Take Expert Advice Judiciously: Combining Groupwise Calibrated Model Probabilities with Expert Predictions

Sumeet Gupta[a], **Shweta Jain**[a], **Shashi Shekhar Jha**[a], **Pao-Ann Hsiung**[b] **and Ming-Hung Wang**[b]

[a]Indian Institute of Technology Ropar
[b]National Chung Cheng University, Taiwan
ORCiD ID: Shweta Jain https://orcid.org/0000-0002-2666-9058, Shashi Shekhar Jha https://orcid.org/0000-0002-1375-2266, Pao-Ann Hsiung https://orcid.org/0000-0002-3639-1467

Abstract. Training the machine learning (ML) models require a large amount of data, still the capacity of these models is limited. To enhance model performance, recent literature focuses on combining ML models' predictions with that of human experts, a setting popularly known as the human-in-the-loop or human-AI teams. Human experts can complement the ML models as they are well-equipped with vast real-world experience and sometimes have access to private information that may not be accessible while training the ML model. Existing approaches for combining an expert and ML model either require end-to-end training of the combined model or require expert annotations for every task. End-to-end training further needs a custom loss function and human annotations, which is cumbersome, results in slower convergence, and may adversely impact the ML model's accuracy. On the other hand, using expert annotations for every task is also cost-ineffective. We propose a novel technique that optimizes the cost of seeking the expert's advice while utilizing the ML model's predictions to improve accuracy. Our model considers two intrinsic parameters: the expert's cost for each prediction and the misclassification cost of the combined human-AI model. Further, we present the impact of group-wise calibration on the combined model that improves the overall model's performance. Experimental results on our combined model with group-wise calibration show a significant increase in accuracy with limited expert advice against different established ML models for the image classification task. In addition, the combined model's accuracy is always greater than that of the ML model, irrespective of the expert's accuracy, the expert's cost, and the misclassification cost.

Keywords: K-way Classification, Human-AI Team, Model Probabilities, Human-in-the-loop model, Deferred Model, Calibration

1 Introduction

We live in a world flooded with data. Analyzing patterns in data would be an exhaustive or impossible task for humans; that is why we need machine learning. Machine learning (ML) aims to develop algorithms that learn from data and operate robustly without human intervention. Modern ML algorithms are fast, scalable, and can attain high accuracy on real-world datasets. Data collection and annotation are at the core of training ML models; however, these are tedious and always incomplete. Due to such issues, how well an ML model will perform in the real world is always concerning. In order to improve the ML model's accuracy, there exist ensemble approaches [5, 11, 12, 18] that discuss different methods and algorithms for combining classifiers. However, all ML models are data-dependent and may tend to make the same mistakes. Moreover, blindly trusting an ML model is still not practical for critical business scenarios such as cyber-attacks or medical diagnoses, even if the ML model has very high accuracy.

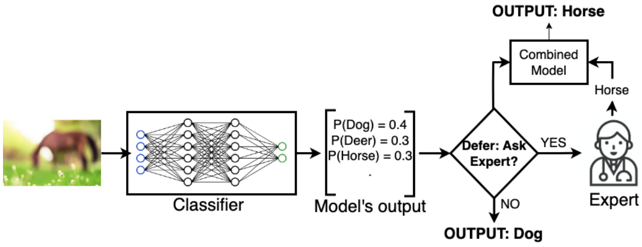

Figure 1. Overall architecture of the proposed human-in-the-loop combined model

On the other hand, human experts are inherently competent for tasks such as classification and only need a small dataset for training. It can also be assumed that humans can access sensitive or private information, which can not be translated or shared while training ML models due to privacy issues. For example, in the case of medical image analysis, patients' complete medical records or attributes can not be shared publicly. Thus, experts can make informed decisions on unseen data based on the preliminary or private information that they may have. However, human experts are scarce, may only be available sometimes, and are insufficient to classify a large number of tasks. In addition, even experts may not be able to classify correctly due to the complexity of the tasks that ML models can learn.

Hence in the real world, neither ML models nor experts are perfect. Some recently published literature [9, 15, 16] suggests that efficiently combining the human-predicted labels with an ML model's probabilistic output improves the accuracy of the combined model. The primary motivation behind combining the outputs of human and ML models is their respective strengths, as they do not make the same mistakes. The authors in [9] showed that the combined model increases the accuracy compared to the human or the model alone. However, their combined model considers human labels for all tasks,

* Corresponding Author. Email: shashi@iitrpr.ac.in

leading to higher costs since the experts are scarce.

To balance the cost and speed of decision-making on all tasks, recent approaches [15, 16, 20] learn a deferred model that treads the cost and performance trade-off. The deferred model provides a decision module that decides whether the classification task should be deferred to a human expert. One of the significant disadvantages of these approaches is that the deferred model is learned from scratch leading to high computation costs. Moreover, even for tasks requiring human expertise, it may not be a good idea to always defer to humans and completely ignore the ML model's output. This paper aims to combine two essential concepts in human-in-the-loop based decision modeling: deferring to humans [16] and combining model probabilities with human-predicted label [9].

During prediction, an ML model predicts each label with a confidence score. The confidence of ML models reflects the ground truth corresponding to each class. However, literature [6] shows that modern ML models are overconfident in their predictions compared to traditional ML models. This confidence becomes crucial when we defer to the human experts based on the model's confidence value. For example, in self-driving cars, if the ML model is not confident in their prediction, it should relinquish the controls to humans instead of making false predictions with high confidence. To address this overconfidence issue, we employ a calibration technique where the model's confidence is calibrated using post-hoc calibration methods. We use temperature scaling (a single parameter variant of Patt Scaling [6,9]) calibration technique and show that the calibration significantly improves the system's overall accuracy. While calibration on the complete dataset helps, our experimental results suggest that the ML model has different accuracy on different classes. To further improve the accuracy, we apply calibration on individual groups in the dataset with a clustering-based approach to learn group-wise calibration parameters. Our results show that this group-wise calibration enhances the accuracy of the overall approach.

In summary, this paper focuses on the problem of k-way classification using a trained ML model's probabilistic output. It judiciously takes the human expert advice that provides hard classification labels. Our proposed model consists of two major modules: the classifier and the defer. Based on the calibrated probabilistic output of the classifier module, the defer module decides whether to consider the classifier's output as the final output or defer to a human. If the model defers to a human, then the calibrated probabilistic classifier output is combined with the human's hard classification labels via Bayes' rule [9]. The overall architecture of our proposed approach is shown in Figure 1. We show that a well-calibrated model can choose better instances to defer, leading to higher accuracy. We further show that even with a naïve deferred model (contrary to training an end-to-end deferred model), calibration and combining the model's experience with deferred instances to humans significantly improve accuracy. In particular, we put forward the following research questions in this work:

- Does a simple deferred model exist that does not require end-to-end training but improves the model accuracy?
- What is the impact of a calibrated classifier on the proposed deferred model?
- Does learning a different calibration parameter for each group within the dataset help as opposed to applying the calibration on the entire dataset?
- If we defer to an expert, can we combine model and human output such that the combined model's accuracy is better than the individual accuracy?

2 Related Work

We discuss the relevant related work in two parts. First is the work on learning to defer [16, 20], which essentially learns a deferred model which learns whether to defer to a human expert for classification or not. Second, the work describes combining model probabilities with the expert's output [9].

2.1 *Deferred Approaches in Human-Machine Collaboration*

As discussed earlier, humans and ML models make mistakes. In order to improve the accuracy of the overall system, many researchers suggest deferring to human experts when the ML model has a high misclassification cost on a certain task. One naïve way to achieve this is to defer to a human expert when the model is more likely to make a mistake [7]. While this naïve approach seems to work well, it makes an essential assumption that experts are highly accurate. This assumption may not be true, and [13] shows that the overall model may lead to low accuracy when humans are not highly accurate. To avoid such issues, many approaches suggest training a deferred model [3, 4, 8, 15–17, 20], which trains a rejector module and a classifier module. The trained rejector module decides whether to use an ML model for the given instance or to defer to a human expert based on a custom loss function. The main drawback of these approaches is that they assume the availability of human annotations in training data. If the expert accuracy is higher than the model, the combined model always results in deferring to the human and could be highly cost-ineffective.

In another complementary approach, in an AI-assisted setting, human makes decisions and can defer to the AI's recommendation [1, 2, 19]. An AI-assisted framework is extended to multiple humans in [14], which defines personalized loss functions for each user to improve the team's overall compatibility. The AI model's performance is optimized in all these approaches to achieve the optimal performance-compatibility trade-off.

2.2 *Combinations of ML Models*

A considerable amount of research talks about combining predictions from multiple classifiers. For the machine learning model, which gives a non-probabilistic or hard classification, the most common method of combining outputs is weighted majority voting or its variant [5,18]. Another approach is where each predictor's confusion matrix is combined through the Naïve Bayes method [10, 12, 21]. However, all the machine models are data-dependent and may make the same mistakes. These methods do not consider uncertainty based on the instance; instead, they focus on individual classes. If a classifier assigns the same class label to two different instances then treating them equally might not be correct because they might have different confidence levels.

Kerrigan et al. [9] suggest combining human output with probabilistic machine output for the k-way classification problem. They also consider a calibrated ML model over the whole dataset. They show that the combined model leads to much higher accuracy than baselines. The combined model has a major drawback: it requires human and classifier outputs for every input, which can be costly if expert labels are unavailable. Another issue with this approach is that if there is a significant gap between the classifier and human accuracies, one can dominate the other on all the classification tasks. Hence deferring to less accurate humans or classifiers becomes redundant.

This paper uses an approach where the groupwise calibrated ML model output is combined with humans only when the defer module defers to a human.

3 APPROACH

Dataset: We have utilized the CIFAR-10H dataset to assess various combination strategies along with our proposed model. The CIFAR-10H dataset comprises ten categories of images and associated probabilistic classifier output for 10,000 images. Furthermore, the dataset includes several human annotations for each test image. We opted for this dataset because human annotations are available, enabling us to demonstrate how humans operate in real-world scenarios and compare our findings with those of previous studies such as [9]. In addition, we also compare the results on CIFAR-10 dataset with simulated human annotations to compare our findings with those of previous studies such as [16, 20].

Table 1. Accuracy of ML models and expert on CIFAR-10H dataset

Dataset	Model	Model Accuracy(%)	Expert Accuracy(%)
CIFAR-10H	ResNet-110	88.9	95
	ResNet-164	93.5	
	PreResNet-164	94.8	
	DenseNet-BC	96.3	

Baseline ML Models: As our baseline ML models, we have selected four different CNN models: ResNet-100, Resnet-164, PreResNet-164, and DenseNet. Each model exhibits varying CIFAR-10H image classification task accuracies, as shown in Table 1. These baseline models will help us understand our approach's behavior in different scenarios, such as when the model accuracy is higher or lower, or close to expert accuracy. More details are provided in Section 3.1.

Human Expert: The CIFAR-10H dataset includes several human annotations for each test image. To emulate an expert and enable a fair comparison with [9], we utilized the same code used in the [9] to produce expert annotations. The final accuracy achieved by the expert was 95%.

In the upcoming sections, namely Sections 3.1 to 3.6, we will build intuitions behind our combined model interleaved with empirical results gathered on the CIFAR-10H dataset. Sections 3.1 and 3.2 will delve into our model's classifier and defer modules, respectively. Section 3.3 will introduce calibration, which is crucial given that modern neural networks are typically overconfident in their predictions. Further, to make decision based on model confidence, calibration becomes inevitable. In Section 3.4, we will elaborate on why calibrating the whole dataset is not an optimal strategy and introduce the concept of group-wise calibration. Until Section 3.4, we inherently assume that the expert performs better on all instances that are deferred by the defer module for expert's advice, but that might not always be true. Section 3.5 will address such situations with our proposed human-in-the-loop combined approach. Finally, in Section 3.6, we will go the extra mile to improve the accuracy of our combined model with a group-wise confusion matrix estimation of the human expert.

3.1 Classifier module

A $k-$way classification problem consists of an input space denoted by $\mathcal{X}$ and the output space denoted by $\mathcal{Y} = \{1, 2, \ldots, K\}$. Let $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{R}^K$ denote the classifier that provides the normalized probabilistic output on each of the possible classes. That is, for a given instance $x \in \mathcal{X}$, $\mathcal{M}(x) = [c_1, c_2, ..., c_K]$ where c_i represents classifier's probabilistic output corresponding to i^{th} class. Since classifier output is always normalized, we also have $\sum_{i=1}^{K} c_i = 1$.

Algorithm 1 Algorithm of defer module

```
Input: Expert cost $C_h$, misclassification cost $C$, classifier $\mathcal{M}$, Instance $x$
Output: Predicted label $\hat{y}$ and total cost
Initialize $Cost \leftarrow 0$
Let $\mathcal{M}(x) = [c_1, c_2, \ldots, c_K]$ and $c_{max} \leftarrow \max_i c_i$
if $C_h < \{C * (1 - c_{max})\}$ then
    $\hat{y} = AskHuman()$
    $Cost += C_h$
    if Human is wrong then
        $Cost += C$
    end
else
    $\hat{y} = arg \max_i c_i$
    if Machine is wrong then
        $Cost += C$
    end
end
```

We evaluated multiple classifiers, including ResNet-110, ResNet-164, PreResNet-164, and DenseNet-BC, on the CIFAR-10H dataset. Each classifier generates a probabilistic output for a given input image. Table 1 displays the accuracy of each model on the test data and the human accuracy on the CIFAR-10H dataset. Notably, all ML models are pre-trained, and human annotations were not used during the classifier training process, implying that the annotations did not affect the training of the classifiers.

Note that the information in Table 1 highlights an essential point: humans and ML models are not perfect at making predictions. Furthermore, the varying accuracies of the different classifiers allow us to make insightful observations for our proposed approach. In the following section, we will introduce a defer module based on cost, which aims to defer to an expert when the classifier's confidence on a particular instance is low.

3.2 Defer Module

This section describes our deferred module, which given an instance, decides whether to defer to the expert or not. This module takes three inputs: the expert cost denoted by C_h, the misclassification cost denoted by C, and the classifier probabilistic output denoted by $\mathcal{M}$. Since we are experimenting in a single human setting, for simplicity, we assume that the expert cost C_h and the misclassification cost C is instance independent and thus is the same across all the instances. This assumption holds for all cases, irrespective of whether the tasks are similar. For example, in medical diagnosis, each patient is equally essential, thus leading to the same misclassification cost. The expert (doctor in this case) is equally likely to make a mistake for any of the patients.

Algorithm 1 presents the pseudocode of our defer module. The intuition behind this algorithm is to find when to defer to an expert through cost minimization. Expert cost defines the cost paid each time the defer module defers to an expert, and misclassification cost defines the cost paid each time a wrong prediction is made. Fur-

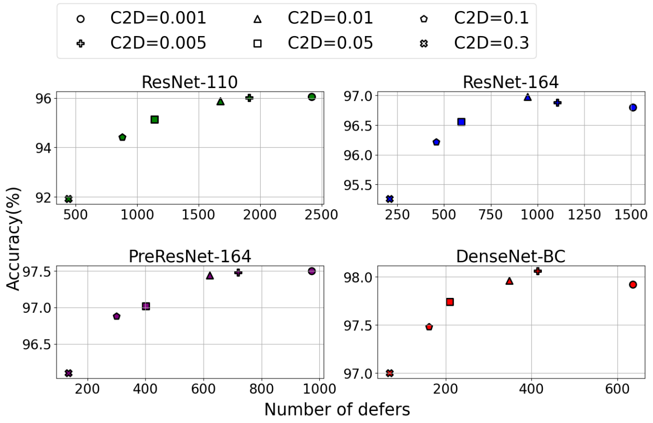

Figure 2. The impact of cost to defer (C2D) on deferred model accuracy and the number of defers on four different ML models.

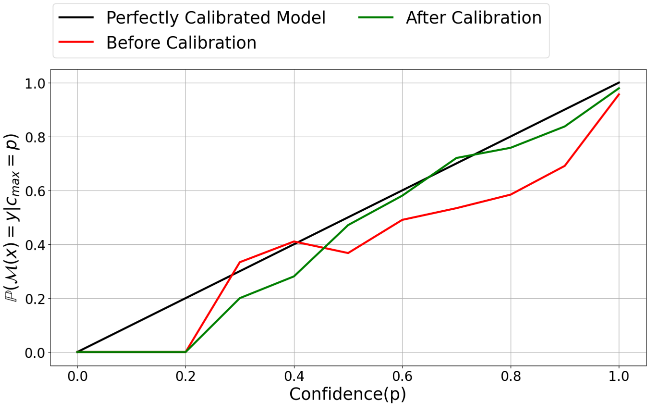

Figure 3. Confidence vs accuracy before and after calibration on CIFAR-10H dataset and ResNet-110 ML model

ther, suppose the defer module defers to an expert, and the expert makes a wrong prediction. In that case, the cost associated with that instance is the total of expert and misclassification costs. The final algorithm defers to the expert when the expected misclassification cost given by $C(1 - c_{max})$ is higher than that of expert cost, i.e., $C_h < C(1 - c_{max})$. Here, c_{max} denote the highest confidence that the model $\mathcal{M}(x)$ associate with a class, for instance, x. It is important to note that if $C <= C_h$, then defer module never defers for predictions. Hence to have an efficient and cost-effective combination, C is always assumed to be higher than C_h in all the results.

Note that the decision to defer depends on the cost ratio denoted by C_h/C. We refer to this ratio as the cost to defer (C2D) ratio, which essentially depicts the ratio of the cost of the expert to that of misclassification cost. It is easy to see that if C2D is higher, then defer module should defer fewer instances to the experts and vice versa. Since, we assume C to be higher than C_h, we have 0 $<$ C2D $<$ 1. If the model's confidence is low, i.e., $c_{max} < 1-$C2D, then the model defers to the expert; otherwise it considers the classifier's output. It should be further noted that in the practical scenario C2D value is generally fixed (or given). For instance, in healthcare applications, C_h will denote the cost of the doctor, and C will quantify the repercussions of a misclassification. To show the efficacy of our proposed approach, we show all our results by considering multiple C2D ratios. However, comparisons must always be made across a single C2D ratio for noticing the improvements.

Figure 2 shows accuracy vs. the number of defers corresponding to four different ML models on the CIFAR10H dataset. Each subplot shows how the cost to defer (C2D) impacts deferred model accuracy and the number of times the expert is consulted. The plot shows that increasing C2D reduces the number of times the combined model defer to the expert and reduces the deferred model's accuracy. As the C2D reaches close to 1, the deferred model accuracy becomes equal to ML's model accuracy. Creating such plots can be very helpful in finding optimal C2D, which corresponds to higher accuracy and fewer expert defers. For example, at C2D=0.1, ResNet-110 accuracy is 94.4% with 877 defers.

The proposed defer module uses the confidence of the classifier to decide whether to defer to an expert. If the classifier's confidence does not reflect the true distribution of the classifier accuracy, then the proposed defer module can lead to more/less number of deferred instances. For example, if the ML model is under-confident about an instance compared to its true accuracy, this will lead to unnecessary costs due to deferring the instance to the human. On the other hand, if the ML model is overconfident, then not deferring to the expert can lead to high misclassification costs. Hence it becomes crucial that the classifier's confidence must reflect the ground truth accuracy associated with each class. The following section elaborates on model calibration and the method to calibrate model probabilities.

3.3 *Effect of Calibrated Classifier on Defer Module*

For supervised learning tasks such as classification, it becomes crucial for classifiers to output confidence that reflects the ground truth corresponding to each class for each instance. If an ML model predicts with 50% confidence on a subset of examples, then ideally, it should classify at least 50% of the examples correctly. ML models exhibiting such behavior are called calibrated models.

A perfectly calibrated model is one which satisfies the following:

$$\mathbb{P}(\mathcal{M}(x) = y|c_{max} = p) = p \tag{1}$$

The above equation implies that if a model is confident with probability p, then it achieves the accuracy of p on the instance x. In Figure 3, the diagonal (black) line represents the relationship between the confidence and accuracy of a perfectly calibrated model. The red line represents the quantity $\mathbb{P}(\mathcal{M}(x) = y|c_{max} = p)$ for the classifier model ResNet-110. This quantity is computed by taking the average accuracy of all the examples with maximum confidence of p. The gap between the red and black lines is the *calibration error* which needs to be minimized. The figure clearly highlights that this calibration error can be high even for ML models with very high accuracy on the dataset. We found similar observations in the plots for other classifiers as well.

Hence, an ill-calibrated model (with high calibration error) can affect the deferred model adversarily. Note that our deferred model defers to an expert only if the model's confidence for the given instance is low. Therefore, the examples with low accuracy and high confidence (due to calibration error) will not be deferred by our model resulting in higher rate of misclassification. Next, we show how a

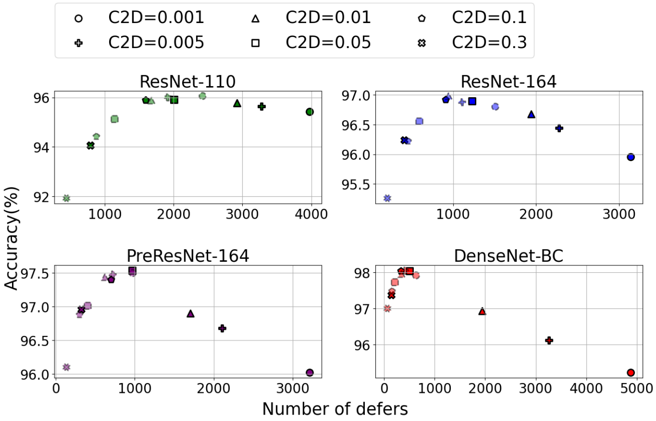

Figure 4. The impact of cost to defer(C2D) on combined model accuracy and the number of defers on four different ML models compared with sections 3.2 results

well-calibrated model can significantly improve the system's overall accuracy.

In literature, model calibration is a well-known problem [6, 14]. The authors in [6] discuss various techniques for model calibration that include histogram binning, Isotonic regression, Bayesian binning into quantiles, and Platt scaling methods. For calibration, we used Platt scaling method, which outputs a calibrated probability $\hat{q}_i = \max_j \sigma_{SM}(z_i/T)^{(j)}$, where z_i is the model's logit vector of instance x_i and σ_{SM} is the softmax function. For calibration, parameter T is optimized using the negative log-likelihood loss function. It is worth noting that since T is just a scaling factor, it does not change the maximum of the softmax function, and hence, this method does not affect the accuracy of the model $\mathcal{M}$. In Figure 3, the green line represents the quantity $\mathbb{P}(\mathcal{M}(x) = y|c_{max} = p)$ after applying model calibration. As can be seen that the calibration error significantly reduces, more importantly, for the examples with higher model confidence. Examples with higher confidence are more important for the deferred model because these examples are not deferred to human experts. Therefore, the model must show the same accuracy in highly confident instances. The calibration achieves this, thus, playing a crucial role in improving the accuracy of the model.

Figure 4 shows the efficacy of calibration on the deferred model. The points in lighter shade show results from the previous Section 3.2, while the points in darker shade show the results after calibration. As can be seen from the figure, for large values of C2D, the calibration leads to improved accuracy without affecting the number of defers much. The reason is that with higher expert costs, the deferred model intelligently chooses the tasks which need to be deferred hence leading to improved combined accuracy. On the other hand, with lower C2D values, the deferred module defers a large number of tasks to the expert, thus eventually reaching that expert accuracy and completely ignoring the model predictions. Therefore, lower C2D values lead to reduced accuracy and high defers. It is important to note that high defers to expert do not result in higher combined accuracy. On all defers to the expert, the ML model may perform much better compared to the expert. However after calibration, the ML model's probabilistic output starts reflecting ground truth, hence the count of expert defers increases. For instance, at C2D=0.1, ResNet-110 accuracy before calibration is 94.4% with 877 defers to expert, whereas after calibration, the accuracy increase up to 95.9% with 1595 defers to expert. Similar performance can be noted across other three ML models as well.

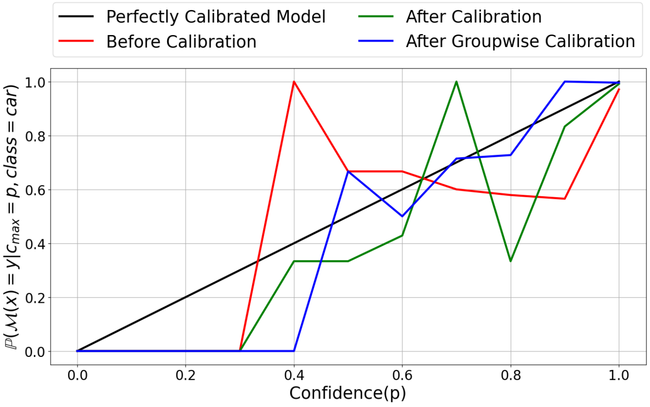

Figure 5. Confidence vs accuracy of car class before calibration on CIFAR-10H dataset and ResNet-110 ML model

The current calibration on the considered ML models is performed considering the accuracy of the whole dataset. However, in k-way classification problems, the overall accuracy of the ML model over the whole dataset might be very different from the accuracies shown on individual classes. This necessitates looking into a more granular form of model calibration in k-way classification problems.

The next section discusses how group-wise calibration can further boost the deferred module's overall accuracy.

3.4 Group-wise Calibration

We begin this section with a simple observation shown in Figure 5, which shows that a calibrated model can be highly uncalibrated on a particular class. Figure 5 compares the calibration error of the ResNet-110 model with that of calibrated ResNet-110 proposed in the previous section for a single class ('car'). The goal of calibration is to get perfect calibration even on the subset of the dataset, which in this case, all instances belong to the car class. The red line represents ill-calibrated model probabilities before calibration, and similar results are generated even after calibration shown in the green line. The figure suggests the need to calibrate the model on finer groups instead of the complete dataset. One possibility is to use class-wise calibration. The blue line in Figure 5 shows how groupwise calibration was able to calibrate the class probabilities efficiently and hence, in turn, calibrates the whole dataset. Our experimental results suggest a finer group-wise calibration produces better results. We find similar groups using the k-means clustering algorithm and find the optimal number of clusters using the elbow method, as shown in Figure 6.

Once the groups are identified, a separate calibration parameter is learned corresponding to each group. We present the accuracy of the deferred module after applying group-wise calibration on classifiers in Figure 7. As can be seen, the accuracy further improves compared to single parameter calibration shown in Figure 4. Results show for C2D=0.1, the ResNet-110 accuracy slight decrese to 95.75% with decrese in defer count to 1563. But for other C2D values, the results are improved as shown in Figure 7.

So far, we have illustrated a completely deferred setting, where for each instance, either model's or the expert's output gets considered

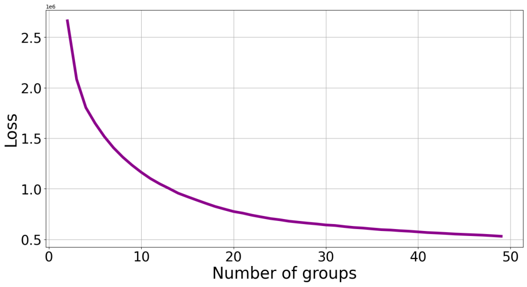

Figure 6. Graph representing relation between number of groups and loss function value in case of ResNet-110 on CIFAR-10H dataset

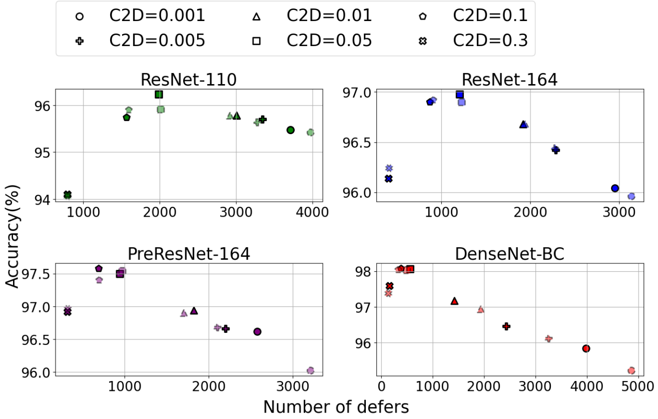

Figure 7. The impact of cost to defer (C2D) on combined model accuracy and the number of defers on four different ML models compared with Section 3.3 results

to make the final decision. In the next section, we show that even when the deferred module defers an instance to an expert, considering the model's output for that instance can lead further improvement in performance. This also happens because the expert, though highly accurate, may not be perfect. Therefore, for deferred instances (i.e. tough tasks), model and expert as a team work better as compared to the expert alone.

3.5 *Human-in-the-Loop Combined Model*

As discussed previously, even for the deferred instances, it may not be a good idea to ignore the classifier's output completely. Recently [9] discusses the method using Bayes' rule to combine the hard label provided by the human with the probabilistic output of the classifier. To use the Bayes' rule, the authors in [9] consider the confusion matrix of the expert denoted by $\phi \in \mathbb{R}^{K \times K}$, where each entry ϕ_{ij} denotes the probability that for a given instance with true label j, human labels the instance as i, i.e., $p(h(x) = i|y = j)$. The confusion matrix is learned from the training data, assuming that the human annotations are known for the training. Then, the combined predicted probability for class j can be computed as:

$$p(y = j|h(x) = i, \mathcal{M}(x)) = \frac{\phi_{ij} c_j}{\sum_{k=1}^{K} \phi_{ik} c_k} \tag{2}$$

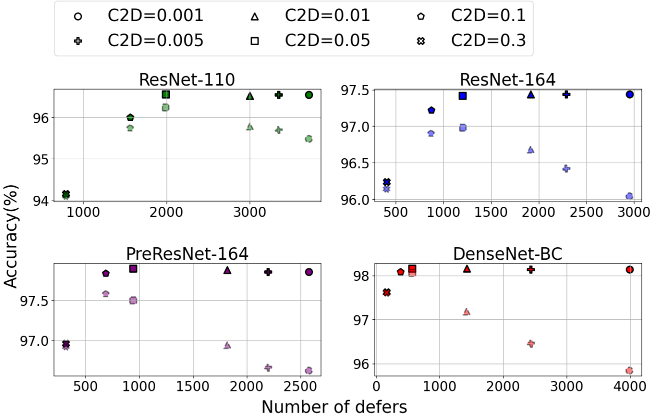

Figure 8. The impact of cost to defer (C2D) on combined model accuracy and the number of defers on four different ML models compared with Section 3.4 results

Here, c_j represent the confidence of model $\mathcal{M}(x)$, for instance, x on class j.

In order to use the human-in-the-loop approach, we also assume that the training data contains the human annotated labels that can be used to learn the human confusion matrix. Once the confusion matrix is learned, we use the Bayes' approach for calculating the combined predicted probability of instances deferred to experts. This is different from the approach used in [9] wherein they use a combination approach for all the instances and hence require the human outputs for all the test instances. On the other hand, we combine the outputs for only those instances that are defered to human. Apart from improving the accuracy, this approach also results in resorting to fewer asks from the experts, resulting in cost efficiency. Figure 8 shows the results with our combined model. Human-in-the-loop show good increase in accuracy for low value of C2D, because the number of expert defers are high. In addition, the combination of ML model's and expert's output always produces better results than either ML model or expert alone. For instance, at C2D=0.1, the combined model accuracy of ResNet-110 increses to 96% with exactly same number of defers as in previous section.

3.6 *Going Extra Mile with Group-wise Confusion Matrices (GCM)*

To combine the probabilistic outputs of the classifier with that of the expert's output, [9] uses the single confusion matrix for the expert, which is learned on the complete dataset. However, the accuracy of an expert can also vary across different groups. Therefore, instead of creating a single confusion matrix for the expert, we propose calculating the human confusion matrix corresponding to each cluster generation in Section 3.4. Figure 9 shows the results of the group-wise human-in-the-loop model using four different classifiers (ResNet-110, Resnet-164, PreResNet-164, DenseNet-54) on the CIFAR-10H dataset. As can be seen, the group-wise confusion matrices (GCM) further improve the accuracy of our proposed model for all the classifiers. For C2D value of 0.1, the ResNet-110 model accuracy further increases to 96.24% with exactly same number of defers.

Table 2. Comparison of our proposed human-in-the-loop combined model with two end-to-end trained defer models from the literature [16, 20]. Results are from CIFAR-10 dataset with test size of 5,000.

Model		C2D	# defers	Accuracy(%)	Classifier Accuracy(%)
Learning to Defer to an Expert [16]		-	2320	97.3 (+3.18)	94.3
Calibrated Learning to Defer [20]			1501	97.2 (+5.39)	92.23
Our Combined Model with Pretrained ML Models	ResNet-110	0.01	3002	97.14 (+9.26)	88.9
	ResNet-164		**1916**	**98.66 (+5.16)**	**93.5**
	PreResNet-164		1820	98.74 (+ 4.16)	94.8
	DenseNet-BC		1426	99.4 (+3.22)	96.3

Table 3. Comparison of our proposed human-in-the-loop combined model with that of the combination model presented in [9]. Results are from CIFAR-10H dataset with test size of 5,000.

Model Name	Combination Model from [9]		Our Combination Approach		
	Pretrained Model Accuracy(%)	Combined Model Accuracy	C2D	Our Combined Model Accuracy(%)	# defers
ResNet-110	88.9	96.1 (+8.1)	0.01	97.08 (+9.25)	3002
ResNet-164	93.5	96.8 (+3.5)		97.74 (+4.54)	1916
PreResNet-164	94.86	97.1 (+2.4)		98.12 (+3.44)	1820
DenseNet-BC	96.36	97.8 (+1.5)		98.6 (+2.32)	1426

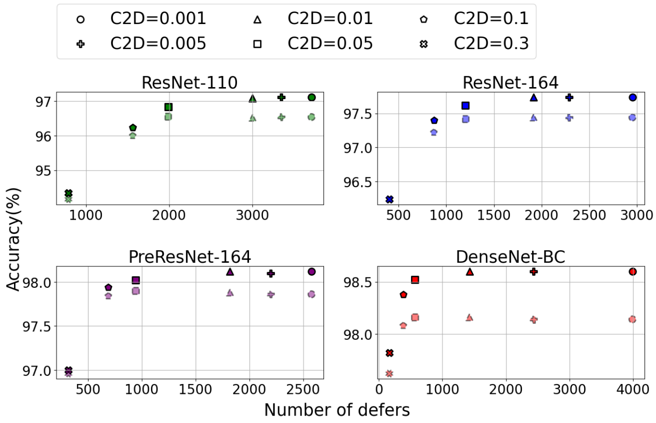

Figure 9. The impact of cost to defer (C2D) on combined model accuracy and the number of defers on four different ML models compared with Section 3.5 results

4 Final Results

This section compares the results obtained by our model with that of existing literature. Defer models proposed in [16, 20] require an end-to-end training of the combination of classifier and defer module. Our combined model uses an existing trained classifier to make predictions. The results are compared on the CIFAR-10 dataset with a synthetic expert. CIFAR-10 dataset is considered to perform equivalent comparisons as the same dataset is used to generate results in [16, 20]. Further, the authors in [16, 20] consider that the expert has 100% accuracy in classes from one to six and generates random values for other classes. We also simulate humans in the same way to compare our results. Table 2 compares the accuracies of these end-to-end defer models [16,20] to our human-in-the-loop combined model. As can be noted, our approach's accuracy considering multiple classifiers shows that our model always achieves better accuracies with fewer defers on comparable classifier accuracies. Moreover, our model does not require end-to-end training of either the classifier or defer model; hence it is much faster to implement compared to [16, 20]. Figure 9 shows how increasing expert cost reduces combined model accuracy and the number of expert defers.

Combining each test instance with the expert's predictions can generate good results, but it is not cost-effective. The results from Table 3 show the comparison between [9] and our model. The percentages in the brackets show the percentage improvement in the accuracy of the respective combined models as compared to the accuracy of the classifier. As noted, our model's accuracy is much better compared to [9] with fewer defers to expert (with cost savings in the range of 20% to 80% depending upon the classifier). These results highlight the cost-effectiveness and performance improvement of our proposed approach. Tables 2 and 3 show that our combined model outperforms [16, 20] and [9] in terms of accuracy for a fixed C2D.

5 Conclusions

This paper proposes a human-in-the-loop based classification approach considering a simple deferred model with two parameters, i.e., expert and misclassification costs. The proposed deferred model does not require end-to-end training and human annotations for the training set, thus saving computational cost and expert costs on training. We showed that the current ML models are generally ill-calibrated, with over-confidence in some instances as opposed to the achieved accuracy. We then studied the effects of well-calibrated ML models on the accuracy of the deferred model. We showed that calibration leads to increased accuracy and robustness with respect to the expert cost. We then extend the ideas where the calibration is done at each group, leading to further accuracy improvement. Finally, we showed that if the expert's annotations are present in the training set, these can be used to learn the confusion matrices of the humans, which in turn can be used to combine the model's output on the deferred instances during testing. The overall approach leads to better accuracy and fewer experts deferred compared to existing deferred and combined models.

References

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld, 'Is the most accurate ai the best teammate? optimizing ai for teamwork', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11405–11414, (2021).

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

[3] Peter L Bartlett and Marten H Wegkamp, 'Classification with a reject option using a hinge loss.', *Journal of Machine Learning Research*, **9**(8), (2008).

[4] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri, 'Learning with rejection', in *International Conference on Algorithmic Learning Theory*, pp. 67–82. Springer, (2016).

[5] Thomas G Dietterich, 'Ensemble methods in machine learning', in *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pp. 1–15. Springer, (2000).

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, 'On calibration of modern neural networks', in *International conference on machine learning*, pp. 1321–1330. PMLR, (2017).

[7] Dan Hendrycks and Kevin Gimpel, 'A baseline for detecting misclassified and out-of-distribution examples in neural networks', *arXiv preprint arXiv:1610.02136*, (2016).

[8] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta, 'To trust or not to trust a classifier', *Advances in neural information processing systems*, **31**, (2018).

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

[10] Hyun-Chul Kim and Zoubin Ghahramani, 'Bayesian classifier combination', in *Artificial Intelligence and Statistics*, pp. 619–627. PMLR, (2012).

[11] 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).

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

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

[14] Jonathan Martinez, Kobi Gal, Ece Kamar, and Levi HS Lelis, 'Improving the performance-compatibility tradeoff with personalized objective functions', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5967–5974, (2021).

[15] Hussein Mozannar, Arvind Satyanarayan, and David Sontag, 'Teaching humans when to defer to a classifier via exemplars', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5323–5331, (2022).

[16] Hussein Mozannar and David Sontag, 'Consistent estimators for learning to defer to an expert', in *International Conference on Machine Learning*, pp. 7076–7087. PMLR, (2020).

[17] Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal, 'Consistent algorithms for multiclass classification with an abstain option', *Electronic Journal of Statistics*, **12**(1), 530–554, (2018).

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

[19] Frederik Träuble, Julius Von Kügelgen, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Peter Gehler, 'Backward-compatible prediction updates: A probabilistic approach', volume 34, pp. 116–128, (2021).

[20] Rajeev Verma and Eric Nalisnick, 'Calibrated learning to defer with one-vs-all classifiers', in *International Conference on Machine Learning*, pp. 22184–22202. PMLR, (2022).

[21] 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).