Combining Priors with Experience: Confidence Calibration Based on Binomial Process Modeling

Jinzong Dong, Zhaohui Jiang*, Dong Pan, Haoyang Yu

School of Automation, Central South University
{dongjinzong, jzh0903, pandong, yu.haoyang}@csu.edu.cn

Abstract

Confidence calibration of classification models is a technique to estimate the true posterior probability of the predicted class, which is critical for ensuring reliable decision-making in practical applications. Existing confidence calibration methods mostly use statistical techniques to estimate the calibration curve from data or fit a user-defined calibration function, but often overlook fully mining and utilizing the prior distribution behind the calibration curve. However, a well-informed prior distribution can provide valuable insights beyond the empirical data under the limited data or low-density regions of confidence scores. To fill this gap, this paper proposes a new method that integrates the prior distribution behind the calibration curve with empirical data to estimate a continuous calibration curve, which is realized by modeling the sampling process of calibration data as a binomial process and maximizing the likelihood function of the binomial process. We prove that the calibration curve estimating method is Lipschitz continuous with respect to data distribution and requires a sample size of $3/B$ of that required for histogram binning, where B represents the number of bins. Also, a new calibration metric (TCE_{bpm}), which leverages the estimated calibration curve to estimate the true calibration error (TCE), is designed. TCE_{bpm} is proven to be a consistent calibration measure. Furthermore, realistic calibration datasets can be generated by the binomial process modeling from a preset true calibration curve and confidence score distribution, which can serve as a benchmark to measure and compare the discrepancy between existing calibration metrics and the true calibration error. The effectiveness of our calibration method and metric are verified in real-world and simulated data. We believe our exploration of integrating prior distributions with empirical data will guide the development of better-calibrated models, contributing to trustworthy AI.

Code — https://github.com/NeuroDong/TCEbpm
Extended version — https://arxiv.org/abs/2412.10658

1 Introduction

The prediction accuracy of modern machine learning classification methods such as deep neural networks is steadily increasing, leading to adoption in many safety-critical fields such as intelligent transportation (Lu, Lin, and Hu 2024), industrial automation (Jiang et al. 2023), and medical diagnosis (Luo et al. 2024). However, decision-making systems in these fields not only require high accuracy but also require signaling when they might be wrong (Munir et al. 2024). For example, in an automatic disease diagnosis system, when the confidence of the diagnostic model is relatively low, the decision-making should be passed to the doctor (Jiang et al. 2012). Specifically, along with its prediction, a classification model should offer accurate confidence (matching the true probability of event occurrence). In addition, accurate confidence also provides more detailed information than the no-confidence or class label (Huang et al. 2020). For example, doctors can gather more information to make more reliable decisions in "there is a 70% probability that the patient has cancer" than just a class label of "cancer". Furthermore, accurate confidence facilitates the incorporation of classification models into other probabilistic models. For instance, accurate confidence allows active learning to select more representative samples (Han et al. 2024) and improves the generalization performance of knowledge distillation (Li and Caragea 2023). Therefore, pursuing more accurate confidence for classification models is a significant work (Penso, Frenkel, and Goldberger 2024; Wang et al. 2024).

However, modern classification neural networks often suffer from inaccurate confidence (Guo et al. 2017), which means that their confidence does not match the true probabilities of predicted class. For example, if a deep neural network classifies a medical image as "benign" with a confidence score of 0.99, the true probability of the medical image being "benign" could be significantly lower than 0.99, and even its true class may be "malignant". Therefore, in recent years, this problem has been attracting increasing attention (Dong et al. 2024; Geng et al. 2024), and many confidence calibration methods, which aim to obtain more accurate confidence through additional processing, have been proposed (Silva Filho et al. 2023; Zhang et al. 2023).

Previous works calibrate confidence mostly from three directions: 1) Performing calibration during the classifier's training (train-time calibration), usually modifying the classifier's objective function (Liu et al. 2023; Müller, Kornblith, and Hinton 2019; Fernando and Tsokos 2021); 2) Binning

*corresponding author.

confidence scores and estimating the calibrated confidence using the average accuracy inside the bins (binning-based calibration) (Zadrozny and Elkan 2001; Naeini, Cooper, and Hauskrecht 2015; Patel et al. 2020); 3) Fitting a function on logit or confidence score so that the outcome of the function is calibrated (fitting-based calibration) (Platt et al. 1999; Guo et al. 2017; Zadrozny and Elkan 2002; Zhang, Kailkhura, and Han 2020). Despite the existence of the three valuable methods mentioned above, these methods mainly focus on how to estimate calibrated confidence from data or fit a user-defined calibration function (e.g., temperature scaling (Guo et al. 2017) is a scaling function of logit, and platt scaling (Platt et al. 1999) is a sigmoid function of logit), without systematically and principledly utilizing the prior distributions behind the calibration curve. However, in the field of statistics, especially bayesian statistics, the utilization of prior distributions is crucial. In particular, when data size is insufficient, such as in low-density regions of confidence scores, a correct prior distribution can be more informative than the data.

Therefore, a natural but ignored question is studied: how to integrate the prior distribution behind the calibration curve with the empirical data to achieve a better calibration and develop a more accurate calibration metric? To address this, this paper conducts binomial process modeling on the sampling process of calibration data, which cleverly integrates the prior distribution of the calibration curve with the empirical data. By maximizing the likelihood function of the binomial process, a continuous calibration curve can be estimated. A general and effective prior is suggested as the prior distribution behind calibration curves, which is a principled function family derived from beta distributions. We prove that the estimated calibration curve is Lipschitz continuous with respect to data distribution and requires only a sample size of $3/B$ of that required for histogram binning, where B represents the number of bins. Furthermore, using the estimated calibration curve, a new calibration metric is proposed, named TCE_{bpm}. TCE_{bpm} is proved to be a consistent calibration measure (Błasiok et al. 2023). Finally, by modeling the sampling process of the calibration data as a binomial process, we can sample realistic calibration data from a preset true calibration curve and confidence score distribution, which can serve as a benchmark to measure and compare the discrepancy between existing calibration metrics and the true calibration error.

Our contributions can be summarized as follows:

- A new calibration curve estimating method is proposed, which integrates prior distributions behind the calibration curve with empirical data through binomial process modeling. By maximizing the likelihood function of the binomial process, a continuous calibration curve can be estimated. We prove that the new calibration curve estimation method is Lipschitz continuous with respect to the data distribution and requires only a sample size of $3/B$ of that required for histogram binning, where B represents the number of bins.
- A new calibration metric (TCE_{bpm}) is proposed, which leverages the estimated calibration curve to estimate the true calibration error (TCE). Theoretically, TCE_{bpm} is proven to be a consistent calibration measure.
- A realistic calibration data simulation method based on binomial process modeling is proposed, which can serve as a benchmark to measure and compare the discrepancy between existing calibration metrics and the true calibration error.

2 Background and Related Work

For a K-class classification problem, let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be jointly distributed random variables, where $\mathcal{X} \subset R^d$ denotes the feature space and $\mathcal{Y} = \{1, 2, ..., K\}$ is the label space. The classification model can be expressed as $f(X) : \mathcal{X} \to \mathcal{S}$, where $S = (S_1, S_2, ..., S_K) \in \mathcal{S} \subset \Delta_{K-1}$ and Δ_{K-1} represents a simplex with free-degree $K - 1$. The predicted class $\hat{Y} = \underset{k}{\operatorname{argmax}} \{S_k\}_{1 \leq k \leq K}$, and the confidence score of predicted class is $\hat{S} = \max \{S_k\}_{1 \leq k \leq K}$.

Typically, we just care about the confidence of the predicted class. In this case, a multi-classification problem can be formally unified into a binary classification problem. Let the "hit" variable $H = I[Y = \hat{Y}]$, where I is the indicative function, that is, when $Y = \hat{Y}$, $I[Y = \hat{Y}] = 1$, otherwise $I[Y = \hat{Y}] = 0$. Therefore, the data samples become observations of $(\hat{S}, H)$.

2.1 Confidence Calibration

The purpose of confidence calibration is to make the confidence of predicted class match the true posterior probability of the predicted class. Formally, we state:

Definition 1. (Perfect calibration) *A classification model is perfectly calibrated if the following equation is satisfied:*

$$P(Y = \hat{Y} | \hat{S} = \hat{s}) = \hat{s}, \tag{1}$$

where $\hat{s}$ is the observed confidence score of predicted class, $\hat{Y}$ is the predicted class.

Obviously, Eq. 1 can also be written as $P(H = 1 | \hat{S} = \hat{s}) = \hat{s}$. Typically, we call $P(H = 1 | \hat{S})$ as the true calibration curve.

2.2 Estimates of Calibration Curve

Currently, confidence calibration methods can be mainly divided into two groups: **train-time** calibration (Liu et al. 2023; Müller, Kornblith, and Hinton 2019; Fernando and Tsokos 2021) and **post-hoc** calibration (Guo et al. 2017; Kull et al. 2019a; Zhang, Kailkhura, and Han 2020; Rahimi et al. 2020). Train-time calibration usually performs calibration during the training of the classifier by modifying the objective function, which may increase the computational cost of the classification task (Naeini, Cooper, and Hauskrecht 2015) and affect the classification effect (Joy et al. 2023). Post-hoc calibration learns a transformation (referred to as a calibration map) of the trained classifier's predictions on a calibration dataset in a post-hoc manner (Zhang, Kailkhura, and Han 2020), which does not change the weights of the classifier and usually performs simple operations.

Pioneering work along the post-hoc calibration direction can be divided into two subgroups: **binning-based** calibration and **fitting-based** calibration. Binning-based calibration methods divide the confidence scores into multiple bins and estimate the calibrated value using the average accuracy inside the bins. The classic methods include Histogram binning (Zadrozny and Elkan 2001), Bayesian binning (Naeini, Cooper, and Hauskrecht 2015), Mutual-information-maximization-based binning (Patel et al. 2020). Fitting-based calibration methods fit a function on logit or confidence score so that the outcome of the function is calibrated. The classic methods include Platt scaling (Platt et al. 1999), Temperature scaling (Guo et al. 2017), Isotonic regression (Zadrozny and Elkan 2002), Mix-n-Match (Zhang, Kailkhura, and Han 2020).

2.3 Estimates of Calibration Error

True Calibration Error The true calibration error is described as the l_p norm difference between the confidence score of the predicted class and the true likelihood of being correct (Kumar, Liang, and Ma 2019):

$$TCE = (E_{\hat{S}}[|\hat{S} - P(H = 1|\hat{S})|^p])^{\frac{1}{p}}. \quad (2)$$

The true calibration curve $P(H = 1|\hat{S})$ and the distribution of confidence scores $\hat{S} \sim \hat{\mathcal{S}}$ determine the value of TCE. TCE is not computable since the ground truth of $P(H = 1|\hat{S})$ and the true distribution of $\hat{S}$ cannot be obtained in practice. Therefore, statistical methods are needed to estimate $P(H = 1|\hat{S})$ and the distribution of $\hat{\mathcal{S}}$, and then estimate the true calibration error.

Binning-Based Calibration Metrics Binning-based calibration metrics use the average accuracy of each bin to approximate $P(H = 1|\hat{S})$ and use the sample size proportion of the bin to approximate the $\hat{\mathcal{S}}$. Formally, assume that all confidence scores are partitioned into M equally-spaced non-overlapping bins, and the i-th bin is represented by B_i, then the binning-based expected calibration error (ECE_{bin}) is calculated as follows:

$$ECE_{bin} = (\sum_{i=1}^{M} \frac{|B_i|}{N} |\text{acc}(B_i) - \text{conf}(B_i)|^p)^{\frac{1}{p}}, \quad (3)$$

where N represents the total number of samples, $|B_i|$ represents the element count of B_i, $\text{acc}(B_i)$ represents the average accuracy on B_i, and $\text{conf}(B_i)$ represents the average confidence on B_i. Typically, the binning scheme is divided into equal width binning (Guo et al. 2017; Naeini, Cooper, and Hauskrecht 2015) and equal mass binning (Kumar, Liang, and Ma 2019; Zadrozny and Elkan 2001). Recently, $Nixon\ et\ al.$ (Nixon et al. 2019) and $Roelofs\ et\ al.$ (Roelofs et al. 2022) observed that ECE_{bin} with equal mass binning produces more stable calibration effect. ECE_{bin} is sensitive to the binning scheme (Kumar, Liang, and Ma 2019; Nixon et al. 2019). Therefore, some improvements to ECE_{bin} have been proposed. $Ferro$ and $Fricker$ (Ferro and Fricker 2012) and $Brocker$ (Bröcker 2012) propose a debiased estimator, $ECE_{debiased}$, which employs a jackknife technique to estimate the per-bin bias in the standard ECE_{bin}. This bias is then subtracted to estimate the calibration error better. $Roelofs\ et\ al.$ (Roelofs et al. 2022) propose ECE_{sweep}, which introduces the monotonically increasing property of the calibration curve into ECE_{bin}.

Binning-Free Calibration Metrics In recent years, confidence calibration evaluation methods that are not based on binning have also been proposed. $Gupta\ et\ al.$ (Gupta et al. 2020) proposed $KS - error$, which uses the Kolmogorov-Smirnov statistical test to evaluate the calibration error. $Zhang\ et\ al.$ (Zhang, Kailkhura, and Han 2020) and $Blasiok\ et\ al.$ (Blasiok and Nakkiran 2023) propose smoothed Kernel Density Estimation (KDE) methods for evaluating calibration error. $Chidambaram\ et\ al.$ (Chidambaram et al. 2024) smooth the logit and then use the smoothed logit to build calibration metric.

2.4 Combining Prior with Experience

In statistics, integrating prior distribution with experience data to estimate the distribution behind the data is a classic and practical tradition (Zellner 1996; Lavine 1991). Priors refer to initial inference on the form or value of model parameters before observing data. Experience refers to the knowledge a model learns from data. When there is enough data, the model can learn well from experience, and the role of the prior may not be reflected. However, when data size is insufficient, a well-informed prior is often more effective than experience data.

In confidence calibration, most existing calibration methods predominantly focus on how to estimate calibrated confidence from data, fit a user-defined calibration function on logit or confidence score, or use naive fitting (e.g., least square method, minimizes cross-entropy loss) to combine priors (e.g., beta prior (Kull, Silva Filho, and Flach 2017b), dirichlet prior (Kull et al. 2019b)) with experience. However, although fitting a user-defined calibration function may also imply some user-observed priors, such as the choice of function shape, these priors are too empirical, and their universality needs to be considered. In addition, naive fitting is prone to be overly affected by data with larger statistical biases (e.g., sparse data). Therefore, it is necessary to study a principled method that better integrates a well-informed prior distribution with empirical data to estimate the calibration curve.

3 Method

In this section, the following questions are studied: 1) How to introduce priors to estimate calibration curve $P(H = 1|\hat{S})$ better? 2) How to choose an appropriate prior? 3) How to build a calibration metric using the estimated calibration curve? 4) How about the theoretical properties of the proposed method?

In Section 3.1, to solve the first problem, the sampling process of the calibration data is modeled as a binomial process, and then the calibration curve can be estimated by maximizing the likelihood function of this binomial process. In Section 3.2, a general and effective prior function family is suggested. In Section 3.3, a new calibration metric TCE_{bpm} is proposed. Section 3.4 analyzes the theoretical guarantee of the proposed calibration method and metric.

3.1 Estimating Calibration Curve

Binomial Process For any fixed $\hat{S}$, the repeated sampling process of H is a binomial distribution on $N_{\hat{S}}^{pos}$, where $N_{\hat{S}}^{pos} = \sum_{i=1}^{N_{\hat{S}}} H_i^{(\hat{S})}$ represents the number of "hit", $H_i^{(\hat{S})}$ represents the i-th "hit" label at $\hat{S}$, $N_{\hat{S}}$ represents the number of samples at $\hat{S}$. Specifically, $N_{\hat{S}}^{pos} \sim \mathcal{BI}(N_{\hat{S}}, P(H = 1|\hat{S}))$, where $\mathcal{BI}(\cdot)$ represents binomial distribution. Formally, the following equation is satisfied:

$$\begin{aligned} &P(N_{\hat{S}}^{pos}) = \mathcal{BI}(N_{\hat{S}}, P(H=1|\hat{S})) \\ &= C_{N_{\hat{S}}}^{N_{\hat{S}}^{pos}} P(H=1|\hat{S})^{N_{\hat{S}}^{pos}} (1-P(H=1|\hat{S}))^{N_{\hat{S}}-N_{\hat{S}}^{pos}}, \end{aligned} \tag{4}$$

where $C_{N_{\hat{S}}}^{N_{\hat{S}}^{pos}}$ is the binomial coefficient.

Furthermore, for all $\hat{S} \in [0, 1]$, the repeated sampling process of H is a binomial process. Binomial process is a random process (Grimmett and Stirzaker 2020) with binomial distribution in a continuous domain. Specifically, in the binomial process, the distribution of random variable $N_{\hat{S}}^{pos}$ under every point $\hat{S}$ in the continuous domain $[0, 1]$ is a binomial distribution. Since we are interested in $P(H = 1|\hat{S})$, $P(H = 1|\hat{S})$ is modeled as a prior function family $g(\hat{S}; \theta)$. Formally, the following equation is satisfied:

$$\begin{aligned} &P(N_{\hat{S}}^{pos}) = \mathcal{BP}(N_{\hat{S}}, g(\hat{S};\theta)) \\ &= C_{N_{\hat{S}}}^{N_{\hat{S}}^{pos}} g(\hat{S};\theta)^{N_{\hat{S}}^{pos}} (1-g(\hat{S};\theta))^{N_{\hat{S}}-N_{\hat{S}}^{pos}}, \end{aligned} \tag{5}$$

where $\mathcal{BP}(\cdot)$ represents binomial process, and $\hat{S} \in [0, 1]$.

Maximum Likelihood Estimation Our purpose is to estimate the calibration curve $g(\hat{S}; \theta)$. Usually, solving $g(\hat{S}; \theta)$ requires a combination of prior and experience. Prior defines that $g(\hat{S}; \theta)$ follows a function family of fixed parameter form or structure. Experience represents seeking the optimal parameter θ from the data. Maximum likelihood estimation is a classic and effective solution to this problem. Formally, the following equation needs to be solved:

$$\underset{\theta}{\text{Argmax}}\, P[D|g(\hat{S};\theta)], \tag{6}$$

where D is the calibration data set, and $D = \{\hat{s}_i, h_i\}_{1\leq i\leq N}$. Equivalently, $D = \{(\hat{s}_j, N_{\hat{s}_j}, N_{\hat{s}_j}^{pos})\}_{1\leq j\leq N'}$, where N' is the number of sampling locations of $\hat{S}$. Therefore:

$$\begin{aligned} &\underset{\theta}{\text{Argmax}}\, P[D|g(\hat{S};\theta)] \\ &= \underset{\theta}{\text{Argmax}}\, E_{\hat{S}\in D} P[N_{\hat{S}}^{pos}|g(\hat{S};\theta), N_{\hat{S}}] \\ &= \underset{\theta}{\text{Argmax}}\, E_{\hat{S}\in D}[\mathcal{BP}(N_{\hat{S}}, g(\hat{S};\theta))] \\ &= \underset{\theta}{\text{Argmax}} \sum_{\hat{S}\in D} \mathcal{BP}(N_{\hat{S}}, g(\hat{S};\theta))P(\hat{S}). \end{aligned} \tag{7}$$

Therefore, Eq. 7 can be solved just by knowing $P(\hat{S})$. Typically, estimating $P(\hat{S})$ requires discretization methods. In order to make $P(\hat{S})$ as accurate as possible, the idea of bayesian averaging is adopted here. Formally, Eq. 7 becomes that:

$$\begin{aligned} &\underset{\theta}{\text{Argmax}}\, E_B[E_b[\mathcal{BP}(N_{\hat{S}_b}, g(\hat{S}_b;\theta))|b]|B], \\ &\Downarrow \\ &\underset{\theta}{\text{Argmax}} \sum_{B\in\mathcal{B}} P(B) \sum_{b\in B} P(b)\mathcal{BP}(N_{\hat{S}_b}, g(\hat{S}_b;\theta)), \end{aligned} \tag{8}$$

where $B \in \mathcal{B}$ represents B-th binning scheme in D, $\mathcal{B}$ represents the binning scheme space, b represent the b-th bin in B, $\hat{S}_b$ represent the average confidence score in the b-th bin, $N_{\hat{S}_b}$ represent the sample size in the b-th bin. In this paper, a uniform prior for modeling $P(B)$ is used, and $P(b) = |b|/|D|$, where $|\cdot|$ the element count function.

Equivalence Optimization Directly solving Eq. 8 can obtain a feasible solution. However, due to the non-convexity of the likelihood function with respect to $g(\hat{S}; \theta)$ in Eq. 8, the feasible solution may not be the global optimal solution. This non-global optimal solution will affect the estimation of the calibration curve. To solve this problem, an equivalent new optimization problem is proposed whose objective function is convex on $g(\hat{S}; \theta)$, as shown below:

$$\underset{\theta}{\text{Argmin}} \sum_{B\in\mathcal{B}} P(B) \sum_{b\in B} P(b) \cdot e^{(g(\hat{S}_b;\theta)-\frac{N_{\hat{S}_b}^{pos}}{N_{\hat{S}_b}})^2}. \tag{9}$$

The proof of equivalence between Eq. 8 and Eq. 9 is provided in Appendix A.1. In general, Eq. 9 can be solved well using a common optimizer (e.g., Gradient Descent (Andrychowicz et al. 2016), Quasi-Newton Methods (Mokhtari and Ribeiro 2020), Nelder-Mead (Nelder and Mead 1965)).

3.2 Selecting Prior

The appropriate prior is crucial to solving Eq. 9. If the prior function family is not correctly selected, the true calibration curve will not be in the feasible region (i.e., solution space). If the prior function family is correctly selected, a good calibration curve will be found quickly and efficiently. Next, this paper suggests a general and effective prior.

It is well known that the prior distribution of confidence scores can be modeled by beta distribution (Kull, Silva Filho, and Flach 2017b; Roelofs et al. 2022; Kull, Silva Filho, and Flach 2017a). Specifically, $P(\hat{S}|H = 0) \sim \mathcal{B}eta(\alpha_0, \beta_0)$ and $P(\hat{S}|H = 1) \sim \mathcal{B}eta(\alpha_1, \beta_1)$. By Bayes' theorem:

$$\begin{aligned} &P(H=1|\hat{S}) = \frac{P(\hat{S}|H=1)P(H=1)}{P(\hat{S}|H=1)P(H=1)+P(\hat{S}|H=0)P(H=0)} \\ &= \frac{1}{1+\frac{P(\hat{S}|H=0)}{P(\hat{S}|H=1)}\cdot\frac{P(H=0)}{P(H=1)}} \\ &= \frac{1}{1+\frac{beta(\alpha_0,\beta_0)^{-1}\hat{S}^{\alpha_0-1}(1-\hat{S})^{\beta_0-1}}{beta(\alpha_1,\beta_1)^{-1}\hat{S}^{\alpha_1-1}(1-\hat{S})^{\beta_1-1}}\cdot\frac{P(H=0)}{P(H=1)}} \\ &= \frac{1}{1+\frac{beta(\alpha_1,\beta_1)}{beta(\alpha_0,\beta_0)}\cdot\hat{S}^{\alpha_0-\alpha_1}(1-\hat{S})^{\beta_0-\beta_1}\cdot\frac{P(H=0)}{P(H=1)}}, \end{aligned} \tag{10}$$

Algorithm 1: Estimating calibration curve.

Initialize:
$P(D|g) = 0$; $D = \{\hat{s}_i, h_i\}_{1 \le i \le N}^{N}$; $\mathcal{B}$; α; β; c.
for B in $\mathcal{B}$:
Initialize $P(D|g, B) = 0$.
for b in B:
$\hat{S}_{list} = \{\hat{s}_i \in b|D\}$; $H_{list} = \{h_i \in b|D\}$,
$\hat{S}_b = \text{mean}(\hat{S}_{list})$,
$N_{\hat{S}_b}^{pos} = \text{sum}(H_{list})$; $N_{\hat{S}_b} = \text{len}(H_{list})$,
$P_b = e^{(g(\hat{S}_b;\alpha,\beta,c) - \frac{N_{\hat{S}_b}^{pos}}{N_{\hat{S}_b}})^2}$,
$P(D|g, B) = P(D|g, B) + P_b \cdot \frac{|b|}{|D|}$,
$P(D|g) = P(D|g) + P(D|g, B)$,
$\alpha, \beta, c = \underset{\alpha,\beta,c}{\text{argmin}}\, P(D|g)$,
Return $P(H = 1|\hat{S}) = \frac{1}{1+\hat{S}^{-\alpha}(1-\hat{S})^{\beta} \cdot e^{c}}$.

where $beta(\cdot)$ represent beta function, $\frac{beta(\alpha_1,\beta_1)}{beta(\alpha_0,\beta_0)} \cdot \frac{P(H=0)}{P(H=1)}$ is a positive constant independent of $\hat{S}$. Let $\frac{beta(\alpha_1,\beta_1)}{beta(\alpha_0,\beta_0)} \cdot \frac{P(H=0)}{P(H=1)}$ be equal to e^c. In order to maintain the monotonically increasing of the calibration curve (Roelofs et al. 2022; Blasiok and Nakkiran 2023), the following two constraints need to be satisfied:

$$\begin{cases} \alpha_0 - \alpha_1 \le 0, \\ \beta_0 - \beta_1 \ge 0. \end{cases} \tag{11}$$

Therefore, $g(\hat{S}; \theta)$ can be selected as:

$$g(\hat{S}; \alpha, \beta, c) = \frac{1}{1+\hat{S}^{-\alpha}(1-\hat{S})^{\beta} \cdot e^{c}}, \tag{12}$$

where $\hat{S} \in [0, 1]$, $g(\hat{S}; \alpha, \beta, c) \in [0, 1]$, $\alpha \ge 0$, $\beta \ge 0$, $c \in (-\infty, +\infty)$.

To sum up, the computational steps for estimating the calibration curve $P(H = 1|\hat{S})$ are shown in Algorithm 1.

3.3 Estimating TCE

As can be seen from Algorithm 1, the estimated calibration curve is a continuous function. This allows us to estimate the true calibration error (see Eq. 2) instead of just the expected calibration error (see Eq. 3).

According to Eq. 2, the calculation formula of TCE is:

$$TCE = [\int_0^1 |P(H = 1|\hat{S}) - \hat{S}|^p \xi(\hat{S}) d\hat{S}]^{\frac{1}{p}}, \tag{13}$$

where $\xi(\hat{S})$ is the probability density function of $\hat{S}$. Typically, p can be set to 1. Since it is already known that the prior distribution of $\xi(\hat{S})$ is the beta distribution, the parameters of the beta distribution can be estimated using the moment estimation method (Wang and McCallum 2006), as shown in the Estimating TCE part of Algorithm 2. Therefore, TCE can be estimated by computing a definite integral on the interval [0,1], as shown in Algorithm 2.

Algorithm 2: Estimating TCE.

$P(H = 1|\hat{S}) = g(\hat{S}; \theta)$ from Algorithm 1.
$m = \text{mean}(\{\hat{s}_i|\hat{s}_i \in D\})$,
$v = \text{var}(\{\hat{s}_i|\hat{s}_i \in D\})$,
$a_1 = \frac{m^2(1-m)}{v} - m$,
$a_2 = a_1 \cdot \frac{(1-m)}{m}$,
$\xi(\hat{S}) = beta(a_1, a_2)^{-1} \cdot \hat{S}^{(a_1-1)} \cdot (1 - \hat{S})^{a_2-1}$,
Return $TCE_{bpm} = \int_0^1 |P(H = 1|\hat{S}) - \hat{S}| \cdot \xi(\hat{S}) d\hat{S}$.

3.4 Theoretical Guarantee

Continuity Continuity with respect to data distribution is an important property for a calibration method and metric. It tells us whether a slight change in data distribution will lead to a drastic jump in the calibration curve and the calibration metric. Before conducting continuity analysis, a distance measure of the data distributions needs to be defined. It tells us how far away the two distributions are. In this paper, Wasserstein distance is used, as shown in Definition 2. Wasserstein distance measures the minimum cost of transforming one distribution into another.

Definition 2. (Wasserstein distance) *For two data distribution D_1, D_2 over $[0, 1] \times \{0, 1\}$, let Γ be the family of all couplings of distributions D_1 and D_2, Wasserstein distance is defined as follows:*

$$W(D_1, D_2) = \inf_{\gamma \in \Gamma} \underset{(\hat{S}_1,\hat{S}_2)\sim\gamma}{E} [|\frac{N_{\hat{S}_1}^{pos}}{N_{\hat{S}_1}} - \frac{N_{\hat{S}_2}^{pos}}{N_{\hat{S}_2}}| + |\hat{S}_1 - \hat{S}_2|]. \tag{14}$$

Specially, when $N_{\hat{S}_1} = N_{\hat{S}_2} = 1$, then:

$$W(D_1, D_2) = \inf_{\gamma \in \Gamma} \underset{(\hat{S}_1,\hat{S}_2)\sim\gamma}{E} [|H_1 - H_2| + |\hat{S}_1 - \hat{S}_2|]. \tag{15}$$

Next, this paper first proves that $g(\hat{S}; \theta)$ obtained by Algorithm 1 is Lipschitz continuity $w.r.t.$ data distributions, as shown in Theorem 1. Then, this paper proves that TCE_{bpm} is Lipschitz continuity $w.r.t.$ data distribution when certain conditions are met, as shown in Theorem 2. The proofs of Theorem 1 and Theorem 2 are given in Appendix A.3.

Theorem 1. *For two distribution D_1, D_2 over $[0, 1] \times \{0, 1\}$, let Γ be the family of all couplings of distributions D_1 and D_2, and $g(\hat{S}; \theta_D)$ represents the calibration curve learned from D via Eq. 9, then $\forall \gamma \in \Gamma$, it holds that:*

$$\begin{aligned} &\underset{(\hat{S}_1,\hat{S}_2)\sim\gamma}{E} |g(\hat{S}_1; \theta_{D_1}) - g(\hat{S}_2; \theta_{D_2})| \\ &\le L \cdot \underset{(\hat{S}_1,\hat{S}_2)\sim\gamma}{E} [|\frac{N_{\hat{S}_1}^{pos}}{N_{\hat{S}_1}} - \frac{N_{\hat{S}_2}^{pos}}{N_{\hat{S}_2}}| + |\hat{S}_1 - \hat{S}_2|], \end{aligned} \tag{16}$$

where $L \ge 0$. Therefore:

$$\inf_{\gamma \in \Gamma} \underset{(\hat{S}_1,\hat{S}_2)\sim\gamma}{E} |g(\hat{S}_1; \theta_{D_1}) - g(\hat{S}_2; \theta_{D_2})| \le L \cdot W(D_1, D_2). \tag{17}$$

Theorem 2. *$\forall \hat{S} \in [0,1]$, if $g(\hat{S};\theta_D)$ and $\xi_D(\hat{S})$ are Lipschitz continuous $w.r.t.$ D, then for two distribution D_1, D_2 over $[0,1] \times \{0,1\}$, TCE_{bpm} satisfies:*

$$|TCE_{bpm}(D_1) - TCE_{bpm}(D_2)| \leq L \cdot W(D_1, D_2), \quad (18)$$

where $L \geq 0$.

Consistency To theoretically prove the effectiveness of a calibration metric, $Blasiok\ et\ al.$ (Błasiok et al. 2023) have proposed a unified theoretical framework: consistent calibration measure. Consistent calibration measure means two things: 1) When the true distance to calibration is small, the calibration metric should also be small(i.e., Robust completeness); 2) When the true distance to calibration is large, the calibration metric should also be large (i.e., Robust soundness).

Before defining the consistent calibration measure, we need to define how far a data distribution D is from its nearest perfect calibration distribution, as shown in the Definition 3. Then, the consistent calibration measure can be defined as shown in Definition 4.

Theorem 3 proves that TCE_{bpm} is a consistent calibration measure when certain conditions are met. Corollary 1 proves that when $P(\hat{S}|H=0)$ and $P(\hat{S}|H=1)$ follow beta distribution, TCE_{bpm} calculated using Eq. 12 is a consistent calibration measure. The proof of Theorem 3 is given in Appendix A.4, and the proof of Corollary 1 is given in Appendix A.5.

Definition 3. (True distance to calibration) *$\forall D$ over $[0,1] \times \{0,1\}$, let $\mathcal{P}$ be the family of all perfectly calibrated distributions, the true distance to calibration is:*

$$\underline{dCE}(D) = \inf_{\mathcal{D} \in \mathcal{P}} W(D, \mathcal{D}). \quad (19)$$

Definition 4. (Consistent calibration measure) *For $q, t, L_1, L_2 > 0$, calibration metric μ, and data distribution D over $[0,1] \times \{0,1\}$, if:*

$$\mu(D) \leq L_1 \cdot (\underline{dCE}(D))^q, \quad (20)$$

then μ satisfies q-robust completeness. If:

$$\mu(D) \geq L_2 \cdot (\underline{dCE}(D))^t, \quad (21)$$

then μ satisfies t-robust soundness. If satisfying robust completeness and soundness, μ is a consistent calibration measure.

Theorem 3. *TCE_{bpm} is a consistent calibration measure if the following two conditions hold:*

- *The hypothesis set $\mathcal{G}$ (The set of all possible $g(\hat{S};\theta_D)$) includes the true calibration curve;*
- *$\forall \hat{S} \in [0,1]$, if $g(\hat{S};\theta_D) \in \mathcal{G}$ and $\xi_D(\hat{S})$ are Lipschitz continuous $w.r.t.$ D.*

Corollary 1. *$\forall \hat{S} \in [0,1]$, if:*

$$g(\hat{S};\alpha,\beta,c) = \frac{1}{1 + \hat{S}^{-\alpha}(1-\hat{S})^{\beta} \cdot e^c},$$

and $\xi_D(\hat{S})$ are Lipschitz continuous $w.r.t.$ D, and $P(\hat{S}|H=0)$ and $P(\hat{S}|H=1)$ follow beta distribution, TCE_{bpm} is a consistent calibration measure.

Algorithm 3: Simulating dataset with binomial process.

Initialize:
  $P(H=1|\hat{S}) = g(\hat{S};\theta)$; a_1; a_2; N.
Sampling:
  $i = 1$; $D = \{\}$.
  while $i \leq N$**:**
    $\hat{S}$ = Sampling from $Beta(a_1, a_2)$,
    $P(H=1|\hat{S}) = g(\hat{S};\theta)$,
    H = Sampling from $\mathcal{BI}(1, P(H=1|\hat{S}))$,
    Adding $(\hat{S}, H)$ into D.
Return D.

Sample Efficiency Sample efficiency tells us how many samples a method requires to keep the error small enough. According to Hoeffding's inequality, the sample size required for histogram binning is $N \geq \frac{B \cdot \text{In}(1/\delta)}{2\varepsilon^2}$, where B represents the number of bins. Specifically, for the error of the histogram binning to be lower than ε, $\frac{\text{In}(1/\delta)}{2\varepsilon^2}$ samples are required at each bin. Next, we will prove that the sample efficiency of Algorithm 1 is $N \geq \frac{3 \cdot \text{In}(1/\delta)}{2\varepsilon^2}$, as shown in Theorem 4. This is due to the fact that the prior function of the calibration curve has three parameters, which ensures that the calibration curve can be uniquely determined by three distinct observation points. Therefore, in theory, as long as selecting three most representative bins, a good calibration curve can be estimated by Algorithm 1. The proof of Theorem 4 is given in Appendix A.6.

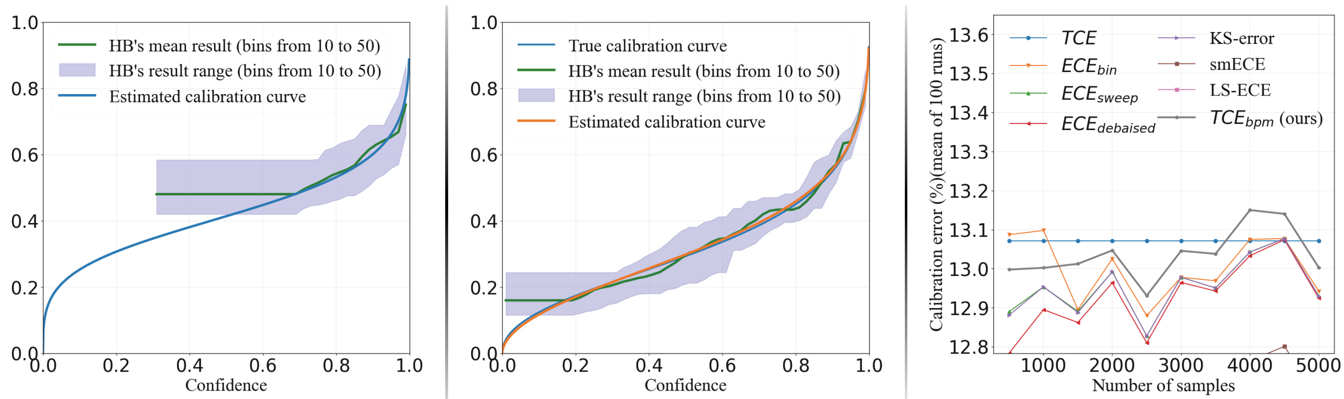
Theorem 4. *For the function family in Eq. 12, if $P(\hat{S}|H=0)$ and $P(\hat{S}|H=1)$ follow beta distribution, then $\forall \delta \in (0,1)$, when $N \geq \frac{3 \cdot \text{In}(1/\delta)}{2\varepsilon^2}$, satisfy the following result with $1-\delta$ probability:*

$$E_{\hat{S}}|g(\hat{S};\alpha_D,\beta_D,c_D) - P(H=1|\hat{S})| \leq \varepsilon.$$

4 Simulating Datasets to Compare Evaluation Metrics

A key challenge in developing calibration metrics is the lack of ground truth for calibration curves and confidence scores, hindering the measurement of discrepancies between metrics and actual calibration errors. $Roelofs\ et\ al.$ (Roelofs et al. 2022) use the fitted function on the publicly available logit datasets as the true distribution behind the data, then use the fitted function to calculate TCE and compare TCE with existing calibration metrics. In this paper, an opposite operation is proposed, i.e., first preseting the true calibration distribution and then obtaining realistic calibration data through binomial process sampling.

Specifically, in Section 3.1, we model the process of sampling calibration data as a binomial process. Another important role of this modeling is that realistic calibration data sets can be sampled through using known calibration curves and confidence distributions, as shown in Algorithm 3. The confidence score $\hat{S}$ are first sampled, and then the calibration value $P(H=1|\hat{S})$ is calculated. Then, $P(H=1|\hat{S})$



(a) ResNet110 + Cifar10 (b) Simulated distribution and data (c) Calibration error comparison

Figure 1: Experimental results of our method. HB represents Histogram binning (Zadrozny and Elkan 2001). In (a), the estimated calibration curve on real data aligns well with histogram binning results from various binning schemes and closely matches the mean result. In (b), the calibration curve estimated by our method closely approximates the true calibration curve in simulated data. In (c), our calibration metric is closest to the true calibration error (TCE) in many times (e.g., when the number of samples is 1500, 2000, 2500, 3000, 3500, and 5000).

is used as the probability of a single event success in the binomial distribution, and binomial distribution sampling is performed to sample H. Since the true calibration curve and confidence distribution are known, TCE can be calculated accurately. The sampled calibration data is then used to calculate other calibration metrics, and by comparing these metrics with the accurately calculated TCE, it can be determined which calibration metrics are better.

5 Results

The effectiveness of the proposed method is verified from four perspectives: 1) On the real-world datasets, the calibration curve estimated by our method is compared with the results of histogram binning under various binning schemes; 2) On the datasets simulated by Algorithm 3, the discrepancy between the calibration curve estimated by our method and the true calibration curve is compared; 3) On the datasets estimated by Algorithm 3, the discrepancy between TCE_{bpm} and the true calibration error is compared; 4) On the real-world datasets, multiple calibration metrics comparison between our calibraiton method with other calibration methods is performed. Due to space limitations, details of data selection and implementation details are given in Appendix B.1 and Appendix B.2. In Appendix B.1, ten publicly available logit datasets (i.e., real-world datasets) and five true distributions (named D1, D2,$\cdots$, and D5, respectively) were selected for the experiments.

5.1 Estimated Results of Calibration Curves

Results in Real Datasets The estimated results of the calibration curve on the public ResNet110's logit dataset trained on Cifar10 is shown in (a) of Fig. 1. The results on other datasets are shown in Appendix B.3. In order to intuitively show the effect of the calibration curve estimated by our method, the means and ranges of the calibration values estimated by the histogram binning under various binning schemes are simultaneously visualized. The calibration curve estimated by our method is close to the mean result of the histogram binning under various binnings and always falls within the result range of histogram binning under various binnings, indicating a relatively accurate and robust performance. In addition, Appendix B.3 shows that our method can achieve such performance under various sharpness, meaning that our method has certain versatility. Furthermore, in the regions of low confidence scores, which are also the regions of low density of confidence scores, the calibration confidence estimated by the binning method fluctuates greatly (e.g., the result range is broad when the confidence score is lower than 0.8 in (a) of Fig. 1) and sometimes even non-monotonic (e.g., (e), (g), and (h) in Fig. 3 of Appendix B.3). The two situations are obviously unreasonable (Kumar, Liang, and Ma 2019; Roelofs et al. 2022). Thanks to the continuous and monotonic prior distribution, the two unreasonable situations can be well avoided by adopting our method.

Results in Simulating Datasets The estimated result of the calibration curve on the dataset simulated by the true distribution D1 is shown in (b) of Fig. 1. The results on other true distributions are shown in Appendix B.4. The calibration curve estimated by our method is closely aligned with the true calibration curve, with a mean absolute error (see Eq. 51 in Appendix B.2) of 0.0099, which is lower than 0.0233 of the mean result of histogram binning. This verifies the effectiveness of our method. In addition, in the regions of low confidence scores (i.e., the low-density regions of confidence scores), the broad result range indicates that the results estimated by the histogram binning method under a specific binning scheme may deviate from the true calibration value. Even the mean result of histogram binning under various binning schemes sometimes deviates significantly from the true calibration curve (e.g., when the confidence score is around 0.8 in D2 of Fig. 4 in Appendix B.4).

Network and Dataset	Calibration methods	$ECE_{bin}\downarrow$	$ECE_{debaised}\downarrow$	$ECE_{sweep}\downarrow$	KS-error↓	smECE↓	$TCE_{bpm}\downarrow$
ResNet110 Cifar10	Uncalibration	0.04755	0.04752	0.04750	0.04750	0.04271	0.05511
	Temperature scaling	0.00745	0.00568	0.00576	0.00590	0.00901	0.00825
	Isotonic regression	0.00753	0.00550	0.00586	0.00601	0.00920	0.00649
	Mix-n-Match	0.00745	0.00574	0.00576	0.00590	0.00901	0.00825
	Spline calibration	0.01322	0.01181	0.00347	0.00430	0.01280	0.01136
	TPM calibration (Ours)	**0.00458**	**0.00312**	**0.00146**	**0.00162**	**0.00756**	**0.00214**

Table 1: Comparison with other calibration methods on real data. Bold represents the best result, and underline represents the second-best result.

However, in our method, benefiting from the excellent integration between the well-informed prior distribution and the empirical data, the calibration curve can be well estimated even in the low-density regions of confidence scores. This good estimation effect makes us believe that fully integrating well-informed prior distribution with empirical data is a promising future direction for confidence calibration.

5.2 Estimated Results of Calibration Metrics

Six state-of-the-art calibration metrics are compared with TCE_{bpm}. The details of these six metrics are shown in Appendix B.2. Among these calibration metrics, the one closer to TCE is more accurate. In (c) of Fig. 1, the calculation results of these metrics on the dataset simulated by the true distribution D1 are shown. The results on other true distributions are shown in Appendix B.4. It can be seen that our calibration metric TCE_{bpm} is closest to the true calibration error (TCE) in many times (e.g., when the number of samples is 1500, 2000, 2500, 3000, 3500, and 5000 in (c) of Fig. 1). This shows that the calibration metrics estimated by our method are competitive. Even though we use a common confidence score estimation method (Moment estimation method (Wang and McCallum 2006)), the estimated TCE_{bpm} is so competitive, which once again verifies that the calibration curve estimated by our method is quite accurate.

5.3 Comparison with Other Calibration Methods

Table 1 shows the comparison results of calibration metrics between our calibration method and other calibration methods on the public ResNet110's logit dataset trained on Cifar10, where the last column is our calibration metric. All considered calibration methods can significantly improve confidence. The calibration error of our method on six metrics is 50.15% less on average than the second place. The comparison results on Wide-ResNet32's logits dataset of Cifar100 and DenseNet162 logits dataset of ImageNet are shown in Appendix B.5.

6 Conclusion and Discussion

In this paper, we focus on how to effectively incorporate prior distributions behind calibration curves with empirical data to calibrate confidence better. To address this, we propose a new calibration curve estimation method via binomial process modeling and maximum likelihood estimation and perform theoretical analysis. Furthermore, using the estimated calibration curve, a new calibration metric is proposed, which is proven to be a consistent calibration measure. In addition, this paper proposes a new simulation method for calibration data through binomial process modeling, which can serve as a benchmark to measure and compare the discrepancy between existing calibration metrics and the true calibration error. Extensive empirical studies on real-world and simulated data support our findings and showcase the effectiveness of our method.

Potential Impact, Limitations and Future Work We explore the impact of the prior distribution behind the calibration curve on the confidence calibration effect. We also provide a solution as a starting point to utilize the prior distribution of the calibration curve, which we believe has the potential to inspire more rich and subsequent works, ultimately leading to improved decision-making in real-world applications, especially for underrepresented populations and safety-critical scenarios. However, our study also has several limitations. First, the Bayesian average strategy is used to construct the likelihood function of the binomial process, which increases computational cost. Although the increased computational cost may be negligible, further reduction in computational cost is promising. Because Theorem 4 tells us that just three most representative bins are selected, a good parameter estimation can be achieved. Future research can investigate how to select the three most representative bins to achieve efficient parameter estimation. Additionally, we only focus on the confidence calibration in the closed-set classification problem. Future work can generalize this to multi-label, open-set, or generative settings.

Acknowledgements

This work was supported in part by the Science and Technology Innovation Program of Hunan Province (Grant No.2024RC1007), in part by the Young Scientists Fund of the National Natural Science Foundation of China (Grant 62303491), and in part by the Project of State Key Laboratory of Precision Manufacturing for Extreme Service Performance (Grant No.ZZYJKT2023-14).

References

Andrychowicz, M.; Denil, M.; Gomez, S.; Hoffman, M. W.; Pfau, D.; Schaul, T.; Shillingford, B.; and De Freitas, N. 2016. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29.

Błasiok, J.; Gopalan, P.; Hu, L.; and Nakkiran, P. 2023. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 1727–1740.

Blasiok, J.; and Nakkiran, P. 2023. Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing. In *The Twelfth International Conference on Learning Representations*.

Bröcker, J. 2012. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, 39: 655–667.

Chidambaram, M.; Lee, H.; McSwiggen, C.; and Rezchikov, S. 2024. How Flawed Is ECE? An Analysis via Logit Smoothing. In *Forty-first International Conference on Machine Learning*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, J.; Jiang, Z.; Pan, D.; Chen, Z.; Guan, Q.; Zhang, H.; Gui, G.; and Gui, W. 2024. A survey on confidence calibration of deep learning under class imbalance data. *Authorea Preprints*.

Fernando, K. R. M.; and Tsokos, C. P. 2021. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7): 2940–2951.

Ferro, C. A.; and Fricker, T. E. 2012. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668): 1954–1960.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6577–6595.

Grimmett, G.; and Stirzaker, D. 2020. *Probability and random processes*. Oxford university press.

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. PMLR.

Gupta, K.; Rahimi, A.; Ajanthan, T.; Mensink, T.; Sminchisescu, C.; and Hartley, R. 2020. Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800*.

Han, Y.; Liu, D.; Shang, J.; Zheng, L.; Zhong, J.; Cao, W.; Sun, H.; and Xie, W. 2024. BALQUE: Batch active learning by querying unstable examples with calibrated confidence. *Pattern Recognition*, 110385.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 646–661. Springer.

Huang, L.; Zhao, J.; Zhu, B.; Chen, H.; and Broucke, S. V. 2020. An experimental investigation of calibration techniques for imbalanced data. *Ieee Access*, 8: 127343–127352.

Jiang, X.; Osl, M.; Kim, J.; and Ohno-Machado, L. 2012. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2): 263–274.

Jiang, Z.; Dong, J.; Pan, D.; Wang, T.; and Gui, W. 2023. A novel intelligent monitoring method for the closing time of the taphole of blast furnace based on two-stage classification. *Engineering Applications of Artificial Intelligence*, 120: 105849.

Joy, T.; Pinto, F.; Lim, S.-N.; Torr, P. H.; and Dokania, P. K. 2023. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14919–14926.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019a. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Kull, M.; Perello-Nieto, M.; Kängsepp, M.; Song, H.; Flach, P.; et al. 2019b. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv preprint arXiv:1910.12656*.

Kull, M.; Silva Filho, T.; and Flach, P. 2017a. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, 623–631. PMLR.

Kull, M.; Silva Filho, T. M.; and Flach, P. 2017b. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11: 5052–5080.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Lavine, M. 1991. Sensitivity in Bayesian statistics: the prior and the likelihood. *Journal of the American Statistical Association*, 86(414): 396–399.

Li, Y.; and Caragea, C. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6316–6329.

Liu, B.; Rony, J.; Galdran, A.; Dolz, J.; and Ben Ayed, I. 2023. Class Adaptive Network Calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16070–16079.

Lu, Z.; Lin, R.; and Hu, H. 2024. Disentangling Modality and Posture Factors: Memory-Attention and Orthogonal Decomposition for Visible-Infrared Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*.

Luo, X.; Wu, J.; Yang, J.; Chen, H.; Li, Z.; Peng, H.; and Zhou, C. 2024. Knowledge Distillation Guided Interpretable Brain Subgraph Neural Networks for Brain Disorder Exploration. *IEEE Transactions on Neural Networks and Learning Systems*.

Mokhtari, A.; and Ribeiro, A. 2020. Stochastic quasi-newton methods. *Proceedings of the IEEE*, 108(11): 1906–1922.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Munir, M. A.; Khan, S. H.; Khan, M. H.; Ali, M.; and Shahbaz Khan, F. 2024. Cal-DETR: Calibrated Detection Transformer. *Advances in Neural Information Processing Systems*, 36.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nelder, J. A.; and Mead, R. 1965. A simplex method for function minimization. *The computer journal*, 7(4): 308–313.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *CVPR workshops*, volume 2.

Patel, K.; Beluch, W. H.; Yang, B.; Pfeiffer, M.; and Zhang, D. 2020. Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning. In *International Conference on Learning Representations*.

Penso, C.; Frenkel, L.; and Goldberger, J. 2024. Confidence Calibration of a Medical Imaging Classification System that is Robust to Label Noise. *IEEE Transactions on Medical Imaging*.

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

Rahimi, A.; Shaban, A.; Cheng, C.-A.; Hartley, R.; and Boots, B. 2020. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33: 13456–13467.

Roelofs, R.; Cain, N.; Shlens, J.; and Mozer, M. C. 2022. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, 4036–4054. PMLR.

Silva Filho, T.; Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; and Flach, P. 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9): 3211–3260.

Wang, M.; Yang, H.; Huang, J.; and Cheng, Q. 2024. Moderate Message Passing Improves Calibration: A Universal Way to Mitigate Confidence Bias in Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21681–21689.

Wang, X.; and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433.

Zadrozny, B.; and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, 609–616.

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

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zellner, A. 1996. Models, prior information, and Bayesian analysis. *Journal of Econometrics*, 75(1): 51–68.

Zhang, J.; Kailkhura, B.; and Han, T. Y.-J. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, 11117–11128. PMLR.

Zhang, X.-Y.; Xie, G.-S.; Li, X.; Mei, T.; and Liu, C.-L. 2023. A survey on learning to reject. *Proceedings of the IEEE*, 111(2): 185–215.