

UC Irvine

UC Irvine Previously Published Works

Title

Active Bayesian Assessment for Black-Box Classifiers

Permalink

<https://escholarship.org/uc/item/30v539zh>

Authors

Ji, Disi
IV, Robert L Logan
Smyth, Padhraic
[et al.](#)

Publication Date

2020-02-16

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Active Bayesian Assessment for Black-Box Classifiers

Disi Ji

*Department of Computer Science
University of California, Irvine*

DISIJ@UCI.EDU

Robert Logan

*Department of Computer Science
University of California, Irvine*

RLOGAN@UCI.EDU

Padhraic Smyth

*Department of Computer Science
University of California, Irvine*

SMYTH@ICS.UCI.EDU

Mark Steyvers

*Department of Cognitive Science
University of California, Irvine*

MARK.STEYVERS@UCI.EDU

Abstract

Recent advances in machine learning have led to increased deployment of black-box classifiers across a wide variety of applications. In many such situations there is a crucial need to assess the performance of these pre-trained models, for instance to ensure sufficient predictive accuracy, or that class probabilities are well-calibrated. Furthermore, since labeled data may be scarce or costly to collect, it is desirable for such assessment be performed in an efficient manner. In this paper, we introduce a Bayesian approach for model assessment that satisfies these desiderata. We develop inference strategies to quantify uncertainty for common assessment metrics (accuracy, misclassification cost, expected calibration error), and propose a framework for active assessment using this uncertainty to guide efficient selection of instances for labeling. We illustrate the benefits of our approach in experiments assessing the performance of modern neural classifiers (e.g., ResNet and BERT) on several standard image and text classification datasets.

1. Introduction

Complex machine learning models, particularly deep learning models, are now being applied to a variety of practical prediction problems ranging from diagnosis of medical images (Kermany et al., 2018) to autonomous driving (Du et al., 2017). As a result, software systems with embedded machine learning components are becoming increasingly common. Many of these models will be black boxes from the perspective of downstream users, such as models developed remotely by commercial entities and hosted as a service in the cloud (Yao et al., 2017). For a variety of reasons (legal, economic, competitive), users will often have no direct access to the detailed workings of the model, how the model was trained, or the training data.

In this context it is increasingly important for the user of a model to have accurate and robust assessments of the quality of a model’s predictions. However, as an example, “self-confidence” estimates provided by machine learning predictors can often be quite unreliable and miscalibrated (Zadrozny and Elkan, 2002; Kull et al., 2017; Ovadia et al., 2019). In particular, complex models such as deep networks with high-dimensional inputs (e.g., images and text) can be significantly

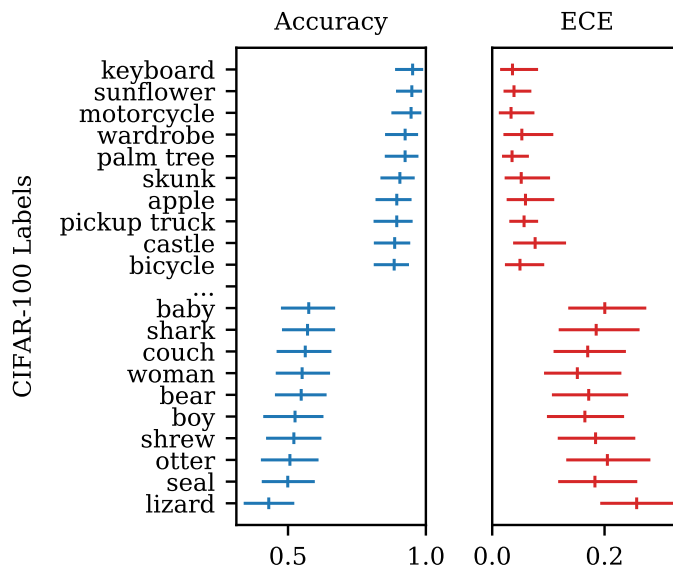


Figure 1: Mean posterior estimates and 95% credible intervals for the classwise accuracy and expected calibration error (ECE) of a ResNet-110 image classifier on the CIFAR-100 test set, using our Bayesian assessment framework. For further details see Section 3.

overconfident in practice (Gal and Ghahramani, 2016; Guo et al., 2017; Lakshminarayanan et al., 2017). Furthermore, test set performance metrics from train-test splits might not be an accurate reflection of downstream performance due to factors such as label and distribution shift (e.g., Lipton et al. (2018); Hendrycks and Gimpel (2017); Ovadia et al. (2019) and/or implicit optimistic bias (Recht et al., 2019).

Thus, downstream users of black-box predictors will need the capability to carry out assessment separately and independently from the training and evaluation procedures used when fitting the model. This assessment could for example be conducted by organizations not involved in training the model, in a manner similar to the assessments of commercial products carried out by regulatory agencies. Additional motivations for independent assessment include legal requirements that may mandate independent assessment of models, the need to build trust on the part of a human consumer of model predictions, or situations where the predictor is being deployed in an environment with a different distribution over inputs and outputs compared to the environment the model was trained on.

In this paper we develop a Bayesian framework for performing this assessment that requires no internal access to the model being evaluated, only its output probabilities. We focus on classification models in particular, although the ideas are broadly applicable to prediction models in general. Figure 1 provides an illustrative example applying our approach to assess performance of a ResNet classifier on the CIFAR-100 dataset, with detailed assessments of the classwise accuracy and calibration properties, along with posterior credible intervals measuring how certain we are in our assessment given the data available.

Assessment of model properties such as accuracy and calibration requires labeled data. In real-world deployment scenarios, this data is likely to be scarce and costly to collect, e.g., consider a pre-trained classification model being deployed in a diagnostic imaging context in a particular hospital. With this in mind we also develop a framework for active assessment of black-box classifiers,

	Mode	Size	Classes	Model
CIFAR-100	Image	10K	100	ResNet-110
ImageNet	Image	50K	1000	ResNet-152
SVHN	Image	100K	10	ResNet-152
20 Newsgroups	Text	7.5K	20	BERT _{BASE}
DBpedia	Text	70K	14	BERT _{BASE}

Table 1: Assessment datasets and models used in our experiments. Size refers to the maximum number of labeled instances available for assessment.

using techniques from active learning to efficiently select instances to label so that model deficiencies such as low accuracy, or high cost mistakes can be quickly identified.

In summary, our primary contributions are:

- We propose a general framework for black-box classifier assessment, using a Bayesian approach that is applicable to a range of performance-related metrics.
- We illustrate the utility of the framework via Bayesian inference with posterior uncertainty for quantities such as classwise accuracy, expected calibration error (ECE), and reliability diagrams.
- We develop a new framework called *active assessment* and demonstrate how this framework can be used to identify extreme classes in an online label-efficient manner with significant gains over traditional random sampling methods.

2. Preliminaries

2.1 Notation

We consider classification problems with a feature vector \mathbf{x} and a class label $y \in \{1, \dots, K\}$, e.g., classifying image pixels \mathbf{x} into one of K classes. We assume access to a trained prediction model M that makes predictions of y given a feature vector \mathbf{x} . In particular we assume that the model produces numerical scores per class, reflecting its confidence, typically in the form of a set of estimates of class-conditional probabilities $p_M(y = k|\mathbf{x}), k = 1, \dots, K$. Such probability estimates can be obtained from a logistic classifier, from the softmax output layer of a neural network, from averages over leaf nodes in tree-based models, and so on. A notational aside: for probabilities that are being generated by the model we use subscript M , e.g., $p_M(y = k|\mathbf{x})$. When we refer to the actual true probability with respect to the underlying true distribution $p(\mathbf{x}, y)$ we drop the subscript, e.g., when using terms like $p(y = k|\mathbf{x})$ and $p(\mathbf{x})$ in computing expectations.

Under 0-1 classification loss, $\hat{y} = \arg \max_k p_M(y = k|\mathbf{x})$ will be the classifier’s label prediction for a particular input \mathbf{x} . We can define $s(\mathbf{x}) = p_M(y = \hat{y}|\mathbf{x})$ as the **score** of a model, as a function of \mathbf{x} , i.e., the class probability that the model produces for its predicted class $\hat{y} \in \{1, \dots, K\}$ given input \mathbf{x} . This is sometimes also referred to as a model’s **confidence** in its prediction and can be viewed a model’s own estimate of its accuracy when it predicts \hat{y} given \mathbf{x} . The model’s scores in general need not be perfectly calibrated, i.e., they need not match the true probabilities $p(y = \hat{y}|\mathbf{x})$.

2.2 Datasets and Classification Models

Assessment Datasets We assess performance characteristics of neural models on several standard image and text classification datasets. The image datasets we use are: *CIFAR-100* (Krizhevsky and Hinton, 2009), *SVHN* (Netzer et al., 2011) and *ImageNet* (Russakovsky et al., 2015). The text datasets we use are: *20 Newsgroups* (Lang, 1995) and *DBpedia* (Zhang et al., 2015). Detailed

statistics are provided in Table 1. The assessment datasets are based on standard test sets used for each dataset in the literature.

Prediction Models For image classification we use ResNet (He et al., 2016) architectures with either 110 layers (CIFAR-100) or 152 layers (SVHN and ImageNet). For ImageNet we use the pretrained model provided by PyTorch, and for CIFAR and SVHN we use the pretrained model checkpoints provided at: <https://github.com/bearpaw/pytorch-classification>. For text classification tasks we use fine-tuned BERT_{BASE} (Devlin et al., 2019) models. These models were all trained on standard training sets in the literature, independent from the datasets used for assessment. To facilitate reproducing our results we provide all of the model predictions used in our experiments at: <https://github.com/disiji/bayesian-blackbox>.

3. Bayesian Assessment of Classification Metrics

We focus on the problem of assessing the performance of a model on data drawn from some unknown distribution $p(\mathbf{x}, y)$ representing the environment where the model is being used. This joint distribution in general need not necessarily be the same as the distribution that the model was trained on. We are interested in particular in the situation where the model is a black box, where we can observe the inputs \mathbf{x} and the outputs $p_M(y = k|\mathbf{x})$, but don't have any other information about its inner workings (for example about any internal parameters of the model). Specifically, in this paper, rather than learning a model itself we want to learn about the characteristics of a fixed model that is making predictions in a particular environment.

A natural approach to assessing the performance of a black-box classifier is to adopt a Bayesian framework where we treat the metrics of interest (classification accuracy, calibration error) as unknown parameters that we estimate from (limited) labeled data drawn from a distribution $p(\mathbf{x}, y)$.

3.1 Assessing Classwise Accuracy

Beginning with classification accuracy, the *marginal accuracy* of a classification model is defined as $A(\mathbf{x}) = p(y = \hat{y}|\mathbf{x})$. We also define *regional accuracy* over local regions of the input space. For any region \mathcal{R} in the input space, regional accuracy is the marginal probability that the predicted label matches with the true label, conditioned on $\mathbf{x} \in \mathcal{R}$:

$$\begin{aligned} A_{\mathcal{R}} &= \mathbb{E}_{p(\mathbf{x}, y|\mathbf{x} \in \mathcal{R})}[A(\mathbf{x})] \\ &= \int_{\mathcal{R}} p(y = \hat{y}|\mathbf{x})p(\mathbf{x}|\mathbf{x} \in \mathcal{R})d\mathbf{x}. \end{aligned} \quad (1)$$

We will use this as one of our main assessment tools.

In particular, we will focus on assessment of *classwise accuracy*, $A_{\mathcal{R}_k}, k = 1, \dots, K$, the expected accuracy of the model whenever it predicts class k . This corresponds to having the input region be the classifier's decision region $\mathcal{R}_k = \{\mathbf{x}|\hat{y} = k\}$. To estimate the classwise accuracies $A_{\mathcal{R}_k}$ from data, a standard approach would be to empirically approximate the integral above by sampling \mathbf{x}, y pairs from the conditional distribution $p(\mathbf{x}, y|\mathbf{x} \in \mathcal{R}_k)$. Equivalently, $A_{\mathcal{R}_k}$ can be modeled as an unknown Bernoulli parameter θ_k , with draws $(\mathbf{x}^{(i)}, y^{(i)})$, conditioned on $\mathbf{x} \in \mathcal{R}_k$, leading to binary outcomes $\mathbb{1}(y^{(i)} = \hat{y}^{(i)}) \in \{0, 1\}$, with a frequency-based (maximum likelihood) estimate:

$$\hat{\theta}_k = \frac{1}{S} \sum_{i=1}^S \mathbb{1}(y^{(i)} = \hat{y}^{(i)}). \quad (2)$$

It is natural to consider Bayesian estimation in this context, especially in situations where there is little labeled data available for assessment and/or where the number of regions K is large. In particular, we can put a Beta(α_k, β_k) prior on θ_k , model the draws $\mathbb{1}(y^{(i)} = \hat{y}^{(i)})$ with a binomial

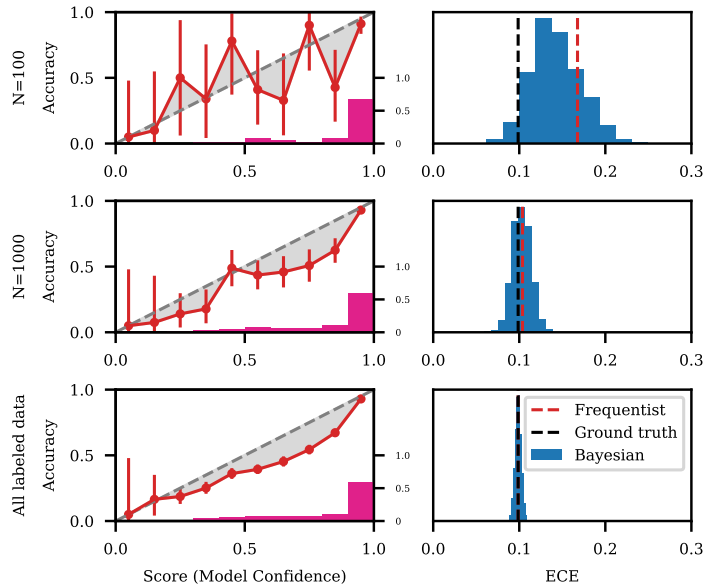


Figure 2: Bayesian reliability diagrams (left) and posterior densities for ECE (right) for CIFAR-100 as the amount of data used for estimation increases. Vertical lines in the right plots depict the ground truth ECE (black, evaluated with all available assessment data) and frequentist estimates (red).

likelihood, and produce Beta posteriors for each $\theta_k, k = 1, \dots, K$. The Bayesian approach allows for uncertainty in our inferences about quantities such as $\theta_k = A_{\mathcal{R}_k}$ as well as providing a basis for supporting techniques such as active selection of examples for labeling (discussed later in the paper). In situations where we have no *a priori* information about θ_k we can use a weak uninformative prior with $\alpha_k = \beta_k = 1$; alternatively, we can use strong prior information (e.g., if the assessor believes the performance metrics reported by the model-builder) when available.

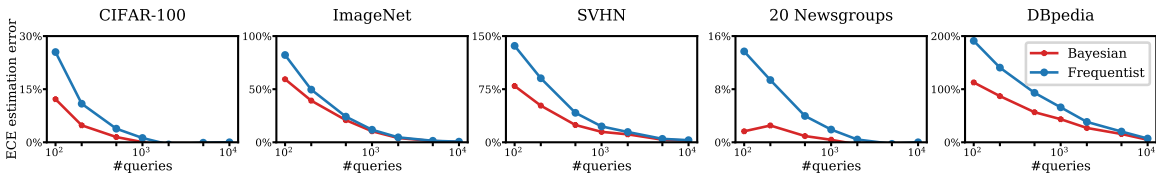


Figure 3: Percentage error in estimating expected calibration error (ECE) as a function of dataset size, for Bayesian (red) and frequentist (blue) estimators, across five datasets.

3.2 Assessing Calibration Performance

We can also assess calibration-related metrics for a classifier in a Bayesian fashion using any of well-known various calibration metrics (Kumar et al., 2019; Nixon et al., 2019). Here we focus on expected calibration error (ECE) given that it is among the widely-used calibration metrics in the machine learning literature (e.g., Guo et al. (2017); Ovadia et al. (2019)). We use the standard ECE

binning procedure, with scores aggregated into equal-width bins, denoting the b -th bin or region as:

$$\mathcal{R}_b = \{\mathbf{x} | s(\mathbf{x}) \in [(b-1)/B, b/B)\}, \quad (3)$$

where $b = 1, \dots, B$ ($B = 10$ is often used in practice). Note that these are not the decision regions induced by the model that we discussed earlier but instead are a partition of the input space determined by the score of the predicted class $s(\mathbf{x})$. The marginal ECE is defined as a weighted average of the absolute distance between the true accuracy θ_b and the average score s_b per bin:

$$\text{ECE} = \sum_{b=1}^B p_b |\theta_b - s_b| \quad (4)$$

where p_b is the probability of a score lying in bin b . The unknown accuracy of the model per bin is θ_b , which can be viewed as a marginal accuracy over the region \mathcal{R}_b in the input space corresponding to $s(\mathbf{x}) \in \text{Bin}_b$, i.e., $\theta_b = \int_{\mathcal{R}_b} p(y = \hat{y}_M | \mathbf{x}) p(\mathbf{x} | \mathbf{x} \in \mathcal{R}_b) d\mathbf{x}$.

To assess the marginal ECE (Eqn. 4), we put Beta priors over the θ_b 's, $b = 1, \dots, B$. Our default setting for the priors is a weak prior ($\alpha + \beta = 2$) with the mean of the prior on the diagonal for each bin, i.e., we assume *a priori* that the model is calibrated but allow the data to easily overwhelm the prior if there is evidence that the model is not well-calibrated. The posterior distribution over ECE is a weighted average of the absolute value of B shifted Beta posterior distributions corresponding to the individual θ_b 's. The posterior is not available in closed form but Monte Carlo samples are straightforward to obtain.

As for accuracy, we can also model *classwise ECE*, $\text{ECE}_k = \sum_{b=1}^B p_{b,k} |\theta_{b,k} - s_b|$, by modifying the model described above to use regions $R_{b,k} = \{\mathbf{x} | \hat{y} = k, s(\mathbf{x}) \in \text{Bin}_b\}$ that partition the input space by predicted class in addition to the model score.

3.3 Experiments with Accuracy and Calibration Assessment

A simple illustration of our approach is provided in Figure 1. We plot mean posterior estimates and 95% credible intervals of classwise accuracy and ECE produced using predictions from a ResNet-110 model on the entire CIFAR-100 test set. The assessment shows that (a) model accuracy and calibration varies substantially across classes and (b) that classes with low classwise accuracy also tend to be less calibrated. We discuss this further in the supplemental material where we show that negative correlation between classwise accuracy and ECE is observed across all five datasets.

Another example of where we can apply Bayesian assessment is in assessing reliability diagrams for classifiers, a widely used tool for visually diagnosing model calibration (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005) (e.g., Figure 2). These diagrams plot the empirical sample accuracy $A(\mathbf{x}) = p(y = \hat{y} | \mathbf{x})$ as a function of the model's confidence $s(\mathbf{x}) = p_M(y = \hat{y} | \mathbf{x})$. If the model is perfectly calibrated, then $A(\mathbf{x}) = s(\mathbf{x})$ and the diagram should plot the identity function on the diagonal. Any deviation away from the diagonal reflects miscalibration of the model. For a particular value $s(\mathbf{x}) = s \in [0, 1]$ along the x-axis, the corresponding y value is defined as $\mathbb{E}_{\mathbf{x} | s(\mathbf{x})=s} [A(\mathbf{x})]$. As we did in the previous section, we model the marginal accuracy within each bin as an unknown quantity $\theta_j = \mathbb{E}_{p(\mathbf{x}, y | \mathbf{x} \in \mathcal{R}_j)} [A(\mathbf{x})]$. We use a Beta prior over each $\theta_j \in [0, 1]$, then update this prior using a binomial likelihood over binary observations $y \in [0, 1]$.

In Figure 2, the prior distribution of marginal accuracy within each bin is a Beta distribution with its mean on the diagonal and pseudocount $\alpha + \beta = 2$. As the amount of data used increases, the credible intervals of the Bayesian reliability diagram (left column) get narrower, the posterior density of ECE (right column) converges to ground truth, and the uncertainty about ECE decreases. When the number of samples is small, with the same set of randomly selected samples 100 samples (row 1), the Bayesian estimation of ECE puts non-negative probability mass on ground truth marginal ECE, where "ground truth" refers to the marginal ECE computed with all labeled assessment data, while the frequentist method significantly overestimates ECE without any notion of uncertainty.

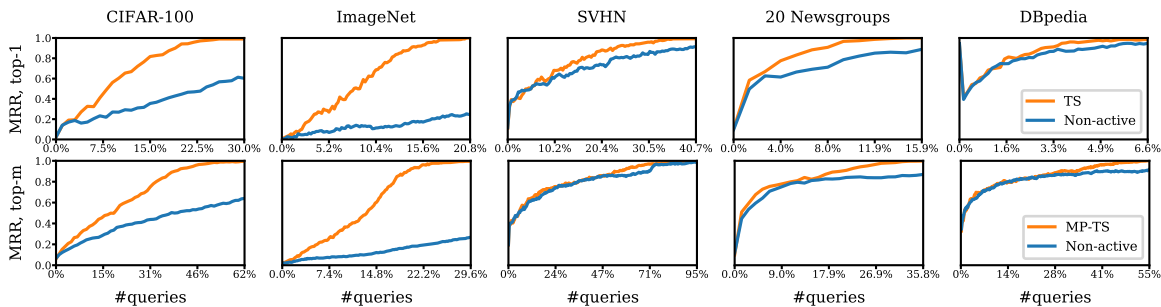


Figure 4: Mean reciprocal rank (MRR) of the m lowest accuracy predicted classes, comparing active learning (with Thompson sampling (TS)) with no active learning, across five datasets. In the top row $m = 1$, and in the bottom row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

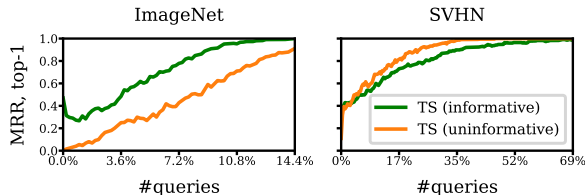


Figure 5: Comparison of the effect of informative (green) and uninformative (orange) priors on identifying the least accurate predicted class on ImageNet and SVHN.

In Figure 3 we show the percentage error obtained for Bayesian mean posterior estimates (MPE) and frequentist estimates of marginal ECE as a function of the number of labeled data points (“queries”) across five datasets. The percentage is computed relative to the ground truth marginal $ECE = ECE^*$. The MPE is computed with Monte Carlo samples from the posterior distribution (an example of histograms of such samples are shown in Figure 2). At each step, we randomly draw and label N queries from the pool of unlabeled data, and compute both a Bayesian and frequentist estimate of marginal calibration error with these labeled data. We run the simulation 100 times, and report the average ECE_N over the N samples. Figure 2 plots $(ECE_N - ECE^*)/ECE^*$ as a percentage. The Bayesian method consistently has lower ECE estimation error, especially when the number of queries is small.

4. Active Bayesian Assessment

The results in the previous section illustrate how the Bayesian approach can be useful in obtaining assessments with uncertainty quantification, e.g., for practical deployment situations where labeled data is likely to be sparse, in contrast to typical results in the literature which assume the availability of large test datasets for assessing performance metrics. In this section we illustrate how we can further improve performance by extending our Bayesian framework to *active assessment*, allowing for model assessment to be performed by actively selecting examples \mathbf{x} for labeling in a data-efficient manner. This scenario is particularly relevant to problems where we have a potentially large pool of unlabeled examples \mathbf{x} available, and have limited resources for labeling (e.g., a human labeler). The question we address here is, if we can only select N classes from a larger pool of unlabeled examples, which examples should we select. We illustrate below how active data selection can be performed to

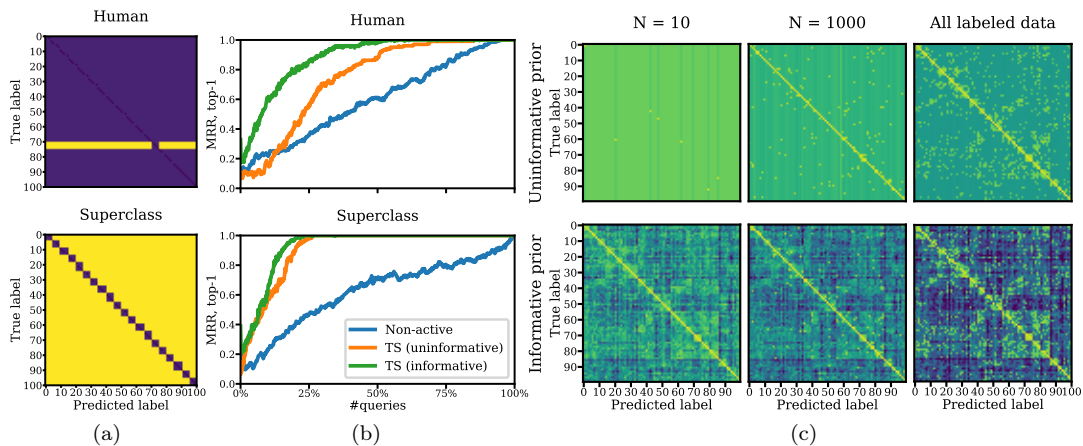


Figure 6: (a) Cost matrices used in our experiments. (top): misclassifying humans (the row in yellow) is 10x more costly than other mistakes. (bottom): confusing a class with another superclass (e.g., a *vehicle* with a *fish*) is 10x more expensive than a mistake within the same superclass (e.g., confusing *shark* with *trout*). (b) Success rates of active assessment using Thompson sampling (TS) (with both informative and uninformative priors) vs. non-active assessment for determining the predicted class with highest expected cost. Active learning with the informative prior detects the highest cost predicted class using far fewer queries than the other methods. (c) Comparison of the mean posterior estimates of the confusion probabilities θ (in the superclass experiment) under informative and uninformative priors.

solve the problem of identifying extreme classes (e.g., classes that rank first or the last according to a given metric such accuracy, calibration error, or expected cost)

4.1 Best Arm Identification

Thompson Sampling The problem of identifying extreme classes can be formulated as a multi-armed bandit problem where the “arms” are predicted classes \hat{y} , and the “reward” signal is a function of the model’s prediction \hat{y} and true label y (e.g., +1 every time $y \neq \hat{y}$ if the goal is to determine the least accurate predicted class). The popular Thompson sampling algorithm used to solve this class of problems (Thompson, 1933; Russo et al., 2018) readily lends itself to our Bayesian framework. The basic idea is, at each query, to sample from the posterior distribution of the evaluation metric, and select one data point \mathbf{x} to be labeled (from a pool of unlabeled examples) for the predicted class (or arm) with highest/lowest sampled value.

We also experimented with a modified version of Thompson sampling called top-two Thompson sampling (TTTS) which has theoretical advantages for identifying the best arm in a pure exploration mode (Russo, 2016). This algorithm adds a re-sampling process to encourage more exploration. At each step, the second-most optimal arm may be randomly selected in place of the most optimal arm. We found that for the problems and datasets we investigated in this paper that TS and TTTS gave very similar performance, so we for simplicity we just present results for TS in this paper. A formal description of these algorithms is provided in the supplemental material.

There are a variety of other active learning algorithms (such as epsilon-greedy and UCB methods) that could also be used for active assessment. We found Thompson sampling to be more reliable and consistent in terms of efficiency than these methods across all five datasets (results and sensitivity analysis for prior strength in the supplementary material). We focus on the Bayesian/TS approach in

our results below since our primary aim is to demonstrate the utility of active assessment compared to no active assessment.

4.2 Best- m Arms Identification

Another variant of an active Bayesian assessment framework is best- m arms identification.¹ This can be motivated for example by task allocation, e.g., finding the m predicted classes that a model is least accurate on, so that whenever the model predicts one of these classes the prediction decision is handed instead to a more accurate predictor (e.g., a human). For example, suppose we have a dataset with $K = 100$ equally-likely classes (e.g., CIFAR-100) and a budget where we can send 10% of our examples to a human to make predictions (and the other 90% are made by our black-box model). One way to address this is to find the set of 10 predicted classes that the model is least accurate for and use the human to make predictions when \hat{y} is in this set.

Identification of the best- m arms can be formulated as a multiple-play multi-armed bandit (MAB) problem. Komiyama et al. (2015) proposed the multiple-play Thompson sampling (MP-TS) algorithm and proved that MP-TS has the optimal regret upper bound when the reward is binary. This algorithm differs from standard Thompson sampling in that, at each step, data points for the top- m arms (according to a sample from the posterior) are labeled as opposed to just the top arm. A detailed algorithm description is provided in the supplemental material.

4.3 Active Bayesian Assessment of Accuracy

We apply the active Bayesian assessment framework to the problem of determining the predicted classes with the lowest classwise accuracies, using the beta-binomial model described in Section 3.1. Figure 4 compares our active Bayesian assessment method to a traditional non-active assessment method (i.e., evaluating on a test set of uniformly drawn data points). Each algorithm was run 100 different times on each of the datasets listed in Table 1. For evaluation, for each run, at each step, we identify the least accurate classes, according to the MPE of the posterior distribution for the Bayesian method and the frequency-based estimate for the non-active method.

The x-axis measures the number of queries made to the oracle. The y-axis measures the mean reciprocal rank (MRR) of the predicted top- m classes:

$$MRR = \frac{1}{m} \sum_{i=1}^m \frac{1}{\text{rank}_i} \quad (5)$$

where rank_i is the predicted rank of the i th best class. Following standard practice, other classes in the best- m are ignored when computing rank so that $MRR = 1$ if the predicted top- m classes match ground truth.

Our results demonstrate that the active learning approach is much more effective at identifying the least accurate class or least accurate top- m classes relative to working with a randomly sampled test set. For example, for CIFAR-100 and ImageNet, in all of the trials, the correct class is identified after querying 30% and 20% of the pool of unlabeled data respectively. In contrast, the non-active strategy only gets MRR around 0.5 and 0.2 on these two datasets with the same amount of labeled data. In general, active assessment appears to achieve the largest gains in efficiency on datasets where the number of classes is large (e.g., CIFAR-100 and ImageNet).

Figure 4 shows the results when the prior distribution of accuracy is uninformative. However, since a model’s confidence reflects a model’s self-assessment of accuracy, we could also use an informative prior by placing a prior distribution for accuracy that is centered around the model’s confidence per class. Figure 5 shows the results for two data sets comparing active assessment for an informative prior (green) and an uninformative prior (orange). We set the uninformative prior for classwise

1. This is typically referred to as best- k arms identification in the literature. We use the symbol m to avoid overloading k .

accuracy to be Beta(1, 1) for each predicted class, the informative prior to be Beta($2s_k, 2(1 - s_k)$), where s_k is the average model confidence (score) of all data points (which we can obtain using unlabeled data) for the predicted class k . The results in Figure 5 illustrate that the informative prior can be helpful when the prior captures the relative ordering of classwise accuracy well (e.g., ImageNet), but less helpful when the difference in classwise accuracy across classes is small and the classwise ordering reflected in the ‘‘self-assessment prior’’ is more likely to be in error (e.g., SVHN; classwise accuracies provided in the supplementary material).

4.4 Active Bayesian Assessment of Confusion Matrices and Misclassification Costs

Accuracy assessment can be viewed as implicitly assigning a binary cost to model mistakes, i.e. a cost of 1 to incorrect predictions and a cost of 0 to correct predictions. In this sense, identifying the predicted class with lowest accuracy is equivalent to identifying the class with greatest expected cost. However, in real world applications, costs of different types of mistakes can vary drastically. For example, in autonomous driving applications, misclassifying a pedestrian as a crosswalk can have much more severe consequences than other misclassifications.

To deal with such situations, we extend our approach to incorporate an arbitrary cost matrix $\mathbf{C} = [c_{jk}]$, where c_{jk} is the cost of predicting class $\hat{y} = k$ for a data point whose true class is $y = j$. Conditioned on a predicted class \hat{y} , the true class label y has a categorical distribution $\theta_{jk} = p(y = j | \hat{y} = k)$. We will refer to θ_{jk} as confusion probabilities since they resemble the elements of a confusion matrix. The **classwise expected cost** for predicted class k is given by:

$$C_{\mathcal{R}_k} = \mathbb{E}_{p(\mathbf{x}, y | \mathbf{x} \in \mathcal{R}_k)}[c_{jk} \mathbb{1}(y = j)] = \sum_{j=1}^K c_{jk} \theta_{jk}. \quad (6)$$

Similar to how we use a beta-binomial distribution to model accuracy, we can model these confusion probabilities using a Dirichlet-multinomial distribution: $\theta_{\cdot k} \sim \text{Dirichlet}(\alpha_{\cdot k})$. The same active querying approach described in the previous section can then be used to identify the class with the highest classwise expected cost (e.g., $k^* = \arg \max_k C_{\mathcal{R}_k}$).

For prior distributions we evaluate two options. The first is an uninformative prior with $\alpha_{jk} = \alpha, \forall j, k$. The second is an informative prior based on the model’s own prediction scores, $\alpha_{jk} \propto \sum_{\mathbf{x} \in \mathcal{R}_k} p_M(y = j | \mathbf{x})$. This informative prior is likely to be more useful for problems such as CIFAR-100 where the number of confusion probabilities to estimate is large and observations are relatively sparse.

We experiment with two different cost matrices on the CIFAR-100 dataset: (i) the cost of misclassifying a person (e.g., predicting *fish* when the true class is a *woman, girl, boy*, etc.) is more expensive than other mistakes, (ii) the cost of confusing a class with another superclass (e.g., a *vehicle* with a *fish*) is more expensive than the cost of mistaking labels within the same superclass (e.g., confusing *shark* with *trout*). We visualize these cost matrices in Figure 6(a). In Figure 6(b), we compare the performance of active and non-active assessment at identifying the class with highest cost, averaged over 100 trials. We set the pseudocount of both priors to be 1, and the cost of expensive mistakes to be 10x the cost of other mistakes. For both of the cost matrices we find that the active approach is more effective than the non-active approach, and that the informative prior is more effective than the uninformative prior. Even though the model is not well-calibrated (e.g., see Figures 1 and 2) there is nonetheless valuable information about confusion probabilities available from the model’s estimates of class-conditional probabilities.

To illustrate how the informative prior helps deal with sparsity, we plot samples from the posterior of θ when the number of queries is 10, 1000, 10000 in Figure 6(c). For the uninformative prior (top panel), even when all of the available data is used, there is still considerable uncertainty about the magnitude of the off-diagonal confusion probabilities. However, this is not the case for the informative prior (bottom panel) since the prior for the confusion probabilities more closely resembles the true confusion matrix.

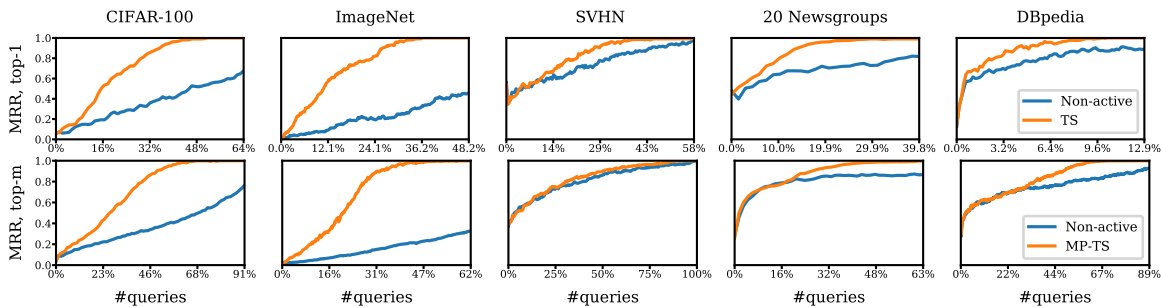


Figure 7: Mean reciprocal rank (MRR) of the m highest ECE predicted classes, comparing active learning (with Thompson sampling (TS)) with no active learning, across five datasets. In the top row $m = 1$, and in the bottom row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

Lastly, we investigate sensitivity of varying the relative cost of mistakes. We consistently observe that active assessment with an informative prior performs best, followed by active assessment with an uninformative prior and finally random sampling. Results are provided in the supplemental material.

4.5 Active Bayesian Assessment of Calibration

To identify classes for which the model is least calibrated (i.e., has lowest classwise ECE), we use the Bayesian model we proposed in Section 3 to actively assess classwise ECE of a model across different predicted classes. In Figure 7 we plot the average mean reciprocal rank (MRR) from 100 independent runs. As with classwise accuracy, active assessment can identify the correct top- m predicted classes much more efficiently than the non-active approach, as a function of the number of label queries. The improvement in efficiency is particularly significant when the classwise calibration performance has large variance across the classes, e.g., CIFAR-100, ImageNet and 20 Newsgroups (additional details in the supplemental material).

5. Related Work

Prior work on using Bayesian ideas in the context of classifier assessment has tended to focus on very specific types assessment. Goutte and Gaussier (2005) propose a framework for Bayesian estimation of precision, recall, and F-score in an information retrieval context, and Johnson et al. (2019) use Bayesian mixture models to provide posterior distributions of diagnostic metrics (such as true positive rates) for medical tests. Benavoli et al. (2017) develop a general Bayesian framework for comparing multiple classifiers as an alternative to more traditional null hypothesis significance testing. We contribute to this body of work in two significant ways. Firstly, we expand the set of Bayesian diagnostics to a broader range of metrics, such as classwise accuracy and calibration metrics such as ECE and reliability diagrams. In addition, we develop approaches for the previously unstudied task of *active assessment* that are more label-efficient (and cost-effective) than traditional approaches which use fixed-size test sets or uniform sampling.

Other work has proposed frequentist methods for uncertainty quantification in an assessment context, e.g., resampling approaches such as the bootstrap for generating confidence intervals on calibration performance (Bröcker and Smith, 2007; Vaicenavicius et al., 2019). Our focus in this work is not to supplant these existing techniques, but instead to supplement them by providing an approach that includes the ability to incorporate of prior knowledge, and which readily lends itself to be used for active assessment.

While there is a large literature on active learning and multi-armed bandits (e.g., Settles (2012); Russo et al. (2018)), our paper is the first that applies these ideas to classifier assessment. In particular, our work builds on multi-armed bandit (MAB) inspired, pool-based active learning algorithms for data selection (Thompson, 1933; Russo, 2016; Komiyama et al., 2015). The techniques we use in this paper can in principle be replaced by any Bayesian active learning algorithms designed for MAB problems—determining the optimal active learning approach for model assessment is an interesting avenue for future research.

6. Conclusions

In this paper we described a Bayesian framework for assessing performance metrics of black-box classifiers, developing inference procedures for classwise accuracy, expected cost, and calibration metrics such as ECE. In addition, we proposed a new framework called *active assessment* for label-efficient assessment of classifier performance, and demonstrated its performance across five well-known datasets for identification of extreme classes such as the least accurate, least calibrated, or highest cost.

There are a number of interesting and useful directions for future work, such as Bayesian estimation of continuous functions related to accuracy and calibration (rather than over regions). The framework can also be extended to assess a particular model operating in multiple environments using a Bayesian hierarchical approach, or to comparatively assess multiple models operating in the same environment. A related direction is to consider environments where humans are in the loop where, given a constraint on the number of problems that can be allocated to humans, the goal is to identify for which types of prediction problems human accuracy will most likely exceed model accuracy.

Appendix A: Classwise ECE and Accuracy are Negatively Correlated

Figure 8 shows scatter plots of classwise accuracy and ECE assessed with our proposed Bayesian method for five datasets used in the paper. The assessment shows that model accuracy and calibration vary substantially across classes. For CIFAR-100, ImageNet and 20 Newsgroups, the variance of classwise accuracy and ECE among all predicted class is considerably greater than the variance of two other datasets. Figure 8 also illustrates that there is significant negative correlation between classwise accuracy and ECE across all 5 datasets, i.e. classes with low classwise accuracy also tend to be less calibrated.

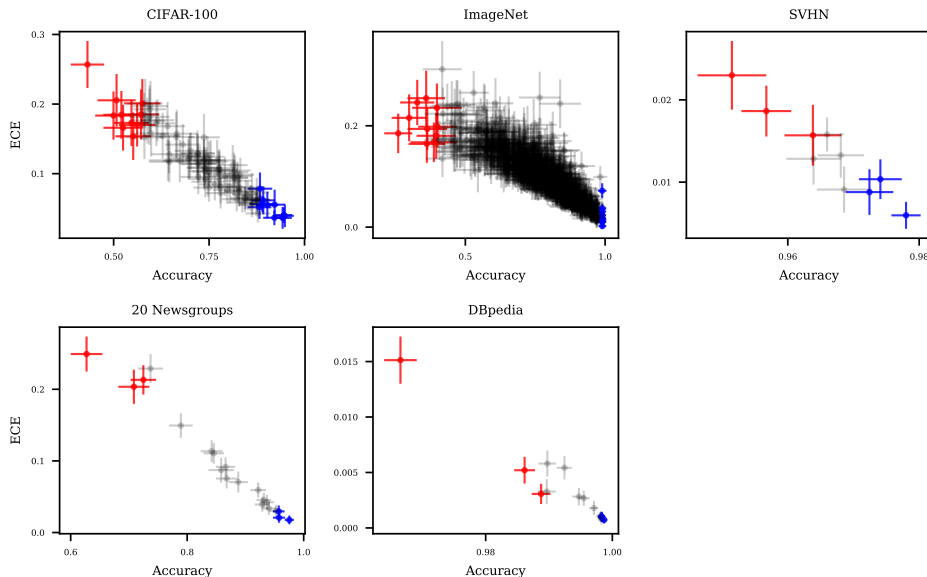


Figure 8: Scatter plots of classwise accuracy and ECE for 5 datasets. Each marker represents posterior means and 95% credible intervals of posterior accuracy and ECE for each predicted class. Markers in red and blue represent the top- m least and most accurate predicted classes, markers in gray represent the other classes, with $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

Appendix B: Bayesian Reliability Diagrams

Figure 9 shows Bayesian reliability diagrams for five datasets, based on different amounts of labeled data. We used a Beta prior for each bin with $\alpha_j = j/10$ and $\beta_j = 1 - j/10, j = 1, \dots, 10$, i.e., a weak prior with pseudocount $\alpha_j + \beta_j = 1$ centered on the diagonal. Rows 1 and 2 display reliability diagrams estimated using $N = 100$ and $N = 1000$ randomly selected examples (respectively). Row 3 displays diagrams estimated using the full set of available labeled examples for each dataset (e.g., the *size* column in Table 1).

With the full set of examples (row 3), the posterior means and the posterior 95% credible intervals are generally below the diagonal, i.e., we can infer with high confidence that the models are miscalibrated (and overconfident, to varying degrees) across all five datasets. For some bins where the scores are less than 0.5, the credible intervals are wide due to little data, and there is not enough information to determine with high confidence if the corresponding models are calibrated or not in these regions. With $N = 100$ examples (row 1), the posterior uncertainty captured by the 95% credible intervals indicates that there is not yet enough information to determine whether the models are miscalibrated given only $N = 100$ labeled examples. With $N = 1000$ examples (row 2) there is enough information to reliably infer that the CIFAR-100 model is overconfident in all bins for scores above 0.3. For the remaining datasets the credible intervals are generally wide enough to encompass 0.5 for most bins, meaning that we do not have enough data to make reliable inferences about calibration, i.e., the possibility that the models are well-calibrated cannot be ruled out without acquiring more data.

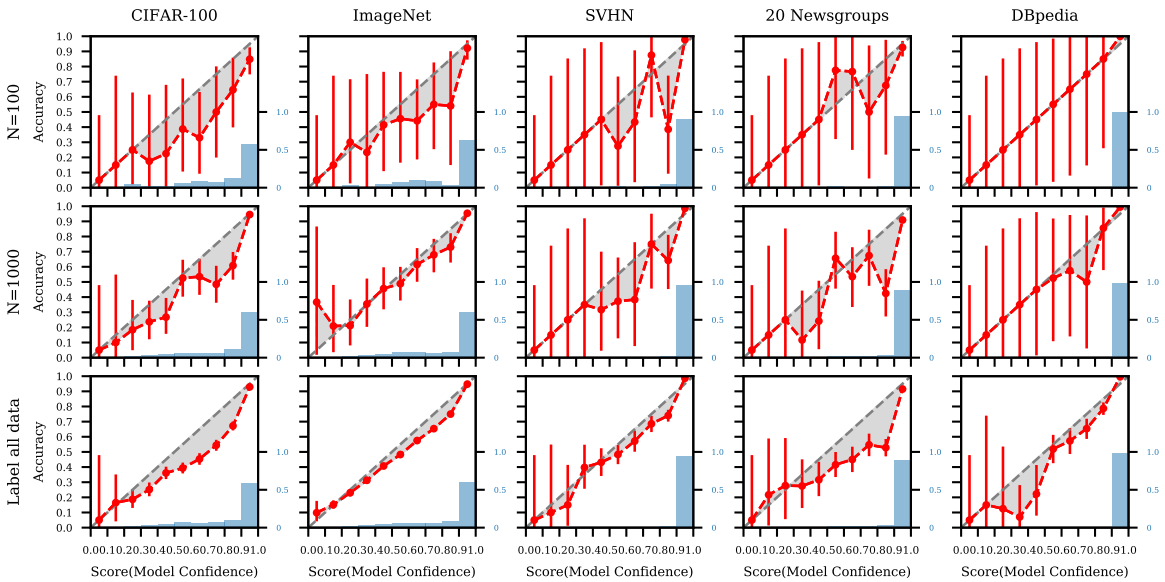


Figure 9: Bayesian reliability diagrams for five datasets (columns) estimated using varying amounts of test data (rows). The red circles plot the posterior mean for θ_j under our Bayesian model. Red bars display 95% credible intervals. Shaded gray areas indicate the estimated magnitudes of the calibration errors, relative to the Bayesian estimates. The blue histogram shows the distribution of the scores for N randomly drawn samples.

Appendix C: Inferring Statistics of Interest via Monte Carlo Sampling

An additional benefit of the Bayesian framework is that we can draw samples from the posterior to infer other statistics of interest. Here we illustrate this method with two examples.

Bayesian Ranking via Monte Carlo Sampling We can infer the Bayesian ranking of classes in terms of classwise accuracy or expected calibration error (ECE), by drawing samples from the posterior distributions. For instance, we can estimate the ranking of classwise accuracy of a model for CIFAR-100, by sampling \hat{A}_k 's (from their respective posterior Beta densities) for each of the classes and then compute the rank of each class with the sampled accuracy. We run this experiment 10,000 times and then for each class we can empirically estimate the distribution of its ranking. The MPE and 95% credible interval of ranking per predicted class for top 10 and bottom 10 are provided in Figure 10a for CIFAR-100.

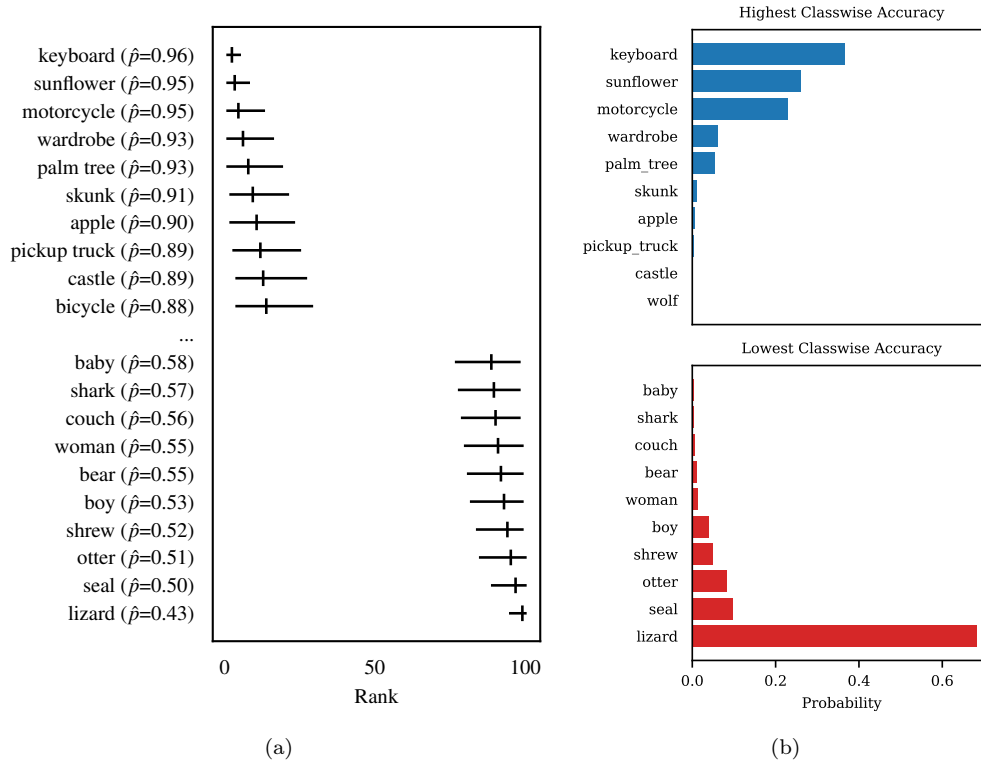


Figure 10: (a) MCMC-based ranking of accuracy across predicted classes for CIFAR-100 (where 1 corresponds to the class with the highest accuracy). (b) Posterior probabilities of the most and least accurate predictions on CIFAR-100. The class with the highest classwise accuracy is somewhat uncertain, while the class with the lowest classwise accuracy is very likely *lizard*.

Posterior probabilities of the most and least accurate predictions We can estimate the probability that a particular class such as *lizard* is the least accurate predicted class of CIFAR-100 by sampling \hat{A}_{k^*} 's (from their respective posterior Beta densities) for each of the classes and then measuring whether \hat{A}_{lizard} is the minimum of the sampled values. Running this experiment 10,000 times and then averaging the results, we determine that there is a 68% chance that lizard is the least

accurate class predicted by ResNet-110 on CIFAR-100. The posterior probabilities for other classes are provided in Figure 10b, along with results for estimating which class has the highest classwise accuracy.

Appendix D: Different Multi-Armed Bandit Algorithms

Below we provide brief descriptions and pseudocode for the different variants of multi-armed bandit algorithm for the best arm or top- m arms identification we investigated in this paper, including Thompson Sampling(TS), Top-Two Thompson Sampling(TTTS), and multiple-play Thompson sampling(MP-TS)(Thompson, 1933; Russo, 2016; Komiyama et al., 2015).

Best Arm Identification Thompson sampling(TS) is a widely used method for online learning of multi-armed problems. The algorithm samples actions according to the posterior probability that they are optimal. Top-two Thompson sampling(TTTS)is a modified version of TS that is tailored for best-arm identification, and has some theoretical advantages. This algorithm adds a re-sampling process to encourage more exploration. Algorithm 1 and 2 describe the sampling process for identifying the most accurate predicted class with TS and TTTS. K is the number of classes and $\text{Beta}(\alpha, \beta)$ is the prior distribution of A_k .

Top- m Arms Identification Multiple-play Thompson sampling(MP-TS)is an extension of TS to the multiple-play multi-armed bandit problem and it has a theoretical optimal regret guarantee with binary rewards. Algorithm 3 is the sampling process to identify the top- m arms with MP-TS, where m is the number of best arms to identify.

Algorithm 1 Thompson Sampling (TS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ to K
repeat
 for $k = 1$ to K **do**
 $\hat{A}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
 end for
 $k^* = \arg \min_k \hat{A}_{1:K}$
 select data point $(x, \hat{y} = k^*)$, query oracle for true label y
 if $y = k^*$ **then**
 $n_{k^*,0} \leftarrow n_{k^*,0} + 1$
 else
 $n_{k^*,1} \leftarrow n_{k^*,1} + 1$
 end if
until all data labeled

Algorithm 2 Top Two Thompson Sampling (TTTS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ to K
repeat
 for $k = 1$ to K **do**
 $\hat{A}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
 end for
 $I = \arg \min_k \hat{A}_{1:K}, B \sim \text{Bernoulli}(\beta)$
 if $B = 1$ **then**
 $k^* = I$
 else
 repeat
 for $k = 1$ to K **do**
 $\hat{A}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
 end for
 $J = \arg \min_k \hat{A}_{1:K}$
 until $J \neq I$
 $k^* = J$
 end if
 select data point $(x, \hat{y} = k^*)$, query oracle for true label y
 if $y = k^*$ **then**
 $n_{k^*,0} \leftarrow n_{k^*,0} + 1$
 else
 $n_{k^*,1} \leftarrow n_{k^*,1} + 1$
 end if
until all data labeled

Algorithm 3 Multiple-play Thompson sampling (MP-TS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ to K
repeat
 for $k = 1$ to K **do**
 $\hat{A}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
 end for
 $I^* = \text{top-}m \text{ arms ranked by } \hat{A}_k$
 for $k^* \in I^*$ **do**
 select data point $(x, \hat{y} = k^*)$, query oracle for true label y
 if $y = k^*$ **then**
 $n_{k^*,0} \leftarrow n_{k^*,0} + 1$
 else
 $n_{k^*,1} \leftarrow n_{k^*,1} + 1$
 end if
 end for
until all data labeled

Appendix E: Derivation of Classwise Expected Cost

Suppose we are given a model M producing probability estimates $p_M(y|\mathbf{x})$, and cost-matrix $\mathbf{C} = \{c_{jk}\}$ where c_{jk} is the cost of predicting class $\hat{y} = k$ for a data point whose true class is $y = j$. Conditioned on a predicted class \hat{y} , the true class label y has a categorical distribution $p(y = j|\hat{y} = k) = \theta_{jk}$. The *classwise expected cost* for predicted class k is given by,

$$C_{\mathcal{R}_k} = \mathbb{E}_{p(\mathbf{x}, y|\mathbf{x} \in \mathcal{R}_k)}[c_{jk}\mathbb{I}(y = j)].$$

We compute:

$$\begin{aligned} C_{\mathcal{R}_k} &= \mathbb{E}_{p(\mathbf{x}, y|\mathbf{x} \in \mathcal{R}_k)}[c_{jk}I(y = j)] \\ &= \sum_{j=1}^K \int_{\mathbf{x} \in \mathcal{R}_k} c_{jk}I(y = j)p(y|\mathbf{x})p(\mathbf{x}|\mathbf{x} \in \mathcal{R}_k) d\mathbf{x} \\ &= \sum_{j=1}^K c_{jk} \int_{\mathbf{x} \in \mathcal{R}_k} I(y = j)p(y|\mathbf{x})p(\mathbf{x}|\mathbf{x} \in \mathcal{R}_k) d\mathbf{x} \\ &= \sum_{j=1}^K c_{jk} \int_{\mathbf{x} \in \mathcal{R}_k} p(y = j|\mathbf{x})p(\mathbf{x}|\mathbf{x} \in \mathcal{R}_k) d\mathbf{x}. \end{aligned}$$

Since $\hat{y} = k, \forall \mathbf{x} \in \mathcal{R}_k$,

$$\begin{aligned} &= \sum_{j=1}^K c_{jk} \int_{\mathbf{x} \in \mathcal{R}_k} p(y = j|\hat{y} = k)p(\mathbf{x}|\mathbf{x} \in \mathcal{R}_k) d\mathbf{x} \\ &= \sum_{j=1}^K c_{jk}p(y = j|\hat{y} = k) \\ &= \sum_{j=1}^K c_{jk}\theta_{jk}. \end{aligned}$$

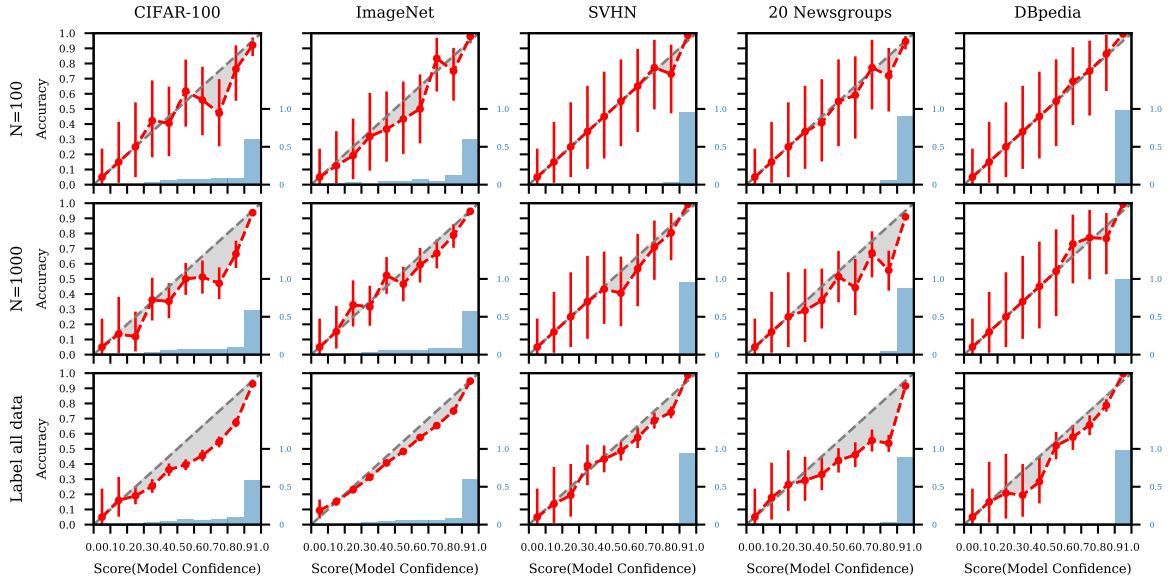
Appendix F: Sensitivity Analysis for Hyperparameters

In Figure 11, we show Bayesian reliability diagrams for five datasets as the strength of the prior increases from 10 to 100. As the strength of the prior increases, it takes more labeled data to overcome the prior belief that the model is calibrated. In Figure 12, we show MRR of the m lowest accurate predicted classes as the strength of the prior increases from 2 to 10 to 100. And in Figure 13, we show MRR of the m least calibrated predicted classes as the strength of the prior increase from 2 to 5 and 10. From these plots, the proposed approach appears to be relatively robust to the prior strength.

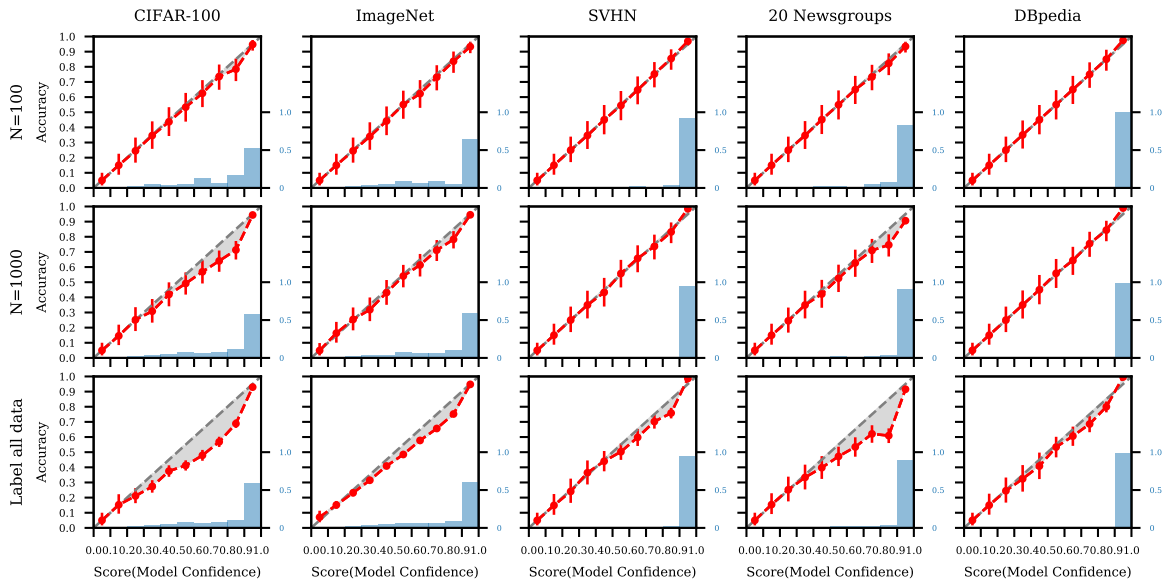
We also investigate the sensitivity of varying the relative cost of mistakes. Results are provided in Table 2. We consistently observe that active assessment with an informative prior performs the best, followed by active assessment with an uninformative prior and finally random sampling.

	Cost	Non-active	Uninformative Prior	Informative Prior
Human	2	7.7K	2.0K	1.6K
	5	8.9K	3.1K	2.1K
	10	8.6K	5.5K	3.4K
	20	7.1K	5.2K	2.5K
Superclass	2	9.2K	2.3K	1.7K
	5	9.9K	2.4K	2.3K
	10	9.7K	2.1K	1.8K
	20	9.6K	2.5K	2.2K

Table 2: Number of queries required by different methods to achieve a 0.95 mean reciprocal rank(MRR) identifying the class with highest classwise expected cost. A pseudocount of 1 is used in the priors for Bayesian models.



(a)



(b)

Figure 11: Bayesian reliability diagrams for five datasets (columns) estimated using varying amounts of test data (rows) with prior strength ($\alpha_j + \beta_j$ for each bin) set to be (a) 10 and (b) 100 respectively. The red circles plot the posterior mean for θ_j under our Bayesian approach. Red bars display 95% credible intervals. Shaded gray areas indicate the estimated magnitudes of the calibration errors, relative to the Bayesian estimates. The blue histogram shows the distribution of the scores for N randomly drawn samples.

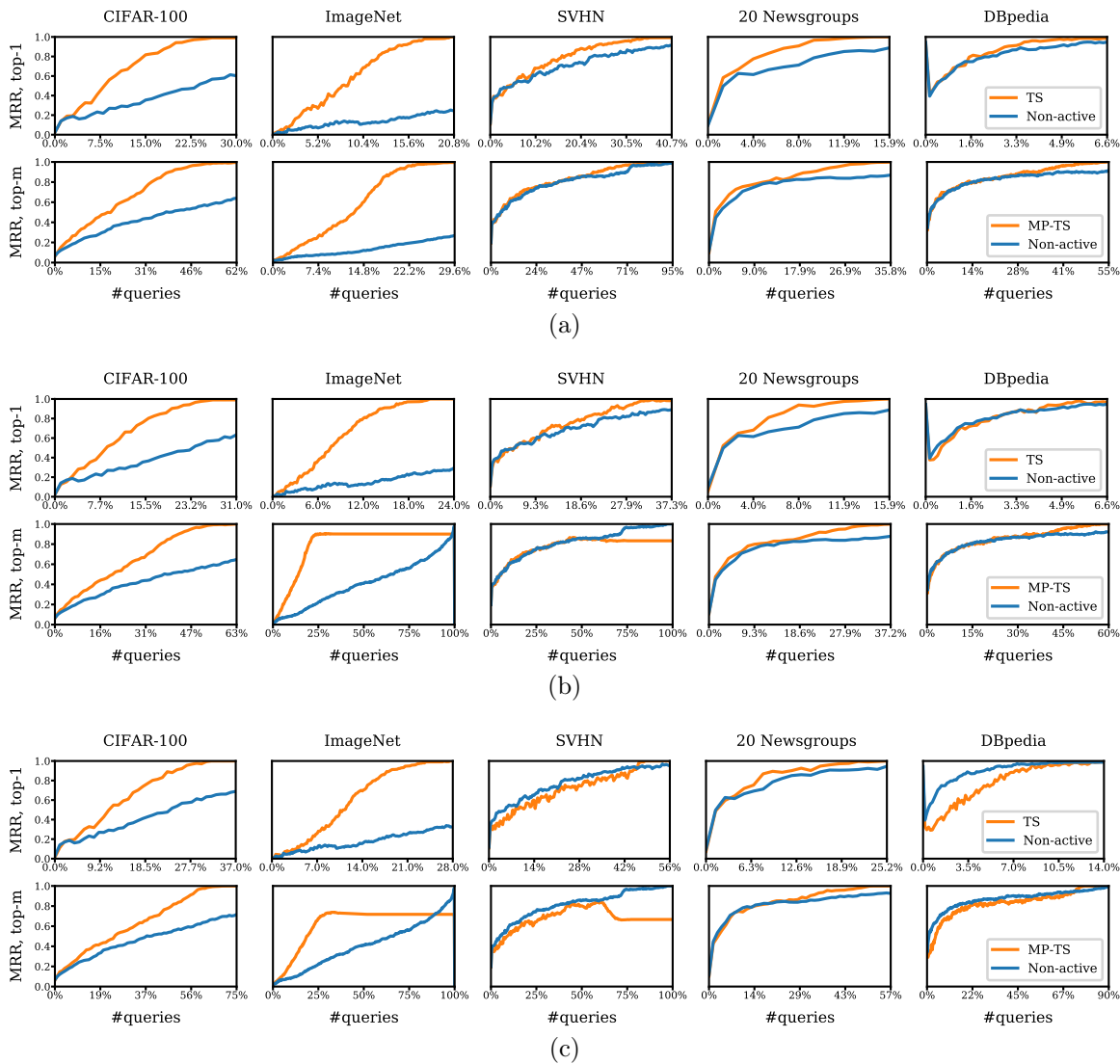


Figure 12: Mean reciprocal rank (MRR) of the m classes with the estimated lowest classwise accuracy as the strength of the prior varies from (a) 2 to (b) 10 and (c) 100, comparing active learning (with Thompson sampling (TS)) with no active learning, across five datasets. For each of (a), (b) and (c), in the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

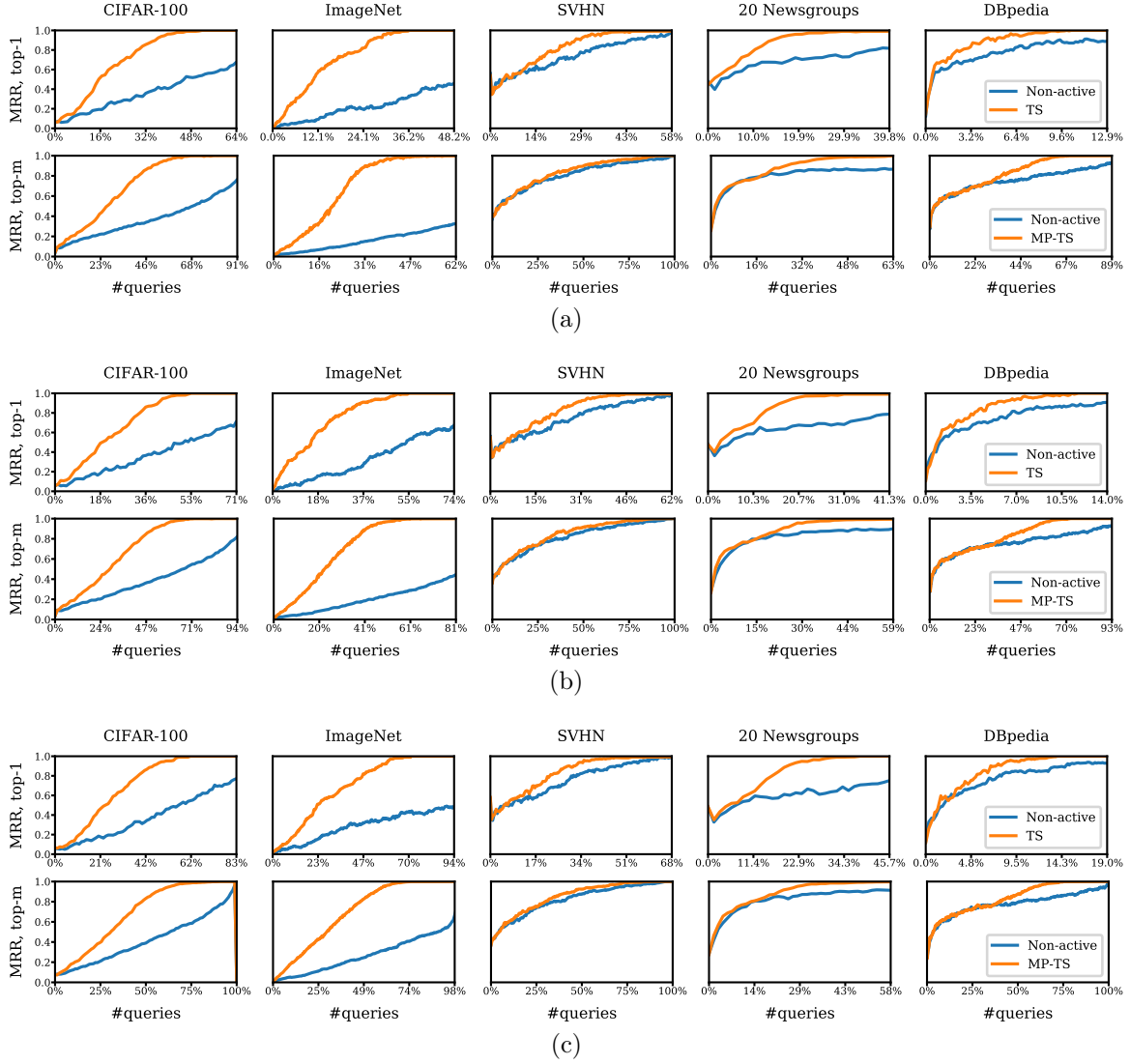


Figure 13: Mean reciprocal rank (MRR) of the m classes with the estimated highest classwise ECE as the strength of the prior varies from (a) 2 to (b) 5 and (c) 10, comparing active learning (with Thompson sampling (TS)) with no active learning, across five datasets. For each of (a), (b) and (c), in the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

References

- Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.
- Jochen Bröcker and Leonard A Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007.
- Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, volume 1, pages 4171–4186, 2019.
- Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry Davis. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *Winter Conference on Applications of Computer Vision*, pages 953–961, 2017.
- Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.
- Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359, 2005.
- 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, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.
- Wesley O Johnson, Geoff Jones, and Ian A Gardner. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Preventive Veterinary Medicine*, 167:113–127, 2019.
- Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.
- Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161, 2015.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631, 2017.
- Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3787–3798, 2019.

- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.
- Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings*, pages 331–339. Elsevier, 1995.
- Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3128–3136, 2018.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pages 625–632, 2005.
- Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- Daniel Russo. Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.
- Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.
- Burr Settles. *Active Learning*. Synthesis Lectures on AI and ML. Morgan Claypool, 2012.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, pages 3459–3467, 2019.
- Yuanshun Yao, Zhujun Xiao, Bolun Wang, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Complexity vs. performance: empirical analysis of machine learning as a service. In *Internet Measurement Conference*, pages 384–397, 2017.
- Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015.