We assume the classifier is pre-trained with true labels on a separate training set $\mathcal{D}_T$.

Confusion Matrix Estimation. Recall that φ denotes a a confusion matrix for the human of shape $K \times K$, where φ_{ij} is $p(h(x) = i|y = j)$. The most straightforward estimate for this quantity is the maximum likelihood estimate, where φ_{ij} is estimated by the number of datapoints in $\mathcal{D}_C$ where the human labeler predicts $h(x) = i$ when the ground truth is $y = j$, normalized by the number of points in $\mathcal{D}_C$ where $y = j$.

However, as the size of the confusion matrix is quadratic in K, this estimate will have high variance for small amounts of labeled data and collecting enough human labels to overcome this variance could be prohibitively expensive. We can instead take a Bayesian approach and incorporate informative

Table 1: Summary of combination methods studied in this work. Except for logistic regression, parameter counts correspond to calibration using MAP temperature scaling (one parameter), and confusion matrices are fit with MAP inference. The human output is always a label.

Method Name	Acronym	Parameters	Model Output	Ground Truth?	Label Efficient?
Logistic Regression	LR	$k^2 + 2k$	Probabilities (P)	✓	X
Calibrated Machine Probs & Single-Parameter Confusion	SP	2	Probabilities (P)	✓	✓
Machine Labels & Human Labels	L+L	$2k^2$	Labels (L)	✓	X
Calibrated Machine Probs. & Human Labels	P+L	$k^2 + 1$	Probabilities (P)	✓	✓
Calibrated Machine Probs. & Human Labels (EM)	P+L-EM	$k^2 + 1$	Probabilities (P)	X	✓

prior information. Given the true label $y = j$, the human label $h|y = j \sim \text{Discrete}(\varphi_{*j})$ is assumed to be drawn from a discrete distribution with parameters corresponding to the jth row in the confusion matrix. We place a conjugate Dirichlet prior $\varphi_{*j} \sim \text{Dirichlet}(\alpha_j)$ over each column with parameters $\alpha_j \in \mathbb{R}^k$. The prior parameters α_j are chosen such that

$$(\alpha_j)_i = \begin{cases} \beta & i \neq j \\ \gamma & i = j \end{cases}$$

That is, the prior matrix is $\gamma \in \mathbb{R}_{>0}$ along the diagonal and $\beta \in \mathbb{R}_{>0}$ on the off-diagonal. We choose β and γ such that the resulting Dirichlet distribution has mode equal to the train-set accuracy of the classifier, which can be obtained without additional human labels. This choice of prior reflects our belief that the confusion matrix will have a diagonally dominant structure. Posterior estimates of the confusion matrix can then be obtained straightforwardly by conjugacy.

Calibration Parameter Estimation. Scaling-based calibration maps are typically fit by optimizing the log-likelihood [29, 45, 46]. In this section, we detail a Bayesian version of temperature scaling [29], allowing us to incorporate informative prior information. A temperature $T \in (0, 1)$ indicates underconfidence, and $T \in (1, \infty)$ indicates overconfidence. To account for this difference in scale, we place a Gaussian prior on the log-temperature $\log T = \tau \sim \mathcal{N}(\mu, \sigma^2)$. As this is a non-conjugate prior, the maximum a posteriori (MAP) temperature is estimated via gradient-based optimization. In our experiments, we choose $\sigma = 0.5$ for the CIFAR-10 models and $\sigma = 0.75$ for the ImageNet models, and we use $\mu = 0.5$ throughout. These parameters were chosen to reflect our belief that deep models tend to be overconfident and to concentrate the prior on reasonable temperature values. In Appendix G, we derive a fully Bayesian approach where we marginalize over the posterior distribution over temperature (e.g. using Monte Carlo methods [47]). However, we find empirically that the simpler MAP approach is more effective and, as a result, focus on MAP estimation in this paper.

Learning without Ground Truth. Requiring both human and ground truth labels in $\mathcal{D}_C$ can be a potentially limiting assumption in domains where ground truth labels are unavailable or expensive to obtain. To avoid this, we propose an unsupervised approach that is able to learn both calibration parameters and confusion matrix parameters from a combination dataset of the form $\mathcal{D}_C = \{(h(x_\ell), m(x_\ell))\}$, consisting only of human labels and machine probabilities. We treat the ground truth labels as latent and fit the required parameters using Expectation-Maximization (EM) [48] (details in Appendix I). This approach can be seen as a novel extension of the Dawid-Skene model [49], where calibration parameters are fit for the model rather than a confusion matrix. We perform both maximum likelihood and MAP estimation with this method, using the same priors as above.

For clarity of exposition, we focus on three types of combinations in our results:

- **Machine Labels & Human Labels (L+L)**, a baseline where the instance-level probabilities from the model are discarded and instead a confusion matrix is fit for the model, as well as a confusion matrix for the human. The confusion matrices are estimated via supervised MAP inference. This can be viewed as a naive Bayes' combination [10, Chapter 4] for non-probabilistic predictors.
- **Calibrated Machine Probabilities & Human Labels (P+L)** combined via Equation (3) using *supervised* MAP estimates for both the calibration parameters and confusion matrix parameters.
- **Calibrated Machine Probabilities & Human Labels (P+L-EM)** combined via Equation (3) using *unsupervised* MAP estimates fit with our EM algorithm, i.e. using human labels but no ground truth.

In terms of complexity, these methods sit between a simple model with only one parameter for the human confusion matrix **(SP)** (where the diagonal entry corresponds to the human's marginal accuracy), and a full multinomial logistic regression model (**LR**).[2] We find that **LR** can obtain slightly lower error rates in some cases, but requires significantly more labeled data than the other methods to do so. At the other extreme, **SP** is highly data efficient, but underfits compared to our preferred methods. We provide additional discussion of these methods to Appendix H. The various estimation methods discussed in this section are summarized in Table 1.

5 Experiments

Datasets and Models. We evaluate various combination strategies on two pre-existing image classification datasets that include human annotations: CIFAR-10H [27] and ImageNet-16H [28]. CIFAR-10H contains 10-way human classifications for 10,000 images from the standard CIFAR10 test set [50]. ImageNet-16H contains 16-way human classifications for noisy images from the ImageNet test set [51], distorted by phase noise at each spatial frequency based on four levels of phase noise (80, 95, 110, and 125). The human classifications for CIFAR-10H and ImageNet-16H come from the Amazon Mechanical Turk platform.

For both CIFAR-10H and ImageNet-16H, we select a single human label for each image by randomly sampling from the available human annotations. We experiment with four CNN models on CIFAR-10H (ResNet-110, Resnet-164 [52], PreResNet-164 [53], DenseNet [54]), and eight models (four VGG-19 [55], four GoogLeNet [56]) of varying accuracy on ImageNet-16H. Our models are chosen to span a range of performance, from below human accuracy to exceeding human accuracy. See Appendix E for details regarding our model architectures and training procedures.

Both datasets are partitioned into three disjoint subsets: (i) a model training set $\mathcal{D}_T$, (ii) a combination training set $\mathcal{D}_C$, and (iii) an evaluation set $\mathcal{D}_E$. The model training set is the same as the suggested training split for the original CIFAR-10 and ImageNet datasets, and is used to fit the classification models. The combination training set is used to estimate any calibration parameters and confusion matrices, and the held-out evaluation set is used solely for testing. The combination training set and evaluation set are subsets of the original evaluation sets, where 70% of the data is used for fitting the combinations and 30% is used for evaluation. The true labels (ground truth) for both CIFAR-10 and ImageNet correspond to the originally-provided labels for each of these datasets. In our experiments, (i) is fixed and we average over randomly selected splits for (ii) and (iii). The sampled human label for each image is fixed across all trials.

Calibration Methods. In our experiments, model calibration is done by MAP temperature scaling (TS) [29]. In Appendix C, we experiment with two additional calibration methods: ensemble temperature scaling [45] and I-Max binning [57]. We find that the combination performance is robust to the choice of calibration map, and hence restrict our focus to TS in this section. In addition, we find in general that combinations using uncalibrated model probabilities produce less accurate combinations than combinations using calibrated model probabilities (see Appendix C).

Learning Curves. Given that it is highly desirable to learn human-machine combinations from small amounts of data, we empirically study the data efficiency of the various inference methods previously described. In Figure 2, we plot the combination error rate on the evaluation data as a function of dataset size for the four CIFAR-10 models (first row) and the VGG-19 model on ImageNet (second row). For supervised methods, the dataset size corresponds to the number of $(h(x), m(x), y)$ triples used for learning, whereas for unsupervised methods the dataset size corresponds to the number of $(h(x), m(x))$ pairs without ground truth y.

As Figure 2 demonstrates, the P+L method is able to learn a human-model combination that outperforms both the human and model alone, with very few datapoints. While the P+L-EM method requires more human labels than the P+L method, it is able to learn such a combination without any ground-truth labels. The baseline L+L method fails to learn an effective combination on the CIFAR-10 dataset, and only does so on the VGG-19 ImageNet dataset with a large number of ground

[2] In fact, the CI combination with temperature scaling can be seen as a special case of multinomial logistic regression taking $m(x)$ and $h(x)$ as inputs (see Appendix F).

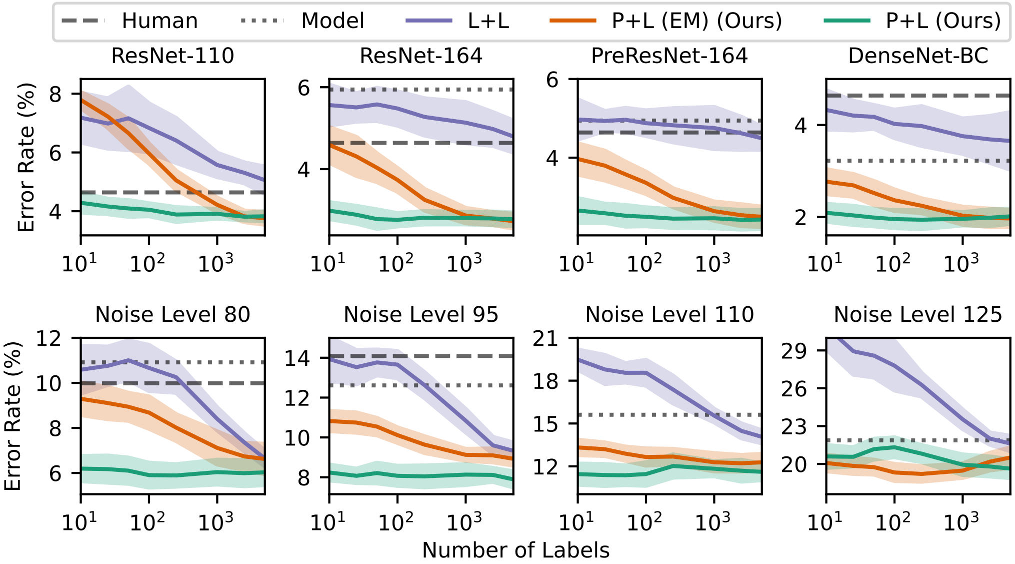

Figure 2: Learning curves for various models on CIFAR-10H (top) and VGG-19 on ImageNet-16H (bottom). For the supervised methods (L+L, P+L), the x-axis corresponds to the number of human labels, machine predictions, and ground truth labels. For the unsupervised method (P+L (EM)), the x-axis corresponds to the number of human labels and machine predictions (but no ground truth).

truth labeled datapoints. This demonstrates that the instance level probabilities from the model are a key component in efficiently learning human-model combinations with high accuracy.

In Appendix H, we provide similar plots for GoogLeNet and for maximum likelihood based inference procedures. In general, we find that maximum likelihood estimation requires more data than MAP estimation, and hence we focus our presentation of results on MAP.

Calibration Properties of Combinations. In addition to the error rate, we study the calibration properties of P+L combinations. In Table 2, we report the ECE [29], classwise-ECE (cwECE) [46, 57], and negative log-likelihood (NLL) for our various CIFAR-10 models (Model) and the resulting human-machine combinations (Comb.) on the held-out evaluation set. The ECE and cwECE are evaluated using 15 bins containing an equal number of data points. In addition to P+L combinations fit with 10 or 5000 labeled datapoints, we evaluate a combination consisting of the uncalibrated classifier probabilities with the MAP human confusion matrix estimated with 5000 labeled datapoints (No Calibration).

Combining classifier probabilities with human labels generally results in a combination that is better calibrated than the model alone. Moreover, MAP TS can be fit using a very small number of labeled datapoints. Our results show that the calibration properties (of both the classifier alone and the resulting human-machine combination) significantly improve with only ten labeled examples. However, increasing the number of labeled examples to 5000 does not result in further calibration gains. We provide similar results for our ImageNet-16H models in Appendix D.

6 Theoretical Analysis

Confidence Ratios. The key quantity in our analysis is the *confidence ratio* of a predictor, which is a random variable representing a predictor's confidence for the correct class relative to the predictor's confidence for other classes. This quantity can be thought of as the predictor's instance-level odds

		No Calibration		10 Datapoints		5000 Datapoints	
Metric	Model Name	Model	Comb.	Model	Comb.	Model	Comb.
ECE (10^{-2})	ResNet-110	5.23 ± 0.35	2.08 ± 0.25	3.03 ± 0.58	1.30 ± 0.23	2.99 ± 0.36	1.76 ± 0.18
	ResNet-164	2.98 ± 0.34	1.63 ± 0.23	1.95 ± 0.33	1.25 ± 0.18	1.89 ± 0.32	1.39 ± 0.18
	PreResNet-164	3.03 ± 0.29	1.87 ± 0.22	2.31 ± 0.33	1.40 ± 0.26	2.27 ± 0.31	1.43 ± 0.21
	DenseNet-BC	2.18 ± 0.27	1.53 ± 0.20	1.76 ± 0.28	1.34 ± 0.14	1.73 ± 0.28	1.27 ± 0.13
cwECE (10^{-2})	ResNet-110	0.81 ± 0.07	0.23 ± 0.05	0.58 ± 0.07	0.24 ± 0.05	0.58 ± 0.06	0.19 ± 0.06
	ResNet-164	0.39 ± 0.06	0.15 ± 0.03	0.31 ± 0.05	0.15 ± 0.04	0.31 ± 0.05	0.13 ± 0.03
	PreResNet-164	0.29 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03
	DenseNet-BC	0.23 ± 0.03	0.11 ± 0.02	0.24 ± 0.02	0.12 ± 0.02	0.24 ± 0.02	0.11 ± 0.02
NLL	ResNet-110	0.40 ± 0.02	0.16 ± 0.01	0.35 ± 0.02	0.15 ± 0.01	0.35 ± 0.02	0.14 ± 0.01
	ResNet-164	0.24 ± 0.02	0.11 ± 0.01	0.20 ± 0.01	0.10 ± 0.01	0.20 ± 0.01	0.10 ± 0.01
	PreResNet-164	0.23 ± 0.02	0.13 ± 0.02	0.19 ± 0.02	0.11 ± 0.01	0.19 ± 0.02	0.10 ± 0.01
	DenseNet-BC	0.17 ± 0.01	0.10 ± 0.01	0.14 ± 0.01	0.09 ± 0.01	0.14 ± 0.01	0.08 ± 0.01

Table 2: Calibration metrics for various CIFAR-10H models with P+L combinations. Even a small amount of labeled data (10 labels) reduces the calibration error of both the classifier and combination. For all metrics, lower is better.

for making a correct prediction. More specifically, the confidence ratios r_m and r_h for machine and human labelers are defined as

$$r_m(x) = \frac{m_y^\theta(x)}{1 - m_y^\theta(x)} \qquad r_h(x) = \frac{\varphi_{h(x)y}}{1 - \varphi_{h(x)y}} \tag{4}$$

We note that unlike the machine classifier, the human does not directly output such confidences – rather, this quantity is estimated empirically through the human's confusion matrix.

If the model has a confidence ratio of $r_m(x) > 1$ (indicating that the model has confidence greater than 0.5 for the correct class), then the model is guaranteed to correctly label x. On the other hand, $r_m(x) > 1$ is not sufficient for the combination to correctly label x – instead, the model must be sufficiently confident in its prediction as well. The following theorem formalizes this notion by a lower bound on the accuracy of the combination in terms of the confidence ratios of the individual predictors. For binary classification tasks, this lower bound is achieved, i.e. we have equality.

Theorem 1 (Combination Model Accuracy.)**.** *The accuracy of the P+L combination $c(x)$ is at least the probability that the confidence ratio for m exceeds the inverse confidence ratio for h.*

$$\mathbb{E}\left[\mathbb{1}\left(c(x) = y\right)\right] \geq p(r_m(x) \geq (r_h(x))^{-1}) \tag{5}$$

A detailed proof is provided in Appendix A. An analogous result holds for the combination of two probabilistic classifiers or two non-probabilistic classifiers. As our focus is on combining human predictions with model probabilities, we discuss these cases in Appendix A.

Theorem (1) is further illustrated in Figure 3 for a ResNet-164 classifier on CIFAR-10H (first row) and a VGG-19 classifier on ImageNet-16H (second row). For each row, we create a family of classification models where each model makes the same class-level predictions (and hence has the same error rate) but with different confidence ratio distributions. This is achieved by tempering the output probabilities of a base classifier via the map $m(X) \mapsto (m_1(X)^{1/T}, \ldots, m_k(X)^{1/T}) / \sum_{i=1}^{K} m_i(X)^{1/T}$ with temperature $T > 0$ [29, 45]. In the first column of each row, the solid curve plots the error rate of the combination of a fixed human with the various classification models. Despite each classification model having the same accuracy, the accuracy of the resulting human-model combination varies.

This behavior can be explained by Theorem (1), which tells us that the combination accuracy is driven not only by the human and machine accuracies but by their confidence ratios as well. At large temperatures (purple), the classifier becomes underconfident in its predictions, and the combination error rate approaches the human error rate. At small temperatures (green), the model becomes overconfident in its predictions, and the combination error rate approaches the model error rate. When the combination is fit by our P+L method (orange), the classifier is well-calibrated (reflected by its low ECE), and the resulting combination obtains a lower error rate than each of the under- and over-confident classifier combinations.

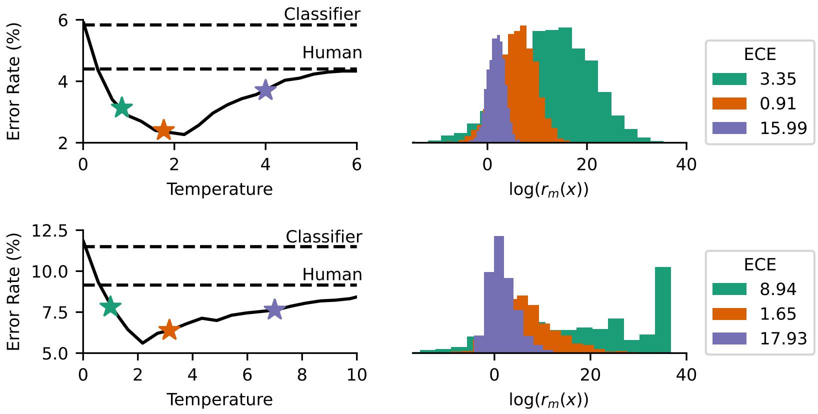

Figure 3: Left: Single human labeler combined with various classifiers of equal accuracy, resulting in combinations with varied accuracies. The orange point corresponds to the P+L combination. Right: First row: ResNet-164 on CIFAR-10H. Second row: VGG-19 on ImageNet-16H at noise level 80.

Relationship between combination and calibration/confusion error. We additionally quantify the estimation error in our P+L method, incurred by empirically estimating the human confusion matrix and calibration parameters. The result below shows we can upper bound our estimation error by the estimation error for the confusion matrix and the ℓ_1 marginal calibration error (MCE) [58].

Theorem 2 (Estimation Error Upper Bound.)**.** *Let* $\eta(x,y) = |p(h(x)|y)p(y|m(x)) - m^\theta_y(x)\widehat{\varphi}_{h(x)y}|$ *be the estimation error (up to normalizing constants), where* $\widehat{\varphi}_{ij}$ *represents an estimate of* $p(h(x) = i|y = j)$*. Under the CI assumption, in expectation over random* (x, y) *pairs,*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\eta(X,Y)\right] \leq ||\varphi - \widehat{\varphi}||_1 + \mathit{MCE}(m^\theta) \tag{6}$$

(Proof in Appendix A). With a sufficient amount of labeled data, the confusion matrix error $||\varphi - \widehat{\varphi}||_1$ can be made arbitrarily small (i.e. the MAP confusion matrix estimate is unbiased). This is not necessarily the case for MCE(m^θ). However, if an asymptotically unbiased calibration method is used, this result guarantees that our posterior estimation error will converge to zero.

7 Limitations, Societal Impact, and Conclusions

Limitations. One limitation of our work is that our experiments only involve two datasets and both involve image classification. Thus, there is no guarantee that similar results (in terms of combined improvements) are achievable for other tasks, such as question-answering from text data or more general problem-solving tasks. Another potential limitation is our reliance on conditional independence in our approach. A reverse view of this is that both our theoretical and experimental results demonstrate that there is ample room for improving human and machine performance by combining their predictions, even without taking dependence into account.

Potential Societal Impacts. Combining human and machine predictions to improve overall classification accuracy has the potential for positive societal benefit, particularly for example in high-stakes applications such as medical image diagnosis and autonomous driving. However, there are also potential negative societal impacts. For example, if there is a lack of transparency in terms of how the system operates (e.g., how predictions are being combined to arrive at a final result), augmenting an individual's predictions with a machine's could have negative psychological consequences for the individual, such as decreasing trust, reducing individual autonomy, and eventual disengagement.

Conclusions. We investigated methods for combining predictions using instance-level confidence from a model and class-level information from a human. Across a variety of image classification experiments our proposed combination framework leads to systematic increases in accuracy over both the model and human alone while requiring few human labels. Supporting theory illustrates how combined human-model performance is affected by calibration properties of the model.

References

[1] Ece Kamar. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *Proceedings of IJCAI*, pages 4070–4073, 2016.

[2] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071, 2017.

[3] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1): 237–293, 2018.

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

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

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

[7] Zahra Zahedi and Subbarao Kambhampati. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*, 2021.

[8] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[9] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[10] Ludmila I Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2014.

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

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

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

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

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

[16] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019.

[17] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, volume 12, pages 467–474, 2012.

[18] Melanie R Beck, Claudia Scarlata, Lucy F Fortson, Chris J Lintott, BD Simmons, Melanie A Galloway, Kyle W Willett, Hugh Dickinson, Karen L Masters, Philip J Marshall, et al. Integrating human and machine intelligence in galaxy morphology classification tasks. *Monthly Notices of the Royal Astronomical Society*, 476(4):5516–5534, 2018.

[19] Yashesh Gaur, Florian Metze, Yajie Miao, and Jeffrey P Bigham. Using keyword spotting to help humans correct captioning faster. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[21] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Medicine*, 15(11):e1002699, 2018.

[22] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, 2 (1):1–10, 2019.

[23] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ Digital Medicine*, 3(1):1–8, 2020.

[24] Gideon Keren. Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273, 1991.

[25] Daniel Kahneman and Amos Tversky. On the reality of cognitive illusions. *Psychological Review*, pages 582–591, 1996.

[26] Joshua Klayman, Jack B Soll, Claudia Gonzalez-Vallejo, and Sema Barlas. Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247, 1999.

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

[28] Anonymous. Imagenet-16H: a large-scale behavioral data set of noisy image classifications. To be publicly released on Open Science Foundation, 2021.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[30] Christian Genest, James V Zidek, et al. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[31] Robert A Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, 1995.

[32] Ludmila I Kuncheva. On the optimality of naive Bayes with dependent binary features. *Pattern Recognition Letters*, 27(7):830–837, 2006.

[33] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435, 1992.

[34] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR, 2012.

[35] Clintin P Davis-Stober, David V Budescu, Stephen B Broomell, and Jason Dana. The composition of optimally wise crowds. *Decision Analysis*, 12(3):130–143, 2015.

[36] Michael D Lee and Megan N Lee. The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, 12(4):328, 2017.

[37] Darryl E Wright, Chris J Lintott, Stephen J Smartt, Ken W Smith, Lucy Fortson, Laura Trouille, Campbell R Allen, Melanie Beck, Mark C Bouslog, Amy Boyer, et al. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, 472(2):1315–1323, 2017.

[38] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2121–2131, 2015.

[39] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017.

[40] Maithra Raghu, Katy Blumer, Greg Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *CoRR*, abs/1903.12220, 2019. URL `http://arxiv.org/abs/1903.12220`.

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

[42] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

[43] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Dan Weld. Is the most accurate AI the best teammate? optimizing AI for teamwork. In *AAAI 2021*, February 2021.

[44] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. *arXiv preprint arXiv:2103.08902*, 2021.

[45] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pages 11117–11128. PMLR, 2020.

[46] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

[47] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[48] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[49] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[50] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[51] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[53] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[54] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[57] Kanil Patel, William Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. *arXiv preprint arXiv:2006.13092*, 2020.

[58] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. *CoRR*, abs/1909.10155, 2019. URL `http://arxiv.org/abs/1909.10155`.

Appendix A Proofs of Theorem (1) and Theorem (2)

We provide proofs for our theoretical claims in Section 6.

A.1 Confidence Ratios

Proof of Theorem (1). Recall that $c(x)$ is the prediction output by Equation (3). The accuracy is then bounded as follows:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(c(x)=y\right)\right] = \mathbb{P}\left\{y = \arg\max_k \varphi_{h(x)k} m_k^\theta(x)\right\} \\
&= \mathbb{P}\left\{\varphi_{h(x)y} m_y^\theta(x) > \max_{k\neq y} \varphi_{h(x)k} m_k^\theta(x)\right\} \\
&\geq \mathbb{P}\left\{\varphi_{h(x)y} m_y^\theta(x) > \max_{k\neq y} \varphi_{h(x)k} \max_{k\neq y} m_k^\theta(x)\right\} \\
&\geq \mathbb{P}\left\{\varphi_{h(x)y} m_y^\theta(x) > \left(1-\varphi_{h(x)y}\right)\left(1-m_y^\theta(x)\right)\right\} \\
&= \mathbb{P}\left\{r_m(x) > \left(r_h(x)\right)^{-1}\right\}
\end{aligned}$$

In fact, we have proved the stronger but somewhat less interpretable inequality:

$$\mathbb{E}\left[\mathbb{1}\left(c(x)=y\right)\right] \geq \mathbb{P}\left\{\frac{m_y^\theta(x)}{\max_{k\neq y} m_k^\theta(x)} > \left(\frac{\varphi_{h(x)y}}{\max_{k\neq y} \varphi_{h(x)k}}\right)^{-1}\right\}$$

□

We note further that the same argument can be used to analyze the combination of two probabilistic predictors when the combination is done by pointwise multiplying their calibrated probabilities. In particular, if we have two probabilistic classifiers m and $\widetilde{m}$,

$$\mathbb{E}\left[\mathbb{1}\left(c(x)=y\right)\right] \geq \mathbb{P}\left\{r_m(x) > \left(r_{\widetilde{m}}(x)\right)^{-1}\right\}$$

where $r_{\widetilde{m}}$ is defined analogously to r_m. The proof for this statement is exactly analogous to that of Theorem (1), where $\widetilde{m}_y^\theta$ now plays the role of $\varphi_{h(x)y}$. This same argument can again be adapted for the combination of two non-probabilistic combiners, combined by parameterizing Equation (2) with their confusion matrices.

A.2 Estimation Error

We begin with a useful lemma that will play a key part in our estimation error analysis.

Lemma 1. *For scalars $a_1, a_2, b_1, b_2 \in [0,1]$, the difference of the products is at most the sum of the differences:*

$$|a_1 b_1 - a_2 b_2| \leq |a_1 - a_2| + |b_1 - b_2| \tag{7}$$

Proof.

$$\begin{aligned}
|a_1 b_1 - a_2 b_2| &= |a_1 b_1 - a_2 b_2 + a_1 b_2 - a_1 b_2| \\
&= |a_1(b_1 - b_2) + b_2(a_1 - a_2)| \\
&\leq |a_1| \cdot |b_1 - b_2| + |b_2| \cdot |a_1 - a_2| \qquad \text{(triangle inequality)} \\
&\leq |b_1 - b_2| + |a_1 - a_2|
\end{aligned}$$

□

We now proceed to the proof of Theorem (2).

Proof of Theorem (2). Recall that $\eta(x, y) = |p(h(x)|y)p(y|m(x)) - m_y^\theta(x)\widehat{\varphi}_{h(x)y}|$ is the estimation error for Equation (3) (up to normalizing constants), where $\widehat{\varphi}_{ij}$ represents an estimate of $p(h(x) = i|y = j)$.

By the law of total expectation, we can condition on a particular value of y and $h(x)$:

$$\mathbb{E}\left[\eta(x, y)\right] = \sum_{i=1}^{K}\sum_{j=1}^{K} p(y = j)\varphi_{ij}\mathbb{E}\left[\eta(x, y)|y = j, h(x) = i\right] \tag{8}$$

We now apply Lemma (1) to the conditional expectation above:

$$\begin{aligned}
&\mathbb{E}\left[\eta(x, y)|y = j, h(x) = i\right] \\
&= \mathbb{E}\left[|\varphi_{ij}p(y = j|m(x)) - \widehat{\varphi}_{ij}m_j^\theta(X)|\middle| y = j, h(x) = i\right] \\
&\leq \mathbb{E}\left[|\varphi_{ij} - \widehat{\varphi}_{ij}| + |p(y = j|m(x)) - m_j^\theta(x)|\middle| y = j, h(x) = i\right] \\
&= |\varphi_{ij} - \widehat{\varphi}_{ij}| + \mathbb{E}\left[|p(y = j|m(x)) - m_j^\theta(x)|\middle| y = j\right]
\end{aligned}$$

We additionally employ the conditional independence assumption to arrive at the last line.

Plugging this back in to Equation (8), we obtain

$$\begin{aligned}
\mathbb{E}\left[\eta(x, y)\right] \leq &\sum_{i=1}^{K}\sum_{j=1}^{K} P(y = j)\varphi_{ij}|\varphi_{ij} - \widehat{\varphi}_{ij}| \\
&+ \sum_{j=1}^{K} p(y = j)\mathbb{E}\left[|p(y = j|m(x)) - m_j^\theta(x)|\middle| y = j\right]
\end{aligned}$$

Since $\varphi_{ij}, p(y = 1) \leq 1$, the first summand is at most $\sum_{i=1}^{K}\sum_{j=1}^{K}|\varphi_{ij} - \widehat{\varphi}_{ij}| = ||\varphi - \widehat{\varphi}||_1$. In fact, the first summand is typically much smaller than $||\varphi - \widehat{\varphi}||_1$ – for example, if all classes are equally likely, the first summand is at most $\frac{1}{K}||\varphi - \widehat{\varphi}||_1$.

The second summand is readily recognized as the ℓ_1 marginal calibration error [58]. □

Table 3: Conditional and unconditional mutual information for various datasets and models.

Dataset	Model	Noise	CMI($M; H|Y$)	MI($M; H$)
CIFAR-10H	Densenet		0.030	2.829
CIFAR-10H	PreResnet-164		0.043	2.770
CIFAR-10H	Resnet-110		0.037	2.404
CIFAR-10H	Resnet-164		0.038	2.707
ImageNet-16H	VGG-19	80	0.119	2.954
ImageNet-16H	VGG-19	95	0.174	2.816
ImageNet-16H	VGG-19	110	0.230	2.277
ImageNet-16H	VGG-19	125	0.314	1.527
ImageNet-16H	GoogLeNet	80	0.121	2.825
ImageNet-16H	GoogLeNet	90	0.161	2.643
ImageNet-16H	GoogLeNet	110	0.260	2.182
ImageNet-16H	GoogLeNet	125	0.364	1.421

Appendix B Assessing Conditional Independence/Dependence in CIFAR-10H and Imagenet-16H Datasets

We investigate the degree to which our conditional independence assumption is satisfied empirically in the datasets used in the paper. Specifically, of interest is the assumption of conditional independence of $m(x)$ and $h(x)$, given y. Assessing conditional independence is not straightforward given that $m(x)$ is a K-dimensional real-valued vector and $h(x)$ and y each take one of K categorical values, with $K = 10$ for CIFAR-10H and $K = 16$ for ImageNet-16H. While there exist statistical tests for assessing conditional independence for categorical random variables, with real-valued variables the situation is less straightforward and there are multiple options such as different non-parametric tests involving different tradeoffs [59, 60, 61, 62].

Given these issues we investigate the degree of conditional dependence using two relatively simple approaches. The first approach looks at the conditional mutual information (CMI) between the predicted label from the model and the predicted label from the human, conditioned on the true label. While this is indirect, in that it does not use the real-valued scores, it does allow us to measure CMI in a straightforward manner given that all the variables involved are categorical. The CMI is defined as

$$\text{CMI}(M;H|Y) \;=\; \sum_{y} p(y) \sum_{m,h} p(m,h|y) \log \frac{p(m,h|y)}{p(m|y)p(h|y)}$$

where M, H, Y are the K-ary random variables for the model, human, and true labels respectively (taking values m, h, y). The inner sum over m, h is the mutual information between M and H conditioned on a particular value of $Y = y$. All probabilities were estimated using relative frequencies (maximum likelihood) from the evaluation sets for each dataset.

Table 3 shows the results for the 4 different models for CIFAR-10H and the 2 × 4 different combinations of models and noise for ImageNet-16H. To put the CMI numbers on an interpretable scale, we also compute the (unconditional) mutual information between M and H in each case. If M and H are truly independent conditioned on Y, then the true CMI values should be 0.

The broad conclusion from Table 3 is that for the CIFAR-10H there appears to be little to no conditional dependence (of model labels and human labels, given true labels) given that the CMI values are very close to 0. For the ImageNet-16H data the CMI values are higher, suggesting evidence for weak conditional dependence in this dataset, particularly at high noise levels where neither the human or the model are very accurate.

Figures 4, 5, 6 show the results of another assessment, now using model probabilities, for the 4 models for the CIFAR-10H data, for VGG-19 on ImageNet-16H, and for GoogLeNet on ImageNet-16H, respectively. The x-axis in each plot is the mean probability from the model for the true label y, conditioned on $Y = y$. The y-axis shows the mean probability (in red) from the model for the true label y, conditioned now on **both** $Y = y$ **and** $H = y$, i.e., conditioned on the event that the human also predicts the true label.

If the model's probabilities for the true labels are independent of $H = y$, then the x and y values should be the same (i.e., on the diagonal). The degree to which these points (in red) are not on the diagonal is an indication of some conditional dependence of the model's probabilities on the human labels h. The red points are generally close to the diagonal, or slightly above (indicating, not surprisingly, that if the human predicts the true label, the model's probability for the true label tends to increase slightly (if at all) rather than decrease.) To put these values on an appropriate scale we also compute (empirically from the data) the maximum possible increase that could occur, when additionally conditioning on the human label h being correct (the black points). The conclusions are similar to what we found with conditional mutual information, namely, that there is little indication of conditional dependence in the CIFAR-10H data, and some indication of dependence in the ImageNet-16H data, particularly for higher noise levels.

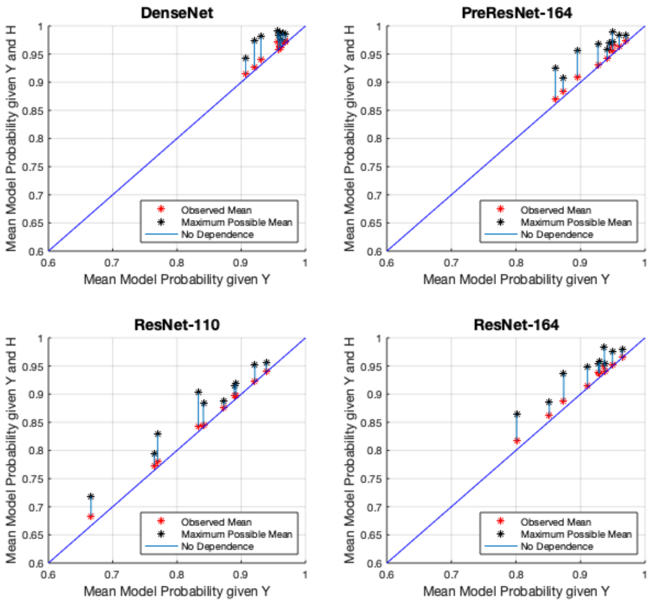

Figure 4: Change in expected values of model probabilities on CIFAR-10H data for the true class y, conditioning on just y (x-axis), versus conditioning on both y and $h(x) = y$ (y-axis, in red).

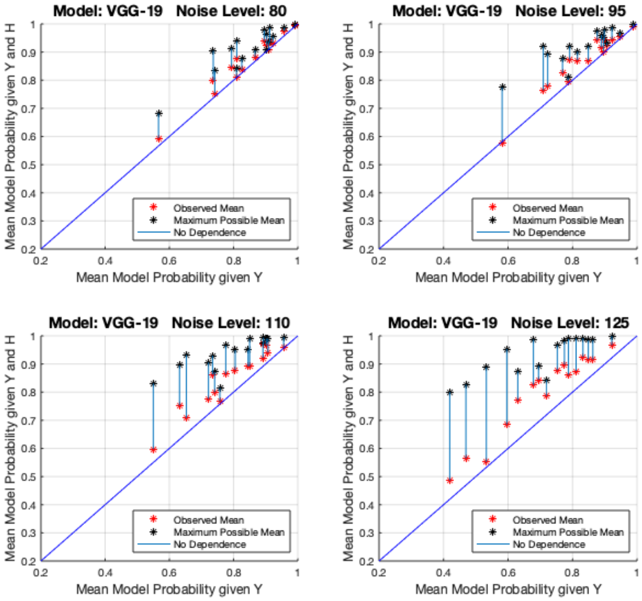

Figure 5: Same as Figure 4 but for VGG-19 models on ImageNet-16H data.

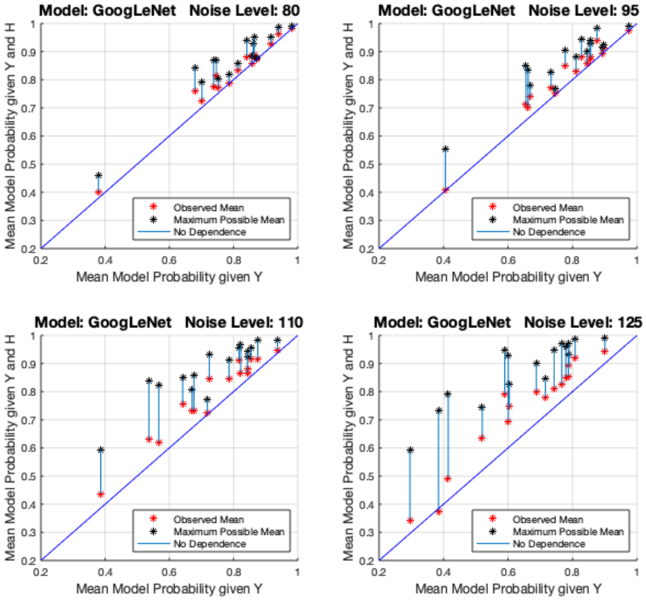

Figure 6: Same as Figure 4 but for GoogLeNet models on ImageNet-16H data.

Appendix C Calibration Methods and Uncalibrated Combinations

In this section we provide additional empirical results on CIFAR-10H and ImageNet-16H. In particular, we evaluate several different calibration methods (MAP TS (as used in the main paper) [29], Ensemble TS [45], IMax Binning [57]). We also compare to the L+L combination, and the P+L combination of the uncalibrated model probabilities with the human labels (Uncalibrated). The error rate of the human alone (Human) and model alone (Model) are provided for context.

In most cases, human-machine combinations using calibrated probabilities outperform those using uncalibrated probabilities. Moreover, in some cases we obtain small gains in performance by using a more complex calibration map (IMax Binning), but it is not clear how to incorporate prior information with this method. As prior information is useful in increasing the label efficiency and decreasing the error rate of the combination, our focus in the main paper is on MAP TS as our calibration method.

All tables in this section correspond to error rates (± one standard deviation) averaged across 25 different random seeds. The combinations (P+L, Equation (3)) are fit using 5000 labeled data points on CIFAR-10H and using between 5067 and 5152 data points on ImageNet-16H (varies by noise level). The combinations are evaluated using 3000 data points on CIFAR-10H and using between 2171 and 2208 on ImageNet-16H.

			Combination				
Model Name	Human	Model	L+L	Uncalibrated	TS	ETS	IMax
ResNet-110	4.62 ± 0.33	11.28 ± 0.44	4.70 ± 0.36	4.40 ± 0.25	3.83 ± 0.15	3.76 ± 0.25	$\mathbf{3.80 \pm 0.24}$
ResNet-164	—	6.10 ± 0.38	4.71 ± 0.37	3.05 ± 0.23	$\mathbf{2.78 \pm 0.15}$	2.82 ± 0.23	2.85 ± 0.23
PreResNet-164	—	5.00 ± 0.36	4.36 ± 0.39	2.90 ± 0.22	$\mathbf{2.43 \pm 0.22}$	2.46 ± 0.25	$\mathbf{2.43 \pm 0.26}$
DenseNet-BC	—	3.25 ± 0.30	3.39 ± 0.32	2.22 ± 0.21	$\mathbf{2.01 \pm 0.15}$	2.17 ± 0.17	2.04 ± 0.18

Table 4: Error rates (%, ± one standard deviation) averaged over 25 seeds on CIFAR-10H for various classifiers.

		Combination				
Human	Model	L+L	Uncalibrated	TS	ETS	IMax
9.99 ± 0.48	11.10 ± 0.60	6.78 ± 0.42	7.52 ± 0.52	$\mathbf{6.03 \pm 0.54}$	6.79 ± 0.44	6.31 ± 0.46
14.07 ± 0.70	12.58 ± 0.53	9.01 ± 0.57	9.02 ± 0.44	$\mathbf{7.89 \pm 0.37}$	9.32 ± 0.49	8.62 ± 0.47
22.99 ± 0.71	15.51 ± 0.62	14.07 ± 0.82	12.59 ± 0.53	$\mathbf{11.62 \pm 0.54}$	13.18 ± 0.59	12.30 ± 0.57
39.76 ± 0.75	22.07 ± 0.69	21.89 ± 0.66	$\mathbf{19.45 \pm 0.62}$	19.63 ± 0.70	20.74 ± 0.62	20.47 ± 0.63

Table 5: Error rates (%, ± one standard deviation) averaged over 25 seeds, VGG-19 on ImageNet-16H. Each row corresponds to a different noise level (80, 95, 110, 125).

		Combination				
Human	Model	L+L	Uncalibrated	TS	ETS	IMax
9.99 ± 0.48	14.48 ± 0.70	7.33 ± 0.39	8.06 ± 0.50	$\mathbf{6.80 \pm 0.47}$	7.73 ± 0.41	7.66 ± 0.44
14.07 ± 0.70	17.22 ± 0.72	9.93 ± 0.73	10.66 ± 0.52	$\mathbf{9.67 \pm 0.40}$	10.23 ± 0.51	10.05 ± 0.49
22.99 ± 0.71	19.09 ± 0.75	15.43 ± 0.71	14.78 ± 0.64	$\mathbf{14.09 \pm 0.55}$	14.76 ± 0.53	14.53 ± 0.53
39.76 ± 0.75	27.06 ± 0.47	25.64 ± 0.43	23.06 ± 0.72	$\mathbf{22.60 \pm 0.63}$	24.38 ± 0.64	23.91 ± 0.69

Table 6: Error rates (%, ± one standard deviation) averaged over 25 seeds, GoogLeNet on ImageNet-16H. Each row corresponds to a different noise level (80, 95, 110, 125).

Appendix D Calibration Properties of Combinations

We further study the calibration properties of human-machine (P+L) combinations. The results in this Appendix are analogous to the results in Table 2 for our ImageNet-16H models, where we show various calibration metrics as we vary the number of labeled datapoints used for fitting the combination. In general, we find that using only a small number of labeled datapoints (10 in our experiments) is sufficient, and we do not observe further improvements in calibration by using more labeled data (5000 points in our experiments) to fit the combination.

In addition, we investigate whether the resulting human-machine combination can be further calibrated. We calibrate the resulting human-machine combinations (with MAP TS) using the same data used to fit the combination, i.e. 5000 labeled datapoints (Recal. Comb.). We find that it is possible to further reduce the ECE of the combinations, but other metrics only see small improvements. However, we note that this does not affect the error rate of the combination, as MAP TS is accuracy-preserving.

		No Calibration		10 Datapoints		5000 Datapoints		
Metric	Model Name	Model	Comb.	Model	Comb.	Model	Comb.	Recal. Comb.
ECE (10^{-2})	ResNet-110	5.23 ± 0.35	2.08 ± 0.25	3.03 ± 0.58	1.30 ± 0.23	2.99 ± 0.36	1.76 ± 0.18	0.85 ± 0.22
	ResNet-164	2.98 ± 0.34	1.63 ± 0.23	1.95 ± 0.33	1.25 ± 0.18	1.89 ± 0.32	1.39 ± 0.18	0.84 ± 0.20
	PreResNet-164	3.03 ± 0.29	1.87 ± 0.22	2.31 ± 0.33	1.40 ± 0.26	2.27 ± 0.31	1.43 ± 0.21	1.06 ± 0.21
	DenseNet-BC	2.18 ± 0.27	1.53 ± 0.20	1.76 ± 0.28	1.34 ± 0.14	1.73 ± 0.28	1.27 ± 0.13	0.95 ± 0.18
cwECE (10^{-2})	ResNet-110	0.81 ± 0.07	0.23 ± 0.05	0.58 ± 0.07	0.24 ± 0.05	0.58 ± 0.06	0.19 ± 0.06	0.19 ± 0.04
	ResNet-164	0.39 ± 0.06	0.15 ± 0.03	0.31 ± 0.05	0.15 ± 0.04	0.31 ± 0.05	0.13 ± 0.03	0.14 ± 0.03
	PreResNet-164	0.29 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03	0.28 ± 0.04	0.13 ± 0.03	0.13 ± 0.03
	DenseNet-BC	0.23 ± 0.03	0.11 ± 0.02	0.24 ± 0.02	0.12 ± 0.02	0.24 ± 0.02	0.11 ± 0.02	0.10 ± 0.02
NLL	ResNet-110	0.40 ± 0.02	0.16 ± 0.01	0.35 ± 0.02	0.15 ± 0.01	0.35 ± 0.02	0.14 ± 0.01	0.12 ± 0.01
	ResNet-164	0.24 ± 0.02	0.11 ± 0.01	0.20 ± 0.01	0.10 ± 0.01	0.20 ± 0.01	0.10 ± 0.01	0.09 ± 0.01
	PreResNet-164	0.23 ± 0.02	0.13 ± 0.02	0.19 ± 0.02	0.11 ± 0.01	0.19 ± 0.02	0.10 ± 0.01	0.08 ± 0.01
	DenseNet-BC	0.17 ± 0.01	0.10 ± 0.01	0.14 ± 0.01	0.09 ± 0.01	0.14 ± 0.01	0.08 ± 0.01	0.07 ± 0.01

Table 7: Calibration metrics on CIFAR-10H.

		No Calibration		10 Datapoints		5000 Datapoints		
Metric	Noise Level	Model	Comb.	Model	Comb.	Model	Comb.	Recal. Comb
ECE (10^{-2})	80	8.54 ± 0.54	5.17 ± 0.49	7.30 ± 0.69	3.91 ± 0.53	7.15 ± 0.60	4.01 ± 0.42	3.17 ± 0.42
	95	8.96 ± 0.48	5.72 ± 0.39	7.49 ± 0.77	4.93 ± 0.36	7.26 ± 0.51	4.56 ± 0.37	3.23 ± 0.35
	110	9.76 ± 0.53	7.81 ± 0.48	7.81 ± 0.91	6.31 ± 0.67	7.24 ± 0.56	6.07 ± 0.53	3.92 ± 0.49
	125	11.81 ± 0.64	10.89 ± 0.52	7.34 ± 1.45	10.21 ± 0.64	7.29 ± 0.56	8.46 ± 0.68	4.49 ± 0.60
cwECE (10^{-2})	80	1.10 ± 0.07	0.68 ± 0.05	1.01 ± 0.07	0.59 ± 0.06	1.01 ± 0.06	0.54 ± 0.05	0.56 ± 0.04
	95	1.18 ± 0.06	0.82 ± 0.05	1.13 ± 0.06	0.73 ± 0.06	1.12 ± 0.06	0.72 ± 0.04	0.69 ± 0.04
	110	1.44 ± 0.06	1.14 ± 0.06	1.38 ± 0.07	1.04 ± 0.08	1.36 ± 0.07	1.03 ± 0.07	0.96 ± 0.05
	125	1.98 ± 0.06	1.73 ± 0.07	1.86 ± 0.05	1.54 ± 0.07	1.85 ± 0.04	1.52 ± 0.06	1.45 ± 0.06
NLL	80	0.71 ± 0.05	0.49 ± 0.04	0.53 ± 0.05	0.37 ± 0.04	0.52 ± 0.03	0.34 ± 0.03	0.27 ± 0.02
	95	0.70 ± 0.03	0.52 ± 0.03	0.55 ± 0.04	0.41 ± 0.03	0.54 ± 0.03	0.39 ± 0.02	0.32 ± 0.02
	110	0.73 ± 0.03	0.61 ± 0.04	0.60 ± 0.04	0.51 ± 0.05	0.57 ± 0.03	0.49 ± 0.04	0.41 ± 0.02
	125	0.89 ± 0.03	0.83 ± 0.04	0.75 ± 0.03	0.77 ± 0.03	0.74 ± 0.02	0.71 ± 0.03	0.64 ± 0.02

Table 8: Calibration metrics for VGG-19 on ImageNet-16H.

		No Calibration		10 Datapoints		5000 Datapoints		
Metric	Noise Level	Model	Comb.	Model	Comb.	Model	Comb.	Recal. Comb.
ECE (10^{-2})	80	7.39 ± 0.58	4.10 ± 0.47	4.60 ± 0.62	3.04 ± 0.32	4.52 ± 0.58	3.07 ± 0.40	1.97 ± 0.39
	95	9.23 ± 0.58	5.68 ± 0.40	5.91 ± 0.55	4.19 ± 0.54	5.76 ± 0.59	4.32 ± 0.42	2.51 ± 0.45
	110	9.04 ± 0.68	7.60 ± 0.46	5.34 ± 1.15	6.66 ± 0.71	5.34 ± 0.37	5.98 ± 0.50	3.00 ± 0.43
	125	11.98 ± 0.40	10.95 ± 0.62	6.78 ± 1.40	11.33 ± 0.36	6.54 ± 0.43	7.98 ± 0.60	3.37 ± 0.42
cwECE (10^{-2})	80	1.33 ± 0.07	0.67 ± 0.04	1.30 ± 0.07	0.63 ± 0.04	1.30 ± 0.07	0.57 ± 0.03	0.57 ± 0.03
	95	1.47 ± 0.08	0.87 ± 0.05	1.46 ± 0.05	0.78 ± 0.06	1.47 ± 0.05	0.75 ± 0.04	0.74 ± 0.05
	110	1.70 ± 0.08	1.23 ± 0.06	1.66 ± 0.03	1.12 ± 0.07	1.65 ± 0.04	1.10 ± 0.06	1.03 ± 0.06
	125	2.31 ± 0.06	1.92 ± 0.06	2.19 ± 0.06	1.70 ± 0.06	2.19 ± 0.06	1.68 ± 0.05	1.59 ± 0.06
NLL	80	0.59 ± 0.03	0.34 ± 0.02	0.50 ± 0.02	0.29 ± 0.02	0.50 ± 0.03	0.28 ± 0.02	0.25 ± 0.02
	95	0.65 ± 0.03	0.43 ± 0.03	0.56 ± 0.01	0.37 ± 0.02	0.56 ± 0.01	0.36 ± 0.02	0.33 ± 0.02
	110	0.75 ± 0.03	0.60 ± 0.03	0.66 ± 0.03	0.56 ± 0.03	0.66 ± 0.02	0.53 ± 0.02	0.47 ± 0.02
	125	0.97 ± 0.02	0.88 ± 0.03	0.85 ± 0.02	0.89 ± 0.02	0.85 ± 0.01	0.79 ± 0.02	0.74 ± 0.02

Table 9: Calibration metrics for GoogLeNet on ImageNet-16H.

Appendix E Dataset, Model Training, and Code Details

E.1 CIFAR-10H

The CIFAR-10H dataset [27] consists of the $10,000$ images in the standard CIFAR-10 test set, but each image is labeled by approximately 50 individual human labelers. There are ten classes in this dataset.

We study four CNN model architectures on CIFAR-10H:

- ResNet-110 and ResNet-164 [52]: Deep residual networks with 110 and 164 layers respectively.
- PreResNet-164 [53]: A deep residual network with identity mappings as skip connections, with 164 layers.
- DenseNet-BC [54]: A densely connected CNN with $L = 190$ layers and a growth-rate of $k = 40$, using bottleneck layers.

For each model, we use pre-trained weights available at `https://github.com/bearpaw/pytorch-classification` (MIT License). These models were trained on the standard CIFAR-10 training split.

E.2 ImageNet-16H

The ImageNet-16H dataset [28] consists of noisy images from the ImageNet test set [51], distorted by phase noise at each spatial frequency based on four levels of phase noise (80, 95, 110, and 125). Approximately 7200 images were classified at each noise level (with slight variability per noise level). The number of classes is reduced to 16 (as compared to 1000 in the original ImageNet dataset).

We study two model architectures on ImageNet-16H: VGG-19 [55] and GoogLeNet [56]. Our training procedure is detailed as follows. We first load a pre-trained ImageNet model (trained on the original 1000 class ImageNet dataset) from the PyTorch model library [63]. We remove the final linear layer and replace it with a randomly initialized linear layer with a 16-dimensional output. We then fine-tune all model weights (using the cross-entropy loss) on noisy images from the ImageNet-16H training set (261,168 images). The models are fine-tuned to all levels of noise simultaneously by randomly assigning a different degree of phase noise (ranging from 0 to 130 degrees) to each training image in a batch.

E.3 Additional Code Details

Our experiments are implemented in Python 3.8, and make use of the following libraries:

- Scikit-Learn [64] (BSD License)
- PyTorch [63] (BSD License)
- Pyro [65] (Apache 2.0 License)
- NumPy [66] (BSD License)
- IMax Calibration [57] (AGPL-3.0 License)
- Ensemble Temperature Scaling [45] (MIT License)

E.4 Compute Resources

All of our experiments were conducted on a standard desktop computer (AMD Ryzen 5 6-core @ 3.6GHz, 16GB memory).

Other than the fully Bayesian combination, all combination methods studied in this work do not require significant computational resources and can be fit on the order of seconds. The fully Bayesian method (Appendix G) is more computationally intensive as it requires the use of MCMC to sample from the posterior distribution over calibration parameters, but can still be fit in approx. 2 minutes with 5000 labeled datapoints. However, we focus on MAP estimation in our main results (which does not require MCMC), and only compare to the fully Bayesian setup as a baseline comparison. In addition, we find the fully Bayesian setup to be less label efficient than the MAP counterpart (see Appendix H).

In terms of model training, our ImageNet-16H models were trained on an internal GPU server with 8x GTX 2080ti GPUs and 2 x Intel Xeon Gold 5218 (16 core) processors. On our hardware, fine-tuning for 50 epochs requires approximately 6 hours of training per model.

Appendix F Conditional Independence Combination as a Special Case of Logistic Regression

We demonstrate that the conditional independence combination (Equation (3)) can be seen as a special case of logistic regression taking $m(x)$ and $h(x)$ as inputs, but only when the calibration map takes a particular functional form. Calibration maps such as temperature scaling and Dirichlet calibration [46] satisfy this requirement.

F.1 Logistic Regression

In the logistic regression (LR) model, for input x we have features $z \in \mathbb{R}^{2k}$, $z(x) = m(x) \oplus H(x)$, where $H(x)$ is the one-hot version of $h(x)$ and $\oplus$ is the direct sum. A weight matrix $W \in \mathbb{R}^{k \times 2k}$ and a bias $b \in \mathbb{R}^k$ are to be learned. The probabilistic output is given by an element-wise softmax:

$$x \mapsto \text{SoftMax}(Wz(x) + b) \in \mathbb{R}^k \tag{9}$$

We can write $W = [W_m | W_h]$ as a block matrix, where $W_m, W_h \in \mathbb{R}^{k \times k}$ are the model and human weights respectively. In log-space, the LR model is then

$$\log p(y|m(x), h(x)) = W_m m(x) + W_h H(x) + b - \log(C) \tag{10}$$

where C is a normalizing constant. Since $H(x)$ is one-hot, the term $W_h H(x)$ corresponds to a column in W_h, e.g. if $H(x) = [1, 0, \ldots, 0]^\mathsf{T}$, then $W_h H(x)$ is the first column of W_h. The above is the full vector of probabilities. To make it clearer, for an index i, let W_m^i be the ith row of W_m (resp. for W_h).

$$\begin{aligned} p(y = i|m(x), h(x)) &= W_m^i m(x) + W_h^i H(x) + b_i - \log(C) &(11)\\ &= W_m^i m(x) + (W_h)_{ih(x)} + b_i - \log(C) &(12) \end{aligned}$$

F.2 CI Model

In the CI model,

$$p(y|m(x), h(x)) \propto p(y|m(x))p(h(x)|y) \tag{13}$$

In log-space for a single index i:

$$\begin{aligned} \log p(y = i|m(x), h(x)) &= \log p(y|m(x)) + \log p(h(x)|y) - \log(C) &(14)\\ &= \log m_i^\theta(x) + \log \varphi_{h(x)y} - \log(C) &(15) \end{aligned}$$

F.3

From this, we see that W_h is analogous to the log-confusion matrix of h. Similarly, W_m can be thought of as a linear operator mapping the model probabilities to log-calibrated model probabilities.

If we use $\log m(x)$ (pointwise) for the input feature $z(x)$, the LR model is

$$p(y = i|m(x), h(x)) = W_m^i \log m(x) + (W_h)_{ih(x)} + b_i - \log(C) \tag{16}$$

In the special case $W_m = \frac{1}{T} I$, $b_i = 0$, and $W_h = \log \varphi^\mathsf{T}$, we recover temperature scaling CI. In fact, the equation $W_m \log m(x) + b$ is the same as Dirichlet calibration – vector scaling / matrix scaling are special cases as well.

Appendix G Derivation of Fully Bayesian Model for TS

In this section, we derive a fully Bayesian method for combining classifier probabilities with human labels. In summary, we place a Gaussian prior on the log-temperature (for calibration) and independent Dirichlet priors over the columns of the human confusion matrix. The posterior human confusion matrix is available in closed-form (due to conjugacy), and we sample from the posterior distribution over calibration parameters using MCMC. To predict on a new datapoint, we marginalize over the calibration and confusion parameters using the sampled temperatures and closed-form posterior confusion parameters. This marginalization is only approximate due to the required sampling step.

In more detail, let $\varphi_{*i} \sim$ Dirichlet(α_i) for $i = 1, 2, \ldots, k$ be priors over the columns of the confusion matrix, and let $\log T = \tau \sim \mathcal{N}(\mu_0, \sigma_0^2)$ be a prior over the log-temperature. We use φ to denote the confusion matrix with columns $\varphi_{*1}, \ldots, \varphi_{*K}$. We assume a fully labeled dataset is available, and of the form $\mathcal{D} = \{(h_\ell, m_\ell, y_\ell)\}$. Take the calibration and confusion parameters to be conditionally independent given the data:

$$p(\tau, \varphi|\mathcal{D}) = p(\tau|\varphi, \mathcal{D})p(\varphi|\mathcal{D}) = p(\tau|\mathcal{D})p(\varphi|\mathcal{D}) \tag{17}$$

The confusion parameters have a conjugate prior, but the calibration parameters do not – hence, suppose that we have sampled $\{\tau_1, \ldots, \tau_{n_s}\}$ from the posterior $p(\tau|\mathcal{D})$. To do inference on a new datapoint (h, m), we marginalize over φ and τ for a particular choice of y:

$$p(y|h, m, \mathcal{D}) = \iint p(y, \tau, \varphi|h, m, \mathcal{D})d\varphi d\tau \tag{18}$$

$$= \iint p(y|\tau, \varphi, h, m, \mathcal{D})p(\tau|\mathcal{D})p(\varphi|\mathcal{D})d\varphi d\tau \tag{19}$$

The second line is obtained by conditioning on τ, φ and using the fact that τ and φ are independent given $\mathcal{D}$. We now use Equation (2) to re-write the first term, obtaining (up to a constant):

$$\propto \iint p(h|y, \varphi)p(y|m, \tau)p(\tau|\mathcal{D})p(\varphi|\mathcal{D})d\varphi d\tau \tag{20}$$

We now split the integral into its independent components, and use our parametric assumptions to replace $p(h|y, \varphi)$ with φ_{hy} and $p(y|m, \tau)$ with $m_y^{(\tau)}$:

$$= \left[\int m_y^{(\tau)} p(\tau|\mathcal{D})d\tau\right] \left[\int \varphi_{hy} p(\varphi|\mathcal{D})d\varphi\right] \tag{21}$$

The second integral is the posterior mean of φ_{hy}, which is available in closed-form by conjugacy. However, as we do not have a closed-form posterior for $p(\tau|\mathcal{D})$, we estimate the first integral using our samples. In all, we obtain

$$\approx \left[\frac{1}{n_s}\sum_{j=1}^{n_s} m_y^{(\tau_j)}\right] \cdot \frac{\alpha'_{hy}}{\sum_{\ell=1}^{K} \alpha'_{\ell j}} \tag{22}$$

where α'_{ij} is the posterior Dirichlet parameter for entry (i, j) in the confusion matrix φ. Note that the resulting probabilities will be un-normalized, but normalization is straightforward as we are considering a set of discrete outcomes.

In practice, we use HMC [68] in the Pyro probabilistic programming language [65] to sample from the posterior over log-temperatures.

Appendix H   Learning Curves

In addition to those in Figure 2, we provide learning curves that include additional baseline models: logistic regression (LR), the single-parameter confusion matrix method (SP), and the fully Bayesian P+L method (P+L Fully Bayesian). We report only the mean error rate averaged over 10 random seeds for the sake of visual clarity. All methods (other than LR) are fit using MAP inference. We do not present the maximum likelihood (ML) variants for these methods, as the MAP methods outperform their ML counterparts in our experiments.

While the SP method is label efficient given its low parameter count, it often underfits to the data and converges to an error rate worse than the P+L method. In some cases, the fully Bayesian obtains a lower error rate than the P+L method, but requires more labeled data to be fit, as well as being more computationally intensive. On the CIFAR-10 data, the logistic regression method is label inefficient, and while it outperforms the L+L method, converges to a worse error rate than the P+L method. In contrast, LR is able to outperform the P+L method on the ImageNet-16H datasets, but only when fit several hundred datapoints.

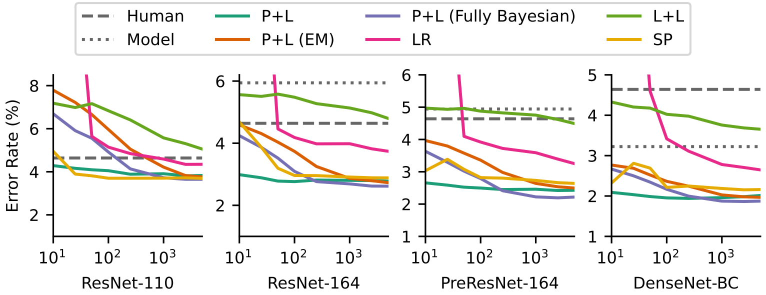

Figure 7: Learning curves for various models on CIFAR-10H.

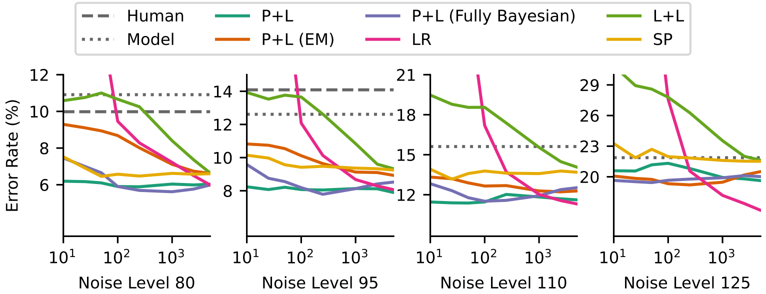

Figure 8: Learning curves for VGG-19 on ImageNet-16H at various noise levels.

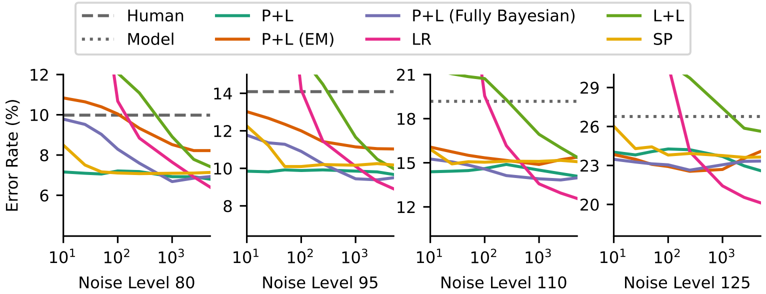

Figure 9: Learning curves for GoogLeNet on ImageNet-16H at various noise levels.

Appendix I EM Algorithm Details

In this section, we provide a detailed derivation and description of our EM algorithm.

Let $\mathcal{D}_C = \{(m(x_\ell), h(x_\ell))\}_{\ell=1}^n$ be an unlabeled dataset used for fitting combination parameters, consisting of classifier probabilities and human labels but no ground truth labels. Our goal is to infer classifier calibration parameters θ and the human confusion matrix φ from $\mathcal{D}_C$. We use m_ℓ as a shorthand for $m(x_\ell)$ throughout (respectively for h).

We can fit this model via EM, where the ground truth is treated as latent. For simplicity, we derive the maximum likelihood variant, and discuss the necessary changes for the MAP variant at the end of this section. In the E-step, $p(y|m_\ell, h_\ell, \varphi, \theta)$ is estimated from Equation (3).

For the M-step, we maximize the expected log-likelihood, where we use $\Theta = \{\theta, \varphi\}$ to denote the set of all parameters:

$$\begin{aligned}
\Theta_{t+1} &= \arg\max_\Theta \sum_i \mathbb{E}_{y\sim p(y|h,m,\theta_t)} \left[\log p(y, h_\ell, m_\ell|\Theta)\right] \\
&= \arg\max_\Theta \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(y, h_\ell, m_\ell|\Theta) \\
&= \arg\max_\Theta \Bigg[\sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(h_\ell|y, \Theta) \\
&+ \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(y|m_\ell, \Theta) + C \Bigg]
\end{aligned}$$

where C is a constant not depending on Θ. Assuming further that the calibration and confusion parameters are independent, the M-step becomes two independent optimizations (i.e. one for θ and one for φ):

$$\theta_{t+1} = \arg\max_\theta \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(y|m_\ell, \theta) \tag{23}$$

$$\varphi_{t+1} = \arg\max_\varphi \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(h_\ell|y, \varphi) \tag{24}$$

In Equation (23), $\log p(y|m_\ell, \theta)$ depends on the calibration method we choose, and the update for θ_{t+1} does not have a closed-form update. We use gradient methods to maximize this term.

Equation (24) is maximum likelihood for the confusion matrix and hence φ_{t+1} can be solved for in closed-form. In particular, the value for φ_{t+1} at entry i, j is

$$\varphi_{i,j} = \frac{\sum_{\ell:h_\ell=a} p(y=j|h_\ell, m_\ell, \Theta_t)}{\sum_\ell p(y=j|h_\ell, m_\ell, \Theta_t)} \tag{25}$$

For the MAP variant of our EM algorithm, our optimizations become

$$\theta_{t+1} = \arg\max_\theta \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(y|m_\ell, \theta) + \log p(\theta) \tag{26}$$

$$\varphi_{t+1} = \arg\max_\varphi \sum_\ell \sum_y p(y|h_\ell, m_\ell, \Theta_t) \log p(h_\ell|y, \varphi) + \log p(\varphi) \tag{27}$$

The first optimization (Equation (26)) is still fit using gradient methods. As we choose independent Dirichlet priors for each column of φ, the closed-form estimate for φ becomes

$$\varphi_{i,j} = \frac{\alpha_{ji} - 1 + \sum_{\ell:h_\ell=a} p(y=j|h_\ell, m_\ell, \Theta_t)}{\gamma + (K-1)\beta - K + \sum_\ell p(y=j|h_\ell, m_\ell, \Theta_t)} \tag{28}$$

which is analogous to the typical Dirichlet-multinomial posterior.

Appendix References

[59] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

[60] Alexander Marx and Jilles Vreeken. Testing conditional independence on discrete data using stochastic complexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 496–505. PMLR, 2019.

[61] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. CCMI: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pages 1083–1093. PMLR, 2020.

[62] Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

[63] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[65] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[66] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[67] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

[68] Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.