ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230327

Towards a Rigorous Calibration Assessment Framework: Advancements in Metrics, Methods, and Use

Lorenzo Famiglini[a;*], Andrea Campagner[b] and Federico Cabitza[a,b]

[a]Universitá degli Studi di Milano-Bicocca, Milan, Italy
[b]IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
ORCiD ID: Lorenzo Famiglini https://orcid.org/0000-0002-1934-5899,
Andrea Campagner https://orcid.org/0000-0002-0027-5157,
Federico Cabitza https://orcid.org/0000-0002-4065-3415

Abstract. Calibration is paramount in developing and validating Machine Learning models, particularly in sensitive domains such as medicine. Despite its significance, existing metrics to assess calibration have been found to have shortcomings in regard to their interpretation and theoretical properties. This article introduces a novel and comprehensive framework to assess the calibration of Machine and Deep Learning models that addresses the above limitations. The proposed framework is based on a modification of the Expected Calibration Error (ECE), called the Estimated Calibration Index (ECI), which grounds on and extends prior research. ECI was initially formulated for binary settings, and we adapted it to fit multiclass settings. ECI offers a more nuanced, both locally and globally, and informative measure of a model's tendency towards over/underconfidence. The paper first outlines the issues related to the prevalent definitions of ECE, including potential biases that may arise in the evaluation of their measures. Then, we present the results of a series of experiments conducted to demonstrate the effectiveness of the proposed framework in supporting a more accurate understanding of a model's calibration level. Additionally, we discuss how to address and potentially mitigate some biases in calibration assessment.

1 Introduction

Calibration [22], which is a Machine Learning (ML) model's ability to provide confidence scores that accurately represent the true likelihood of the outcomes (which, hence, can be interpreted as probabilities), is a multifaceted concept that demands a thorough evaluation, encompassing the assessment of average performance, identification of miscalibrated regions within the probability space, and determination of the model's over- or under-confidence. Such a comprehensive evaluation is of paramount importance in practical applications, particularly in decision-making and critical settings [8, 4]: indeed, inaccurate calibration can result in significant prediction errors and mislead decision-makers who rely on the confidence scores associated with these predictions to make decisions [17, 12]. Machine Learning and Deep Learning (DL) models greatly benefit from adequate calibration for several reasons:

1. For the instance-level accurate estimation of the class precision or positive predictive value for each class (the probability that each class would actually be the correct class if the model chose it for its classification), especially when this information can affect/prime human decision making [2];
2. For accurate prediction (classification and regression) itself, especially in settings of automatic decision making [2, 22] and whenever, in classification, low calibration regards the neighborhood of the cutoff point and, in regression, confidence scores approximate risk estimates in risk stratification and alternatives ranking;
3. For the correct reputation of the model or the assessment of its actual capacities in normative and regulatory settings to demonstrate trustworthiness and transparency [26];
4. For an accurate recalibration process [12], aimed at improving the model calibration level.

Despite the impact of this quality dimension on model performance and actual reliability, still, relatively low attention has been paid to its comprehensive assessment, which is usually accomplished by means of a number of alternative metrics (chiefly among them, the Brier score [6] and the Expected Calibration Error (ECE) [16]), that, despite their popularity, present several shortcomings [19, 21]. These mainly concern their interpretability [11, 23] (in terms of non-linear scales or measurand factors, as for the Brier score), consistency [19, 21] (undermining comparisons and benchmarking) and comprehensiveness [3] (when they do not account for local calibration, that is for levels of calibration in the surroundings of relevant portions of the probability space or bins).

In this article, we present a set of complementary metrics, namely estimated calibration indices, that address these shortcomings. These are incorporated into a comprehensive framework and associated online tool[1]. In what follows, we will describe this framework and the encompassed metrics by presenting their formal derivation and then illustrating their strengths (w.r.t. state-of-the-art metrics) in experiments in binary and multiclass classification tasks. In particular, our main contributions are as follows:

- We introduce a modification of the ECE, called Estimated Calibration Index (ECI), which provides a more informative and nuanced measure of a model's calibration performance for both the binary and multiclass settings;

[*] Corresponding Author. Email: l.famiglini@campus.unimib.it

[1] http://calibrationassessment.pythonanywhere.com

- We compare the proposed ECI with the two main variants of ECE, and we empirically show that our proposed measure provides a better estimation of the true calibration error.
- Finally, we present a series of experiments that showcase the benefits of our proposed framework in supporting a more accurate understanding of a model's calibration level.

Thus, by providing a more comprehensive view of a model's calibration and ensuring a more accurate representation of the model's performance across all classes and regions of the probability space, our proposed framework facilitates better-informed decision-making based on the model's predictions, especially in a multiclass setting where a deeper understanding of the calibration behavior for each class is crucial.

2 Measuring Calibration

In this section, we first formally define the notion of calibration in ML; then, we introduce the main state-of-the-art metric for the assessment of a model's level of calibration, namely the Expected Calibration Error (ECE), and discuss its shortcomings.

2.1 Definition of Calibration

In the literature, two definitions of calibration can be distinguished [7, 23, 21]: *weak* and *strong* calibration. Formally, let $g : \mathcal{X} \to \Delta(\mathcal{Y})$ be a probabilistic model, where $\Delta(Y)$ denotes the probability simplex over the finite set of classes $\mathcal{Y}$: we denote with $g_y(x)$ the predicted probability of class y for a given input X given by the model g. Then, we define:

- *Weak Calibration*: A model is considered weakly calibrated if it satisfies the following equation:

$$P\left[Y = \arg\max_y g_y(X) \mid \max_y g_y(X)\right] = \max_y g_y(X) \quad (1)$$

  In Equation (1), the confidence scores are represented by $\max_y g_y(X)$.
- *Strong Calibration*: A model is considered strongly calibrated if it fulfills the following equation:

$$P[Y = y \mid g(X)] = g_y(X) \quad (2)$$

Intuitively, a strongly calibrated probabilistic model ensures that the predicted confidence scores $g_y(x)$ accurately represent the probabilities of all classes y [7]. Unlike weak calibration, which only guarantees the reliability of the highest predicted confidence score, strong calibration allows users to trust the confidence scores for all classes, as they accurately reflect the corresponding outcome probabilities.

2.2 Expected Calibration Error

The Expected Calibration Error (ECE) is one of the most common metrics used for quantifying a model's level of calibration. This metric was initially proposed in [16] and is based on the binning of predictions. Consider a dataset S partitioned into H bins, denoted as $\mathcal{S}_H = \{S_h\}_{h=1}^H$. In the binary setting, the ECE is defined as:

$$ECE = \sum_{h=1}^{H} P(h) \cdot |o_h - e_h| \quad (3)$$

In Equation 3, H represents the number of bins used for discretizing the continuous $[0, 1]$ range of probabilities, $P(h)$ denotes the frequency of h-th bin's w.r.t. the total observations, o_h is the frequency of the positive class in the h-th bin, while $e_h = \frac{1}{|S_h|}\sum_{x \in S_h} g(x)$ is the average predicted confidence score within the h-th bin. These two values are used to compute the difference between the predicted probabilities and the observed (true) probabilities for each bin. Formally, the ECE ranges in the domain $[0, 1]$, where the value 0 denotes perfect calibration. The ECE metric was extended to multiclass settings by [12], through the following formula, that we term *accuracy-based* ECE:

$$ECE_{acc} = \sum_{h=1}^{H} P(h)|acc(S_h) - conf(S_h)| \quad (4)$$

Here, S_h denotes the set of examples in the h-th bin, $acc(S_h)$ denotes the model accuracy on the instances in the h-th bin (i.e., $acc(S_h) = \frac{|\{(x,y) \in S_h \mid \arg\max_{y' \in \mathcal{Y}} g(y') = y\}|\}}{|S_h|}$), and $conf(S_h)$ represents the average maximum confidence score for examples in the h-th bin (i.e., $conf(S_h) = \frac{1}{|S_h|}\sum_{x \in S_h} \max_{y \in \mathcal{Y}} g_y(x)$). The main drawback of ECE_{acc} is that it refers to the definition of weak calibration due to its exclusive reliance on the predicted class's probability and its disregard for the model's accuracy for the remaining K-1 class probabilities. To address this, [17] introduced an alternative formulation of the multiclass ECE that is more aligned with the notion of strong calibration, called Static Calibration Error (SCE):

$$\text{SCE} = \frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H} P(h)|acc(h,k) - conf(h,k)| \quad (5)$$

In Equation 5, $acc(h, k)$ represents the accuracy of the model on instances in bin h that belong to class k, while $conf(h, k)$ indicates the average confidence for the instances in the same bin and class. In contrast with ECE_{acc}, SCE reflects the notion of strong calibration more accurately. As such, it ensures that a value of zero would be met only by perfectly calibrated models w.r.t. all considered classes. As the SCE is based (similarly to the binary ECE) on the actual classes' frequencies, in the following, we will refer to this metric (as well as to related ones) with the term *frequency-based* ECE, denoted with ECE_{freq}.

2.2.1 Limitations of the ECE

The ECE is a widely used method for evaluating the calibration of a classifier, but it has several limitations. As we will show later, as it is currently constructed, one limitation of the ECE is that it does not analyze the problem *locally*, nor does it differentiate between over-confident or under-confident predictions. By *locally*, we refer to an analysis that delves into specific regions of the predicted probabilities: since the ECE, by definition, does not allow such an analysis, it does not adequately assess calibration performance across distinct probability intervals, potentially overlooking disparities across those intervals. As a consequence, the ECE is not informative enough to identify localized mis-calibrations, which could prove critical in some applications or domains. Most remarkably, however, we note two other inherent limitations of the ECE. First, there is a conflict (also noted in [21]) in the definitions of the original formulation of the ECE_{freq} and the currently more widely adopted formulations of the ECE_{acc}: indeed, these two formulations refer to two different properties of models (namely, empirical frequency and accuracy), as well as to two different notions of calibration (namely, strong and weak). Second, the estimation of ECE is known to be severely affected by the selection of the binning scheme [19]. Let the bias of the ECE be defined as $\Delta = \mathcal{E} - \mathbb{E}(ECE)$, where $\mathcal{E}$ is the true calibration error. Previous work [19] has shown that Δ is generally different from 0, meaning that ECE is a biased estimator of the true calibration error. The error is strongly dependent on the choice of the number of bins used to compute the ECE[2].

3 A Comprehensive Framework for Calibration Assessment

As we noted in the previous section, despite its popularity, the ECE is affected by several limitations. As we will show in the following by means of an extensive experimental assessment, these imply that the ECE is only able to offer a partial understanding of a model's calibration, particularly in multiclass settings.

In this section, we introduce our main proposal, which consists of an encompassing framework that, by incorporating five different indices that reflect as many aspects of the calibration of an ML model, aims at addressing the above limitations.

Our framework starts from an alternative formulation of the ECE for binary settings, called *Estimated Calibration* (EC). This formulation, which we recently proposed in [3], considers the normalized L2 distance (rather than the L1 employed in the ECE) between the true positive rate and the mean confidence scores. The main advantage of this choice is that the L2 distance assigns more weight to larger deviations and less weight to smaller deviations [5]: this property can be useful in certain applications where minimizing the impact of outliers on the overall calibration error is important [13]. Here we discuss the extension of this metric to the multiclass setting. We propose, in particular, five different indices (called, Estimated Calibration Indices (ECI)) for the evaluation of a model's calibration: a local index that evaluates the model's calibration w.r.t. a specific, relevant portion of the probability space; an under-estimation index (and, complementarily, an over-estimation one), that measures the model's general tendency toward underconfidence (resp., overconfidence); an index of the general trend toward overconfidence or underconfidence.; and, finally, a global calibration index that allows assessing the overall level of calibration of an ML model by piecing together the above indices. By considering all these factors, our proposed framework facilitates a deeper understanding of the model's calibration and enables targeted adjustments to improve its performance. We share our code as open-source and make it available on GitHub [3].

3.1 Estimated Calibration Index

In this section, we define the proposed framework and the metrics it encompasses, namely: a local calibration index, ECI_l, which can be applied to analyze different regions of the probability space separately; an over-confidence, ECI_{over}, and under-confidence, ECI_{under} calibration index, that measures the global tendency of the model to over- or under-estimate the real probability distribution of labels; and two global calibration indices, ECI_{global} and $ECI_{balance}$, that measure the overall calibration of an ML model, either in absolute terms or as an indicator of an overall tendency towards over- or under-confidence. More in particular, first, we recall the formulation of the Estimated Calibration Index (ECI) for the binary setting, as originally introduced in [3]; then we discuss the extension of the ECI to the multiclass setting.

3.1.1 Local Estimated Calibration Index

Consider a dataset S, and a partition of S into H bins, denoted as $\mathcal{S}_H = \{S_h\}_{h=1}^{H}$. A classifier g is employed to determine a collection of *calibration points*, i.e., the sequence $\{(e_h, o_h)\}_{h=1}^{H}$, where e_h and o_h are defined as in the previous sections. Intuitively, a calibration point corresponds to a point on the reliability curve determined by classifier g: the more calibrated g is, the closer the calibration points will be to the main diagonal of the reliability diagram and the lower the calibration error. The definition of the ECI as a global calibration measure is based on the computation of the calibration level at the level of individual bins, referred to as local ECI for bin h (denoted as ECI_l^h). Intuitively, ECI_l^h is defined as the normalized L2 distance between the calibration point $p_h = (e_h, o_h)$ and the point p_h^* on the bisector line that is closest to p_h: This distance is then normalized by the maximum distance from the bisector and the corresponding distance between the x-axis of the reliability diagram [9] and the bisector passing through the predicted confidence score.

The bisector line is mathematically represented as a one-dimensional vector space, denoted by $b = \frac{1}{\sqrt{2}}(1,1)$. The projection of the point p_h onto this bisector line, represented by p_h^*, is computed utilizing the equation:

$$p_h^* = \langle b, p_h \rangle b \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $| \cdot |_2$ represents the Euclidean norm. The distance between p_h and p_h^* is defined as:

$$d_h = |p_h - p_h^*|_2 \tag{7}$$

To normalize the d_h values, we define a point $\tilde{p}_h = (e_h, \tilde{o}_h)$ where

$$\tilde{o}_h = \begin{cases} 1 & e_h \leq 0.5, \\ 0 & otherwise \end{cases} \tag{8}$$

Let $\tilde{p}_h$ be the point with its first component equal to e_h and the maximum distance from the bisector line. We then define $\tilde{p}_h^* = \langle b, \tilde{p}_h \rangle b$, and d_h^{max} as the maximum distance between $\tilde{p}_h$ and $\tilde{p}_h^*$:

$$d_h^{max} = |\tilde{p}_h - \tilde{p}_h^*|_2 \tag{9}$$

The local ECI_l on bin h can thus be formally defined as:

$$ECI_l^h = 1 - \frac{d_h}{d_h^{max}} \tag{10}$$

The local ECI quantifies calibration within a bin by computing the distance of a calibration point from the bisector line.

3.1.2 Global Estimated Calibration Index

We can extend the local ECI to the whole H-binning $\mathcal{S}_H$ to obtain a global calibration measure. First, we partition $\mathcal{S}_H$ into $\mathcal{S}_H^- = S_h : e_h > o_h$ and $\mathcal{S}_H^+ = S_h : e_h \leq o_h$. $\mathcal{S}_H^-$ represents the set of bins for which the corresponding calibration point lies below the bisector line (corresponding to overconfidence). In contrast, $\mathcal{S}_H^+$ represents the set of bins for which the corresponding calibration point lies above the bisector line (corresponding to underconfidence). We can then define two ECIs which refer, respectively, to under- and

[2] On the other hand, and remarkably *variance* of the ECE has been observed to be relatively insensitive to the formulation (i.e., ECE_{freq} vs. ECE_{acc}) and the number of bins used, see, e.g., [19].

[3] The code and the corresponding Appendix can be accessed at: https://github.com/lorenzofamiglini/CalFram

over-confidence, as well as a summative index of over- vs. under-confidence, as:

$$ECI_{under}(\mathcal{S}_H) = \frac{\sum_{S_h \in \mathcal{S}_H^+} P(h) * ECI_l^h}{\sum_{S_h \in \mathcal{S}_H^+} P(h)} \tag{11}$$

$$ECI_{over}(\mathcal{S}_H) = \frac{\sum_{S_h \in \mathcal{S}_H^-} P(h) * ECI_l^h}{\sum_{S_h \in \mathcal{S}_H^-} P(h)} \tag{12}$$

$$ECI_{balance}(\mathcal{S}_H) = ECI_{over}(\mathcal{S}_H) - ECI_{under}(\mathcal{S}_H) \tag{13}$$

Finally, based on the u and o indices, we can define an overall metric of calibration that we call the Global ECI:

$$ECI_{global}(\mathcal{S}_H) = \frac{\sum_{S_h \in \mathcal{S}_H} P(h) * ECI_l^h}{\sum_{S_h \in \mathcal{S}_H} P(h)} \tag{14}$$

The ranges of ECI_{global} and $ECI_{balance}$ are, respectively, $[0, 1]$ and $[-1, 1]$. The global ECI_{global} measures a classifier's calibration on a dataset relative to an H-binning, representing the average normalized deviation from perfect calibration. In contrast, $ECI_{balance}$ indicates the model's tendency to over- or under-estimate probabilities. A positive $ECI_{balance}$ suggests overestimation, a negative value implies underestimation, and zero means balanced confidence. We remark that the ECI_{global} is related to the ECE, indeed it can easily be shown that $ECE = \sum_{S_h \in \mathcal{S}_H} \sqrt{2} \cdot d_h^{max} \cdot P(h) \cdot (1 - ECI_l^h)$. Thus, the ECI_{global} acts as a normalized, alternative definition of the ECE, offering enhanced interpretability through uniform scaling across bins. This consistency allows for easier interpretation and potentially better understanding, unlike the ECE's variable range.

3.1.3 Multiclass Generalization

In this work, we extend the ECI to the multiclass settings by following the class-wise method [12, 17]: the local ECI (ECI_l) is defined for each separate class K. Formally, let k be a class:

$$ECI_l^{h,k} = 1 - \frac{d_{k,h}}{d_{k,h}^{max}}, \tag{15}$$

where $d_{k,h}$ and $d_{k,h}^{max}$ are defined as in Eqs. (7) and (9), but relative to class k, obtained by performing a one-vs-rest transformation of the original dataset. This formulation allows us to identify which bins and classes the model performs poorly and hence may help implement localized re-calibration strategies.

Based on the above formulation, the class-wise global $ECI_{CW,global}(S_H)$ is defined as follows:

$$ECI_{CW,global}(S_H) = \frac{1}{K}\sum_{k=1}^{K} ECI_{global}^k, \tag{16}$$

while the class-wise $ECI_{CW,bal}(S_H)$ is defined as:

$$ECI_{CW,bal}(S_H) = \frac{1}{K}\sum_{k=1}^{K} ECI_{balance}^k, \tag{17}$$

where ECI_{global}^k (resp., $ECI_{balance}^k$) denotes the global (resp., balanced) ECI for class k.

As we will show in the following section, the proposed framework offers a more detailed analysis of model calibration, particularly in multiclass settings, allowing for a more granular evaluation of the model's calibration: this allows to gain additional insight into the model's performance for each class, enabling the identification of poorly calibrated regions for further investigation and improvement. Thus, through the proposed approach, we aim to provide a more robust and informative calibration framework for assessing the performance of machine learning models in various applications.

4 Evaluation Experiments

In this section, we present the results of two experiments conducted to evaluate our proposed measure (ECI) compared to the state-of-the-art metrics (ECE_{acc} and ECE_{freq}) and to illustrate the usefulness of the proposed calibration framework. One set of experiments is aimed at showcasing that ECI has a lower bias than ECE. The second series is aimed at demonstrating that ECI allows the discovery of relevant but opposite local miscalibrations, which cancel out and are hidden in global scores but may undermine the model performance or interpretability. Thus, in the first set of experiments, we exploited a synthetic dataset to estimate the bias of both metrics, that is, their deviation from the true calibration error. In the second set of experiments, instead, we utilized two real datasets to investigate the impact of global versus local evaluation of model calibration.

4.1 Bias Comparison: ECI vs. ECE

The following section outlines the experiments aimed at assessing the bias present in the ECI_{global} and ECE scores by calculating them on synthetic data that had been generated by Algorithm 1. This method has enabled us to determine the bias of either metrics under different scenarios, such as class imbalance and noise.

We conducted two experiments: one for binary settings and one for multiclass settings (with five classes). We generated 1,000 to 3,000 synthetic data points with 5 to 15 features: we considered, in particular, different levels of feature noise, class imbalance, and model accuracy to simulate real-world scenarios and evaluate the proposed metric performance under various conditions. We executed 1,000 experiments for each setting and model, resulting in a total of 8,000 evaluations of the metrics against the true calibration error.

We trained multiple classifiers on the synthetic datasets, including Logistic Regression (Logit), Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), Multi-Layer Perceptron (MLP), X Gradient Boosting (XGBoost), Naive Bayes (NB), and K-Nearest Neighbors (KNN), to analyze calibration measures' performance across different model architectures.

The use of synthetic data allows us to establish a theoretical probability relationship between the characteristics, X, and the objective variable, Y, which in turn enables us to assess the bias of calibration measures accurately. We analyzed the metrics systematically by grouping the evaluation based on task type (binary and multiclass) and further classifying them according to each model.

As previously stated, both ECI and ECE rely on the discretization of the confidence scores space (i.e., bins). To determine the ideal number of bins for these experiments, we used the *Monotonic Sweep Method (with equal mass)* [19] to minimize any potential bias and obtain a more unbiased comparison of the metrics.

4.1.1 Evaluation Criteria

To evaluate the bias of the proposed metric, compared to ECE, we present multiple assessment criteria to accurately quantify its behavior across various scenarios. This approach allows us to comprehend better the extent to which the metric approximates the true empirical calibration error.

- **Estimated True Error** ($\mathcal{E}$): The estimated true error is the average absolute difference between the true class probabilities $P_{\text{true}}(Y_i|X_i)$ and the predicted class probabilities $P_{\text{pred}}(Y_i|X_i)$:

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}\left|P_{\text{true}}(Y_i|X_i) - P_{\text{pred}}(Y_i|X_i)\right|. \tag{18}$$

- **Expected Calibration Error** (ECE): we consider two versions of the ECE, calculated, respectively, based on ECE_{acc} and ECE_{freq}.
- **Estimated Calibration Index** (ECI_{global}): To facilitate comparison with the estimated error $\mathcal{E}$, we employed $1 - ECI_{global}$ as the metric for assessing calibration performance.

Our primary goal was to determine which metric most accurately approximates the true calibration error $\mathcal{E}$ under varying performance conditions (in terms of balanced accuracy). To accomplish this, we implemented two distinct evaluation methodologies: The first method involved quantifying the mean and 95% confidence interval for the difference between $\mathcal{E}$ and each metric (i.e., *bias* denoted as Δ). The second evaluation method, instead, provides a qualitative assessment by projecting the empirical error onto a diagram that plots the true error $\mathcal{E}$ on the y-axis against the empirical error on the x-axis[4].

4.1.2 Synthetic Data Generation & Evaluation

This section outlines the procedure for generating synthetic data to assess the proposed metric's performance under various conditions. The steps involve creating datasets with different characteristics, such as noise and class imbalance, as detailed below:

1. Generate a synthetic dataset $X \in \mathbb{R}^{n_{samples}\times m_{features}}$ by sampling from a multivariate normal distribution:

$$X \sim \mathcal{N}(\mu = \mathbf{0}, \Sigma), \tag{19}$$

   where $\Sigma \in \mathbb{R}^{m_{features}\times m_{features}}$ is a covariance matrix with diagonal elements equal to 1 and off-diagonal elements equal to $correlation$. The Cholesky decomposition is used to create correlated samples [18].
2. Calculate the conditional probabilities of the classes Y given the features X:
   - For the binary task:

$$P(Y = 1|X) = \frac{1}{1 + \exp(-XW)}, \tag{20}$$

     where $W \in \mathbb{R}^{m_{features}}$ is a random weight vector sampled from a multivariate normal distribution, i.e., $W \sim \mathcal{N}(\mu, \Sigma)$.
   - For the multiclass task:

$$P(Y = k|X) = \frac{\exp(XW_k)}{\sum_{i=1}^{k_{classes}}\exp(XW_i)}, \quad k = 1, \ldots, k_{classes}, \tag{21}$$

     where $W \in \mathbb{R}^{m_{features}\times k_{classes}}$ is a random weight matrix $\sim \mathcal{N}(\mu, \Sigma)$.
3. Create the ground truth Y:

$$Y_i = \arg\max_k P(Y = k|X_i), \quad i = 1, \ldots, n_{samples}. \tag{22}$$

4. Add noise to the features X to create a noisy dataset X_{noisy}:

$$X_{noisy} = X + noise \times \mathcal{N}(0, I), \tag{23}$$

   where $I \in \mathbb{R}^{m_{features}\times m_{features}}$ is the identity matrix.
5. Create an imbalanced dataset $X_{rebalance}$ and $Y_{rebalanced}$ by separately sampling each class according to the specified imbalance ratios $r_1, \ldots, r_{k_{classes}}$.

Building upon the synthetic data generation process detailed earlier, we now present the algorithm that integrates the various steps and evaluates the performance of different models on the generated data. We introduce the following notation: $G(n, m, \mu, \Sigma)$ generates a $n \times m$ synthetic dataset X from a multivariate normal distribution with mean μ and covariance matrix Σ. $C(X, k)$ calculates the conditional probabilities p of classes Y given features X. $I(X, Y, p, r)$ creates re-balanced dataset $X_{rebalanced}$ and $Y_{rebalanced}$ based on imbalance ratio r. $S(X, Y, p)$ splits the dataset into training and test sets $X_{tr}, X_{ts}, Y_{tr}, Y_{ts}$, and corresponding class probabilities p_{tr} and p_{ts}. Having detailed the steps involved in synthetic data generation, we now present an algorithm that encapsulates the entire process, from data generation to model evaluation.

Algorithm 1 Synthetic Data Generation and Evaluation

```
procedure SyntheticData($n, m, \Sigma, \nu, N$)
    for $j = 1, \ldots, N$ do
        $X_j \leftarrow G(n, m, \mu, \Sigma)$
        $Y_j, p_j \leftarrow C(X_j, k)$
        $X_j \leftarrow X_j + \nu_j \cdot \mathcal{N}(0, 1)$
        $r_j \sim U(.05, .5)$                                   ▷ Sample $r$ from uniform distribution
        $X_{j,i}, Y_{j,i}, p_{j,i} \leftarrow I(X_j, Y_j, p_j, r_j)$
        $X^{tr}_{j,i}, X^{ts}_{j,i}, Y^{tr}_{j,i}, Y^{ts}_{j,i}, p^{tr}_{j,i}, p^{ts}_{j,i} \leftarrow S(X_{j,i}, Y_{j,i}, p_{j,i})$
        for m $\in M$ do                                       ▷ M models
            $\theta_{j,m} \leftarrow \underset{\theta_{j,m}}{\text{argmin}}\mathcal{L}(X^{tr}_{j,i}, Y^{tr}_{j,i})$        ▷ Training
            $\hat{Y}^{ts}_{j,i}, \hat{p}^{ts}_{j,i} \leftarrow g_{\theta_{j,m}}(X^{ts}_{j,i})$                 ▷ Prediction
            $E_{j,i,\theta_{j,m}} \leftarrow e(Y^{ts}_{j,i}, \hat{Y}^{ts}_{j,i}, \hat{p}^{ts}_{j,i}, p^{ts}_{j,i})$      ▷ Evaluation
            store $E_{j,i,\theta_{j,m}}$                          ▷ Store metrics for each model
        end for
    end for
    return $E$                              ▷ Return the computed metrics for all models
end procedure
```

4.2 Local and Global Miscalibration Assessment

In this section, we outline the second set of experiments that we performed to investigate the usefulness of the proposed calibration framework for understanding local vs. global calibration and identifying under vs. over-confidence. In particular, we carried out two experiments utilizing benchmark datasets from the biomedical field. These datasets are encompassed within the MedMNIST benchmark suite of datasets, as documented in the following sources [24, 25]: **PathMNIST** dataset contains 107,180 colon pathology images, classified into 9 classes: 89,996 training, 10,004 validation, and 7,180 test images. **PneumoniaMNIST** dataset has 5,856 chest X-ray images, categorized as normal or pneumonia, and is divided into 4,708 training, 524 validation, and 624 test images. We assessed SOTA models, such as *ResNet 152* [14] and *Vision Transformer DeiT-base* [20] on multiclass and binary tasks. The models were pre-trained on Imagenet [10] and then fine-tuned by freezing all the layers except

[4] The visualization was categorized by model type and task for a thorough analysis of metric performance in approximating true calibration error. We evaluated ECI_{global} as $1 - ECI_{global}$, representing the error.

the last one. The models were trained for 50 epochs with a learning rate of .001, with the Adam optimizer, and *Reduce on plateau* [1] adjustment, using the validation set for model selection. We used binary-cross entropy and cross-entropy as loss functions for model optimization. The models were then evaluated on the set of test images: in particular, we computed the local calibration (ECI_l) scores for each model and dataset. For simplicity and clarity, we have set the number of bins to 10, making it easier to comprehend the local information and interpret the results.

5 Results and Discussion

As discussed in the previous section, in the present study we have conducted two sets of experiments designed to serve different purposes: the first set aimed at evaluating the bias of our ECI measure (in comparison to the ECE). The second set was intended to illustrate the usefulness of our framework for a thorough assessment of calibration, both at global and local level.

In the first set of experiments, we used synthetic data to evaluate the performance of the ECI scores in estimating the true calibration error: the results of this evaluation are represented (in terms of mean deviation and corresponding 95% confidence intervals), for both binary and multiclass settings, in Figure 1, a and b, respectively.

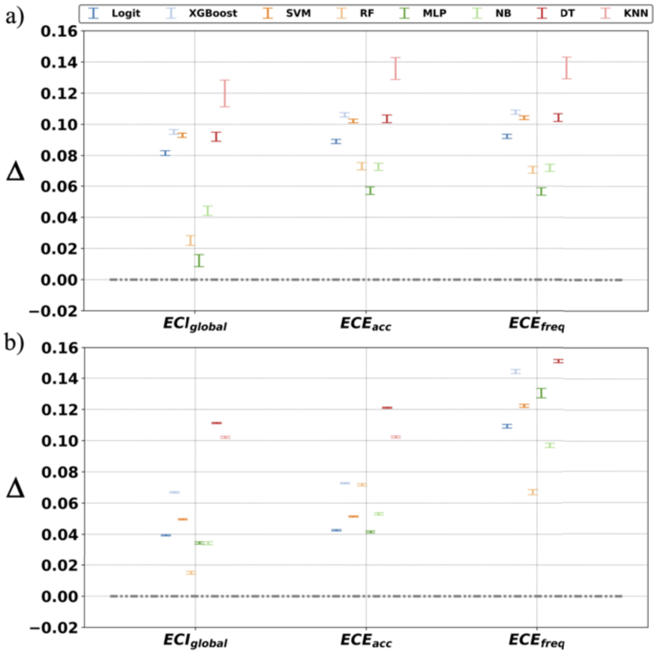

Figure 1. Strip plots of the differences (Δ) between true error ($\mathcal{E}$) and each ECI and ECE estimates, with their 95% confidence intervals, shown for the models of the binary settings (a) and for the models of multiclass settings (b).

As shown in Figure 1, the proposed ECI exhibits a lower bias, on average, compared to both ECE_{acc} and ECE_{freq}, as made clear by the 95% confidence intervals. This finding confirms the limitations of ECE_{acc} and ECE_{freq} in accurately estimating the true calibration error. It highlights the need for more robust calibration error estimation methods, like the proposed ECI metric. Moreover, our ECI metric is generally more conservative in its error estimates (i.e., it overestimates the true error), as shown in Figure 2 (see also the appendix for additional results), which illustrates the differences between error curves for the ECI and the ECE (along with the corresponding 95% confidence bands, which were estimated through bootstrapping to allow for significance assessment). Overestimation of the true calibration error (rather than an underestimation, as in the case of ECE) is a desirable property of a calibration metric: indeed, as we mentioned in the introduction, a model's reputation and reliability cannot be overstated, especially in light of possible normative and regulatory requirements. Thus, adopting a more conservative evaluation metric in model validation can make ensuring compliance with these requirements easier since such a metric would provide a worst-case assessment of the model's performance. Thus, to summarize the results of our first experiment, we proved that our proposed ECI measure is not only closer on average to the true calibration error than ECE_{acc} and ECE_{freq} (i.e., it is less biased as a measure of calibration) but, when it errs, it errs in the desirable direction, making it a more conservative estimate of performance.

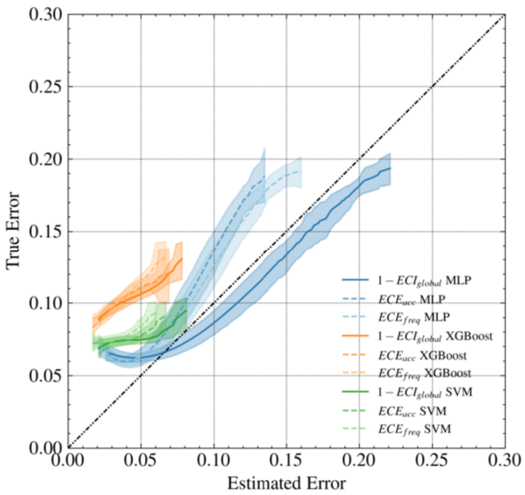

Figure 2. Reliability diagram for multiclass MLP and XGBoost models: true error (y-axis) and estimated error (x-axis) with 95% confidence intervals from bootstrap methods and Spline function smoothing.

In regards to the second set of experiments, which were instead based on real-world datasets and aimed at illustrating the usefulness and informativeness of the proposed framework, the global calibration indices for the considered models and datasets are reported in Table 1, while the local calibration scores and class-wise under-/over-confidence scores (for model DeiT on datasets PneumoniaMNIST and PathMNIST, respectively) are reported in Tables 2 and 3, respectively (see also the appendix for complete results).

Our findings from the real-world dataset tests, as shown in Table 1, indicate that the models, on average, seem well-calibrated. For instance, the ResNet152 in the binary settings has an ECE_{acc} of .937, ECE_{freq} of .935, and an ECI_{global} of .913. However, the analysis based on the over- and under-confidence ECI indices reveals that the model's calibration on the instances associated with an overconfident prediction was less satisfactory compared to that for instances associated with an underconfident prediction, as evidenced by the ECI_{over} metric in Table 1: thus, in general, the model was slightly over-confident in its predictions. Similar conclusions can also be observed for the DeiT model in the binary setting, while the oppo-

Table 1. Model Comparison: Class-wise ECE and ECI for multiclass tasks, with ECEs as 1-ECE for easier comparison.

	PneumoniaMNIST					**PathMNIST**				
Model	$\mathbf{ECE_{acc}}$	$\mathbf{ECE_{freq}}$	$\mathbf{ECI}_{global}$	$\mathbf{ECI}_{over}$	$\mathbf{ECI}_{under}$	$\mathbf{ECE_{acc}}$	$\mathbf{ECE_{freq}}$	$\mathbf{ECI}_{global}$	$\mathbf{ECI}_{over}$	$\mathbf{ECI}_{under}$
ResNet152	.937	.935	.913	.857	.997	.989	.988	.985	.962	.977
DeiT	.887	.885	.850	.850	.999	.991	.990	.988	.973	.946

Table 2. ECI_l scores for DeiT on PneumoniaMNIST test set: Bold values show $ECI_l < ECI_{global}$. Column frequency is the positive class proportion per bin.

	0-.1	.1-.2	.2-.3	.3-.4	.4-.5	.5-.6	.6-.7	.7-.8	.8-.9	.9-1
$\mathbf{ECI}_l$	.952	.897	**.809**	**.579**	**.205**	**.401**	**.580**	**.595**	**.541**	.931
Frequency (%)	13.0	5.0	3.0	2.1	2.2	2.9	2.6	3.2	4.8	61.1

Table 3. Values of $ECI_{balance}$, ECI_{over} and ECI_{under} of the DeiT model for individual classes for the PathMNIST test set.

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
$\mathbf{ECI}_{balance}$	-.111	.009	.014	-.096	-.039	-.051	-.093	.186	-.068
$\mathbf{ECI}_{over}$	.707	.417	.885	1	.962	.931	.999	.786	.992
$\mathbf{ECI}_{under}$	.868	1	1	.797	.918	.954	.859	.958	.887

site conclusion can be observed for both models on the multiclass dataset. Similarly, the DeiT model seems to be well-calibrated on the binary dataset, with an ECE_{acc} of .887, ECE_{freq} of .885, and an ECI_{global} of .85. Nevertheless, as shown in Table 2, the model underperforms within certain confidence score regions: the ECI_l metric reveals that the model is not well-calibrated in the regions close to the cutoff point, where over-confidence is clear in the confidence scores ranging between .3 and .7. Indeed, the ECI_l ranges between .302 to .595, values that deviate significantly from the global calibration measurement, i.e., $ECI_{global} = .850$ as depicted in Table 2. Also, while overall performance appears satisfactory in the multiclass setting, we see that decomposing calibration locally and per class provides valuable cues for a better interpretation of the model performance (see also the appendix). For instance, our analysis reveals that the DeiT model is not well-calibrated for class 0 in the .3-.5 region. From Table 3, in the multiclass settings, both models seem to be calibrated globally (i.e., the DeiT model has ECE_{acc} of .991, ECE_{freq} of .99, and ECI_{global} of .988). However, the model's calibration performance is less satisfactory if we separately consider the different classes or the overconfidence and underconfidence regions of the probability space. Specifically, in regard to classes 0, 1, and 7, the model was more calibrated on the instances associated with underconfident predictions (i.e., ECI_{under} class 0 = .868, ECI_{under} class 1 $\approx$ 1, and ECI_{under} class 7 $=$ .958.), while on the instances associated with an overconfident prediction the model is not well-calibrated for those same classes (i.e., ECI_{over} class 0 = .707, ECI_{over} class 1 = .417, and ECI_{over} class 7 = .786). These findings can be further analyzed by evaluating the $ECI_{balance}$ values in Table 3, which allow us to provide more insight on the under-confidence vs over-confidence behavior of the model. As an example, for class 0, we can easily see that the $ECI_{balance}$ was -.111, which suggests a tendency towards underconfidence, while for class 7, an $ECI_{balance}$ of approximately .2 indicates a stronger (yet still weak) tendency towards overconfidence. To summarize, our experimental findings underscore the necessity of delving deeper into local and class-wise calibration analysis rather than solely relying on global metrics. Drawing on the results presented in this paper, we make the point that global calibration estimates may inadvertently mask potential miscalibrations on a local level (i.e., per confidence score bin or per class). The framework proposed in our study allows us to effectively identify these pitfalls by providing calibration information at different granularity levels. As we highlighted in the introduction, having information about calibration at instance- or class-level can be of paramount importance, particularly when the probabilistic information given by a model can influence, or bias, human decisions [22], e.g. when confidence scores function as estimates for risk stratification and alternative ranking, or in automated decision-making settings, where calibration deficiencies may regard classification cutoff points. Moreover, detecting regions in the probability space where models exhibit suboptimal calibration enables the development of more targeted recalibration strategies that employ local-level calibration information to enhance the overall performance, see e.g. [3], [15].

In conclusion, the observed results emphasize the need to rely on metrics that minimize bias, that is, the confidence estimation error (Δ), and to look beyond global calibration metrics, which can hide miscalibrations, and rather assess instance-level and class-level calibration, as well as over- and under-confidence in models' predictions. In doing so, our proposed calibration framework allows us to holistically evaluate a model's performance and better understand its limitations so as to allow for more informed decisions when using AI in critical applications.

6 Conclusion

This study presents a novel calibration assessment framework for ML models, designed to address the limitations of existing popular metrics, particularly the ECE. Our framework enables a more fine-grained evaluation of calibration by assessing model performance locally, for different confidence regions or classes, providing a comprehensive understanding of the model's behavior. We have demonstrated that our proposed global index, ECI_{global}, offers a less biased estimation of the true calibration error compared to even the most recent versions of ECE. This is attained by delivering a more reliable calibration error estimation while also considering the impact of binning-related biases. Our experiments also highlight the benefits of decomposing calibration locally (i.e., per calibration bin) and per class, allowing for quantification of the impact of class-specific characteristics and sample representation on calibration. Although our experimental approach has some limitations due to the number of models and datasets employed, the results provide evidence that the proposed metrics outperform existing reference metrics, especially in delivering 'instance-level' information. This, in turn, enhances model transparency, user trust, and user satisfaction. Future research should delve into our framework's applications, evaluating calibration effects on accuracy, recalibration and, most importantly, users' trust.

References

[1] Mark Ainsworth and Yeonjong Shin, 'Plateau phenomenon in gradient descent training of relu networks: Explanation, quantification, and avoidance', *SIAM Journal on Scientific Computing*, **43**(5), A3438–A3468, (2021).

[2] Fernando Alarid-Escudero, Richard F MacLehose, Yadira Peralta, Karen M Kuntz, and Eva A Enns, 'Nonidentifiability in model calibration and implications for medical decision making', *Medical Decision Making*, **38**(7), 810–821, (2018).

[3] Anonymous authors, 'Anonymized publication', (2022).

[4] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana, 'Calibration of machine learning models', in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 128–146, IGI Global, (2010).

[5] Christopher M Bishop and Nasser M Nasrabadi, *Pattern recognition and machine learning*, volume 4, Springer, 2006.

[6] Glenn W Brier et al., 'Verification of forecasts expressed in terms of probability', *Monthly weather review*, **78**(1), 1–3, (1950).

[7] Jochen Bröcker, 'Reliability, sufficiency, and the decomposition of proper scores', *Quarterly Journal of the Royal Meteorological Society*, **135**(643), 1512–1519, (2009).

[8] Micah Cearns, Tim Hahn, Scott Clark, and Bernhard T Baune. Machine learning probability calibration for high-risk clinical decision-making, 2020.

[9] Morris H. DeGroot and Stephen E. Fienberg, 'The comparison and evaluation of forecasters', *Journal of the Royal Statistical Society. Series D (The Statistician)*, **32**(1/2), 12–22, (1983).

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, (2009).

[11] Yun Du, Dong Liang, Rong Quan, Songlin Du, and Yaping Yan, 'More than accuracy: An empirical study of consistency between performance and interpretability', in *PRICAI 2022: Trends in Artificial Intelligence: 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022, Shanghai, China, November 10–13, 2022, Proceedings, Part III*, pp. 579–590. Springer, (2022).

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger, 'On calibration of modern neural networks', in *International conference on machine learning*, pp. 1321–1330. PMLR, (2017).

[13] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, volume 2, Springer, 2009.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[15] Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Edward Schmerling, and Marco Pavone, 'Local calibration: Metrics and recalibration', in *The 38th Conference on Uncertainty in Artificial Intelligence*, (2022).

[16] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht, 'Obtaining well calibrated probabilities using bayesian binning', in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, (2015).

[17] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran, 'Measuring calibration in deep learning.', in *CVPR Workshops*, volume 2, (2019).

[18] Mohsen Pourahmadi, Michael J Daniels, and Trevor Park, 'Simultaneous modelling of the cholesky decomposition of several covariance matrices', *Journal of Multivariate Analysis*, **98**(3), 568–587, (2007).

[19] Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer, 'Mitigating bias in calibration error estimation', in *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, (2022).

[20] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou, 'Training data-efficient image transformers and distillation through attention', in *International Conference on Machine Learning*, volume 139, pp. 10347–10357, (July 2021).

[21] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön, 'Evaluating model calibration in classification', in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467. PMLR, (2019).

[22] Ben Van Calster and Andrew J Vickers, 'Calibration of risk prediction models: impact on decision-analytic performance', *Medical decision making*, **35**(2), 162–169, (2015).

[23] David Widmann, Fredrik Lindsten, and Dave Zachariah, 'Calibration tests in multi-class classification: A unifying framework', in *Advances in Neural Information Processing Systems*, eds., H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, volume 32. Curran Associates, Inc., (2019).

[24] Jiancheng Yang, Rui Shi, and Bingbing Ni, 'Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis', in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, (2021).

[25] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, 'Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification', *arXiv preprint arXiv:2110.14795*, (2021).

[26] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy, 'Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making', in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 295–305, (2020).