TCE: A Test-Based Approach to Measuring Calibration Error

Abstract

This paper proposes a new metric to measure the calibration error of probabilistic binary classifiers, called *test-based calibration error* (TCE). TCE incorporates a novel loss function based on a statistical test to examine the extent to which model predictions differ from probabilities estimated from data. It offers (i) a clear interpretation, (ii) a consistent scale that is unaffected by class imbalance, and (iii) an enhanced visual representation alternative to the standard reliability diagram. In addition, we introduce an optimality criterion for the binning procedure of calibration error metrics based on a minimal estimation error of the empirical probabilities. We provide a new computational algorithm for optimal bins under bin-size constraints. We demonstrate properties of TCE through a range of experiments, including multiple real-world imbalanced datasets and ImageNet 1000.

1 INTRODUCTION

In recent years, it has become ubiquitous to deploy complex machine learning models in real-world production systems. Many of these systems rely on probabilistic classifiers that predict the probability that some target outcome occurs. For such systems, it is often crucial that their predictive probabilities are *well-calibrated*, meaning that the predicted probability accurately reflects the true frequency that the outcome occurs. In some contexts, failures to achieve calibration can lead to negative consequences. In applications like medical diagnoses [Topol, 2019] and autonomous driving [Grigorescu et al., 2020], associated risks are often assessed based on model predictions and the consequences of wrong decisions can be severe. In online advertising auctions [Li et al., 2015], it is common to incorporate a prediction of the probability of some outcome of interest (e.g., a click on the advert) when calculating an advertiser's bid.

While a number of metrics—such as log-likelihood, user-specified scoring functions, and the area under the receiver operating characteristic (ROC) curve—are used to assess the quality of probabilistic classifiers, it is usually hard or even impossible to gauge whether predictions are well-calibrated from the values of these metrics. For assessment of calibration, it is typically necessary to use a metric that measures *calibration error*, that is, a deviation between model predictions and probabilities of target occurrences estimated from data. The importance of assessing calibration error has been long emphasised in machine learning [Nixon et al., 2019, Minderer et al., 2021] and for probabilistic forecasting more broadly [Dawid, 1982, Degroot and Fienberg, 1983].

However, existing metrics of calibration error have several drawbacks that in certain scenarios can mean that their values do not appropriately reflect true calibration performance. In particular, we will demonstrate that values of existing calibration error metrics have an inconsistent scale that is influenced by the target class proportion. In applications such as fraud detection [Abdallah et al., 2016, Tax et al., 2021] and advertising conversion prediction [Yang and Zhai, 2022], the prevalence, i.e. the proportion of instances belonging to the target class, is often very low. This leads to situations where one may be unable to identify whether the values of calibration error metrics are small due to a good calibration performance or due to the low prevalence. This is also problematic for monitoring applications aimed at tracking the calibration performance of a model in a production system; prevalence can change over time (i.e., *prior probability shift* [Storkey et al., 2009]), and that makes it difficult to understand whether to attribute changes in the metric to an actual change in calibration performance or to the change in prevalence.

Furthermore, the binning of model predictions—an essential component of any calibration error metric [Naeini et al., 2015]—is often based on heuristics and lacks clear design principles. For calibration error metrics, empirical probab-

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

ilities of the target occurrence are typically estimated by clustering instances into several subsets based on the associated model predictions. The design of this binning is an important factor in the accurate measurement of calibration error, yet no clear design principle has yet to emerge.

In this paper, we elaborate on the issues of existing calibration error metrics in Section 2 and we establish a simple yet novel metric that counterbalances these issues in Section 3. Section 4 demonstrates properties of the proposed metric. Related works are discussed in Section 5, followed by the conclusion in Section 6. This paper focuses on the methodological aspects of the proposed new design of calibration error metrics for binary classification, while theoretical development is left for future research. Our methodological contributions are summarised below:

Contributions

- Our primary contribution is a novel calibration error metric called *test-based calibration error* (TCE) that is based on statistical hypothesis testing and is interpretable as a percentage of model predictions that deviate significantly from estimated empirical probabilities. The value of TCE is normalised to $[0, 100]$ and the metric produces values of comparable range regardless of prevalence.
- We propose a new visual representation of TCE that carries more information than the standard *reliability diagram* and facilitates a better understanding of calibration performance (See Figure 1).
- We formulate a rigorous optimality condition for model prediction binning that is used in calibration error metrics, defining optimal bins as bins that minimise an estimation error of the empirical probabilities.
- We further propose a novel algorithm to compute optimal bins using regularisation based on minimum and maximum bin sizes.

2 BACKGROUND

In this section, we introduce the definition of *calibration* and recap one of the most common *calibration error* metrics. We then outline several critical challenges of existing calibration error metrics. The basic notation used in this paper is introduced below.

$\mathcal{X}$ and $\mathcal{Y}$ respectively denote input and output spaces. We focus on probabilistic binary classification, i.e. $\mathcal{Y} = \{0, 1\}$, in which a probabilistic classifier $P_\theta : \mathcal{X} \to [0, 1]$ models a conditional probability of $Y = 1$ given an input $x \in \mathcal{X}$. The data $\mathcal{D} := \{x_i, y_i\}_{i=1}^N$ are assumed to be i.i.d. realisations from a random variable $(X, Y) \sim \mathbb{P}$. To simplify notation, for any data subset $\mathcal{S} \subseteq \mathcal{D}$, we denote by $\mathcal{S}^x$ a set of all inputs x in $\mathcal{S}$ and by $\mathcal{S}^y$ a set of all outputs y in $\mathcal{S}$. By "a set of bins" or simply "bins", we mean a set of arbitrary disjoint intervals whose union is the unit interval $[0, 1]$. For example, a set $\{\Delta_b\}_{b=1}^B$ of intervals $\Delta_1 = [0.0, 0.4)$ and $\Delta_2 = [0.4, 1.0]$ is a set of bins with $B = 2$.

2.1 CALIBRATION ERROR

A probabilistic classifier $P_\theta : \mathcal{X} \to [0, 1]$ is said to be *calibrated* [Dawid, 1982, Bröcker, 2009] if

$$\mathbb{P}(Y = 1 \mid P_\theta(X) = Q) = Q \tag{1}$$

for all $Q \in [0, 1]$ s.t. the conditional probability is well-defined. Informally, this criterion implies that the model prediction coincides with the actual probability of $Y = 1$ for all inputs. Any deviation between the actual probabilities and the model predictions in eq. (1) is often referred to as *calibration error*, which quantifies to what degree the classifier P_θ is calibrated. The empirical computation of such a deviation involves estimating conditional probability $\mathbb{P}(Y = 1 | P_\theta(X) = Q)$ from data. For given bins $\{\Delta_b\}_{b=1}^B$, define disjoint subsets $\{\mathcal{D}_b\}_{b=1}^B$ of data $\mathcal{D}$ by

$$\mathcal{D}_b := \{(x_i, y_i) \in \mathcal{D} \mid P_\theta(x_i) \in \Delta_b\}. \tag{2}$$

Simply put, $\mathcal{D}_b$ is a subset of data whose model predictions have similar values. The conditional probability $\mathbb{P}(Y = 1 \mid P_\theta(X) = Q)$ for any $Q \in \Delta_b$ can then be estimated by the empirical mean of the output data in subset $\mathcal{D}_b$:

$$\mathbb{P}(Y = 1 \mid P_\theta(X) = Q) \approx \widehat{P}_b := \frac{1}{N_b} \sum_{y_i \in \mathcal{D}_b^y} y_i \tag{3}$$

where we denote by $\widehat{P}_b$ the estimated conditional probability in $\mathcal{D}_b$ and by N_b the sample size of $\mathcal{D}_b$.

One of the most common metrics to measure calibration error is expected calibration error (ECE) [Naeini et al., 2015]. ECE uses equispaced bins $\{\Delta_b\}_{b=1}^B$ over $[0, 1]$ for a given number B and measures an absolute difference between the averaged model predictions and the estimated conditional probability $\widehat{P}_b$ within each data subset $\mathcal{D}_b$. The value of ECE is defined as:

$$\text{ECE} := \sum_{b=1}^{B} \frac{N_b}{N} \left| \widehat{P}_b - \frac{1}{N_b} \sum_{x_i \in \mathcal{D}_b^x} P_\theta(x_i) \right|. \tag{4}$$

ECE has a practical visual representation by the *reliability diagram* [Degroot and Fienberg, 1983, Niculescu-Mizil and Caruana, 2005], which aligns the averaged model prediction and the estimated conditional probability in each $\mathcal{D}_b$ (see Figure 1). The reliability diagram is a powerful tool to intuitively grasp the deviation between the model and the estimated probability in ECE.

2.2 CHALLENGES IN CALIBRATION ERROR

Calibration error metrics, such as ECE are widely used in real-world applications. This section points out that

there nonetheless exist several challenges that may cause a misassessment of calibration. These problems become evident especially when a distribution of model predictions $\{P_\theta(x_i)\}_{i=1}^N$ is not well-dispersed. This scenario often arises in imbalanced classification settings where model predictions tend to be severely skewed towards either 0 or 1. The following paragraphs illustrate challenges of existing calibration error metrics, which we aim to address.

Challenge 1: Scale-Dependent Interpretation In most calibration error metrics, the deviation between the model prediction and the estimated probability $\widehat{P}_b$ in each $\mathcal{D}_b$ is measured by the absolute difference as in eq. (4). However, the use of absolute differences can result in values that have an inconsistent scale influenced by the prevalence. To illustrate this problem, consider an estimated probability $\widehat{P}_b$ and an averaged model prediction, denoted by $\overline{Q}_b$, for some fixed b in eq. (4). If $\widehat{P}_b = 0.50$ and $\overline{Q}_b = 0.49$, their absolute difference is 0.01. On the other hand, if $\widehat{P}_b = 0.01$ and $\overline{Q}_b = 0.0001$, their absolute difference is 0.0099. The comparison under the absolute difference suggests that the probability $\overline{Q}_b = 0.0001$ with respect to $\widehat{P}_b = 0.01$ in the latter case is better calibrated than in the former case. However, one may reasonably argue that the latter is not well-calibrated given the stark difference in the order of magnitude. As illustrated, the values of existing calibration metrics built upon the absolute difference can be proportionally small whenever the scales of $\widehat{P}_b$ and $\overline{Q}_b$ are small. This issue makes it difficult to distinguish whether the metric values are low due to a good calibration performance or due to the small scale of the probabilities as in imbalanced classification.

Challenge 2: Lack of Normalised Range Calibration error metrics that are built on the absolute difference do not have a normalised value range. Their value range can vary depending on the choice of the bins $\{\Delta_b\}_{b=1}^B$. To illustrate the problem, consider a bin Δ_b for some fixed b. If $\Delta_b = [0.4, 0.6]$, the absolute difference between $\widehat{P}_b$ and $\overline{Q}_b$ has a value range $[0.0, 0.6]$ in theory because $\widehat{P}_b$ is the estimated probability in $[0.0, 1.0]$ and the averaged model prediction $\overline{Q}_b$ in the bin $[0.4, 0.6]$ takes its value within $[0.4, 0.6]$. Similarly, a different choice of bin Δ_b leads to a different range of absolute difference. Consequently, the final value range of calibration error metrics is also impacted by the choice of the bins $\{\Delta_b\}_{b=1}^B$. To assure rigorous comparability of the final value of the metric, it is desirable to establish a measurement of the deviation whose value falls into a fixed, normalised range independent of the choice of the bins $\{\Delta_b\}_{b=1}^B$.

Challenge 3: Arbitrary Choice of Bins An appropriate choice of bins is critical in calibration because it meaningfully affects the values of calibration error metrics. Equispaced bins $\{\Delta_b\}_{b=1}^B$ over $[0, 1]$ for a given number B are one of the most common choices in practice, as used in ECE. However, it can often happen that a small number of bins contain the majority of the model predictions when the predictions concentrate only in some local domain. For a bin $\mathcal{D}_b$ that contains the majority of data, the estimated probability $\widehat{P}_b$ simply coincides with the empirical mean of all output data. In contrast, for other bins with a small sample size, the estimated probability $\widehat{P}_b$ becomes unreliable due to the small sample size. A potential solution to this problem is to use bins that adapt based on the model predictions. Nixon et al. [2019] proposed the *adaptive calibration error* (ACE) that computes the same value as eq. (4) using bins based on B-quantiles of the model predictions $\{P_\theta(x_i)\}_{i=1}^N$ for a given B. However, questions remain regarding the optimal number of bins B and the appropriate quantile to use for each bin. To the best of our knowledge, there is no established notion of what makes a bin optimal, nor do clear design principles for bins exist.

3 CALIBRATION ERROR BASED ON TEST AND OPTIMAL BINS

We propose a new calibration error metric that offers a simple yet novel solution to the challenges outlined in Section 2.2. First, in Section 3.1, we present a general formulation of calibration error metrics that encompasses most metrics used in practice. This general formulation allows for a structured understanding of the design of calibration error metrics. In Section 3.2, we derive from the general formulation a new calibration error metric, called TCE, which incorporates a loss based on a statistical test to compare model predictions with estimated empirical probabilities. TCE produces a value that has a clear interpretation as a percentage of model predictions determined to deviate significantly from estimated empirical probabilities, which leads to a normalised range of possible values $[0, 100]$ regardless of the choice of bins $\{\Delta_b\}_{b=1}^B$. In Section 3.3, we consider an optimal criterion of bins $\{\Delta_b\}_{b=1}^B$ from the perspective of minimising an estimation error of the empirical probabilities $\{\widehat{P}_b\}_{b=1}^B$. We further develop a practical regularisation approach to the computation of bins that ensures a minimum and maximum sample size in each subset $\mathcal{D}_b$.

3.1 GENERAL CALIBRATION ERROR

The following definition presents an abstract formulation of calibration error metrics, which we call *general calibration error* (GCE) for terminological convenience. Denote by $2^{\mathcal{D}}$ a power set of $\mathcal{D}$, i.e. a space of all subsets of $\mathcal{D}$ and by $\mathcal{M}$ a space of all probabilistic classifiers below.

Definition 1. (GCE) *Let $L : 2^{\mathcal{D}} \times \mathcal{M} \to \mathbb{R}$ be a loss of any probabilistic classifier evaluated for any data subset. Let $\mathcal{B}$ be a set of bins $\{\Delta_b\}_{b=1}^B$ that define data subsets $\{\mathcal{D}_b\}_{b=1}^B$ as in eq. (2). Let $\|\cdot\|$ be a norm of a B-dimensional vector*

space. For a given probabilistic classifier $P_\theta : \mathcal{X} \to [0,1]$, define a scalar $GCE_b \in \mathbb{R}$ for each $b = 1, \cdots, B$ by

$$GCE_b := L\left(\mathcal{D}_b, P_\theta\right). \tag{5}$$

Then, GCE of the probabilistic classifier P_θ is defined by

$$GCE = \|(GCE_1, \cdots, GCE_B)\|. \tag{6}$$

This formulation translates the problem of designing a calibration error metric into a problem of choosing the tuple $(L, \mathcal{B}, \|\cdot\|)$. Most existing calibration error metrics used in practice can be derived by selecting an appropriate tuple of the loss L, the bins $\mathcal{B}$, and the norm $\|\cdot\|$ in GCE. See Example 1 below for the case of ECE. It is also immediate to show that ACE can be recovered from GCE.

Example 1. *Let $\mathcal{B}$ be equispaced bins $\{\Delta_b\}_{b=1}^B$ over $[0,1]$, let L be $L(\mathcal{D}_b, P_\theta) = |\frac{1}{N_b}\sum_{y\in\mathcal{D}_b^y} y - \frac{1}{N_b}\sum_{x\in\mathcal{D}_b^x} P_\theta(x)|$, and let $\|\cdot\|$ be a weighted 1-norm $\|v\| = \sum_{b=1}^B \frac{N_b}{N} \times |v_b|$. The ECE corresponds to the GCE under this tuple.*

We aim to choose the tuple $(L, \mathcal{B}, \|\cdot\|)$ so that it addresses the aforementioned challenges in Section 2.2. Section 3.2 addresses a choice of the loss L based on a statistical test and presents TCE resulting from the use of the proposed L. Subsequently, Section 3.3 addresses a choice of bins $\mathcal{B}$ that is obtained through optimisation to minimise an estimation error of the empirical probabilities $\{\widehat{P}_b\}_{b=1}^B$. All norms $\|\cdot\|$ are equivalent in finite dimensions, and hence we do not focus on any particular choice. As with ECE, we use the weighted 1-norm $\|\cdot\|$ in Example 1 for TCE.

3.2 TEST-BASED CALIBRATION ERRORS

We present our main contribution, a new calibration error metric TCE, derived from GCE by specifying a novel loss L based on a statistical test. Our proposed loss L summarises the percentage of model predictions that deviate significantly from the empirical probabilities in each subset $\mathcal{D}_b$. We effectively test a null hypothesis "a probability of $Y = 1$ is equal to $P_\theta(x)$" at each $x \in \mathcal{D}_b^x$ using the output data $\mathcal{D}_b^y$. A rigorous formulation of this loss L is provided below, combined with the definition of the TCE. Note that the bins $\{\Delta_b\}_{b=1}^B$ and the norm $\|\cdot\|$ of TCE are arbitrary, while the weighted 1-norm is our default choice of $\|\cdot\|$.

Definition 2. (TCE) *Given a statistical test and its significance level $\alpha \in [0,1]$, let R be a function of any observed dataset of random variable $Y \in \{0,1\}$ and any probability $Q \in [0,1]$, which returns 1 if a hypothesis $P(Y=1) = Q$ is rejected based on the dataset and returns 0 otherwise. In Definition 1, let L be an average rejection percentage s.t.*

$$L(\mathcal{D}_b, P_\theta) = 100 \times \frac{1}{N_b}\sum_{x\in\mathcal{D}_b^x} R\left(\mathcal{D}_b^y, P_\theta(x)\right). \tag{7}$$

GCE in Definition 1 is then called TCE.

In contrast to existing metrics that examine the difference between averaged model predictions and empirical probabilities in each bin, TCE examines each prediction $P_\theta(x)$ and summarises their rejection percentage in each bin, where a test is performed for every prediction in each data subset $\mathcal{D}_b^y$. Our interest is in the gross percentage of rejections, and so multiple testing corrections are not required. In this paper, we use the Binomial test as the *de facto* standard of the statistical test to define R in the TCE. TCE based on other tests, including Bayesian testing approaches, is an open direction for future research. Algorithm 1 summarises a computational procedure to compute TCE. There are multiple advantages of TCE as follows.

Advantage 1: Interpretation The final value of TCE has a clear interpretation as a percentage of model predictions that are determined by the test of choice (here a Binomial test) to deviate significantly from estimated empirical probabilities. The value range is normalised to $[0, 100]$ because it is a percentage.

Advantage 2: Consistent Scale With an appropriate choice of test, TCE evaluates the statistical deviation of data from a model prediction $P_\theta(x)$ adaptively and appropriately for each scale of $P_\theta(x)$ and data size. Therefore TCE produces values that have a consistent scale unaffected by class imbalance.

Advantage 3: Visualisation TCE leads to a new visual representation that shows the distribution of model predictions and the proportion of predictions that deviate significantly from the empirical probability in each bin. See Figure 1 for the description and comparison with the standard reliability diagram.

Given that TCE is based on a statistical testing procedure, it may be possible to apply ideas from power analysis to inform the desired sample size in each $\mathcal{D}_b$. While this is not considered in this paper, there exists room to investigate such an approach.

Algorithm 1 Computation of TCE

Require: data $\mathcal{D}$, model P_θ, norm $\|\cdot\|$, bins $\{\Delta_b\}_{b=1}^B$, function R based on a chosen test and significant level
Ensure: a value TCE $\in \mathbb{R}$
  for $b = 1, \ldots, B$ **do**
    $\mathcal{D}_b \leftarrow \{(x_i, y_i) \in \mathcal{D} \mid P_\theta(x_i) \in \Delta_b\}$ ▷ make subset
    $\text{TCE}_b \leftarrow 0$
    for $x_i \in \mathcal{D}_b^x$ **do**
      $\text{TCE}_b \leftarrow \text{TCE}_b + R(\mathcal{D}_b^y, P_\theta(x_i))$ ▷ test each
    end for
    $\text{TCE}_b \leftarrow 100/N_b \times \text{TCE}_b$
  end for
  $\text{TCE} \leftarrow \|(\text{TCE}_1, \ldots, \text{TCE}_B)\|$

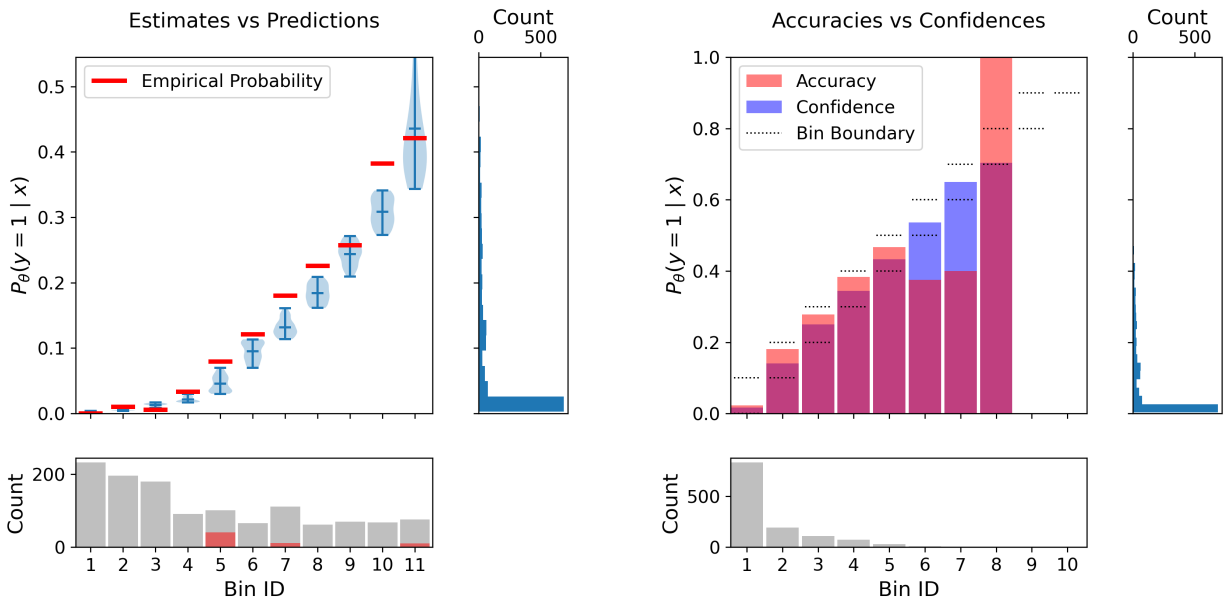

Figure 1: Comparison of two visual representations applied for the same gradient boosting model trained on the *abalone* dataset. (Left) A new visual representation, which we call the *test-based reliability diagram*. The central plot shows a violin plot of model predictions in each bin, whose estimated probability is presented by a red line. The bottom plot shows by grey bar the sample size of each bin and by red bar the percentage of model predictions that deviate significantly from the estimated probability in each bin. The right plot shows a histogram of all model predictions. (Right) The standard *reliability diagram* with the bin-size plot on the bottom and the histogram plot on the right added for comparison.

3.3 OPTIMAL BINS BY MONOTONIC REGRESSOR AND BIN-SIZE CONSTRAINTS

Designing a mechanism to determine how to select the binning scheme used is a fundamental challenge in calibration. We consider an optimal criterion of bins $\{\Delta_b\}_{b=1}^B$ from the perspective of minimising an estimation error of the probability estimates $\{\widehat{P}_b\}_{b=1}^B$ from data $\mathcal{D}$. Recently, Dimitriadis et al. [2021] pointed out that an existing quadratic programming algorithm called pool-adjacent-violators algorithm (PAVA) can be directly applied to compute "optimal" bins in the context of obtaining a better reliability diagram, where the optimality can be associated with the *Brier score* [Brier, 1950]. Moving forward from this observation, we formulate a rigorous definition of the optimal bins under a given measurement of the estimation error of $\{\widehat{P}_b\}_{b=1}^B$.

Definition 3. (Optimal Bins) *Let Π be a space of all sets of bins $\{\Delta_b\}_{b=1}^B$ for any B, with associated data subsets $\{\mathcal{D}_b\}_{b=1}^B$ and probability estimates $\{\widehat{P}_b\}_{b=1}^B$ from $\{\mathcal{D}_b^y\}_{b=1}^B$. Let D be a function to measure a deviation of an observed dataset of random variable $Y \in \{0, 1\}$ around a given probability $Q \in [0, 1]$. Any set of bins that satisfies*

$$\min_{\{\Delta_b\}_{b=1}^B \in \Pi} \sum_{b=1}^{B} W_b \times \mathrm{D}(\mathcal{D}_b^y, \widehat{P}_b) \quad \text{subject to } \widehat{P}_1 \leq \cdots \leq \widehat{P}_B \qquad (8)$$

can be considered an optimal set of bins under the estimation error D, where $W_b := N_b/N$ is the weight associated with each error, which is determined by the bin size.

The monotonic constraint $\widehat{P}_1 \leq \cdots \leq \widehat{P}_B$ of the probability estimates $\{\widehat{P}_b\}_{b=1}^B$ is a natural requirement because the choice of bins becomes trivial otherwise. For example, consider bins $\{\Delta_b\}_{b=1}^B$ with $B = N$ such that Δ_b contains one single point y_b and the probability estimate $\widehat{P}_b = y_b$ for each b. This clearly achieves that $\sum_{b=1}^B W_b \times \mathrm{D}(\mathcal{D}_b^y, \widehat{P}_b) = \frac{1}{N}\sum_{b=1}^N \mathrm{D}(\{y_b\}, y_b) = 0$. Under the monotonic constraint, the choice of bins becomes non-trivial.

For some choices of the estimation error D, the optimisation eq. (8) can be solved by applying a monotonic regression algorithm directly. In a monotonic regression problem, given an ordered dataset $\{y_i\}_{i=1}^N$, the solver algorithm finds N monotonically increasing values $\widehat{y}_1 \leq \cdots \leq \widehat{y}_N$ that minimise some loss between $\{\widehat{y}_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$. There exist algorithms for various loss types, including the l_p loss, the Huber loss, and the Chebyshev loss [de Leeuw et al., 2009]. PAVA is an algorithm to solve a monotonic regression problem under the squared error $\sum_{i=1}^N (\widehat{y}_i - y_i)^2$. If we choose a variance of each $\mathcal{D}_b^y$ as the measurement D, we can compute the optimal bins using PAVA, which corresponds to the case of Dimitriadis et al. [2021]. For clarity, we provide an explicit algorithm to compute the optimal bins using the squared error in Algorithm 2. See Supplement A for the proposition that the bins computed by Algorithm 2 indeed solves the optimisation eq. (8) under variance D.

Algorithm 2 Computation of Optimal Bins with Minimal Total Bin Variance

Input: data $\mathcal{D}$, model P_θ
Output: optimal bins $\{\Delta_b^*\}_{b=1}^B$ under variance
$\{y_i\}_{i=1}^N \leftarrow \text{Sort}(\mathcal{D}, P_\theta)$
$\{\widehat{y}_i\}_{i=1}^N \leftarrow \text{PAVA}(\{y_i\}_{i=1}^N)$
$B^* \leftarrow 1$
$T \leftarrow 0$
for $i = 2, \ldots, N$ **do**
  if $\widehat{y}_{i-1} \neq \widehat{y}_i$ **then**
    $\Delta_{B^*}^* \leftarrow [T, (P_\theta(x_{i-1}) + P_\theta(x_i))/2)$
    $B^* \leftarrow B^* + 1$
    $T \leftarrow (P_\theta(x_{i-1}) + P_\theta(x_i))/2$
  end if
end for
$\Delta_{B^*}^* \leftarrow [T, 1.0]$

Algorithm 2 is a highly appealing solution to the design of the bins in calibration. It achieves a fully-automated design of the bins $\{\Delta_b\}_{b=1}^B$ based on the clear criterion of eq. (8). However, such a fully-automated design can occasionally generate a bin that contains an excessively small or large number of data for the sake of minimising the aggregated estimation error over all $\{\widehat{P}_b\}_{b=1}^B$. For any calibration error metric, imposing a certain minimum and maximum size of each $\mathcal{D}_b$ assists in keeping some baseline quality of the estimation of each individual $\widehat{P}_b$.

Therefore, we propose a modified version of PAVA that regularises based on a minimum and maximum size of each subset $\mathcal{D}_b$. Algorithm 1 in Supplement B summarises the full algorithm, which we call *PAVA with block constraints* (PAVA-BC). By replacing PAVA with PAVA-BC in Algorithm 2, we can obtain bins that satisfy the given minimum and maximum number constraints in each $\mathcal{D}_b$, while benefitting from the automated bin design by PAVA. In general, the introduction of regularisation can cause a mild violation of the monotonic constraint $\widehat{P}_1 \leq \cdots \leq \widehat{P}_B$, where there can exist a few b whose corresponding $\widehat{P}_b$ is smaller than $\widehat{P}_{b-1}$ and the monotonicity is satisfied except for such b. In practice, a mild violation of the monotonic constraint can often be a reasonable cost to achieve better properties of the bins. Tibshirani et al. [2011] studied settings where the monotonic constraint is only "nearly" satisfied.

See Figure 2 for a comparison of the bins computed by three different approaches. It can be seen that the bins produced by PAVA-BC interpolate between the optimal bins produced by PAVA and bins based on quantile that ensure each bin is same size. For each case based on PAVA, PAVA-BC, and quantiles, the total bin variance in eq. (8) is 0.040, 0.042, and 0.048 respectively in Figure 2. On the other hand, the variance within each bin is, in average, 0.132, 0.077, and 0.047 for each case. The bins based on PAVA-BC takes a balance between the total and individual bin variance.

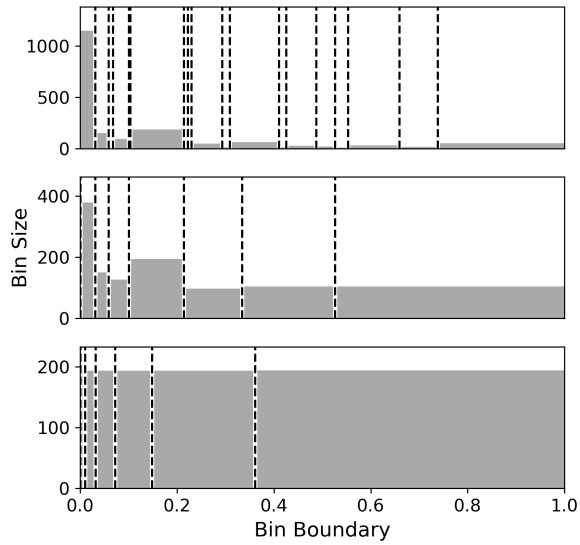

Figure 2: Comparison of three different sets of bins based on (top) PAVA, (middle) PAVA-BC, (bottom) 10-quantiles in the same manner as ACE. The dotted line and grey bar represent each bin boundary over $[0, 1]$ and bin size.

4 EMPIRICAL EVALUATION

In this section, we demonstrate the properties of TCE via three experiments. The first experiment uses synthetic data to demonstrate the properties of TCE under controlled class imbalance. The second experiment involves ten real-world datasets from the University of California Irvine (UCI) machine learning repository [Dua and Graff, 2017], where nine are designed as benchmark tasks of imbalanced classification, and one is a well-balanced classification task for comparison. In the second experiment, we also demonstrate that under class imbalance ECE and ACE may produce misleading assessments of calibration. Given that ECE and ACE are frequently used to evaluate models under class imbalance, TCE has the potential to reduce such misinterpretation risks. The final experiment uses the ImageNet1000 dataset to illustrate that TCE is applicable to large-scale real-world settings.

We compute TCE with bins based on PAVA-BC unless otherwise stated. The minimum and maximum size of each bin for PAVA-BC are set to $N/20$ and $N/5$ for a given dataset size N. With these constraints, the number of bins produced by PAVA-BC falls into a range between 5 and 20. In addition to ECE and ACE, we include the maximum calibration error (MCE) [Naeini et al., 2015] for comparison. MCE is defined by replacing the weighted 1-norm with the supremum norm over $b = 1, \ldots, B$ in Example 1. For any metrics, B-equispaced bins and B-quantiles bins are computed with $B = 10$.

Table 1: Calibration errors by four different metrics. TCE(Q) denotes TCE with bins based on B-quantiles in the same way as the ACE instead of bins based on PAVA-BC.

Prevalence	TCE	TCE(Q)	ECE	ACE
50% vs 50%	7.28%	10.88%	0.0138	0.0150
50% vs 40%	96.10%	96.47%	0.0963	0.0951
50% vs 60%	98.83%	98.93%	0.1097	0.1096
1% vs 1%	3.40%	0.18%	0.0017	0.0031
1% vs 0%	95.50%	68.73%	0.0094	0.0094
1% vs 2%	92.32%	89.73%	0.0139	0.0139

4.1 SYNTHETIC DATA WITH CONTROLLED CLASS IMBALANCE

We first examine TCE using synthetic data from a simulation model considered in Vaicenavicius et al. [2019]. The data are simulated from a Gaussian discriminant analysis model $(x, y) \sim P(x \mid y)P(y)$. Specifically, the output $y \in \{0, 1\}$ is first sampled from a Bernoulli distribution $P(y)$ with parameter π and the input $x \in \mathbb{R}$ is then sampled from a Gaussian distribution $P(x \mid y) = \mathcal{N}(m_y, s_y)$ with mean m_y and scale s_y dependent of y. We set $m_y = (2 \times y - 1)$ and $s_y = 2$, and change the value π for each setting below. By Bayes' theorem, the probability of y for each given x corresponds to a logistic model $P(y \mid x) = 1/(1+\exp(\beta_0+\beta_1 \times x))$ where $\beta_0 = \log(\pi/(1-\pi))$ and $\beta_1 = 4$. A logistic model is therefore capable of reproducing the conditional probability $P(y \mid x)$ of this synthetic data perfectly.

We consider two baseline cases of (i) well-balanced classification and (ii) imbalanced classification in this experiment. We train a logistic model for the training data simulated from the parameter $\pi = 0.5$ i.e. 50% prevalence for case (i) and $\pi = 0.01$ i.e. 1% prevalence for case (ii). In each case (i) and (ii), we generate three different test datasets to create situations where the trained model is (a) well-calibrated, (b) over-calibrated, and (c) under-calibrated. We examine the performance of TCE under these situations. Test datasets for situations (a), (b), and (c) are generated from the simulation model with the prevalence 50%, 40%, and 60% in case (i) and with prevalences 1%, 0%, and 2% in case (ii). We generate 20000 points in total, of which 70% were used as training data and 30% were used as test data.

Table 1 shows the values of four calibration error metrics applied to the logistic regression models under each scenario. Table 1 demonstrates that the values of ECE and ACE for the imbalanced case (ii) can be all smaller than—or very close to—the value of the well-calibrated case (a) in the well-balanced case (i). For example, the ECE value for case (i)-(b) was smaller than one for case (ii)-(a). In contrast, TCE provides values with a consistent scale at any prevalence level.

Table 2: Calibration errors by five different metrics for five different algorithms trained on the abalone dataset.

	TCE	ECE	ACE	MCE	MCE(Q)
LR	7.26%	0.0140	0.0252	0.0946	0.0851
SVM	47.21%	0.0436	0.0473	0.8302	0.1170
RF	33.89%	0.0127	0.0177	0.0670	0.0547
GB	4.86%	0.0182	0.0160	0.2965	0.0418
MLP	3.83%	0.0167	0.0122	0.0806	0.0540

Table 3: Calibration errors by five different metrics for five different algorithms trained on the webpage dataset.

	TCE	ECE	ACE	MCE	MCE(Q)
LR	40.16%	0.0044	0.0034	0.3134	0.0214
SVM	59.83%	0.0043	0.0057	0.5402	0.0239
RF	99.66%	0.0234	0.0241	0.5980	0.1189
GB	71.12%	0.0086	0.0107	0.2399	0.0436
MLP	49.81%	0.0090	0.0018	0.4344	0.0076

4.2 IMBALANCED UCI DATASETS

Next, we compare calibration error metrics in the regime of severe class imbalance. We use nine UCI datasets that were preprocessed by Lemaître et al. [2017] as benchmark tasks of imbalanced classification. We also use one additional UCI dataset with a well-balanced prevalence for comparison. For each dataset, 70% of samples are used for training and 30% of samples are kept for evaluation. We train five different algorithms: logistic regression (LR), support vector machine (SVM), random forest (RF), gradient boosting (GB), and multi-layer perceptron (MLP). We evaluate the calibration of each model using five different calibration error metrics in the following tables. Tables 2 and 3 show the results for the imbalanced dataset *abalone* and *webpage* [Dua and Graff, 2017]. See Supplement C for the results of all the other datasets. In Table 2, the model ranking by TCE and ACE agree with each other while ECE identifies RF as the model with the best calibration. As shown in the corresponding reliability diagram in Supplement C, one bin of ECE contains a large majority of model predictions. Then ECE becomes approximately equivalent to a comparison of global averages of all labels and all model predictions. Table 3 demonstrates a situation where the calibration assessment of ECE and ACE risks misleading. The values of ECE and ACE are all sufficiently small in Table 3, and one may conclude that it is reasonable to use a model with the smallest calibration error. However, TCE indicates that no model has a good calibration performance. In fact, relatively large statistical deviations between model predictions and empirical probabilities can be observed in the corresponding test-based reliability diagram for the webpage dataset in Supplement C.

Table 4: Calibration errors by five metrics for five different deep learning models under ImageNet1000 data.

	TCE	ECE	ACE	MCE	MCE(Q)
AlexNet	42.74%	0.0070	0.0070	0.1496	0.0528
VGG19	23.57%	0.0028	0.0028	0.2148	0.0247
Res18	29.93%	0.0042	0.0042	0.2368	0.0350
Res50	24.60%	0.0020	0.0018	0.1911	0.0152
Res152	16.09%	0.0012	0.0013	0.1882	0.0102
Time (s)	71.78	0.4873	0.4221	0.0046	0.0063

4.3 K-VS-REST ON IMAGENET1000

Finally, we demonstrate that TCE is applicable and produces interpretable values even for a large-scale binary classification task using ImageNet1000 data. We consider a K-vs-rest classification problem, where we use a set of all dog-kind classes (from class 150 to class 275) as a positive class and the rest as a negative class. Under this setting, 12.5% of validation samples belong to the positive class. We used 5 different trained models: AlexNet, VGG19, ResNet18, ResNet50, and ResNet152. Their calibration errors were measured based on 50000 data points of the ImageNet1000 validation dataset. Table 4 demonstrates that TCE produces interpretable values whose model ranking largely agree with other metrics in this setting. The last row of Table 4 shows the average computational time of each metric. Computation of all the procedures used in TCE required only 71.78 seconds on average for 50000 data points with 1 CPU. While the computation of the other metrics is significantly faster, TCE also has a manageable computation time.

5 RELATED WORK

Several calibration error metrics have been proposed, including the aforementioned ECE. MCE is a widely used variant of ECE that replaces the summation over $b = 1, \ldots, B$ in (4) with the supremum over $b = 1, \ldots, B$. [Kumar et al., 2019] introduce a more general l_p calibration error, which includes both ECE and MCE. ACE replaces the equispaced bins in ECE with bins designed based on quantiles of model outputs, which prevents high concentration of data in one bin when data is imbalanced [Nixon et al., 2019]. These calibration error metrics can be extended to multi-class classification [Kumar et al., 2019]. Other than calibration error, scoring functions [Gneiting et al., 2007] are commonly used measurements to evaluate a probabilistic classifier. [Wallace and Dahabreh, 2014] reported a limitation of the Brier score for imbalanced classification in particular, and proposed the *stratified* Brier score that aggregates multiple Brier scores.

In this paper, we designed a new calibration error metric based on a statistical test. While statistical tests have been used in the context of calibration, we are the first to incorporate a statistical test into the design of a calibration error metric. Vaicenavicius et al. [2019] performed a statistical test on whether the ECE computed from synthetic data generated based on model predictive probabilities is significantly different from the ECE computed from actual data. Similarly, Widmann et al. [2019] proposed a statistical test of the value of their calibration error metric built on kernel methods. Kull et al. [2017] used a test to determine the performance of their calibration algorithm mentioned below. In contrast to these existing works which considered a test for final values of calibration error metrics, our approach incorporates a test into the metric itself.

Calibration methods refer to algorithms used to improve the calibration performance of a model P_θ. Usually, they learn some 'post-hoc' function $\varphi : \mathbb{R} \to \mathbb{R}$ to be applied to each model output so that the new output $\varphi(P_\theta(x))$ is better calibrated. Various calibration algorithms have been proposed in parallel to the development of new calibration errors. Platt scaling uses a logistic function for the post-hoc function φ [Platt, 1999]. As an alternative approach, Kull et al. [2017, 2019] proposed to use a beta distribution in binary classification and a Dirichlet distribution in multi-class classification. Isotonic regression is a powerful non-parametric approach to find a monotonically increasing function φ that minimises the Brier score [Zadrozny and Elkan, 2002]. Finally, Bayesian Binning into Quantiles by Naeini et al. [2015] extends a classical histogram-based calibration [Zadrozny and Elkan, 2001] to an ensemble of histogram-based calibrations based on Bayesian model averaging.

6 CONCLUSION

In this paper, we proposed the new calibration error metric TCE that incorporates a novel loss function based on a statistical test. TCE has (i) a clear interpretation as a percentage of model predictions determined to deviate significantly from estimated empirical probabilities, (ii) a consistent scale that is unaffected by class imbalance, and (iii) a rich visual representation that facilitates a better intuitive understanding of calibration of probabilistic classifiers. We further formulated an optimality criterion of bins associated with a minimal estimation error of the empirical probabilities and a new computational algorithm under bin-size constraints.

Our proposal opens up room for new research directions in the context of calibration. This paper focuses on the methodological development of the TCE rather than the theoretical development. There are various directions to investigate in terms of theoretical properties of the TCE. These include the convergence properties of the TCE in the limit of the sample size N, understanding the minimum number of data points that should be contained in each subset $\mathcal{D}_b$, and a rigorous theoretical analysis of the regularised PAV algorithm. By continuing to investigate these areas, we can refine and expand our understanding of the TCE's capabilities.

References

Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113, 2016. ISSN 1084-8045.

Glen 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*, 135(643):1512–1519, 2009.

Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

Morris H. Degroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32: 12–22, 1983.

Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386, 2020.

Meelis Kull, Telmo M. Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052 – 5080, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1959–1968, 2015.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, 2021.

M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901—-2907, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, page 625–632, 2005.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

Amos Storkey et al. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28, 2009.

Niek Tax, Kees Jan de Vries, Mathijs de Jong, Nikoleta Dosoula, Bram van den Akker, Jon Smith, Olivier Thuong, and Lucas Bernardi. Machine learning for fraud detection in e-commerce: A research agenda. In *Proceedings of the KDD International Workshop on Deployable Machine Learning for Security Defense (MLHat)*, pages 30–54. Springer, 2021.

Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

Eric Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25: 44–56, 2019.

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Bo Schön. Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Byron C Wallace and Issa J Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and Information Systems*, 41(1):33–52, 2014.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yanwu Yang and Panyu Zhai. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 59(2):102853, 2022.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, page 609–616, 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*, page 694–699, 2002.