Diagnostic Uncertainty Calibration: Towards Reliable Machine Predictions in Medical Domain

Takahiro Mimori
RIKEN AIP

Keiko Sasada
Kumamoto University Hospital

Hirotaka Matsui
Kumamoto University

Issei Sato
The University of Tokyo, RIKEN AIP, ThinkCyte

Abstract

We propose an evaluation framework for class probability estimates (CPEs) in the presence of label uncertainty, which is commonly observed as diagnosis disagreement between experts in the medical domain. We also formalize evaluation metrics for higher-order statistics, including inter-rater disagreement, to assess predictions on label uncertainty. Moreover, we propose a novel post-hoc method called α-calibration, that equips neural network classifiers with calibrated distributions over CPEs. Using synthetic experiments and a large-scale medical imaging application, we show that our approach significantly enhances the reliability of uncertainty estimates: disagreement probabilities and posterior CPEs.

1 Introduction

The reliability of uncertainty quantification is essential for safety-critical systems such as medical diagnosis assistance. Despite the high accuracy of modern neural networks for a wide range of classification tasks, their predictive probability often tends to be uncalibrated (Guo et al., 2017). Measuring and improving probability calibration, i.e., the consistency of predictive probability for an actual class frequency, has become one of the central issues in machine learning research (Vaicenavicius et al., 2019; Widmann et al., 2019; Kumar et al., 2019). At the same time, the uncertainty of ground truth labels in real-world data may also affect the reliability of the systems. Particularly, in the medical domain, inter-rater variability is commonly observed despite the annotators' expertise (Sasada et al., 2018; Jensen et al., 2019). This variability is also worth predicting for downstream tasks such as finding examples that need medical second opinions (Raghu et al., 2018).

To enhance the reliability of class probability estimates (CPEs), *post-hoc* calibration, which transforms output scores to fit into empirical class probabilities, has been proposed for both general classifiers (Platt et al., 1999; Zadrozny and Elkan, 2001, 2002) and neural networks (Guo et al., 2017; Kull et al., 2019). However, current evaluation metrics for calibration rely on empirical accuracy calculated with ground truth, for which the uncertainty of labels has not been considered. Another problem is that label uncertainty is not fully accounted for by CPEs; *e.g.*, a 50% confidence for class x does not necessarily mean the same amount of human belief, even when the CPEs are calibrated. Raghu et al. (2018) indicated that label uncertainty measures, such as an inter-rater disagreement frequency, were biased when they were estimated with CPEs. They instead proposed directly discriminating high uncertainty instances with input features. This treatment, however, requires training an additional predictor for each uncertainty measure and lacks an integrated view with the classification task.

In this work, we first develop an evaluation framework for CPEs when label uncertainty is indirectly observed through multiple annotations per instance (called *label histograms*). Guided with proper scoring rules (Gneiting and Raftery, 2007) and their decompositions (DeGroot and Fienberg, 1983; Kull and Flach, 2015), evaluation metrics, including calibration error, are naturally extensible to the situation with label histograms, where we derive estimators that benefit from unbiased or debiased property. Next, we generalize the framework to evaluate probabilistic predictions on higher-order statistics, including inter-rater disagreement. This extension enables us to evaluate these statistics in a unified way with CPEs. Finally, we address how the reliability of CPEs and disagreement

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

probability estimates (DPEs) can be improved using label histograms. While the existing post-hoc calibration methods solely address CPEs, we discuss the importance of obtaining a good predictive distribution over CPEs beyond point estimation to improve DPEs. Also, the distribution is expected to be useful for obtaining posterior CPEs when expert labels are provided for prediction. With these insights, we propose a novel method named α-calibration that uses label histograms to equip a neural network classifier with the ability to predict distributions of CPEs. In our experiments, the utility of our evaluation framework and α-calibration is demonstrated with synthetic data and a large-scale medical image dataset with multiple annotations provided from a study of myelodysplastic syndrome (MDS) (Sasada et al., 2018). Notably, α-calibration significantly enhances the quality of DPEs and the posterior CPEs.

In summary, our contributions are threefold as follows:

- Under ground truth label uncertainty, we develop evaluation metrics that benefit from unbiased or debiased property for class probability estimates (CPEs) using multiple labels per instance, *i.e.*, label histograms (Section 3).
- We generalize our framework to evaluate probability predictions on higher-order statistics, including inter-rater disagreement (Section 4).
- We advocate the importance of predicting the distributional uncertainty of CPEs, addressing with a newly devised post-hoc method, α-calibration (Section 5). Our approach substantially improves disagreement probability estimates (DPEs) and posterior CPEs for synthetic and real data experiments (Fig. 1 and Section 7).

2 Background

We overview calibration measures, proper scoring rules, and post-hoc calibration of CPEs as a prerequisite for our work.

Notation Let $K \in \mathbb{N}$ be a number of categories, $e^K = \{e_1, \ldots, e_K\}$ be a set of K dimensional one-hot vectors (*i.e.*, $e_{kl} := \mathbb{I}[k = l]$), and $\Delta^{K-1} := \{\zeta \in \mathbb{R}^K_{\geq 0} : \sum_k \zeta_k = 1\}$ be a $K-1$-dimensional probability simplex. Let (X, Y) be jointly distributed random variables over $\mathcal{X}$ and e^K, where X denotes an input feature, such as image data, and Y denotes a K-way label. Let $Z = (Z_1, \ldots, Z_K)^\top := f(X) \in \Delta^{K-1}$ denote a random variable that represents class probability estimates (CPEs) for input X with a classifier $f : \mathcal{X} \to \Delta^{K-1}$.

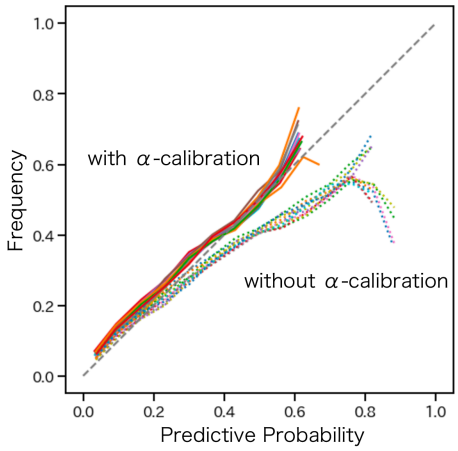

Figure 1: Reliability diagram of disagreement probability estimates (DPEs) in experiments with MDS data: a medical image dataset with multiple labels per instance. The dashed diagonal line corresponds to perfectly calibrated predictions. Calibrations of DPEs were significantly enhanced with α-calibration (solid lines) from the original ones (dotted lines).

2.1 Calibration measures

The notion of calibration, which is the agreement between a predictive class probability and an empirical class frequency, is important for reliable predictions. We reference Bröcker (2009); Kull and Flach (2015) for the definition of calibration.

Definition 1 (Calibration). [1] *A probabilistic classifier $f : \mathcal{X} \to \Delta^{K-1}$ is said to be calibrated if $Z = f(X)$ matches a true class probability given Z, i.e., $\forall k, Z_k = C_k$, where $C_k := P(Y = e_k|Z)$ and $C := (C_1, \ldots, C_K)^\top \in \Delta^{K-1}$, which we call a calibration map.*

The following metric is commonly used to measure calibration errors of binary classifiers:

Definition 2 (Calibration error).

$$\mathrm{CE}_1 := \left(\mathbb{E}[|Z_1 - C_1|^p]\right)^{1/p}, \quad \textit{where} \quad p \geq 1. \tag{1}$$

Note that CE_1 takes the minimum value zero iff $Z = C$. The cases with $p = 1$ and 2 are called the expectation calibration error (ECE) (Naeini et al., 2015) and the squared calibration error (Kumar et al., 2019), respectively. Hereafter, we use $p = 2$ and let CE denote CE_1 for binary cases. For multiclass cases, we denote CE as a commonly used definition of class-wise calibration error (Kumar et al., 2019), i.e., $(\sum_k \mathrm{CE}_k^2)^{1/2}$.

[1] A stronger notion of calibration that requires $Z = C$ is examined in the literature (Vaicenavicius et al., 2019; Widmann et al., 2019) .

2.2 Proper scoring rules

Although calibration is a desirable property, being calibrated is not sufficient for useful predictions. For instance, a predictor that always presents the marginal class frequency $Z = (P(Y = e_1), \dots, P(Y = e_K))^\top$ is perfectly calibrated, but it entirely lacks the sharpness of prediction for labels stratified with Z. In contrast, the strictly proper scoring rules (Gneiting and Raftery, 2007; Parmigiani and Inoue, 2009) elicit a predictor's true belief for each instance and do not suffer from this problem.

Definition 3 (Proper scoring rules for classification)**.** *A loss function $\ell : e^K \times \Delta^{K-1} \to \mathbb{R}$ is said to be proper if $\forall q \in \Delta^{K-1}$ and for all $z \in \Delta^{K-1}$ such that $z \neq q$,*

$$\mathbb{E}_{Y\sim \mathrm{Cat}(q)}[\ell(Y, z)] \geq \mathbb{E}_{Y\sim \mathrm{Cat}(q)}[\ell(Y, q)] \tag{2}$$

holds, where $\mathrm{Cat}(\cdot)$ denotes a categorical distribution. If the strict inequality holds, ℓ is said to be strictly proper. Following the convention, we write $\ell(q, z) = \mathbb{E}_{Y\sim \mathrm{Cat}(q)}[\ell(Y, z)]$ for $q \in \Delta^{K-1}$.

For a strictly proper loss ℓ, the divergence function $d(q, z) := \ell(q, z) - \ell(q, q)$ takes a non-negative value and is zero iff $z = q$, by definition. Squared loss $\ell_{\mathrm{sq}}(y, z) := \|y - z\|^2$ and logarithmic loss $\ell_{\log}(y, z) := -\sum_k y_k \log z_k$ are the most well-known examples of strictly proper loss. For these cases, the divergence functions are given as $d_{\mathrm{sq}}(q, z) = \ell_{\mathrm{sq}}(q, z)$ and $d_{\log}(q, z) = D_{\mathrm{KL}}(q, z)$, a.k.a. KL divergence, respectively.

Let $L := \mathbb{E}[d(Y, Z)] = \mathbb{E}[d(Y, f(X))]$ denote the expected loss, where the expectation is taken over a distribution $P(X, Y)$. As special cases of L,

$$L_{\mathrm{sq}'} := \mathbb{E}\left[\ell_{\mathrm{sq}'}(Y, Z)\right] = \mathbb{E}[(Y_1 - Z_1)^2] \quad (K = 2), \tag{3}$$

$$L_{\mathrm{sq}} := \mathbb{E}[\ell_{\mathrm{sq}}(Y, Z)] = \mathbb{E}[\|Y - Z\|^2] \quad (K \geq 2), \tag{4}$$

are commonly used for binary and multiclass prediction, respectively, where $\ell_{\mathrm{sq}'} := \frac{1}{2}\ell_{\mathrm{sq}}$. When the expectations are taken over an empirical distribution $\widehat{P}(X, Y)$, these are referred to as Brier score (BS) [2] and probability score (PS), respectively (Brier, 1950; Murphy, 1973).

Decomposition of proper losses The relation between the expected proper loss L and the calibration measures is clarified with a decomposition of L as follows (DeGroot and Fienberg, 1983):

$$L = \mathrm{CL} + \mathrm{RL}, \quad \text{where}$$
$$\begin{cases} \mathrm{CL} := \mathbb{E}[d(C, Z)], & \text{(Calibration Loss)} \\ \mathrm{RL} := \mathbb{E}[d(Y, C)]. & \text{(Refinement Loss)} \end{cases} \tag{5}$$

[2] While Brier (1950) originally introduced a multiclass loss that equals PS, we call BS as Brier score, following convention (Bröcker, 2012; Ferro and Fricker, 2012).

The CL term corresponds to an error of calibration because the term will be zero iff Z equals the calibration map $C = \mathbb{E}[Y|Z]$. Relations $\mathrm{CL}_{\mathrm{sq}'} = \mathrm{CE}^2$ and $\mathrm{CL}_{\mathrm{sq}} = \mathrm{CE}^2$ can be confirmed for binary and multiclass cases, respectively. Complementarily, the RL term shows a dispersion of labels Y given Z from its mean $\mathbb{E}[Y|Z]$ averaged over Z.

Under the assumption that labels follow an instance-wise categorical distribution as $Y|X \sim \mathrm{Cat}(Q)$, where $Q(X) \in \Delta^{K-1}$, Kull and Flach (2015) further decompose L into the following terms:

$$L = \underbrace{\mathrm{CL} + \mathrm{GL}}_{\mathrm{EL}} + \mathrm{IL}, \quad \text{where}$$
$$\begin{cases} \mathrm{EL} = \mathbb{E}[d(Q, Z)], & \text{(Epistemic Loss)} \\ \mathrm{IL} = \mathbb{E}[d(Y, Q)], & \text{(Irreducible Loss)} \\ \mathrm{GL} = \mathbb{E}[d(Q, C)]. & \text{(Grouping Loss)} \end{cases} \tag{6}$$

The EL term, which equals zero iff $Z = Q$, is a more direct measure for the optimality of the model than L. The IL term stemming from the randomness of observations is called *aleatoric uncertainty* in the literature (Der Kiureghian and Ditlevsen, 2009; Senge et al., 2014). We refer to Appendix A for details and proofs of the statements in this section.

2.3 Post-hoc calibration for deep neural network classifiers

For deep neural network (DNN) classifiers with the softmax activation, a post-hoc calibration of class probability estimates (CPEs) is commonly performed by optimizing a linear transformation of the last layer's logit vector (Guo et al., 2017; Kull et al., 2019), which minimizes the negative log-likelihood (NLL) of validation data:

$$\mathrm{NLL} = -\mathbb{E}_{X,Y\sim\widehat{P}}[\log P_{\mathrm{obs}}(Y|\widetilde{f}(X))], \tag{7}$$

where $\widehat{P}$, $\widetilde{f} : \mathcal{X} \to \Delta^{K-1}$ and $P_{\mathrm{obs}}(Y|Z) = \prod_k Z_k^{Y_k}$ denote an empirical data distribution, a transformed DNN function from f, and a likelihood model, respectively. More details are described in Appendix D.1. In particular, temperature scaling, which has a single parameter and keeps the maximum confidence class unchanged, was the most successful in confidence calibration. More recent research (Wenger et al., 2020; Zhang et al., 2020; Rahimi et al., 2020) has proposed nonlinear calibration maps with favorable properties, such as expressiveness, data-efficiency, and accuracy-preservation.

3 Evaluation of class probability estimates with label histograms

Now, we formalize evaluation metrics for class probability estimates (CPEs) using label histograms, where multiple labels per instance are observed. We assume that N input samples are obtained in an i.i.d. manner: $\{x_i\}_{i=1}^N \sim P(X)$, and for each instance i, label histogram $y_i \in \mathbb{Z}_{\geq 0}^K$ is obtained from n_i annotators in a conditionally i.i.d. manner, *i.e.*, $\{y_i^{(j)} \in e^K\}_{j=1}^{n_i} | x_i \sim P(Y|X=x_i)$ and $y_i = \sum_{j=1}^{n_i} y_i^{(j)}$. A predictive class probability for the i-th instance is denoted by $z_i = f(x_i) \in \Delta^{K-1}$. In this section, we assume ℓ_{sq} as a proper loss ℓ and omit the subscript from terms: EL and CL for brevity. The proofs in this section are found in Appendix B.

3.1 Expected squared and epistemic loss

We first derive an unbiased estimator of the expected squared loss L_{sq} from label histograms.

Proposition 1 (Unbiased estimator of expected squared loss)**.** *The following estimator of L_{sq} is unbiased.*

$$\widehat{L}_{\mathrm{sq}} := \sum_{i=1}^{N} \frac{w_i}{W} \sum_{k=1}^{K} \left[(\widehat{\mu}_{ik} - z_{ik})^2 + \widehat{\mu}_{ik}(1 - \widehat{\mu}_{ik}) \right], \quad (8)$$

where $\widehat{\mu}_{ik} := y_{ik}/n_i$, $w_i \geq 0$, and $W := \sum_{i=1}^N w_i$.

Note that an optimal weight vector w that minimizes the variance $\mathbb{V}[\widehat{L}_{\mathrm{sq}}]$ would be $w = 1$ if the number of annotators n_i is constant for all instances. Otherwise, it depends on undetermined terms, as discussed in Appendix B. We use $w = 1$ as a standard choice, where $\widehat{L}_{\mathrm{sq}}$ coincides with the probability score PS when every instance has a single label.

In addition to letting $\widehat{L}_{\mathrm{sq}}$ have higher statistical power than single-labeled cases, label histograms also enable us to directly estimate the epistemic loss EL, which is a discrepancy measure from the optimal model. A plugin estimator of EL is obtained as

$$\widetilde{\mathrm{EL}} := \frac{1}{N} \sum_i \sum_k (\widehat{\mu}_{ik} - z_{ik})^2, \quad (9)$$

which, however, turns out to be severely biased. We alternatively propose the following estimator of EL.

Proposition 2 (Unbiased estimator of EL)**.** *The following estimator of* EL *is unbiased.*

$$\widehat{\mathrm{EL}} := \widetilde{\mathrm{EL}} - \frac{1}{N} \sum_i \sum_k \frac{1}{n_i - 1} \widehat{\mu}_{ik}(1 - \widehat{\mu}_{ik}). \quad (10)$$

Note that the second correction term implies that $\widehat{\mathrm{EL}}$ can only be evaluated when more than one label per instance is available. The bias correction effect is significant for a small n_i, which is relevant to most of the medical applications.

3.2 Calibration loss

Relying on the connection between CL and CE, we focus on evaluating CL to measure calibration. The calibration loss is further decomposed into class-wise terms as follows:

$$\mathrm{CL} = \sum_k \mathrm{CL}_k, \quad (11)$$

where $\mathrm{CL}_k := \mathbb{E}[(C_k - Z_k)^2] = \mathbb{E}[\mathbb{E}[(C_k - Z_k)^2 | Z_k]]$. Thus, the case of CL_k is sufficient for subsequent discussion. Note that a difficulty exists in estimating the conditional expectation for Z_k. We take a standard binning-based approach (Zadrozny and Elkan, 2001) to evaluate CL_k by stratifying with Z_k values. Specifically, Z_k is partitioned into B_k disjoint regions $\mathcal{B}_k = \{[\zeta_0 = 0, \zeta_1), [\zeta_1, \zeta_2), \ldots, [\zeta_{B_k - 1}, \zeta_{B_k} = 1]\}$, and CL_k is approximated as follows:

$$\mathrm{CL}_k(\mathcal{B}_k) := \sum_{b=1}^{B_k} \mathrm{CL}_{kb}(\mathcal{B}_k), \quad \text{where}$$
$$\begin{cases} \mathrm{CL}_{kb}(\mathcal{B}_k) := \mathbb{E}[\mathbb{E}[(\bar{C}_{kb} - \bar{Z}_{kb})^2 | Z_k \in \mathcal{B}_{kb}]], \\ \bar{C}_{kb} := \mathbb{E}[Y_k | Z_k \in \mathcal{B}_{kb}], \\ \bar{Z}_{kb} := \mathbb{E}[Z_k | Z_k \in \mathcal{B}_{kb}], \end{cases} \quad (12)$$

in which CL_k is further decomposed into the bin-wise components. A plugin estimator of CL_{kb} is derived as follows:

$$\widetilde{\mathrm{CL}}_{kb}(\mathcal{B}_k) := \frac{|I_{kb}|}{N} (\bar{c}_{kb} - \bar{z}_{kb})^2, \quad (13)$$

where, $I_{kb} = \{i : z_{ik} \in \mathcal{B}_{kb}\}$, $\bar{c}_{kb} := \sum_{i \in I_{kb}} \widehat{\mu}_{ik} / |I_{kb}|$, $\bar{z}_{kb} := \sum_{i \in I_{kb}} z_{ik} / |I_{kb}|$, and $|I_{kb}|$ denotes the size of I_{kb}. We can again improve the estimator by debiasing as follows:

Proposition 3 (Debiased estimator of CL_{kb})**.** *The plugin estimator of CL_{kb} is debiased with the following estimator:*

$$\widehat{\mathrm{CL}}_{kb}(\mathcal{B}_k) := \widetilde{\mathrm{CL}}_{kb}(\mathcal{B}_k) - \frac{|I_{kb}|}{N} \frac{\bar{\sigma}_{kb}^2}{|I_{kb}| - 1}, \quad \textit{where}$$
$$\bar{\sigma}_{kb}^2 := \frac{1}{|I_{kb}|} \sum_{i \in I_{kb}} \widehat{\mu}_{ik}^2 - \left(\frac{1}{|I_{kb}|} \sum_{i \in I_{kb}} \widehat{\mu}_{ik} \right)^2. \quad (14)$$

Note that the correction term against $\widetilde{\mathrm{CL}}_{kb}$ would inflate for small-sized bins with a high label variance $\bar{\sigma}_{kb}^2$. $\widehat{\mathrm{CL}}_{kb}$ can also be computed for single-labeled data,

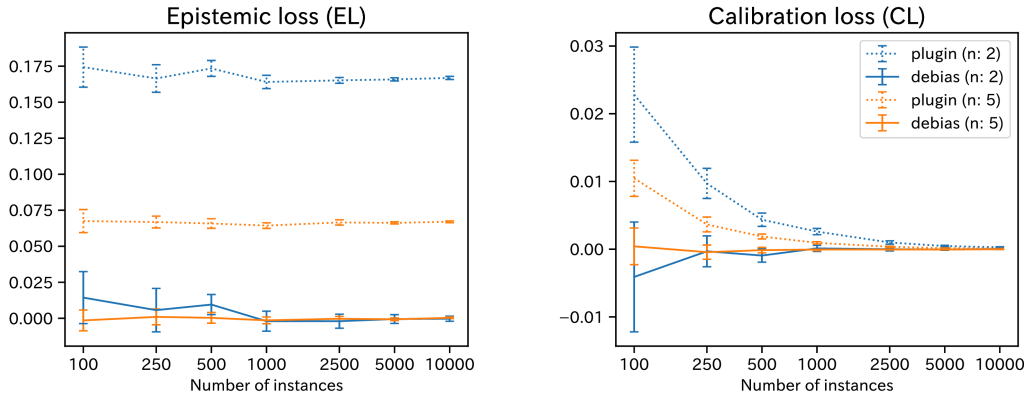

Figure 2: Comparison of the plugin and debiased estimators for synthetic data. For both EL and CL, the debiased estimators (solid lines) are closer to ground truth (which is zero in this experiment) than the plugin estimators (dotted lines). The error bars show 90% confidence intervals for the means of ten runs.

i.e., $\widehat{\mu}_{ik} = y_{ik}$. In this case, the estimator precisely coincides with a debiased estimator for the *reliability* term formerly proposed in meteorological literature (Bröcker, 2012; Ferro and Fricker, 2012).

3.3 Debiasing effects of EL and CL estimators

To confirm the debiasing effect of estimators $\widehat{\mathrm{EL}}$ and $\widehat{\mathrm{CL}}$ against the plugin estimators, we experimented on evaluations of a perfect predictor using synthetic binary labels with varying instance sizes. For each instance, a positive label probability was drawn from a uniform distribution U(0,1); thereby two or five labels were generated in an *i.i.d.* manner. The predictor indicated the true probabilities so that both EL and CL would be zero in expectation. As shown in Fig. 2, the debiased estimators significantly reduced the plugin estimators' biases, even in the cases with two annotators. Details on the experimental setup are found in Appendix B.5.

4 Evaluation of higher-order statistics

Here, we generalize our framework to evaluate predictions on higher-order statistics. As is done for CPEs, the expected proper losses and calibration measures can also be formalized. We focus on a family of symmetric binary statistics $\phi : e^{K\times n} \to \{0,1\}$ calculated from n distinct K-way labels for the same instance. For example, $\phi^{\mathrm{D}} := \mathbb{I}[Y^{(1)} \neq Y^{(2)}]$ represents a disagreement between paired labels $(Y^{(1)}, Y^{(2)})$. The estimator of $\mathbb{E}[\phi^{\mathrm{D}}|X]$ is known as the Gini-Simpson index, which is a measure of diversity.

Given a function $\varphi : \mathcal{X} \to [0,1]$ that represents a predictive probability of being $\phi = 1$, the closeness of $\varphi(X)$ to a true probability $P(\phi = 1|X)$ is consistently evaluated with the expected (one dimensional) squared loss $L_\phi := \mathbb{E}[(\phi - \varphi)^2]$. Then, the calibration loss CL_ϕ is derived by applying equation (5) as follows:

$$L_\phi = \underbrace{\mathbb{E}[(\mathbb{E}[\phi|\varphi] - \varphi)^2]}_{\mathrm{CL}_\phi} + \underbrace{\mathbb{E}[(\phi - \mathbb{E}[\phi|\varphi])^2]}_{\mathrm{RL}_\phi}. \quad (15)$$

An unbiased estimator of L_ϕ and a debiased estimator of CL_ϕ can be derived following a similar discussion as in CPEs. The biggest difference from the case of CPEs is that it requires more careful consideration to obtain an unbiased estimator of $\mu_{\phi,i} := \mathbb{E}[\phi|X = x_i]$ as follows:

$$\widehat{\mu}_{\phi,i} := \binom{n_i}{n}^{-1} \sum_{j \in \mathrm{Comb}(n_i,n)} \phi(y_i^{(j_1)}, \ldots, y_i^{(j_n)}), \quad (16)$$

where $\mathrm{Comb}(n_i, n)$ denotes the distinct subset of size n drawn from $\{1, \ldots, n_i\}$ without replacement. The proof directly follows from the fact that $\widehat{\mu}_{\phi,i}$ is a U-statistic of n-sample symmetric kernel function ϕ (Hoeffding et al., 1948). Details on the derivations for $\widehat{L}_\phi$ and $\widehat{\mathrm{CL}}_\phi$ are described in Appendix C.

5 Post-hoc uncertainty calibration for DNNs with label histograms

We consider post-hoc uncertainty calibration problems using label histograms for a deep neural network (DNN) classifier f that offers CPE with the last layer's softmax activation.

5.1 Class probability calibration

For post-hoc calibration of CPEs using label histograms, existing methods for single-labeled data (Section 2.3) are straightforwardly extensible by replac-

ing the likelihood function P_{obs} in equation (7) with a multinomial distribution.

5.2 Importance of predicting distributional uncertainty of class probability estimates

Although we assume that labels for each input X are sampled from a categorical distribution $Q(X)$ in an i.i.d. manner, it is important to obtain a reliable CPE distribution beyond point estimation to perform several application tasks. We denote such a CPE distribution model as $P(\zeta|X)$, where $\zeta \in \Delta^{K-1}$. In this case, CPEs are written as $Z = \mathbb{E}[\zeta|X]$. Below, we illustrate two examples of those tasks.

Disagreement probability estimation For each input X, the extent of diagnostic disagreement among annotators is itself a signal worth predicting, which is different from classification uncertainty expressed as CPEs. Specifically, we aim at obtaining a disagreement probability estimation (DPE):

$$\varphi^{\mathrm{D}}(X) = \int 1 - \sum_k \zeta_k^2 \, dP(\zeta|X) \tag{17}$$

as a reliable estimator of a probability $P(\phi^{\mathrm{D}} = 1|X)$. When we only have CPEs, *i.e.*, $P(\zeta|X) = \delta(\zeta - f(X))$, where δ denotes the Dirac delta function, we get $\varphi^{\mathrm{D}} = 1 - \sum_k f(X)_k^2$. However, $\varphi^{\mathrm{D}} \simeq 0$ regardless of $f(X)$ would be more sensible if all the labels are given in unanimous.

Posterior class probability estimates We consider a task for updating CPEs of instance X after an expert's annotation Y. Given a CPE distribution model $P(\zeta|X)$, an updated CPEs:

$$Z'(X, Y) := \mathbb{E}[\zeta|X, Y], \tag{18}$$

can be inferred from a Bayesian posterior computation: $P(\zeta|X, Y) \propto P(Y|\zeta)P(\zeta|X)$. If the prior distribution $P(\zeta|X)$ is reliable, Z' would be more close to the true value $Q(X)$ than the original CPEs Z in expectation.

5.3 α-calibration: post-hoc method for CPE distribution calibration

We propose a novel post-hoc calibration method called *α-calibration* that infers a CPE distribution $P(\zeta|X)$ from a DNN classifier f and validation label histograms. Specifically, we use a Dirichlet distribution $\mathrm{Dir}(\zeta|\alpha_0(X)f(X))$ to model $P(\zeta|X)$, and minimize the NLL of label histograms with respect to instance-wise concentration parameter $\alpha_0(X) > 0$. We parameterize α_0 with a DNN that has a shared layer behind the last softmax activation of the DNN f and a successive full connection layer with an exp activation. Details are described in Appendix D.2. Using α_0 is one of the simplest ways to model the distribution over CPEs; hence it is computationally efficient and less affected by over-fitting without crafted regularization terms. In addition, α-calibration has several favorable properties: it is orthogonally applicable with existing CPE calibration methods, will not degrade CPEs since $Z = \mathbb{E}[\zeta|X] = f(X)$ by design, and quantities of interest such as a DPE (17) and posterior CPEs (18) can be computed in closed forms as follows:

$$\varphi^{\mathrm{D}} = \frac{\alpha_0}{\alpha_0 + 1}\left(1 - \sum_k f_k^2\right), \quad Z' = \frac{\alpha_0 f + Y}{\alpha_0 + 1}. \tag{19}$$

Theoretical analysis We consider whether a CPE distribution model $P(\zeta|X) = \mathrm{Dir}(\zeta|\alpha_0(X)f(X))$ is useful for downstream tasks. Let $G = g(X)$ denote a random variable of an output layer shared between both networks f and α_0. We can write $P(\zeta|X) = P(\zeta|G)$ since f and α_0 are deterministic given G. Although it is unclear whether $P(\zeta|G)$ is an appropriate model for the true label distribution $P(Q|G)$, we can corroborate the utility of the model with the following analysis.

To evaluate the quality of DPEs and posterior CPEs dependent on α_0, we analyze the expected loss $L_{\phi^{\mathrm{D}}} = \mathbb{E}_G[L_{\phi^{\mathrm{D}},G}]$ and the epistemic loss $\mathrm{EL}' := \mathbb{E}_G[\mathrm{EL}'_G]$, respectively, where we define $L_{\phi^{\mathrm{D}},G} := \mathbb{E}[(\phi^{\mathrm{D}} - \varphi^{\mathrm{D}})^2|G]$ and $\mathrm{EL}'_G := \mathbb{E}[\sum_k (Z'_k - Q_k)^2|G]$. We denote those for the original model $P_0(\zeta|X) = \delta(\zeta - f(X))$ before α-calibration as $L^{(0)}_{\phi^{\mathrm{D}},G}$ and $\mathrm{EL}'^{(0)}_G$, respectively.

Theorem 1. *There exist intervals for parameter $\alpha_0 \geq 0$, which improve task performances as follows.*

1. *For DPEs, $L_{\phi^{\mathrm{D}},G} \leq L^{(0)}_{\phi^{\mathrm{D}},G}$ holds when $(1 - 2u_Q + s_Z)/2(u_Q - s_Z) \leq \alpha_0$, and $L_{\phi^{\mathrm{D}},G}$ takes the minimum value when $\alpha_0 = (1 - u_Q)/(u_Q - s_Z)$, if $u_Q > s_Z$ is satisfied.*

2. *For posterior CPEs, $\mathrm{EL}'_G \leq \mathrm{EL}'^{(0)}_G$ holds when $(1 - u_Q - \mathrm{EL}_G)/2\,\mathrm{EL}_G \leq \alpha_0$, and EL'_G takes the minimum value when $\alpha_0 = (1 - u_Q)/\mathrm{EL}_G$, if $\mathrm{EL}_G > 0$ is satisfied.*

Note that we denote $s_Z := \sum_k Z_k^2$, $u_Q := \mathbb{E}[\sum_k Q_k^2|G]$, $v_Q := \mathbb{V}[Q_k|G]$, and $\mathrm{EL}_G := \mathbb{E}[\sum_k (Z_k - Q_k)^2|G]$. The optimal α_0 of both tasks coincide to be $\alpha_0 = (1 - u_Q)/v_Q$, if CPEs match the true conditional class probabilities given G, i.e., $Z = \mathbb{E}[Q|G]$.

The proof is shown in Appendix D.3.

6 Related work

Noisy labels Learning classifiers under label uncertainty has also been studied as a noisy label setting, assuming unobserved ground truth labels and label noises. The cases of uniform or class dependent noises have been studied to ensure robust learning schemes (Natarajan et al., 2013; Jiang et al., 2018; Han et al., 2018) and predict potentially inconsistent labels (Northcutt et al., 2019). Also, there have been algorithms that modeled a generative process of noises depending on input features (Xiao et al., 2015; Liu et al., 2020). However, the paradigm of noisy labels requires qualified gold standard labels to validate predictions, while we assume that ground truth labels include uncertainty.

Multiple annotations Learning from multiple annotations per instance has also been studied in crowdsourcing field (Guan et al., 2017; Rodrigues and Pereira, 2017; Tanno et al., 2019), which particularly modeled labelers with heterogeneous skills, occasionally including non-experts. In contrast, we focus on instance-wise uncertainty under homogeneous expertise as in Raghu et al. (2018). Another related paradigm is label distribution learning (Geng, 2016; Gao et al., 2017), which assumes instance-wise categorical probability as ground truth. Whereas they regard the true probability as observable, we assume it as a hidden variable on which actual labels depend.

Uncertainty of CPEs Approaches for predicting distributional uncertainty of CPEs for DNNs have mainly studied as part of Bayesian modeling. Gal and Ghahramani (2016); Lakshminarayanan et al. (2017); Teye et al. (2018); Wang et al. (2019) found practical connections for using ensembled DNN predictions as approximate Bayesian inference and uncertainty quantification (Kendall and Gal, 2017), which however require additional computational cost for sampling. An alternative approach is directly modeling CPE distribution with parametric families. In particular, Sensoy et al. (2018); Malinin and Gales (2018); Sadowski and Baldi (2018); Joo et al. (2020) adopted the Dirichlet distribution for a tractable distribution model and used for applications, such as detecting out-of-distribution examples. However, the use of multiple labels have not been explored in these studies. Moreover, these approaches need customized training procedures from scratch and are not designed to apply for DNN classifiers in a post-hoc manner, as is done in α-calibration.

7 Experiments

We applied DNN classifiers and calibration methods to synthetic and real-world image data with label histograms, where the performance was evaluated with our proposed metrics. Especially, we demonstrate the utility of α-calibration in two applications: predictions on inter-rater label disagreement (DPEs) and posterior CPEs, which we introduced in Section 5.2. Our implementation is available online [3].

7.1 Experimental setup

Synthetic data We generated two synthetic image dataset: Mix-MNIST and Mix-CIFAR-10 from MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky et al., 2009), respectively. We randomly selected half of the images to create mixed-up images from pairs and the other half were retained as original. For each of the paired images, a random ratio that followed a uniform distribution $U(0, 1)$ was used for the mix-up and a class probability of multiple labels, which were two or five in validation set.

MDS data We used a large-scale medical imaging dataset for myelodysplastic syndrome (MDS) (Sasada et al., 2018), which contained over 90 thousand hematopoietic cell images obtained from blood specimens from 499 patients with MDS. This study was carried out in collaboration with medical technologists who mainly belonged to the Kyushu regional department of the Japanese Society for Laboratory Hematology. The use of peripheral blood smear samples for this study was approved by the ethics committee at Kumamoto University, and the study was performed in accordance with the Declaration of Helsinki. For each of the cellular images, a mean of 5.67 medical technologists annotated the cellular category from 22 subtypes, where accurate classification according to the current standard criterion was still challenging for technologists with expertise.

Compared methods We used DNN classifiers as base predictors (Raw) for CPEs, where a three layered CNN architecture for Mix-MNIST and a VGG16-based one for Mixed-CIFAR-10 and MDS were used. For CPE calibration, we adopted temperature scaling (ts), which was widely used for DNNs (Guo et al., 2017). To predict CPE distributions, we used α-calibration and ensemble-based methods: Monte-Calro dropout (MCDO) (Gal and Ghahramani, 2016) and test-time augmentation (TTA) (Ayhan and Berens, 2018), which were both applicable to DNNs at prediction-time. Note

[3] https://github.com/m1m0r1/lh_calib

that TTA was only applied for Mix-CIFAR-10 and MDS, in which we used data augmentation while training. We also combined ts and/or α-calibration with the ensemble-based methods in our experiments, while some of their properties, including the invariance of accuracy for ts and that of CPEs for α-calibration, were not retained for these combinations. The details of the network architectures and parameters were described in Appendix F.1. Considering a constraint of the high labeling costs with experts in the medical domain, we focused on scenarios that training instances were singly labeled and multiple labels were only available for the validation and test set.

7.2 Results

Class probability estimates We observed a superior performance of TTA in accuracy and $\widehat{\mathrm{EL}}$ and a consistent improvement in $\widehat{\mathrm{EL}}$ and $\widehat{\mathrm{CL}}$ with ts, for all the dataset. The details are found in Appendix F.1. By using $\widehat{\mathrm{EL}}$, the relative performance of CPE predictions had been clearer than $\widehat{L}$ since the irreducible loss was subtracted from $\widehat{L}$. We include additional MDS experiments using full labels in Appendix F.3, which show similar tendencies but improved overall performance.

Disagreement probability estimates We compared squared loss and calibration error of DPEs for combinations of prediction schemes (Table 1 [4]). Notably, α-calibration combined with any methods showed a consistent and significant decrease in both $\widehat{L}_{\phi^{\mathrm{D}}}$ and $\widehat{\mathrm{CE}}_{\phi^{\mathrm{D}}}$, in contrast to MCDO and TTA, which had not solely improved the metrics. The improved calibration was also visually confirmed with a reliability diagram of DPEs for MDS data (Fig. 1).

Posterior CPEs We evaluated posterior CPEs, when one expert label per instance was available for test set. This task required a reasonable prior CPE model to update belief with additional label information. We summarize $\widehat{\mathrm{EL}}$ metrics of prior and posterior CPEs for combinations of dataset and prediction methods in Table 2. As we expected, α-calibration significantly decreased losses of the posterior CPEs, i.e., they got closer to the ideal CPEs than the prior CPEs. While TTA showed superior performance for the prior CPEs, the utility of the ensemble-based methods for the posterior computation was limited. We omit experiments on MCDO and TTA combined with α-calibration, as they require further approximation to compute posteriors.

[4] The mechanisms that cause the degradation of DPEs for Raw+ts are discussed in Appendix F.2.

8 Conclusion

In this work, we have developed a framework for evaluating probabilistic classifiers under ground truth label uncertainty, accompanied with useful metrics that benefited from unbiased or debiased properties. The framework was also generalized to evaluate higher-order statistics, including inter-rater disagreements. As a reliable distribution over class probability estimates (CPEs) is essential for higher-order prediction tasks, such as disagreement probability estimates (DPEs) and posterior CPEs, we have devised a post-hoc calibration method called α-calibration, which directly used multiple annotations to improve CPE distributions. Throughout empirical experiments with synthetic and real-world medical image data, we have demonstrated the utility of the evaluation metrics in performance comparisons and a substantial improvement in DPEs and posterior CPEs with α-calibration.

Acknowledgements

IS was supported by JSPS KAKENHI Grant Number 20H04239 Japan. This work was supported by RAIDEN computing system at RIKEN AIP center.

References

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. 2018.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3, 1950.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

Jochen Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, 39(3-4):655–667, 2012.

François Chollet et al. Keras. https://keras.io, 2015.

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.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31 (2):105–112, 2009.

Christopher AT Ferro and Thomas E Fricker. A bias-corrected decomposition of the brier score. *Quar-*

Table 1: Evaluations of disagreement probability estimates (DPEs)

	Mix-MNIST(2)		Mix-MNIST(5)		Mix-CIFAR-10(2)		Mix-CIFAR-10(5)		MDS	
Method	$\widehat{L}_{\phi^D}$	$\widehat{\mathrm{CE}}_{\phi^D}$	$\widehat{L}_{\phi^D}$	$\widehat{\mathrm{CE}}_{\phi^D}$	$\widehat{L}_{\phi^D}$	$\widehat{\mathrm{CE}}_{\phi^D}$	$\widehat{L}_{\phi^D}$	$\widehat{\mathrm{CE}}_{\phi^D}$	$\widehat{L}_{\phi^D}$	$\widehat{\mathrm{CE}}_{\phi^D}$
Raw	.0755	.0782	.0755	.0782	.1521	.2541	.1521	.2541	.1477	.0628
Raw+α	**.0724**	**.0524**	**.0724**	**.0531**	**.0880**	**.0357**	**.0877**	**.0322**	**.1454**	**.0406**
Raw+ts	.0775	.0933	.0773	.0923	.1968	.3310	.1978	.3324	.1482	.0663
Raw+ts+α	**.0699**	**.0344**	**.0702**	**.0379**	**.0863**	**.0208**	**.0861**	**.0164**	**.1445**	**.0261**
MCDO	.0749	.0728	.0749	.0728	.1518	.2539	.1518	.2539	.1470	.0562
MCDO+α	**.0700**	**.0277**	**.0700**	**.0285**	**.0873**	**.0275**	**.0870**	**.0241**	**.1450**	**.0346**
MCDO+ts	.0805	.1062	.0802	.1049	.1996	.3353	.2002	.3362	.1479	.0635
MCDO+ts+α	**.0690**	**.0155**	**.0691**	**.0188**	**.0863**	**.0196**	**.0861**	**.0167**	**.1442**	**.0186**
TTA	NA	NA	NA	NA	.1677	.2856	.1677	.2856	.1441	.0488
TTA+α	NA	NA	NA	NA	**.0860**	**.0245**	**.0857**	**.0231**	**.1428**	**.0334**
TTA+ts	NA	NA	NA	NA	.2421	.3957	.2430	.3968	.1448	.0553
TTA+ts+α	NA	NA	NA	NA	**.0872**	**.0467**	**.0870**	**.0398**	**.1422**	**.0197**

Table 2: Epistemic losses ($\widehat{\mathrm{EL}}$) of prior and posterior class probability estimates (CPEs)

	Mix-MNIST(2)		Mix-MNIST(5)		Mix-CIFAR-10(2)		Mix-CIFAR-10(5)		MDS	
Method	Prior	Post.	Prior	Post.	Prior	Post.	Prior	Post.	Prior	Post.
Raw+α	.0388	**.0292**	.0388	**.0292**	.2504	**.0709**	.2504	**.0693**	.0435	**.0354**
Raw+ts+α	.0379	**.0293**	.0379	**.0298**	.2423	**.0682**	.2423	**.0676**	.0430	**.0352**
MCDO	.0395	**.0391**	.0395	**.0391**	.2473	**.2471**	.2473	**.2471**	**.0437**	.0440
MCDO+ts	.0410	**.0406**	.0404	**.0400**	.2428	**.2425**	.2431	**.2428**	**.0435**	.0438
TTA	NA	NA	NA	NA	.2216	**.2184**	.2216	**.2184**	**.0378**	.0382
TTA+ts	NA	NA	NA	NA	.2452	**.2428**	.2451	**.2427**	**.0379**	.0383

terly Journal of the Royal Meteorological Society, 138(668):1954–1960, 2012.

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.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102 (477):359–378, 2007.

Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. *arXiv preprint arXiv:1703.08774*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Wassily Hoeffding et al. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313, 2018.

Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being bayesian about categorical probability. *arXiv preprint arXiv:2002.07965*, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 68–85. Springer, 2015.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12295–12305, 2019.

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.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Yushan Liu, Markus M Geipel, Christoph Tietz, and Florian Buettner. Timely: Improving labeling consistency in medical imaging for cell type classification. *arXiv preprint arXiv:2007.05307*, 2020.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4): 595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *arXiv preprint arXiv:1911.00068*, 2019.

Giovanni Parmigiani and Lurdes Inoue. *Decision theory: Principles and approaches*, volume 812. John Wiley & Sons, 2009.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. *arXiv preprint arXiv:1807.01771*, 2018.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Byron Boots, and Richard Hartley. Intra order-preserving functions for calibration of multi-class neural networks. *arXiv preprint arXiv:2003.06820*, 2020.

Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *arXiv preprint arXiv:1709.01779*, 2017.

Peter Sadowski and Pierre Baldi. Neural network regression with beta, dirichlet, and dirichlet-multinomial outputs. 2018.

Keiko Sasada, Noriko Yamamoto, Hiroki Masuda, Yoko Tanaka, Ayako Ishihara, Yasushi Takamatsu, Yutaka Yatomi, Waichiro Katsuda, Issei Sato, Hirotaka Matsui, et al. Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome. *Leukemia research*, 69:54–59, 2018.

Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classifica-

tion: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

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

David B Stephenson, Caio AS Coelho, and Ian T Jolliffe. Two extra components in the brier score decomposition. *Weather and Forecasting*, 23(4):752–757, 2008.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11244–11253, 2019.

Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In *International Conference on Machine Learning*, pages 4907–4916, 2018.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. *arXiv preprint arXiv:1902.06977*, 2019.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338: 34–45, 2019.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pages 178–190. PMLR, 2020.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, pages 12236–12246, 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Jize Zhang, Bhavya Kailkhura, and T Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. *arXiv preprint arXiv:2003.07329*, 2020.