Active Bayesian Assessment of Black-Box Classifiers

Disi Ji[1], Robert L. Logan IV[1], Padhraic Smyth[1], Mark Steyvers[2]

[1]Department of Computer Science, University of California, Irvine
[2]Department of Cognitive Sciences, University of California, Irvine
disij@uci.edu, rlogan@uci.edu, smyth@ics.uci.edu, 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 critical need to both reliably assess the performance of these pre-trained models and to perform this assessment in a label-efficient manner (given that labels may be scarce and costly to collect). In this paper, we introduce an active Bayesian approach for assessment of classifier performance to satisfy the desiderata of both reliability and label-efficiency. We begin by developing inference strategies to quantify uncertainty for common assessment metrics such as accuracy, misclassification cost, and calibration error. We then propose a general framework for active Bayesian assessment using inferred uncertainty to guide efficient selection of instances for labeling, enabling better performance assessment with fewer labels. We demonstrate significant gains from our proposed active Bayesian approach via a series of systematic empirical experiments assessing the performance of modern neural classifiers (e.g., ResNet and BERT) on several standard image and text classification datasets.

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). Many of these models are black boxes from the perspective of downstream users, for example, models developed remotely by commercial entities and hosted as a service in the cloud (Yao et al. 2017; Sanyal et al. 2018).

For a variety of reasons (legal, economic, competitive), users of machine learning models increasingly may have no direct access to the detailed workings of the model, how the model was trained, or the training data. In this context, careful attention needs to be paid to accurate, detailed and robust assessments of the quality of a model's predictions, such that the model can be held accountable by users. This is particularly true in the common scenario where the model is being deployed in an environment that does not necessarily distributionally match the data that the model was trained.

In real-world application scenarios, labeled data for assessment is likely to be scarce and costly to collect, e.g., for a

model being deployed in a diagnostic imaging context in a particular hospital where labeling requires expensive human expertise. Thus, it is important to be able to accurately assess the performance black-box classifiers in environments where there is limited availability of labeled data. With this in mind we develop a framework for **active Bayesian assessment** of black-box classifiers, using techniques from active learning to efficiently select instances to label so that uncertainty of assessment can be reduced, and deficiencies of models such as low accuracy, high calibration error or high cost mistakes can be quickly identified.

The primary contributions of our paper are:

- We propose a general Bayesian framework to assess black-box classifiers, providing uncertainty quantification for quantities such as classwise accuracy, expected calibration error (ECE), confusion matrices, and performance comparisons across groups.
- We develop a methodology for active Bayesian assessment for an array of fundamental tasks including (1) estimation of model performance; (2) identification of model deficiencies; (3) performance comparison between groups.
- We demonstrate that our proposed approaches need significantly fewer labels than baselines, via a series of experiments assessing the performance of modern neural classifiers (e.g., ResNet and BERT) on several standard image and text classification datasets.

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 are interested in assessing the performance of a pretrained prediction model M that makes predictions of y given a feature vector $\mathbf{x}$, where M produces K numerical scores per class in the form of a set of estimates of class-conditional probabilities $p_M(y = k|\mathbf{x}), k = 1, \dots, K$. $\hat{y} = \arg\max_k p_M(y = k|\mathbf{x})$ is the classifier's label prediction for a particular input $\mathbf{x}$. $s(\mathbf{x}) = p_M(y = \hat{y}|\mathbf{x})$ is 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 also referred to as a model's **confidence** in its prediction and can be viewed as a model's own estimate of its accuracy. The model's scores in general need not be perfectly calibrated, i.e., they need not match the true probabilities $p(y = \hat{y}|\mathbf{x})$.

We focus in this paper on assessing the performance of a model that 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. 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 characterized by some underlying unknown distribution $p(\mathbf{x}, y)$.

Performance Assessment

Performance Metrics and Tasks: We will use θ to indicate a **performance metric** of interest, such as classification accuracy, true positive rate, expected cost, calibration error, etc. Our approach to assessment of a metric θ relies on the notion of disjoint **groups** (or partitions) $g = 1, \ldots, G$ of the input space $\mathbf{x} \in \mathcal{R}_g$, e.g., grouping by predicted class $\hat{y}$. For any particular instantiation of groups g and metric θ, there are three particular **assessment tasks** we will focus on in this paper: (1) estimation, (2) identification, and (3) comparison.

Estimation: Let $\theta_1, \ldots, \theta_G$ be the set of true (unknown) values for some metric θ for some grouping g. The goal of estimation is to assess the quality of a set of estimates $\hat{\theta}_1, \ldots, \hat{\theta}_G$ relative to their true values. In this paper we will focus on RMSE loss $\left(\sum_g p_g (\theta_g - \hat{\theta}_g)^2 \right)^{1/2}$ to measure estimation quality, where $p_g = p(\mathbf{x} \in \mathcal{R}_g)$ is the marginal probability of a data point being in group g (e.g., as estimated from unlabeled data) and $\hat{\theta}_g$ is a point estimate of the true θ_g, e.g., a maximum a posteriori (MAP) estimate.

Identification: Here the goal is to identify extreme groups, e.g., $g^* = \arg\min_g \theta_g$, such as the predicted class with the lowest accuracy (or the highest cost, swapping max for min). In general we will investigate methods for finding the m groups with highest or lowest values of a metric θ. To compare the set of identified groups to the true set of m-best/worst groups, we can use (for example) ranking measures to evaluate and compare the quality of different identification methods.

Comparison: The goal here is to determine if the difference between two groups g_1 and g_2 is statistically significant, e.g., to assess if accuracy or calibration for one group is significantly better than another group for some black-box classifier. A measure of the quality of a particular assessment method in this context is to compare how often, across multiple datasets of fixed size, a method correctly identifies if a significant difference exists and, if so, its direction.

Groups: There are multiple definitions of groups that are of interest in practice. One grouping of particular interest is where groups correspond to a model's predicted classes, i.e., $g = k$, and the partition of the input space corresponds to the model's decision regions $\mathbf{x} \in \mathcal{R}_k$, i.e., $\hat{y}(\mathbf{x}) = k$. If θ refers to classification accuracy, then θ_k is the accuracy per predicted class. For prediction problems with costs, θ_k can be the expected cost per predicted class, and so on.

Another grouping of interest for classification models is the set of groups g that correspond to bins b of a model's score, i.e., $s(\mathbf{x}) \in \text{bin}_b, b = 1 \ldots, B$, or equivalently $\mathbf{x} \in \mathcal{R}_b$ where $\mathcal{R}_b$ is the region of the input space where model scores lie in score-bin b. The score-bins can be defined in any standard way, e.g., equal width $1/B$ or equal weight $p(s(\mathbf{x}) \in \text{bin}_b) = 1/B$. θ_b can be defined as the accuracy per score-bin, which in turn can be related to the well-known expected calibration error (ECE, e.g., Guo et al. (2017)) as we will discuss in more detail later in the paper[1].

In an algorithmic fairness context, for group fairness (Hardt, Price, and Srebro 2016), the groups g can correspond to categorical values of a protected attribute such as gender or race, and θ can be defined (for example) as accuracy or true positive rate per group.

In the remainder of the paper, we focus on developing and evaluating the effectiveness of different methods for assessing groupwise metrics θ_g. In the two sections below, we first describe a flexible Bayesian strategy for assessing performance metrics θ in the context of the discussion above, then outline a general **active assessment framework** that uses the Bayesian strategy to address the three assessment tasks (estimation, identification, and comparison) in a label-efficient manner.

Bayesian Assessment

We outline below a Bayesian approach to make posterior inferences about performance metrics given labeled data, where the posteriors on θ can be used to support the assessment tasks of estimation, identification, and comparison. For simplicity we begin with the case where θ corresponds to accuracy and then extend to other metrics such as ECE.

The accuracy for a group g can be treated as an unknown Bernoulli parameter θ_g. Labeled observations $(\mathbf{x}_i, y_i), i = 1, \ldots, N_g$ are sampled randomly per group conditioned on $\mathbf{x}_i \in \mathcal{R}_g$, leading to a binomial likelihood with binary accuracy outcomes $\mathbb{1}(y_i, \hat{y}_i) \in \{0, 1\}$. The standard frequency-based estimate is $\hat{\theta}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{1}(y_i, \hat{y}_i)$.

It is natural to consider Bayesian inference in this context, especially in situations where there is relatively little labeled data available per group. With a conjugate prior $\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$ and a binomial likelihood on binary outcomes $\mathbb{1}(y_i, \hat{y}_i)$, we can update the posterior distribution of θ_g in closed-form to $\text{Beta}(\alpha_g + r_g, \beta_g + N_g - r_g)$ where r_g is the number of correct label predictions $\hat{y} = y$ by the model given N_g trials for group g.

For metrics other than accuracy, we sketch the basic idea here for Bayesian inference for ECE and provide additional discussion in the Supplement.[2] ECE is defined as $\sum_{b=1}^{B} p_b |\theta_b - s_b|$ where B is the number of bins (corresponding to groups g), p_b is the probability of each bin b, and θ_b and s_b are the accuracy and average confidence per bin respectively. We can put Beta priors on accuracies θ_b, model the likelihood of outcomes for each bin b as binomial, resulting again in closed form Beta posteriors for accuracy per bin

---

[1] We use ECE for illustration in our results since it is widely used in the recent classifier calibration literature, but other calibration metrics could also be used, e.g., see Kumar, Liang, and Ma (2019).

[2] Link to the Supplement: https://arxiv.org/abs/2002.06532

keyboard
motorcycle
sunflower
wardrobe
palm_tree
skunk
pickup_truck
apple
castle
bicycle

lizard
seal
otter
shrew
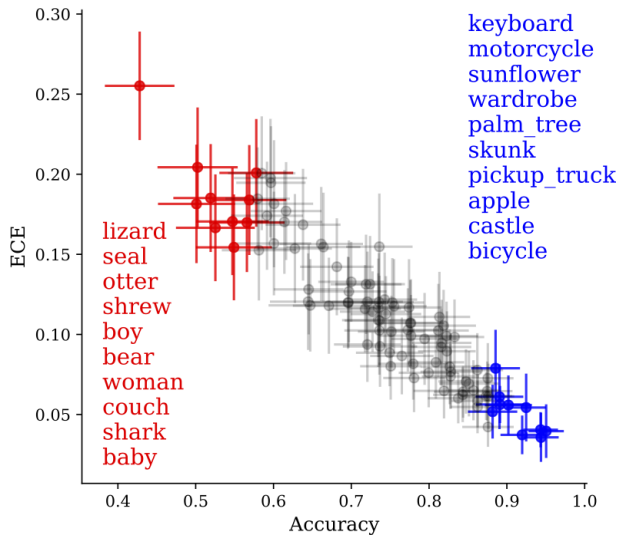boy
bear
woman
couch
shark
baby

Figure 1: Scatter plot of estimated accuracy and expected calibration error (ECE) per class of a ResNet-110 image classifier on the CIFAR-100 test set, using our Bayesian assessment framework, with posterior means and 95% credible intervals per class. Red and blue for the top-10 least and most accurate classes, gray for the other classes.

b. The posterior density for the marginal ECE itself is not available in closed form, but can easily be estimated by direct Monte Carlo simulation from the B posteriors for the B bins. We can also be Bayesian about ECE *per group*, ECE_g (e.g., per class, with $g = k$), in a similar manner by defining two levels of grouping, one at the class level and one at the bin level.

Illustrative Example: To illustrate the general idea of Bayesian assessment, we train a standard ResNet-110 classifier on the CIFAR-100 data set and perform Bayesian inference of the accuracy and ECE of this model using the 10,000 labeled examples in the test set. The groups $g = k$ here correspond to the K predicted classes by the model, $\hat{y} = k \in \{1, \ldots, K\}$. We use Beta priors with $\alpha_k = \beta_k = 1, k = 1, \ldots, K$ for classwise accuracy, and $\alpha_b = 2s_b, \beta_b = 2(1 - s_b), b = 1, \ldots, K$ for binwise accuracy. The prior distributions reflect no particular prior belief about classwise accuracy and a weak prior belief that the confidence of the classifier is calibrated across all predicted classes.

Figure 1 shows the resulting mean posterior estimates (MPEs) and 95% credible intervals (CIs) for accuracy and ECE values for each of the $K = 100$ classes. The accuracies and ECE values of the model vary substantially across classes, and classes with low accuracy tend to be less calibrated than those with higher accuracy. There is also considerable posterior uncertainty for these metrics even using the whole test set of CIFAR-100. For example, while there is confidence that the least accurate class is "lizard" (top left point), there is much less certainty about what class is the most accurate class (bottom right).

Algorithm 1 Thompson Sampling(p, q, r, M)

1: Initialize the priors on metrics $\{p_0(\theta_1), \ldots, p_0(\theta_g)\}$
2: **for** $i = 1, 2, \cdots$ **do**
3:     # Sample parameters for the metrics θ
4:     $\widetilde{\theta}_g \sim p_{i-1}(\theta_g), g = 1, \ldots, G$
5:     # Select a group g (or arm) by maximizing expected reward
6:     $\hat{g} \leftarrow \arg\max_g \mathbb{E}_{q_{\hat{g}}}[r(z|g)]$
7:     # Randomly select an input data point from group $\hat{g}$ group and compute its predicted label
8:     $\mathbf{x}_i \sim \mathcal{R}_{\hat{g}}$
9:     $\hat{y}_i(\mathbf{x}_i) = \arg\max_k p_M(y = k|\mathbf{x}_i)$
10:     # Query to get a true label (pull arm $\hat{g}$)
11:     $z_i \leftarrow f(y_i, \hat{y}_i(\mathbf{x}_i))$
12:     # Update parameters of the $\hat{g}$th metric
13:     $p_i(\theta_{\hat{g}}) \propto p_{i-1}(\theta_{\hat{g}})q(z_i|\theta_{\hat{g}})$
14: **end for**

Figure 2: An outline of the algorithm for active Bayesian assessment using multi-arm bandit Thompson sampling with arms corresponding to groups g.

It is straightforward to apply this type of Bayesian inference to other metrics and to other assessment tasks, such as estimating a model's confusion matrix, ranking performance by groups with uncertainty (Marshall and Spiegelhalter 1998), analyzing significance of differences in performance across groups, and so on. For the CIFAR-100 dataset, based on the test data we can, for example, say that with 96% probability the ResNet-110 model is less accurate when predicting the superclass "human" than it is when predicting "trees"; and that with 82% probability, the accuracy of the model when it predicts "woman" will be lower than the accuracy when it predicts "man." Additional details and examples can be found in the Supplement.

Active Bayesian Assessment

As described earlier, in practice we may wish to assess how well a blackbox classifier performs in a new environment where we have relatively little labeled data. For example, our classifier may have been trained in the cloud by a commercial vendor and we wish to independently evaluate its accuracy (and other metrics) in a particular context. Rather than relying on the availability of a large random sample of labeled instances for inference, as we did in the results in Figure 1, we can improve data efficiency by leveraging the Bayesian approach to support *active assessment*, actively selecting examples $\mathbf{x}$ for labeling in a data-efficient manner. Below we develop active assessment approaches for the three tasks of estimation, identification, and comparison. Efficient active selection of examples for labeling is particularly relevant when we have a potentially large pool of unlabeled examples $\mathbf{x}$ available, and have limited resources for labeling (e.g., a human labeler with limited time).

The Bayesian framework described in the last section readily lends itself to be used in Bayesian active learning algo-

	Assessment Task	Prior: $p(\theta)$	Likelihood: $q_\theta(z\|g)$	Reward: $r(z\|g)$
Estimation	Groupwise Accuracy	$\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$	$z \sim \text{Bern}(\theta_g)$	$p_g \cdot (\text{Var}(\hat{\theta}_g\|\mathcal{L}) - \text{Var}(\hat{\theta}_g\|\{\mathcal{L}, z\}))$
	Confusion Matrix$(g = k)$	$\theta_{\cdot k} \sim \text{Dirichlet}(\alpha_{\cdot k})$	$z \sim \text{Multi}(\theta_k)$	$p_k \cdot (\text{Var}(\hat{\theta}_k\|\mathcal{L}) - \text{Var}(\hat{\theta}_k\|\{\mathcal{L}, z\}))$
Identification	Least Accurate Group	$\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$	$z \sim \text{Bern}(\theta_g)$	$-\widetilde{\theta}_g$
	Least Calibrated Group	$\theta_{gb} \sim \text{Beta}(\alpha_{gb}, \beta_{gb})$	$z \sim \text{Bern}(\theta_{gb})$	$\sum_{b=1}^{B} p_{gb} \left\|\widetilde{\theta}_{gb} - s_{gb}\right\|$
	Most Costly Class$(g = k)$	$\theta_{\cdot k} \sim \text{Dirichlet}(\alpha_{\cdot k})$	$z \sim \text{Multi}(\theta_k)$	$\sum_{j=1}^{K} c_{jk} \widetilde{\theta}_{jk}$
Comparison	Accuracy Comparison	$\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$	$z \sim \text{Bern}(\theta_g)$	$\lambda\|\{\mathcal{L}, (g, z)\}$

Table 1: Thompson sampling configurations for different assessment tasks.

rithms, by considering model assessment as a multi-armed bandit problem where each group g corresponds to an arm or a bandit. In Bayesian assessment in this context, there are two key building blocks: (i) the assessment algorithm's current beliefs (prior or posterior distribution) for the metric of interest $\theta_g \sim p(\theta_g)$, and (ii) a generative model (likelihood) of the labeling *outcome* $z \sim q_\theta(z|g), \forall g$.

Instead of labeling randomly sampled data points from a pool of unlabeled data, we propose instead to actively select data points to be labeled by iterating between: (1) *labeling*: actively selecting a group $\hat{g}$ based on the assessment algorithm's current beliefs about θ_g, randomly selecting a data point $\mathbf{x}_i \sim \mathcal{R}_{\hat{g}}$ and then querying its label; and (2) *assessment*: updating our beliefs about the performance metric θ_g given the outcome z_i. This active selection approach requires defining a **reward function** $r(z|g)$ for the revealed outcome z for the g-th group. For example, if the assessment task is to generate low variance estimates of groupwise accuracy, $r(z|g)$ can be formulated as the reduction in uncertainty about θ_g, given an outcome z, to guide the labeling process.

Our primary goal in this paper is to demonstrate the general utility of active assessment for performance assessment of black-box classifiers rather than comparing different active selection methods. With this in mind, we focus in particular on the framework of Thompson sampling (Thompson 1933; Russo et al. 2018) since we found it to be more reliably efficient compared to other active selection methods such as epsilon-greedy and upper-confidence bound (UCB) approaches (additional discussion is provided in the Supplement).

Algorithm 1 describes a general active assessment algorithm based on Thompson sampling. At each step i, a set of values for metrics $\theta_g, 1 \ldots, G$, are sampled from the algorithm's current beliefs, i.e., $\widetilde{\theta}_g \sim p_{i-1}(\theta_g)$ (line 4). As an example, when assessing groupwise accuracy, $p_{i-1}(\theta_g)$ represents the algorithm's belief (e.g., in the form of a posterior Beta distribution) about the accuracy for group g given $i-1$ labeled examples observed so far. The sampling step is a key difference between Thompson sampling and alternatives that use a point estimate to represent current beliefs (such as greedy approaches).

Conditioned on the sampled $\widetilde{\theta}_g$ values, the algorithm then selects the group $\hat{g}$ that maximizes the expected reward $\hat{g} = \arg\max_g \mathbb{E}_{q_{\widetilde{\theta}_g}}[r(z|g)]$ (line 6) where $r(z|g)$ is task-specific and where $z \sim q_{\widetilde{\theta}_{\hat{g}}}(z|\hat{g})$ is the likelihood for outcome z.

The algorithm then draws an input data point $\mathbf{x}_i$ randomly from $\mathcal{R}_{\hat{g}}$, and uses the model M to generate a predicted label $\hat{y}_i$. The Oracle is then queried (equivalent to "pulling arm $\hat{g}$" in a bandit setting) to obtain a label outcome z_i and the algorithm's belief is updated (line 13) to update the posterior for $\theta_{\hat{g}}$.

Note that this proposed algorithm implicitly assumes that the θ_g's are independent (by modeling beliefs about θ_g's independently rather than jointly). In some situations there may be additional information across groups g (e.g., hierarchical structure) that could be leveraged (e.g., via contextual bandits) to improve inference but we leave this for future work.

We next discuss how specific reward functions r can be designed for different assessment tasks of interest, with a summary provided in Table 1.

Estimation: The MSE for estimation accuracy for G groups can be written in bias-variance form as $\sum_{g=1}^{G} p_g\big(\text{Bias}^2(\hat{\theta}_g) + \text{Var}(\hat{\theta}_g)\big)$. Given a fixed labeling budget the bias term can be assumed to be small relative to the variance (e.g., see Sawade et al. (2010)), by using relatively weak priors for example. It is straightforward to show that to minimize $\sum_{g=1}^{G} p_g \text{Var}(\hat{\theta}_g)$ the optimal number of labels per group g is proportional to $\sqrt{p_g \theta_g(1 - \theta_g)}$, i.e., sample more points from larger groups and from groups where θ_g is furthest from 0 or 1. While the group sizes p_g can be easily estimated from unlabeled data, the θ_g's are unknown, so we can't compute the optimal sampling weights a priori. Active assessment in this context allows one to minimize MSE (or RMSE) in an adaptive sequential manner. In particular we can do this by defining a reward function $r(z|g) = p_g \cdot (\text{Var}(\hat{\theta}_g|\mathcal{L}) - \text{Var}(\hat{\theta}_g|\{\mathcal{L}, z\}))$, where $\mathcal{L}$ is the set of labeled data seen to date, with the goal of selecting examples for labeling that minimize the overall posterior variance at each step. For confusion matrices, a similar argument applies but with multinomial likelihoods and Dirichlet posteriors on vector-valued θ_g's per group (see Table 1). Although we did not develop methods to directly estimate ECE in an active manner, we can nonetheless assess it by actively estimating bin-wise accuracy using the framework above.

Identification: To identify the best (or worst performing) group, $\hat{g} = \arg\max_g \theta_g$, we can define a reward function

using the sampled metrics $\widetilde{\theta}_g$ for each group. For example, to identify the least accurate class, the expected reward of the g-th group is $\mathbb{E}_{q_{\widehat{\theta}}}[r(z_i)|g] = q_{\widehat{\theta}}(y=1)(-\widetilde{\theta}_g) + q_{\widehat{\theta}}(y=0)(-\widetilde{\theta}_g) = -\widetilde{\theta}_g$. Similarly, because the reward functions of other identification tasks (Table 1) are independent of the value of y, when the assessment tasks are to identify the group with the highest ECE or misclassification cost, maximization of the reward function corresponds to selecting the group with the greatest sampled ECE or misclassification cost.

In addition we experimented with a modified version of Thompson sampling (TS) that is designed for best-arm identification, called top-two Thompson sampling (TTTS) (Russo 2016), but found that TTTS and TS gave very similar results— so we just focus on TS in the results presented in this paper.

To extend this approach to identification of the *best-m arms*, instead of selecting the arm with the greatest expected reward, we pull the top-m-ranked arms at each step, i.e. we query the true labels of m samples, one sample $\mathbf{x}$ randomly drawn from each of the top m ranked groups. This best-m approach can be seen as an application of the general best-m arms identification method proposed by Komiyama, Honda, and Nakagawa (2015) for the problem of extreme arm identification. They proposed this multiple-play Thompson sampling (MP-TS) algorithm as a multiple-play multi-armed bandit problem, and proved that MP-TS has the optimal regret upper bound when the reward is binary.

Comparison: For the task of comparing differences in a performance metric θ between two groups, an active assessment algorithm can learn about the accuracy of each group by sequentially allocating the labeling budget between them. Consider two groups g_1 and g_2 with a true accuracy difference $\Delta = \theta_{g_1} - \theta_{g_2}$. Our approach uses the "rope" (region of practical equivalence) method of Bayesian hypothesis testing (e.g., Benavoli et al. (2017)) as follows. The cumulative density in each of three regions $\mu = (P(\Delta < -\epsilon), P(-\epsilon \leq \Delta \leq \epsilon), P(\Delta > \epsilon))$ represents the posterior probability that the accuracy of group g_1 is more than ϵ lower than the accuracy of g_2, that the two accuracies are "practically equivalent," or that g_1's accuracy is more than ϵ higher than that of g_2, where ϵ is user-specified. In our experiments we use $\epsilon = 0.05$ and the cumulative densities μ are estimated with 10,000 Monte Carlo samples.

The assessment task is to identify the region $\eta = \arg\max(\mu)$ in which Δ has the highest cumulative density, where $\lambda = \max(\mu) \in [0,1]$ represents the confidence of the assessment. Using Thompson sampling to actively select labels from g_1 and g_2, at the i-th step, when we get a z_i for a data point from the g-th group we update the Beta posterior of θ_g. The resulting decrease in uncertainty about θ_g depends on the realization of the binary variable z_i and the current distribution of θ_g. We use λ to measure the amount of evidence we gathered from the labeled data from both of the groups. Then we can select the group in a greedy manner that has the greater expected increase $\mathbb{E}_{q_{\widehat{\theta}}}[\lambda|\{\mathcal{L}, (g, z)\} - \mathbb{E}_{q_{\widehat{\theta}}}[\lambda|\mathcal{L}]$, which is equivalent to selecting the arm with the largest $\mathbb{E}_{q_{\widehat{\theta}}}[\lambda|\{\mathcal{L}, (g, z)\}]$. This approach of *maximal expected model change strategy* has also been used in prior work in active

	Mode	Test Set Size	Number of Classes	Prediction Model M
CIFAR-100	Image	10K	100	ResNet-110
ImageNet	Image	50K	1000	ResNet-152
SVHN	Image	26K	10	ResNet-152
20 News	Text	7.5K	20	BERT$_{BASE}$
DBpedia	Text	70K	14	BERT$_{BASE}$

Table 2: Datasets and models used in experiments.

learning for other applications (Freytag, Rodner, and Denzler 2014; Vezhnevets, Buhmann, and Ferrari 2012).

Experimental Settings

Datasets and Prediction Models: In our experiments we use a number of well-known image and text classification datasets, for both image classification (*CIFAR-100* (Krizhevsky and Hinton 2009), *SVHN* (Netzer et al. 2011) and *ImageNet* (Russakovsky et al. 2015)) and text classification (*20 Newsgroups* (Lang 1995) and *DBpedia* (Zhang, Zhao, and LeCun 2015)).

For models M we use ResNet (He et al. 2016) to perform image classification and BERT (Devlin et al. 2019) to perform text classification. Each model is trained on standard training sets used in the literature and assessment is performed on random samples from the corresponding test sets. Table 2 provides a summary of datasets, models, and test sizes. In this paper we focus on assessment of deep neural network models in particular since they are of significant current interest in machine learning—however, our approach is broadly applicable to classification models in general.

Unlabeled data points $\mathbf{x}_i$ from the test set are assigned to groups (such as predicted classes or score-bins) by each prediction model. Values for p_g (for use in active learning in reward functions and in evaluation of assessment methods) are estimated using the model-based assignments of test data-points to groups. Ground truth values for θ_g are defined using the full labeled test set for each dataset.

Priors: We investigate both uninformative and informative priors to specify prior distributions over groupwise metrics. All of the priors we use are relatively weak in terms of prior strength, but as we will see in the next section, the informative priors can be very effective when there is little labeled data available. We set the prior strengths as $\alpha_g + \beta_g = N_0 = 2$ for Beta priors and $\sum \alpha_g = N_0 = 1$ for Dirichlet priors in all experiments, demonstrating the robustness of the settings across a wide variety of contexts. For groupwise accuracy, the informative Beta prior for each group is Beta$(N_0 s_g, N_0(1 - s_g))$, where s_g is the average model confidence (score) of all unlabeled test data points for group g. The uninformative prior distribution is $\alpha = \beta = N_0/2$.

For confusion matrices, there are $\mathcal{O}(K^2)$ prior parameters in total for K Dirichlet distributions, each distribution parameterized by a K-dimensional vector α_j. As an informative prior for a confusion matrix we use the model's own prediction scores on the unlabeled test data, $\alpha_{jk} \propto \Sigma_{\mathbf{x} \in \mathcal{R}_k} p_M(y =$

	N/K	N	UPrior (baseline)	IPrior (ours)	IPrior+TS (ours)
CIFAR-100	2	200	30.7	**15.0**	15.3
	5	500	20.5	**13.6**	13.8
	10	1K	13.3	**10.9**	11.4
ImageNet	2	2K	29.4	13.2	**13.2**
	5	5K	18.8	12.1	**11.6**
	10	10K	11.8	9.5	**9.4**
SVHN	2	20	13.7	5.1	**3.4**
	5	50	7.7	5.1	**3.4**
	10	100	5.4	4.7	**3.1**
20 News	2	40	23.9	12.3	**11.7**
	5	100	15.3	10.8	**10.3**
	10	200	10.4	**8.7**	8.8
DBpedia	2	28	14.9	2.0	**1.5**
	5	70	3.5	2.3	**1.2**
	10	140	2.6	2.1	**1.1**

Table 3: RMSE of classwise accuracy across 5 datasets. Each RMSE number is the mean across 1000 independent runs.

	N/K	N	UPrior (baseline)	IPrior (ours)	IPrior+TS (ours)
CIFAR-100	2	200	1.463	0.077	**0.025**
	5	500	0.071	0.012	**0.004**
	10	1K	0.001	0.002	**0.001**
SVHN	2	20	92.823	0.100	**0.045**
	5	50	11.752	0.022	**0.010**
	10	100	0.946	0.005	**0.002**
20 News	2	40	3.405	0.018	**0.005**
	5	100	0.188	0.004	**0.001**
	10	200	0.011	0.001	**0.000**
DBpedia	2	28	1307.572	0.144	**0.025**
	5	70	33.617	0.019	**0.003**
	10	140	**0.000**	0.004	0.001

Table 4: Scaled mean RMSE for confusion matrix estimation. Same setup as Table 3.

$j|\mathbf{x}$). The uninformative prior for a confusion matrix is set as $\alpha_{jk} = N_0/K, \forall j, k$. In the experiments in the next section we show that even though our models are not well-calibrated (as is well-known for deep models, e.g., Guo et al. (2017); see also Figure 1), the model's own estimates of class-conditional probabilities nonetheless contain valuable information about confusion probabilities.

Experimental Results

We conduct a series of experiments across datasets, models, metrics, and assessment tasks, to systematically compare three different assessment methods: (1) non-active sampling with uninformative priors (UPrior), (2) non-active sampling with informative priors (IPrior), and (3) active Thompson sampling (Figure 2) with informative priors (IPrior+TS). Estimates of metrics (as used for example in computing RMSE

or ECE) correspond to mean posterior estimates $\hat{\theta}$ for each method. Note that the UPrior method is equivalent to standard frequentist estimation with random sampling and weak additive smoothing. We use UPrior instead of a pure frequentist method to avoid numerical issues in very low data regimes.

Best-performing values that are statistically significant, across the three methods, are indicated in bold in our tables. Statistical significance between the best value and next best is determined by a Wilcoxon signed-rank test with p=0.05. Results are statistically significant in all rows in all tables, except for SVHN results in Table 7. Code and scripts for all of our experiments are available at https://github.com/disiji/active-assess.

Our primary goal is to evaluate the effectiveness of active versus non-active assessment, with a secondary goal of evaluating the effect of informative versus non-informative priors. As we will show below, our results clearly demonstrate that the Bayesian and active assessment frameworks are significantly more label-efficient and accurate, compared to the non-Bayesian non-active alternatives, across a wide array of assessment tasks.

Estimation of Accuracy, Calibration, and Confusion Matrices: We compare the estimation efficacy of each evaluated method as the labeling budget N increases, for classwise accuracy (Table 3), confusion matrices (Table 4), and ECE (Table 5). All reported numbers are obtained by averaging across 1000 independent runs, where a run corresponds to a sequence of sampled $\mathbf{x}_i$ values (and sampled θ_g values for the TS method).

Table 3 shows the mean RMSE of classwise accuracy for the 3 methods on the 5 datasets. The results demonstrate that informative priors and active sampling have significantly lower RMSE than the baseline, e.g., reducing RMSE by a factor of 2 or more in the low-data regime of $N/K = 2$. Active sampling (IPrior+TS) improves on the IPrior method in 11 of the 15 results, but the gains are typically small. For other metrics and tasks below we see much greater gains from using active sampling.

Table 4 reports the mean RMSE across runs of estimates of confusion matrix entries for four datasets[3]. RMSE is defined here as $\text{RMSE} = \left(\sum_k p_k \sum_j (\theta_{jk} - \hat{\theta}_{jk})^2 \right)^{1/2}$ where θ_{jk} is the probability that class j is the true class when class k is predicted. To help with interpretation, we scale the errors in the table by a constant θ_0, defined as the RMSE of the confusion matrix estimated with scores from only unlabeled data, i.e. the estimate with IPrior when $N = 0$. Numbers greater than 1 mean that the estimate is worse than using θ_0 (with no labels).

The results show that informed priors (IPrior and IPrior+TS) often produce RMSE values that are orders of magnitude lower than using a simple uniform prior (UPrior). Thus, the model scores on the unlabeled test set (used to construct the informative priors) are highly informative for confusion matrix entries, even though the models themselves are (for the most part) miscalibrated. We see in addition that

---

[3]ImageNet is omitted because 50K labeled samples is not sufficient to estimate a confusion matrix that contains 1M parameters.

	N/K	N	UPrior (baseline)	IPrior (ours)	IPrior+TS (ours)
CIFAR-100	2	20	76.7	**26.4**	28.7
	5	50	40.5	**23.4**	26.7
	10	100	25.7	**21.5**	23.2
ImageNet	2	20	198.7	51.8	**36.4**
	5	50	122.0	55.3	**29.6**
	10	100	66.0	40.8	**22.1**
SVHN	2	20	383.6	86.2	**49.7**
	5	50	155.8	93.1	**44.2**
	10	100	108.2	80.6	**36.6**
20 News	2	20	54.0	**39.7**	46.1
	5	50	32.8	**28.9**	36.6
	10	100	24.7	**22.3**	28.7
DBpedia	2	20	900.3	118.0	**93.1**
	5	50	249.6	130.5	**74.5**
	10	100	169.1	125.9	**60.9**

Table 5: Mean percentage estimation error of ECE with bins as groups. Same setup as Table 3.

active sampling (IPrior+TS) provides additional significant reductions in RMSE over the IPrior method with no active sampling.

For DBpedia, with a uniform prior and randomly selected examples, the scaled mean RMSE is 0.000. One plausible explanation is that the accuracy of the classifier on DBpedia is 99%, resulting in a confusion matrix that is highly diagonally dominant, and this "simple structure" results in an easy estimation problem once there are at least a few labeled examples.

In our bin-wise accuracy experiments samples are grouped into 10 equal-sized bins according to their model scores. The active assessment framework allows us to estimate bin-wise accuracy with actively selected examples. By comparing the bin-wise accuracy estimates with bin-wise prediction confidence, we can then generate estimates of ECE to measure the amount of miscalibration of the classifier. We report error for overall ECE rather than error per score-bin since ECE is of more direct interest and more interpretable. Table 5 reports the average relative ECE estimation error, defined as $(100/R) \sum_{r=1}^{R} |ECE_N - E\hat{C}E_r|/ECE_N$ where ECE_N is the ECE measured on the full test set, and $E\hat{C}E_r$ is the esimated ECE (using MPE estimates of θ_b's), for a particular method on the rth run, $r = 1, \ldots, R = 1000$. Both the IPrior and IPrior+TS methods have significantly lower percentage error in general in their ECE estimates compared to the naive UPrior baseline, particularly on the three image datasets (CIFAR-100, ImageNet, and SVHN). The bin-wise RMSE of the estimated θ_b's are reported in the Supplement and show similar gains for IPrior and IPrior+TS.

Identification of Extreme Classes: For our identification experiments, for a particular metric and choice of groups, we conducted 1000 different sequential runs. For each run, after each labeled sample, we rank the estimates $\hat{\theta}_g$ obtained from

	Top-m	UPrior (baseline)	IPrior (ours)	IPrior+TS (ours)
CIFAR-100	1	81.1	83.4	**24.9**
	10	99.8	99.8	**55.1**
ImageNet	1	96.9	94.7	**9.3**
	10	99.6	98.5	**17.1**
SVHN	1	90.5	89.8	**82.8**
	3	100.0	100.0	**96.0**
20 News	1	53.9	55.4	**16.9**
	3	92.0	92.5	**42.5**
DBpedia	1	8.0	**7.6**	11.6
	3	91.9	90.2	**57.1**

Table 6: Percentage of labeled samples needed to identify the least accurate top-1 and top-m predicted classes across 5 datasets across 1000 runs.

each of the three methods, and compute the mean-reciprocal-rank (MRR) relative to the true top-m ranked groups (as computed from the full test set). The MRR of the predicted top-m classes is defined as $MRR = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\text{rank}_i}$ where rank$_i$ is the predicted rank of the ith best class.

Table 6 shows the mean percentage of labeled test set examples needed to correctly identify the target classes where "identify" means the minimum number of labeled examples required so that the MRR is greater than 0.99. For all 5 datasets the active method (IPrior+TS) clearly outperforms the non-active methods, with large gains in particular for cases where the number of classes K is large (CIFAR-100 and Imagenet). Similar gains in identifying the least calibrated classes are reported in the Supplement.

Figure 3 compares our 3 assessment methods for identifying the predicted classes with highest expected cost, using data from CIFAR-100, with two different (synthetic) cost matrices. In this plot the x-axis is the number of labels L_x (queries) and the y-value is the average (over all runs) of the MRR conditioned on L_x labels. In the left column (Human) the cost of misclassifying a person (e.g., predicting *tree* when the true class is a *woman*, etc.) is 10 times more expensive than other mistakes. In the right column, costs are 10 times higher if a prediction error is in a different superclass than the superclass of the true class (for the 20 superclasses in CIFAR-100). The curves show the MRR as a function of the number of labels (on average, over 100 runs) for each of the 3 assessment methods. The active assessment method (IPrior+TS) is much more efficient at identifying the highest cost classes than the two non-active methods. The gains from active assessment are also robust to different settings of the relative costs of mistakes (details in the Supplement).

Comparison of Groupwise Accuracy: For comparison experiments, Table 7 shows the results for the number of labeled data points required by each method to reliably assess the accuracy difference of two predicted classes, averaged over independent runs for all pairwise combinations of classes. The labeling process terminates when the most
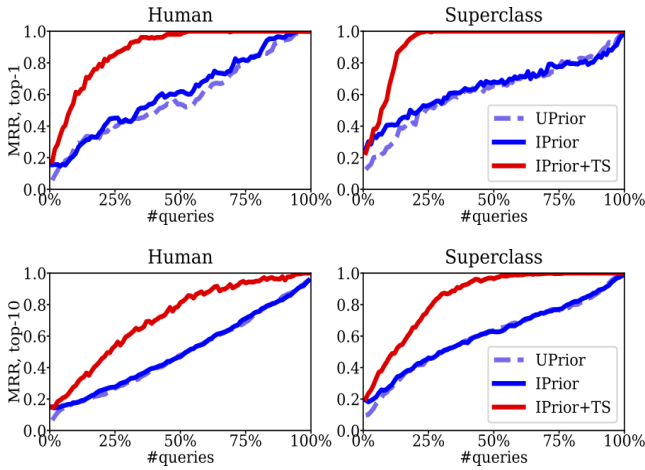
Figure 3: MRR of 3 assessment methods for identifying the top 1 (top) and top 10 (bottom) highest-cost predicted classes, with 2 different cost matrices (right and left), averaged over 100 trials. See text for details.

	UPrior (baseline)	IPrior (ours)	IPrior+TS (ours)
CIFAR-100, Superclass	203.5	129.0	**121.9**
SVHN	391.1	205.2	172.0
20 News	197.3	157.4	**136.1**
DBpedia	217.5	4.3	**2.8**

Table 7: Average number of labels across all pairs of classes required to estimate λ for randomly selected pairs of predicted classes.

probable region η is identified correctly and the estimation error of the cumulative density λ is within 5% of its value on the full test set. The results show that actively allocating a labeling budget and informative priors always improves label efficiency over uniform priors with no active assessment. In addition, active sampling (IPrior+TS) shows a systematic reduction of 5% to 35% in the mean number of labels required across datasets, over non-active sampling (IPrior).

Related Work

Bayesian and Frequentist Classifier Assessment: Prior work on Bayesian assessment of prediction performance, using Beta-Bernoulli models for example, has focused on specific aspects of performance modeling, such as estimating precision-recall performance (Goutte and Gaussier 2005), comparing classifiers (Benavoli et al. 2017), or analyzing performance of diagnostic tests (Johnson, Jones, and Gardner 2019). Welinder, Welling, and Perona (2013) used a Bayesian approach to leverage a classifier's scores on unlabeled data for Bayesian evaluation of performance, and Ji, Smyth, and Steyvers (2020) also used Bayesian estimation with scores from unlabeled data to assess group fairness of blackbox classifiers in a label-efficient manner.

Frequentist methods for label-efficient evaluation of clas-

sifier performance have included techniques such as importance sampling (Sawade et al. 2010) and stratified sampling (Kumar and Raj 2018), and low-variance sampling methods have been developed for evaluation of information retrieval systems (Aslam, Pavlu, and Yilmaz 2006; Yilmaz and Aslam 2006; Moffat, Webber, and Zobel 2007).

The framework we develop in this paper significantly generalizes these earlier contributions, by addressing a broader range of metrics and performance tasks within a single Bayesian assessment framework and by introducing the notion of active assessment for label-efficiency.

Active Assessment: While there is a large literature on active learning and multi-armed bandits (MAB) in general, e.g., (Settles 2012; Russo et al. 2018), our work is the first to our knowledge that applies ideas from Bayesian active learning to general classifier assessment, building on MAB-inspired, pool-based active learning algorithms for data selection. Nguyen, Ramanan, and Fowlkes (2018) developed non-Bayesian active learning methods to select samples for estimating visual recognition performance of an algorithm on a fixed test set and similar ideas have been explored in the information retrieval literature (Sabharwal and Sedghi 2017; Li and Kanoulas 2017; Rahman et al. 2018; Voorhees 2018; Rahman, Kutlu, and Lease 2019). This prior work is relevant to the ideas proposed and developed in this paper, but narrower in scope in terms of performance metrics and tasks.

Conclusions

In this paper we described a Bayesian framework for assessing performance metrics of black-box classifiers, developing inference procedures for an array of assessment tasks. In particular, we proposed a new framework called *active assessment* for label-efficient assessment of classifier performance, and demonstrated significant performance improvements across five well-known datasets using this approach.

There are a number of interesting directions for future work, such as Bayesian estimation of continuous functions related to accuracy and calibration, rather than using discrete groups, and Bayesian assessment in a sequential non-stationary context (e.g., with label and/or covariate shift). The framework can also be extended to assess the same black-box model operating in multiple environments using a Bayesian hierarchical approach, or to comparatively assess multiple models operating in the same environment.

Acknowledgements

Broader Impact

Machine learning classifiers are currently widely used to make predictions and decisions across a wide range of applications in society: education admissions, health insurance, medical diagnosis, court decisions, marketing, face recognition, and more—and this trend is likely to continue to grow. When these systems are deployed in real-world environments it will become increasingly important for users to have the ability to perform reliable, accurate, and independent evaluation of the performance characteristics of these systems and to do this in a manner which is efficient in terms of the need for labeled data. y

Our paper addresses this problem directly, providing a general-purpose and transparent framework for label-efficient performance evaluations of black-box classifier systems. The probabilistic (Bayesian) aspect of our approach provides users with the ability to understand how much they can trust performance numbers given a fixed data budget for evaluation. For example, a hospital system or a university might wish to evaluate multiple different performance characteristics of pre-trained classification models in the specific context of the population of patients or students in their institution. The methods we propose have the potential to contribute to increased societal trust in AI systems that are based on machine learning classification models.

References

Aslam, J. A.; Pavlu, V.; and Yilmaz, E. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 541–548.

Benavoli, A.; Corani, G.; Demšar, J.; and Zaffalon, M. 2017. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research* 18(1): 2653–2688.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, volume 1, 4171–4186.

Du, X.; El-Khamy, M.; Lee, J.; and Davis, L. 2017. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *Winter Conference on Applications of Computer Vision*, 953–961.

Freytag, A.; Rodner, E.; and Denzler, J. 2014. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, 562–577. Springer.

Goutte, C.; and Gaussier, E. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*, 345–359.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

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

Ji, D.; Smyth, P.; and Steyvers, M. 2020. Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference. *Advances in Neural Information Processing Systems* 33.

Johnson, W. O.; Jones, G.; and Gardner, I. A. 2019. 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.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172(5): 1122–1131.

Komiyama, J.; Honda, J.; and Nakagawa, H. 2015. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, 1152–1161.

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

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 3787–3798.

Kumar, A.; and Raj, B. 2018. Classifier risk estimation under limited labeling resources. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 3–15. Springer.

Lang, K. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, 331–339.

Li, D.; and Kanoulas, E. 2017. Active sampling for large-scale information retrieval evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 49–58.

Marshall, E. C.; and Spiegelhalter, D. J. 1998. League tables of in vitro fertilisation clinics: How confident can we be about the rankings. *British Medical Journal* 316: 1701–1704.

Moffat, A.; Webber, W.; and Zobel, J. 2007. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 375–382.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning* .

Nguyen, P.; Ramanan, D.; and Fowlkes, C. 2018. Active testing: An efficient and robust framework for estimating accuracy. In *International Conference on Machine Learning*, 3759–3768.

Rahman, M. M.; Kutlu, M.; Elsayed, T.; and Lease, M. 2018. Efficient test collection construction via active learning. *arXiv preprint arXiv:1801.05605* .

Rahman, M. M.; Kutlu, M.; and Lease, M. 2019. Constructing test collections using multi-armed bandits and active learning. In *The World Wide Web Conference*, 3158–3164.

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

Russo, D. 2016. Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, 1417–1418.

Russo, D. J.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, Z.; et al. 2018. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning* 11(1): 1–96.

Sabharwal, A.; and Sedghi, H. 2017. How good are my predictions? Efficiently approximating precision-recall curves for massive datasets. In *Conference on Uncertainty in Artificial Intelligence*.

Sanyal, A.; Kusner, M. J.; Gascón, A.; and Kanade, V. 2018. TAPAS: Tricks to accelerate (encrypted) prediction as a service. In *International Conference on Machine Learning*, volume 80, 4490–4499. PMLR.

Sawade, C.; Landwehr, N.; Bickel, S.; and Scheffer, T. 2010. Active risk estimation. In *International Conference on Machine Learning*, 951–958.

Settles, B. 2012. *Active Learning*. Synthesis Lectures on AI and ML. Morgan Claypool.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4): 285–294.

Vezhnevets, A.; Buhmann, J. M.; and Ferrari, V. 2012. Active learning for semantic segmentation with expected change. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3162–3169. IEEE.

Voorhees, E. M. 2018. On building fair and reusable test collections using bandit techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 407–416.

Welinder, P.; Welling, M.; and Perona, P. 2013. A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3262–3269.

Yao, Y.; Xiao, Z.; Wang, B.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2017. Complexity vs. performance: Empirical analysis of machine learning as a service. In *Internet Measurement Conference*, 384–397.

Yilmaz, E.; and Aslam, J. A. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 102–111.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 649–657.