**A comprehensive review of classifier probability calibration metrics**

Richard Oliver Lane ([0000-0003-3741-0348](https://orcid.org/0000-0003-3741-0348))

QinetiQ, St. Andrews Road, Great Malvern, Worcestershire, UK

**Abstract**: Probabilities or confidence values produced by artificial intelligence (AI) and machine learning (ML) models often do not reflect their true accuracy, with some models being under or overconfident in their predictions. For example, if a model is 80% sure of an outcome, is it correct 80% of the time? Probability calibration metrics measure the discrepancy between confidence and accuracy, providing an independent assessment of model calibration performance that complements traditional accuracy metrics. Understanding calibration is important when the outputs of multiple systems are combined, to avoid overconfident subsystems dominating the output. Such awareness also underpins assurance in safety or business-critical contexts and builds user trust in models. This paper provides a comprehensive review of probability calibration metrics for classifier models, organizing them according to multiple groupings to highlight their relationships. We identify 94 metrics, and group them into four main families: point-based, bin-based, kernel or curve-based, and cumulative. For each metric, we catalogue properties of interest and provide equations in a unified notation, facilitating implementation and comparison by future researchers. Finally, we provide recommendations for which metrics should be used in different situations.

**Keywords**: Classification, Calibration, Confidence, Multiclass, Uncertainty.

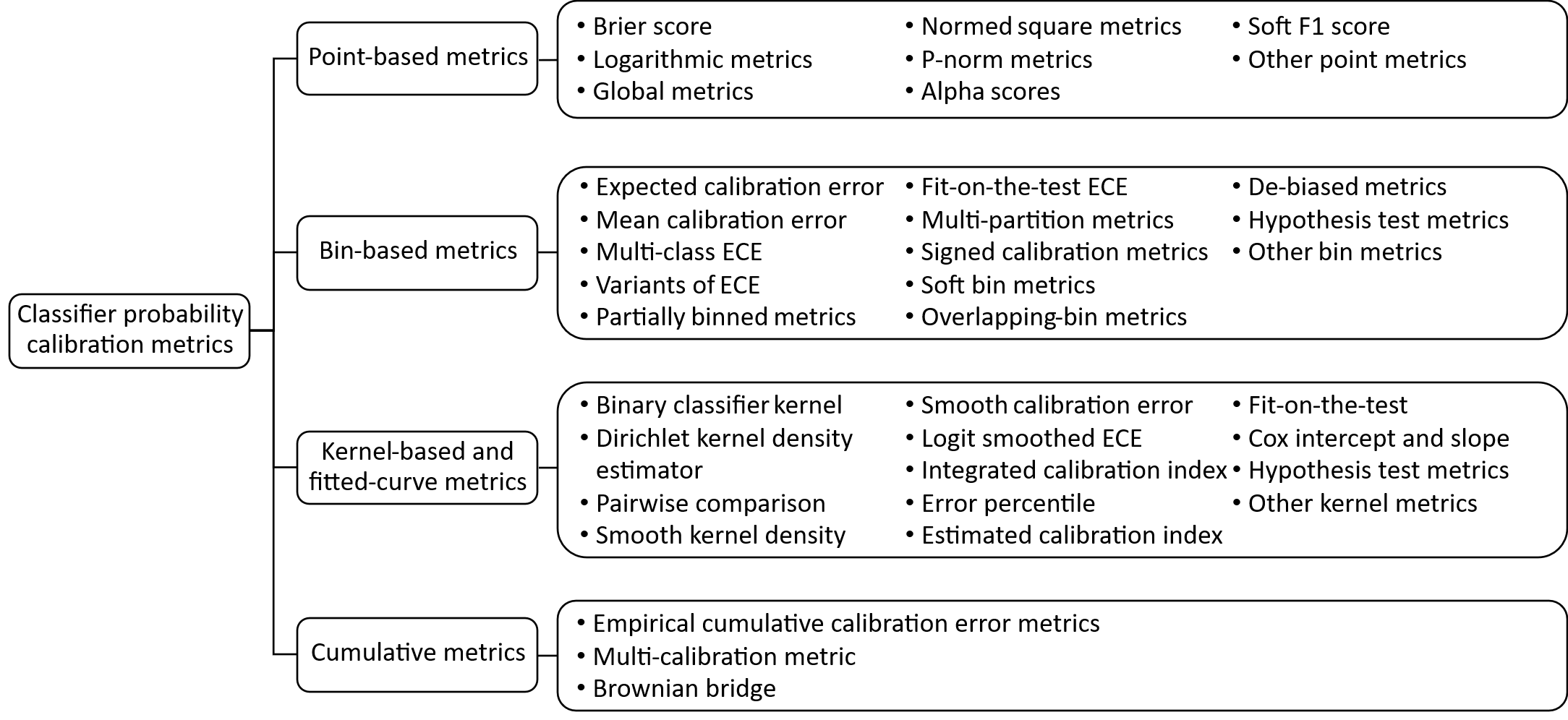
# Introduction

Artificial intelligence (AI) and machine learning (ML) models have seen widespread adoption in recent years. When such models are used in safety or business-critical applications, it is vital to be able to understand and assure their behaviour. Models generate predictions accompanied by confidence scores or probabilities, which are considered calibrated when they accurately reflect the proportion of correct classification decisions. However, confidence scores are not always representative of true probabilities. Assessing model calibration, particularly under any operational conditions that differ from training, requires a robust measure of calibration quality.

The publication of [31] by Guo et al., which highlighted examples of miscalibration in deep neural networks, sparked an intense interest in the concept of calibration and metrics to measure it. This surge in attention has led to numerous publications over the last few years, with approximately ten new metrics being defined and proposed every year since then, and many papers discussing their merits or application. However, as Flach witheringly puts it in the concluding slides of [23] “contrary what recent machine learning literature may lead you to believe, calibration research predates machine learning and has been studied for three-quarters of a century”. Thus, there is a large body of work from which to draw on. Silva Filho et al. [87] were motivated to write a survey on assessing and improving classifier calibration as “the literature on post-hoc classifier calibration in machine learning is now sufficiently rich that it is no longer straightforward to obtain or maintain a good overview of the area”. This lack of understanding has caused an impediment to research, as scientists must examine many individual papers to find relevant information. Indeed, despite the valuable survey of [87], it only covers a subset of the metrics available at the time. A review on more general metrics for image analysis (not just calibration) by a large group of experts describes the field of calibration metric development as “relatively young and currently highly dynamic” [59]. As such, that paper only covers a small number of better-known metrics. Hagopian et al. [36] state that “a summarizing document for calibration metrics is notably lacking in the literature”.

This paper presents a comprehensive review of classifier probability calibration metrics. The aim of the review is to serve as a useful reference for researchers attempting to understand the landscape of such metrics. Contributions of the paper are as follows:

1. A wide-ranging survey of classifier probability calibration metrics for models with a discrete output is provided. This review addresses a gap in the literature for which there is a high demand. Authors of a major review of general classifier metrics [59] originally intended to omit the class of calibration metrics but the decision was reversed due to high demand expressed through crowdsourced feedback. Nevertheless, [59] only covers a small fraction of the calibration metrics described in this paper.
2. Metrics are organized according to several different novel categorizations to help understand the relationships between them. One of the main categorizations is the four families of point-based, bin-based, kernel or fitted-curve, and cumulative metrics. Each family has advantages and disadvantages, as do the individual metrics within the families. These are described in detail in the manuscript.
3. This paper represents the most comprehensive survey of probability calibration metrics to-date, providing descriptions of significantly more metrics than those in the overlapping lists provided by other authors. The five most similar reviews are by Flach et al. [23] (who describe seven metrics), Hagopian et al. [36] (13 metrics), Maier-Hein et al. [59] (eight metrics), Silva Filho et al. [87] (nine metrics), and Tao et al. [91] (11 metrics). Between them, these five reviews discuss 26 metrics, whereas the present paper analyses 94 metrics.
4. Where relevant, equations are provided with a unified notation to facilitate implementation and comparison by future researchers. Original papers that first describe metrics use a wide variety of different notations, making comparisons difficult.
5. Metrics that were previously treated in isolation are brought together and conceptual relationships are highlighted. Several authors have independently invented the same metric but given them a different name. This review treats such metrics as a single entry and consolidates the separate analysis provided by the original authors. Some metrics are special cases of others but have not previously been associated. These connections are noted and integrated discussions follow. Other metrics that are nominally from different families, but have some conceptual similarity, are discussed together, enabling the cross-fertilization of ideas between different research groups.
6. Finally, we recommend specific metrics for general, multiclass, and local calibration scenarios.



**Fig. 1** Hierarchy of probability calibration metric families and sub-families.

The remainder of this paper is organised as follows. Section 2 introduces key concepts relating to calibration metrics and defines the major notation used in this paper. Sections 3 to 6 respectively describe metrics in the point-based, bin-based, kernel or fitted-curve, and cumulative families. A dendrogram showing the hierarchy of families and sub-families, as described in Sections 3 to 6, is shown in Fig. 1. Conclusions and recommendations are given in Section 7. Appendix A describes lesser-used metrics. Appendix B includes further discussion on the topics of skill scores, bootstrapping and consistency sampling, and consistency calibration. Appendix C contains a comprehensive list of symbols used in this paper and their basic definition. Finally, Appendix D contains a table that summarises the main pros and cons of each metric and other information, such as the range of attainable values and alternative names for the same metric.

# Calibration metric concepts

## Notation

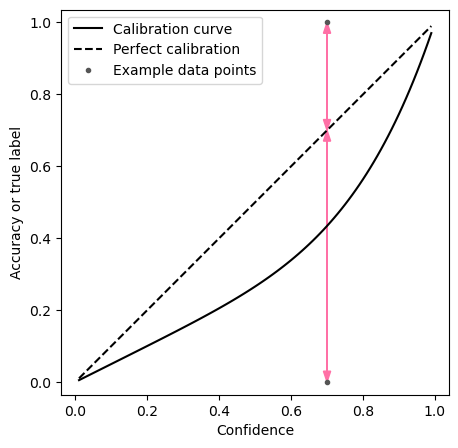
The literature on probability calibration metrics is inconsistent in its terminology and notation. We describe the techniques in a unified way, rather than using the original symbolization. We start by defining the problem of interest to be assigning a piece of data or “datapoint” to one of classes. The data could be a vector of features, an image, a signal, a time series, or any other collection of information. Each datapoint, while potentially representing multiple numbers, is treated here as a single, possibly multi-dimensional, entity. Each datapoint has an associated true label . If the classifier is binary, then by convention. If the classifier is multi-class (with three or more classes), then or is encoded as a one-hot vector, as determined by context. When multi-class problems are decomposed into multiple binary problems, represents making a correct decision for a particular class. The number of datapoints over which the calibration of a model is to be assessed is . The vector is a length- vector of true proportions of each class in the dataset.

For each datapoint, the model produces a predicted probability or confidence . For binary problems, by convention this is a scalar value representing the probability of class 1. In this case, there is no need to specify the probability of class 0 as it is completely determined by . This simplifies the analysis. For multi-class problems, is a vector of probabilities for each class, which sums to unity. When discussing the extension of binary classification to multi-class classification, the multi-class notation may be used for binary classes. That is, may be a two-element vector representing the probabilities of the two classes.

A comprehensive list of symbols used in this paper and their basic definition is given in Table 3 in Appendix C. Each term is described in more detail where it first appears in the paper.

## Calibration curve and reliability diagram

An ideally calibrated classifier outputs confidence scores or predicted probabilities equal to its accuracy, conditional on the score. The term “accuracy” is also named in various sources as “actual positive rate”, “fraction of positives”, “empirical probability”, or “observed relative frequency”. In practice the accuracy for a particular confidence value can be higher or lower than that confidence. The actual accuracy as a function of confidence for a particular class is known as the calibration curve. An example theoretical non-perfect calibration curve is illustrated in Fig. 2, along with the ideal calibration line. In this case, the classifier is over-confident in the target class – the achieved accuracy is lower than the model’s confidence. In binary classification, the calibration curve for class 1 contains all the information necessary to understand the model’s calibration, as the curve for class 0 is its mirror image reflected over the perfect calibration line. For multi-class problems, the situation is more complex; this is explained further below.

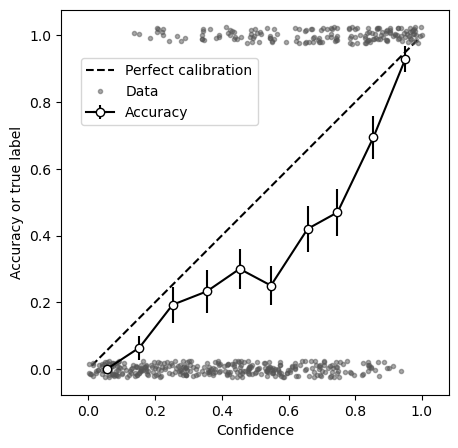


**Fig. 2** Theoretical calibration curve with two example datapoints having the same confidence, but different true labels. The red arrows indicate the calibration errors for those individual datapoints. The curve represents an overconfident classifier.

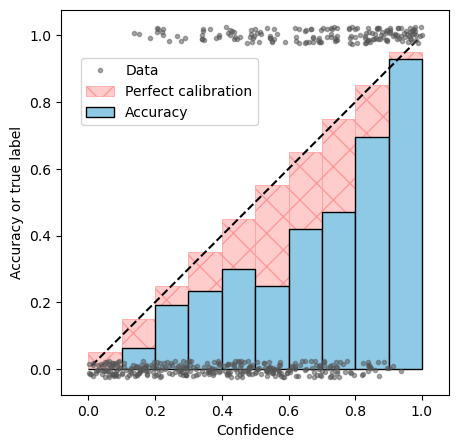
For a single datapoint, the classifier is either correct or incorrect, and the model can only perfectly be calibrated if the confidence is zero or unity. For other confidence values, there is inevitably some calibration error. Fig. 2 demonstrates this for two example datapoints with the same confidence value of 0.7 for class 1. The true label is zero for one datapoint and the error for that datapoint is 0.7 (as indicated by the lower red arrow). The true label for the other datapoint is unity, and the error for that datapoint is 0.3 (as indicated by the higher red arrow).

Although perfect calibration for a single datapoint is impossible without perfect accuracy, a classifier with non-perfect accuracy over a set of datapoints may still be well calibrated. For example, if ten datapoints all have a confidence of 0.7 and seven out of those ten are classified correctly, then this classifier is perfectly calibrated for the dataset. In practice, not all confidence values output by a model are expected to be identical. It is therefore common to group confidence values into non-overlapping bins and analyse the accuracy of datapoints in each bin. When this information is presented visually it is known as a reliability diagram, curve, or plot. This is frequently shown as a bar chart, see [31] for example. However, line plots facilitate better visual inspection of the data because they show trends over the underlying continuous confidence variable. Some works place markers at the centre of each bin on the horizontal confidence axis. However, if the data are not uniformly distributed throughout the bin this can give misleading results. Therefore, it is better to place markers at the mean confidence for each bin [87]. Vasilev et al. [99] distinguish between the bar chart representation being called the reliability diagram and the line representation being called the reliability plot.

An example reliability plot is shown in Fig. 3 as a solid line, and additional information is included. The dataset used to compute this diagram was generated as 500 random samples, with the confidence and true labels determined based on the true calibration curve in Fig. 2. In Fig. 3, true labels are jittered by ±0.025 for improved visualization. The markers on the reliability curve show the mean achieved accuracy in ten equal-width bins, and the error bars represent the standard error of those estimates. The standard error is greater for confidence values near 0.5 than zero or unity and is in general greater when there are fewer datapoints, although this effect is not apparent for this example dataset. The measured reliability curve has approximately the same shape as the true calibration curve. The same information is shown in Fig. 4 using the slightly more common bar plot representation.



**Fig. 3** Reliability line plot for 500 labelled datapoints, with ten bins. The plot shows that, for a finite dataset, empirical accuracy does not always increase with confidence



**Fig. 4** Reliability bar chart for 500 labelled datapoints, with ten bins. The chart emphasises equal width of the bins.

The majority of binary calibration metrics measure some aspect of the data visible in Fig. 2, Fig. 3, or Fig. 4, whether this is the location of the datapoints, or the degree to which the calibration curve (as estimated via binning or other means) deviates from the identity line.

## Multi-class aspects

Most work on probability calibration metrics relates to binary classifiers. However, multi-class versions also exist. There are three ways to define calibration for multiple classes [87]:

* **Top-label**. Top-label calibration considers only the class with the highest predicted probability. Some classifiers only report probabilities for the most likely class. In those situations, this is the only calibration that can be assessed. Top-label calibration is also known as confidence calibration and is the most common form of calibration.
* **Class-wise**. Class-wise calibration assesses the marginal probabilities of all classes. This definition requires all one-vs-rest binary estimators to be calibrated individually. This is a more restrictive definition than confidence as it is not possible for under and overestimates of probability from different classes to cancel out.
* **Multiclass**. Multiclass calibration requires the entire vector of probabilities to be correct in all elements simultaneously. This is also known as joint, full, canonical, or strong calibration.

Vaicenavicius et al. [95] give theoretical examples where a multiclass classifier is either top-label calibrated or class-wise calibrated, but not fully multiclass calibrated. Lack of full multiclass calibration could be important in safety critical applications, especially where the action taken should depend not only on the most likely outcome, but the probabilities of other less-likely outcomes, which may have severe negative consequences.

Multi-class problems are often decomposed into a set of binary sub-problems whose outputs are aggregated to give an overall multi-class score [46]. The advantage of this decomposition is that binary classification theory can be applied to sub-problems without the complication of multi-class issues. Furthermore, some classifiers are inherently designed only to work with binary problems. The decomposition allows such classifiers to be applied in a multi-class setting without further modification. The most common decompositions are one-vs-rest (OVR) or comprehensive pairwise. The decomposition can be represented in code matrices, whose columns represent sub-problems and rows represent classes. Entries are +1 for the positive class, -1 for the negative class, and 0 if the class is not represented in the sub-problem. The matrices are useful for visualizing the binary decomposition and provide a compact representation for automated processing. Example code matrices for one-vs-rest and pairwise decompositions for a four-class problem are respectively shown in (1) and (2).

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

Other code matrix aggregation setups are possible but are far less common.

## Other observations on calibration

Numerous studies have empirically shown that classifiers often exhibit overconfidence. The usual explanation given is that the models are large enough to memorize training data and maximise the confidence. However, Bai et al. [5] show that certain classifiers are inherently overconfident, even when the dataset is large and the number of model parameters is small. Specifically, this applies to logistic regression and other classifiers where the activation function is symmetric and concave in the positive half. Under-confident classifiers can arise when the activation function is convex in the positive half, but convexity over this whole range is not possible, so under-confidence cannot happen at every value of . Munir et al. [66] state that rectified linear unit (ReLU) activation functions, widely used in modern deep neural nets, and similar piece-wise linear functions, are a core reason behind overconfident predictions far away from the training data. However, Minderer et al. [63] show that model structure is more important than model size in understanding probability calibration, and that recent models, especially those not using convolutions, are in fact among the best calibrated.

The metrics described in this paper serve as absolute measures of model calibration. Recalibration techniques attempt to improve the correctness of model confidence values by post processing those values. The ideal aim of such techniques is to invert the effect of the model calibration curve so that the overall process produces an identity calibration curve. In practice, these techniques are not perfect, but it is possible to define their “calibration gain”, which is the improvement in calibration error (by any particular metric) when the technique is applied [109].

## Scope of paper

The systematic and comprehensive review of classifier probability calibration metrics was conducted as follows. First, Google Scholar was used to generate a long list of the top 100 papers that included the words “probability”, “calibration”, and “metric”, starting from 2017, the year of the seminal paper by Guo et al. [31] that spurred recent widespread interest in probability calibration. From this long list, papers that did not describe a new metric or compare existing metrics, were removed, leaving a short list. Metrics that can be used as classifier probability calibration metrics, even if they were originally designed for another purpose, such as a loss function, were defined to be in scope. Forward and backward citation searching, seeded from this short list, was performed to cover any gaps from the initial search. The result of this process was a list of papers that collectively comprehensively describe classifier probability calibration metrics. The metrics were then analysed, grouped into families, and explained in this paper. We include new insights into relationships and properties of these metrics.

Although this review focuses on classifiers, object detection models also produce probabilistic outputs and pose related calibration challenges. Whereas as a classifier only considers the discrete label of a piece of data, an object detector also provides object size and location co-ordinates, often within an image. Although it is possible to assess object detector confidence scores based only on labels of associated ground-truth objects, the additional degrees of freedom allow a more nuanced form of assessment. To maintain focus, analysis of object detection is out of scope for this paper. Example probability calibration metrics specific to object detection are discussed in [19] [50] [51] [68] [71] [79].

This survey concentrates on the assessment of model probabilities rather than methods to improve calibration. Reviews of such methods are covered well by [23] and [87]. The present paper focuses on reporting practical mathematical definitions of metrics, their properties, why they might be used, and a comparison of their performance, where such comparisons already exist in the literature. Experiments that perform numerical comparisons of small subsets of metrics have been conducted – see [28] [42] [61] [75] [84] [104] [109] for example. New experiments to provide a comprehensive comparison would be a significant undertaking requiring thousands of hours of compute time [42] and are not included here for space reasons. A full description of the in-scope probability calibration metrics, arranged by family, now follows.

# Point-based metrics

## Introduction to point-based metrics

Point-based metrics compute a score for each datapoint and aggregate these scores to give an overall calibration measure for the whole dataset. Although easy to define and compute, point-based metrics can only achieve perfect calibration scores when a classifier is 100% confident and is always correct. For lower confidence scores there is always some discrepancy between the confidence and the correct decision for a particular datapoint, as illustrated by the datapoint calibration error in Fig. 2. Point-based metrics often have simpler definitions than the other types of metric. However, some of the individual score functions lack interpretability, despite having potentially useful mathematical properties, which may limit their use. Section 3.2 below describes the abstract concept of “proper scores”, which provides a framework for analysing the properties of point-based metrics. The remainder of this section from 3.3 onwards describes specific point-based metrics.

## Proper scores

Proper scoring rules are point-based evaluation measures for probability estimates that avoid the need for putting confidence scores into bins [87]. Following Bröcker [13], a scoring rule is a function of a probability prediction (or confidence) and an outcome . Let be a length- vector representing the proportion in the data of each class. A scoring function, or simply score, is defined as:

|  |  |
| --- | --- |
|  | (3) |

It is convention that low scores indicate good predictions. A score is “proper” if the divergence is nonnegative, and “strictly proper” if zero divergence implies . Informally, proper scores are optimal when predicted probabilities match the true proportions in the data [87].

Strictly proper scores can be decomposed into three terms: reliability (REL), resolution (RES), and uncertainty (UNC), which facilitates interpretation of the score [13]:

|  |  |
| --- | --- |
|  | (4) |

Reliability is a measure of the degree to which predictions differ from the actual sample relative frequencies [67], with lower values being good. In early works, reliability was referred to as validity [87]. However, it is often now known as calibration loss [87]. As an illustrative example, if a classifier only produces confidence values of exactly 90% or 70%, and is respectively correct 90% or 70% of the time for those specific values, then the classifier is “reliable” and has a reliability value of zero.

Resolution quantifies how much sample proportions for each unique predicted probability differ from the overall sample proportions for the whole dataset [67], with higher values being better. For example, consider a binary dataset where 50% of the datapoints belong to either class. A classifier that always gives a confidence of 50% and chooses randomly between the two classes would have an accuracy of 50% and be well-calibrated globally, but not particularly useful. This classifier has zero resolution. An alternative, reliable classifier that randomly gives confidence scores of 80% or 20% half the time to class 1, has a better, non-zero resolution, despite also having an accuracy of only 50%. Resolution is also known as sharpness [13].

Uncertainty is the score that would be achieved by replacing confidence values with the proportions of the actual samples [67], with lower being better. Thus, this term is inherent to the data and does not relate to the classifier. For example, if 80% of the datapoints are from one class, a theoretical classifier that always gives a confidence value of 80%, would achieve a score equal to the uncertainty value. If all datapoints are from one class, the uncertainty is zero. In a binary classifier, the uncertainty is highest when half of the datapoints belong to each class.

The three-term decomposition of scores is only useful when more than one datapoint has the same predicted probability vector. This may be the case for human forecasters that are prone to specifying probabilities on a discrete scale (10%, 20%, etc.). However, algorithms specify probabilities on a continuous scale from zero to unity and it is unlikely that many predicted probability vectors will have the same exact value. When all probabilities are different, the resolution and uncertainty terms cancel out, and the reliability is the same as the overall score. Thus, the decomposition has less utility in modern algorithmic contexts than traditional human prediction analysis.

An alternative decomposition considers proper scores as a sum of two components: epistemic loss, due to the model not being optimal, and an irreducible or aleatoric loss, which is the loss of the theoretically optimal model, due to randomness in the data [87]. This decomposition helps focus analysis on parts of the problem that are in control of model designers. Other decompositions are also available [80].

Silva Filho et al. [87] recommend that classifiers should be trained using a proper scoring rule as loss function rather than non-proper functions. This is because the resulting models are likely to produce better probabilities, since probability refinement and calibration would be encouraged during training.

## Brier score

The Brier score (BS) is one of the earliest and best-known calibration metrics. It is a common way of measuring how much the accuracy of a model diverges from its confidence, with lower scores meaning that the model is well calibrated. The Brier score is also known as the mean square error (MSE) [22] or the quadratic score [27]. The Brier score can be calculated by equation (5).

|  |  |
| --- | --- |
|  | (5) |

The score can be thought of as the mean of the squares of the arrowed data-point calibration-error line distances in Fig. 2.

Initially developed for weather forecasting, the Brier score is now widely used as a general measure of risk prediction. It has advantages in that it is easily calculated and interpreted and is a strictly proper score. The original definition of Brier score for binary classifiers is in the range . Modern usage divides the classical definition by the number of classes to give a value always in the range [87]. Wallace et al. [101] describe a stratified version of Brier score that captures class-specific calibration. Verhaeghe et al. [100] do not use the Brier score as it does not evaluate the “clinical value” of prediction models. They prefer to use expected calibration error (ECE) or its variants, which are described later in this paper.

The square root Brier score (RBS) is a robust estimator and an upper bound of the canonical calibration error. RBS is compared to bin, kernel, and cumulative metrics in [28]. Of the metrics tested, only RBS and the cumulative Kolmogorov-Smirnoff metric are consistent in value with respect to data size – a desirable property.

## Logarithmic metrics

Negative log likelihood (NLL), also known as binary cross-entropy, ignorance [13], logarithmic score [27], predictive deviance [27], or logistic loss, is calculated for binary problems by the equation:

|  |  |
| --- | --- |
|  | (6) |

The NLL takes on any nonnegative value. When the NLL is small, the model is well-calibrated. Like the Brier Score, NLL is easy to compute but conflates accuracy and calibration [32]. As with the Brier score, NLL is strictly proper and can be decomposed into reliability, resolution, and uncertainty [93]. One issue with NLL is that if the confidence is zero for the correct label for any datapoint, the NLL evaluates to infinity. This reflects the fact that a good prediction system should never assign zero probability to possible events [93]. If the confidence is a small value instead of zero, the NLL can still be very large. Thus, the metric severely penalises highly unlikely predictions, which may be an indicator of lack of calibration. However, this means that single datapoints can have a large effect on the overall metric value, which is an undesirable property [27]. Considering that some datasets have "label noise" where the supposed ground truth labels are incorrect for a small proportion of datapoints, this property of the metric has potential to cause major issues.

The multi-class variant of NLL, also known as multinomial logistic loss, is defined as follows, where is the indicator function, which is unity if and zero otherwise:

|  |  |
| --- | --- |
|  | (7) |

Focal loss (FL) is a modified version of NLL designed to focus on hard-to-classify examples [56]. Focal loss is defined as:

|  |  |
| --- | --- |
|  | (8) |

When , FL down-weights the loss for high-confidence predictions, putting more emphasis on hard examples that are misclassified. The user-defined focussing parameter is usually constant across all datapoints, often taking the value . However, Mukhoti et al. [65] make it depend on the confidence value and Wang et al. [103] set it on a per-data-point basis as part of a meta-learning framework. FL was originally introduced as a loss function for object detection model training, and its use improves calibration, despite it not being strictly proper. However, the loss can be used as a calibration metric. If then this reduces to NLL. Focal loss is usually only computed for the top-confidence class rather than the full class computation in (8). Dual focal loss is a modified version of FL that examines the top two most confident classes to consider the extent to which the top-confidence class leads the other classes [90]. The factor is modified to be , where are the two most confident classes. Using dual FL as a training target results in models that are better calibrated than standard FL. Although focal loss is not strictly proper, it can be made so by adding a scaled version of the Brier score to it to make new metric, focal calibration loss (FCL) [55].

Sumler et al. [88] describe a general metric to measure the consistency of continuous or discrete probabilistic algorithms, including multi-target trackers, classifiers, multi-hypothesis trackers, and particle filters. It is shown that for the single target tracking case this is equivalent to normalised estimation error squared (NEES), a widely used metric in the tracking literature. They derive a simplified version of this metric when applied to binary classifiers and call it the entropic calibration difference (ECD). This is defined as:

|  |  |
| --- | --- |
|  | (9) |

ECD is a signed metric – it can take on any value on the real line. When positive, the classifier is over-confident. When negative, the classifier is under-confident. When zero, the classifier is perfectly calibrated. The signed nature of the metric is an advantage compared to other metrics that only give information on whether a classifier is calibrated but not the direction of miscalibration. The absolute value of ECD is a proper scoring rule.

## Global metrics

Global metrics compute the mean confidence of all datapoints and compare this to the mean accuracy. The two numbers are expected to match for a calibrated system. This type of metric is included as a point based metric due to its use of simple sums over the datapoints. However, it could also be considered to be a bin-based metric with a single bin for the entire dataset. Individual metrics vary in how they compare the global sums. Due to the use of aggregate statistics, global metrics are not proper.

The global squared bias (GSB) is defined by Galbraith et al. [24] as:

|  |  |
| --- | --- |
|  | (10) |

The GSB is a measure of the match between the unconditional mean predicted probability and the unconditional mean probability of the outcome. The metric, as defined in (10), takes on values between zero and unity. However, when comparing GSB with the old definition of Brier score, many authors omit the scale factor in binary problems to keep both metrics on the scale of zero to two. Due to its global nature, the GSB provides only an imprecise measure of calibration error, as the metric can be zero if different parts of the reliability diagram are un-calibrated but cancel out overall. The GSB is equivalent to the mean-square calibration error (see Section 4.3) with a single bin for all probabilities. GSB is also known as reliability-in-the-large [36].

A metric similar to GSB is the multi-class difference of confidence and accuracy (MDCA). It is defined by Hebbalaguppe et al. [37] as:

|  |  |
| --- | --- |
|  | (11) |

MDCA was designed to be used as an additional loss term for mini batches of data in neural net training to encourage calibrated models. However, it can be used to assess calibration of the whole dataset. The metric is differentiable, which allows it to be used as part of gradient descent algorithms. The metric is equivalent to the expected calibration error (see Section 4.2) with a single bin for all probabilities.

Another global metric for binary problems is the ratio of the expected to observed (EO) number of datapoints in the target class:

|  |  |
| --- | --- |
|  | (12) |

The metric takes on non-negative values on the real line. If the model is calibrated, if the model is under-confident, and if , the model is over- confident. This metric is unusual in the sense that it gives a measure of the direction of calibration as well as the magnitude of miscalibration in general. The few other metrics that have this property normally take on any value on the real line, with zero indicating perfect calibration. The inverse of this metric is sometimes reported, in which case it is called the observed to expected (OE) ratio [81].

## Normalised square metrics

The Dawid-Sebastiani score (DSS) is a metric for models that output general probabilistic predictions, using the first two statistical moments of the predictive distribution. DSS is defined in [64] as:

|  |  |
| --- | --- |
|  | (13) |

In (13), is the mean of the probabilistic prediction and is the variance. This general formulation can apply to both continuous and discrete variables. For a binary classifier, and , derived from the Bernoulli distribution. DSS takes on any nonnegative value and is a proper score, although it is not strictly proper. A multivariate version of DSS that takes on any value on the real line is discussed in [105]. This version of the metric can be used to assess multi-class problems.

The normalised squared error score (NSES) is defined by Moore et al. [64] as:

|  |  |
| --- | --- |
|  | (14) |

NSES takes on any nonnegative value. The metric is like the DSS but omits the logarithmic term. However, unlike DSS, NSES is not proper. Despite its impropriety, NSES is considered in [64] to have valuable diagnostic properties, as it can be used to distinguish over-confidence () from under-confidence (). Indeed, NSES is often used in the tracking community as a performance measure for uncertain estimates of target locations where it is called normalized estimation error squared (NEES) [16]. If , which can only happen if both values are zero or unity and the classifier is 100% confident and correct, then the summand is defined as zero to avoid dividing by the zero value of. However, if then the classifier is 100% confident and incorrect, causing the metric value to diverge. As with NLL this is an undesirable property.

## P-norm metrics and variants

The pointwise error (PWE) [54] is a generalised metric based on the p-norm and defined as:

|  |  |
| --- | --- |
|  | (15) |

When , this is the mean absolute error or total variation [95][104]. When , this is the root mean-square error, or the square-root of the Brier score (see Section 3.3). When , this is known as the Chebyshev norm and is equivalent to picking the maximum absolute error for a single datapoint. The Chebyshev norm is unstable, because it relies on a single datapoint and is therefore not recommended as a useful pointwise calibration metric. Values of other than 1, 2, or in the norm are not typically used. The pointwise error takes on values between zero and unity for all values of .

The authors de Leeuw et al. [54] define a modified version of called the L1eps error. This is defined as:

|  |  |
| --- | --- |
|  | (16) |

This is similar to a pseudo Huber loss [7] (see Section A.6 for the standard Huber loss). For small errors L1eps acts like the error, and for large errors it acts like the error. This makes it robust to outliers, while having continuous derivatives for all degrees. In contrast, the standard Huber loss has a continuous first derivative but discontinuous second derivative. The L1eps error takes on values between and

## Alpha scores

Alpha scores raise individual confidence values to the power of an exponent parameterised by . This allows the under- or over-weighting of high confidence values compared to lower ones. The pseudo-spherical score (PSS) is [13]:

|  |  |
| --- | --- |
|  | (17) |

PSS reduces to the spherical score (SS) when and to the ignorance as . The PSS takes on any nonnegative value and is a strictly proper score. The spherical score was found to perform better than the quadratic (Brier) or logarithmic (ignorance) score in a weather forecast application [27]. However, the score was found to be less useful than the logarithmic score and other measures (ranked probability score, DSS, and NSES) in assessing probabilistic predictions of daily deaths from COVID-19 [64].

The power score family of metrics for parameter is defined as:

|  |  |
| --- | --- |
|  | (18) |

This family of scores is strictly proper and takes on values between zero and unity [21]. Typically, , in which case the metric is known as the proper linear score (PLS), or the quadratic score [13]. Although PLS is quadratic, the name is derived from the fact that it is a modification to a naïve linear score. For binary classification problems, PLS is equivalent to the Brier score.

## Soft F1 score

Most of the above metrics are averages of some quantity over all datapoints. If a classifier always produces high confidence values but makes a small number of mistakes, this indicates a bias that will be washed out when computing these metrics, since most datapoints will have a low calibration error. The soft F1 (SF1) score addresses the issue of severe imbalances between mistaken and correct decisions [38]. If , , and respectively denote -length vectors of datapoint confidences, labels, and unity, the soft F1 score is defined as:

|  |  |
| --- | --- |
|  | (19) |

The soft F1 score takes on values between zero and unity.

# Bin-based metrics

## Introduction to bin-based metrics

Point-based metrics such as the Brier score have been used to measure calibration for several decades. However, one issue with such metrics is that it is usually impossible for a classifier to achieve a perfect score under such systems. This is because if any confidence other than pure certainty is predicted for a particular datapoint, there will be a difference between the confidence and the label for the correct class, which by definition is zero or unity. Thus, the only way for a classifier to be assessed to have perfect calibration, is if it only ever assigns a confidence of 100% to the correct class and is always correct. This is impractical for any real dataset.

Bin-based metrics group datapoints with similar confidence values. The proportion of correct classification in each bin can then be compared to the mean confidence of the bin, or some other single value representative of all the datapoints in the bin. This allows imperfect, but well calibrated, classifiers to achieve a low calibration score. An illustration of binned data is shown in Fig. 4. Conceptually, bin-based metrics are based on the difference between the binned accuracy and perfect calibration lines.

The remainder of this section describes several bin-based metrics and some of their advantages and disadvantages. One general disadvantage is that a small change in confidence value of one datapoint can cause it to be assigned to an adjacent bin, resulting in a discrete jump in metric value, which may be undesirable. The following notation specific to binned metrics is used. The number bins is , the number of datapoints in bin is , and the proportion of data in bin is . The mean confidence of bin is and the true proportion of data labels in that bin is . The latter two variables may relate just to class 1 in binary problems or may be vectors for multi-class problems. In an abuse of notation for simplicity, denotes that datapoint is in bin . Unless otherwise stated, bin-based metrics take on values between zero and unity.

## Expected calibration error

Expected calibration error (ECE) is a widely used metric for quantifying the miscalibration of probabilistic classifiers, where lower values indicate better calibration [31]. Several authors use the word “empirical” or “estimated” instead of “expected” to describe the metric, as the ECE is not a true expectation [87]. The binary-ECE is calculated via the formula:

|  |  |
| --- | --- |
|  | (20) |

The standard form of ECE uses bins of equal width (EW). Thus, it is sometimes referred to as ECE-EW [84]. ECE is also known as the mean absolute calibration error [53]. ECE is easy to compute and visualise – it is the absolute area between estimated and perfect calibration bars – see Fig. 4.

Despite its widespread use, the ECE has several disadvantages. First, it is trivially possible to obtain a perfect ECE by randomly estimating examples according to the label distribution [57]. For example, in a binary classification problem, if 60% of the data examples are from class 1, then assigning a confidence of 60% to all items, regardless of the input features, will produce a perfectly calibrated classifier, according to ECE, but one with poor accuracy.

Second, when using fixed bins for the input confidence score, some bins may have very few or no datapoints. Thus, the calibration error for those bins either can’t be computed or has a very high variance, which is then reflected in the overall ECE. The potential sparsity of datapoints for some bins is demonstrated by Guo et al. [31], who provide an example histogram of the number of samples in each bin for a ResNet image classifier. Bins with low confidence values have very few samples, and the lowest two bins have no samples at all.

Third, in multi-class problems, standard ECE is computed just for the highest probability class. It may be the case that it is important to distinguish between second, third, or lower ranking classes where the assessment from the classifier under test is being combined with other information [104].

Fourth, ECE conflates calibration and sharpness when a model is highly accurate [69]. Sharpness is the desire for models to predict with high confidence. This conflation issue is related to the two previous ones. Since only the top-label confidence is analysed, confidence values are naturally concentrated towards unity. As ECE places a low weight on sparsely populated bins with low confidence, the output value relies primarily on the high confidence bins. For these high confidence values, an accurate classifier has tall bin heights that are near to the line of perfect calibration in the reliability diagram. Thus, the accurate classifier naturally has a lower ECE than if confidence values were more uniformly spread.

Fifth, ECE depends on the scale of probabilities. If lots of probabilities are small (e.g. 0.001), this results in a small ECE even if the achieved accuracy is also small but is many factors times the confidence (e.g. 0.01, a factor of 10 difference) [61].

Sixth, ECE is a highly discontinuous function of classifier confidence values due to its fixed width bins. This makes it a difficult metric to use in gradient based optimization schemes [48], and a small change to a single confidence value could have a large effect on the overall ECE.

Seventh, for a model with a certain fixed calibration performance, the value of ECE decreases as the number of datapoints used to compute it increases [28]. This makes it hard to compare performances of models with different size datasets.

Due to these shortcomings, several variations on ECE have been proposed. These are described in subsequent sections. However, we first describe some general definitions for binary and multi-class classifiers that aid discussion of such metrics.

## Mean calibration error

The calibration error (CE), also known as –ECE or mean calibration error, is defined by Lee et al. [53] as:

|  |  |
| --- | --- |
|  | (21) |

This general definition incorporates several well-known specific metrics. Kumar et al. [49] state that the most common value is , and refer to this simply as the calibration error, while Hendrycks et al. [38] call it root mean-square calibration error (RMSCE), and Hagopian et al. [36] call it reliability-in-the-small or just reliability. However, is also commonly used – this defines the standard ECE. Furthermore, , known as the Chebyshev norm, defines the maximum calibration error (MCE), discussed in Section 4.5. In tests on synthetic data with known properties, hypothesis tests based on -ECE are shown to outperform -ECE, in terms of missed detections of miscalibration [53]. Chidambaram et al. [18] consider to be the “right choice” of -ECE, as it is the conditional expectation of the Brier score and is the only -ECE variant induced by a Bregman divergence. Bregman divergence-based metrics have convenient theoretical properties and are closely related to strictly proper scoring rules [18]. Values of other than 1, 2, or are not typically used.

## Binned multi-class calibration error

The mean top-label confidence for a bin is given by (22), where is the maximum confidence over all classes for a particular datapoint .

|  |  |
| --- | --- |
|  | (22) |

The mean top-label outcome for a bin is given by (23), where if the top-label confidence is correct, or if the top-label confidence is incorrect, for a particular datapoint .

|  |  |
| --- | --- |
|  | (23) |

The top-label classification error (TCE) [49], or “confidence-ECE” [87], is then given by (21), with and replaced by and .

The class-specific mean confidence for a bin and class is given by (24), where is the class-wise confidence for a particular datapoint in that bin.

|  |  |
| --- | --- |
|  | (24) |

The class-specific mean outcome for a bin is given by (25), where is the Kronecker delta function that is unity when its arguments are the same and zero otherwise. This represents the presence of that class in a particular datapoint in each bin.

|  |  |
| --- | --- |
|  | (25) |

If denotes the importance of each class, the marginal classification error (MCE), usually with , [49] is:

|  |  |
| --- | --- |
|  | (26) |

If for equally important classes, and , this is also known as class-wise ECE (CWCE) [87], macro subset ECE (MSECE) [70], or static calibration error (SCE) [69]. If is in proportion to the number of datapoints in class , and , it is known as the weighted subset expected calibration error (WSECE) [70]. The unweighted terms of the outer sum in (26), may be considered class-stratified versions of the generalised ECE metric [83].

For full multiclass-ECE, probability vectors could in theory be binned in simplex space. The difference between the mean probability vector and the vector of class proportions in each bin could then be computed. However, most bins would likely be empty or have very few datapoints [87]. This can be seen as follows. If each dimension (class) is subdivided into values, then the total number of bins is . As a concrete example, if a coarse grid is defined with divisions of width 20% each and there are classes, then the total number of bins in the simplex is 125. This is already much higher than the typical number of bins used in the calculation of ECE. Considering many problems have more classes than this, full multiclass binned metrics are not a practical measure for .

## Variants of expected calibration error

Several minor variants to ECE have been proposed. Rather than using equal-width confidence bins, adaptive calibration error (ACE) by Nixon et al. [69] uses bins based on fixed percentiles of the confidence scores in the test dataset so that each bin has the same number of datapoints. ACE is computed class-wise over all classes for multi-class problems, like the static calibration error. ACE is also called the equal-mass (EM) ECE, or ECE-EM [84]. Binning-based estimators with bins of equal mass have lower bias than estimators with bins of equal width [84]. EW and EM methods of binning are also referred to as width binning and frequency or quantile binning, respectively [87]. A disadvantage of ECE-EM is that some parts of the confidence space may have wide bins, preventing the ability to model variation in accuracy in those regions. The equal-area ECE (ECE-EA), or “equiareal ECE”, has bins with approximately equal area, providing a middle ground between ECE-EM and ECE-EW [83]. Thresholded Adaptive Calibration Error (TACE) uses frequency binning but only includes datapoints with a confidence above a certain threshold [69]. The logic behind this is that in situations with many classes, a lot of the class probabilities are low, and this washes out the calibration score. A threshold of 0.01 is used in [69]. When only the top confidence datapoints are selected, the method is called ECE@k [30]. ACE and TACE are both relatively robust to label noise, where lower-rank predictions are more important [69]. Nixon et al. recommend that ACE generally be used in favour of the other metrics or standard ECE. However, if the number of classes exceeds 100, they recommend TACE. Nevertheless, it should be noted that ACE and TACE only measure class-wise calibration, not full multi-class calibration.

Another way of assessing the calibration of confidence values is to measure the area between the reliability plot and the line of perfect calibration, as seen in Fig. 3. This is known as the area between curves (ABC) or the integrated calibration error in [36]. The total absolute area is the standard ECE when the curve is constructed as a binned estimate. It is also possible to provide separate reports of the area above the curve, representing under-confidence in some regions, and the area below the curve, representing over-confidence in other regions [36]. Reporting above and below areas is a useful decomposition of ECE (which is the sum of the two areas) to better understand under and over-confidence. However, the need to analyse two variables makes above/below ABC harder to use when assessing multiple recalibration algorithms automatically, like the Cox intercept and slope (see Section 5.12).

Maximum calibration error (MCE) is a similar metric to ECE but instead of measuring an average of the calibration errors, MCE measures the largest calibration error. This is useful when it is important for the model to be extremely well calibrated across a range of confidence values. It is calculated by the equation:

|  |  |
| --- | --- |
|  | (27) |

MCE can lead to unintuitive results when there is wide variance in calibration between histogram bins, which is more likely to happen when the test set is small. In these situations, the metric is highly sensitive to the placement of bins [87]. The metric may be most suitable for safety-critical applications, where it is important to understand the worst-case calibration at any confidence level. This maximum statistic is also known as the or Cheyshev norm [54].

Region-balanced ECE (RBECE) is defined in [20] as:

|  |  |
| --- | --- |
|  | (28) |

In (28), is the set of eligible bins that each contain a minimum number of datapoints. Note that the metric gives equal weight to all bins rather than weighting them by sample size. The logic behind his metric is that in standard ECE some bins may have a small number of datapoints and thus a high variance. Conversely, in equal mass ECE, the bins may all be concentrated in a particular part of the confidence space, for example if a model outputs many high confidence estimates, the overall ECE-EM will be biased towards confidence in that part of the space. RBECE provides a middle ground between the two extremes [20]. However, the metric does ignore data in sparse regions.

The monotonic sweep calibration error ECE-SWEEP is a bin-based calibration metric. This metric chooses the largest number of bins for which the bin heights, as computed by standard ECE, are monotonic. When tested on data, it is found that the optimal bin count grows with sample size [84]. An efficient implementation of the sweep method for choosing the number of bins is the bin count search method – see Section 4.12 for its application as part of a de-biased metric. ECE-SWEEP has a lower bias than several other metrics, including standard ECE – see Section 4.13 for a discussion. The disadvantage of ECE-SWEEP is that its standard definition is for binary rather than multiclass classification. However, measurement of class-wise calibration can be achieved through a one-vs-rest strategy.

## Partially binned metrics

Distance from calibration error (DCE) is introduced as a theoretical measure by Błasiok et al. [10]. Angelopoulos et al. [1] provide a bin-based empirical estimate of DCE. The bins are of equal width. Let the confidence assigned to a bin be the upper edge of the bin. The estimate of DCE is then:

|  |  |
| --- | --- |
|  | (29) |

The DCE estimate in (29) is shown in [1] to be a tight upper bound on the true DCE for a sensible choice of as the sample size grows. Note that this metric operates on labels as points but confidences as bins.

The Sanders-modified Brier score (SMBS), or Murphy’s Brier score, is a partially binned version of the Brier score [36]. Like DCE, SMBS uses binning to estimate aggregate confidence values in a bin while keeping the labels as individual points. A subtle difference between DCE and SMBS is that DCE uses max aggregation, but SMBS uses mean aggregation to obtain . SMBS is defined as:

|  |  |
| --- | --- |
|  | (30) |

For real datasets, the difference between SMBS and Brier Score is small, and like the Brier score, SMBS can be decomposed into uncertainty, reliability, and resolution [36]. A disadvantage of SMB is that it can only achieve a perfect score of zero when all confidences are zero or unity and the model is always correct. Apart from the difference in aggregation strategy, SMBS is the version of DCE.

Label-binned calibration error (ECE-LB) uses binning to estimate true proportions of labels but operates on individual samples . This contrasts with DCE and SMB, which bin confidences but not labels. ECE-LB is defined by [84] as:

|  |  |
| --- | --- |
|  | (31) |

It can be shown that ECE-LB is at least as large as standard ECE. ECE-LB has the advantage over many other binned metrics that it takes into account the variation of confidence values in each bin rather than relying only on their average. ECE-LB is named probability deviation error (PDE) in [94], where it is shown to have a lower bias than ECE.

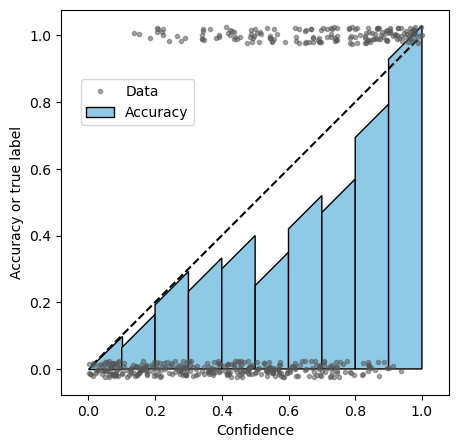
## Fit-on-the-test ECE

Fit-on-the-test (FOTT) calibration error is a metric sub-family that compares a calibration map estimated from data, including bin-based maps, to model confidence values. ECE-FOTT is defined in [42] as:

|  |  |
| --- | --- |
|  | (32) |

Kängsepp et al. [42] define a binning scheme where instead of assuming constant probabilities based on the mean within a bin, bin probabilities may be non-continuous piecewise linear functions of the input confidence. This binning scheme defines . They show that, under the FOTT paradigm, classical ECE is equivalent to the subfamily of such functions where the slope of each bin is unity. This leads to the “tilted-roof” reliability diagram, where the roofs of the bins all have angles of 45 degrees to the horizontal. This makes it easier visually to assess the amount of miscalibration in comparison to the perfect calibration identity line, also at 45 degrees. An example of a tilted-roof reliability diagram is shown in Fig. 5, for the same input data as Fig. 4.

The number of bins to be used for ECE can be selected through cross-validation by optimizing the ECE-FOTT loss. The optimum number of bins for a calibration task containing 5000 datapoints was 14. It is interesting that this is in the range of 10, 15, or 20 bins that are usually arbitrarily used as part of standard ECE calculations. This suggests the number of bins typically used is sensible. However, according to [80], there is no optimal default number of bins, since every scenario has its own bias-variance trade-off.



**Fig. 5** Tilted-roof reliability diagram for 500 labelled datapoints, with ten bins. The diagram facilitates comparison of bar heights with the perfect calibration line.

## Multi-partition metrics

The interval calibration error (ICE) is a theoretical metric that averages the ECE over all possible bin widths and locations. In practice, it cannot be computed directly, as the number of possibilities is huge for large datasets. However, a surrogate ICE (SICE) metric can be computed through Monte Carlo sampling in two stages as outlined by Błasiok et al. [10]. The first stage computes a metric called the random interval calibration error (RICE), which is based on a modification of the bin-based ECE. For a single Monte Carlo run of RICE, all bin widths, except two, are set according to for some integer . The first bin width is chosen randomly in the range and this determines the final bin width, as all bin widths must sum to unity. RICE is then defined for bins, each having an interval denoted by and containing datapoints, and Monte Carlo runs as:

|  |  |
| --- | --- |
|  | (33) |

A maximum number of bins to test is chosen via for a maximum calibration error . SICE is then computed as the optimization over a series of exponentially smaller bin widths:

|  |  |
| --- | --- |
|  | (34) |

The random aspect of SICE averages out the effect of discontinuous jumps in the calibration map at bin edges and thus makes it a consistent estimator but may make it undesirable for repeatable assurance applications. In tests, SICE is better than standard ECE but not as good as the Laplace kernel calibration error (see Section 5.5) or the smooth calibration error (see Section 5.7).

Another metric that aggregates over all possible intervals is the cutoff calibration error (CCE) [85]. This uses the same definition as the maximum calibration error in (27), but with generalized bin widths and locations. Unlike ECE, CCE is “testable”, defined as being able to test the hypothesis that the true theoretical metric for a calibrated system is less than a specified threshold, based on an estimate using finite data [85]. Other advantages of CCE are that it has no adjustable parameters and is a continuous function of confidence values, unlike most bin metrics. A disadvantage is the need to perform search over intervals, which complicates implementation.

## Signed calibration error metrics

Signed calibration error metrics measure both under and overconfidence. This beneficial under/over discrimination property is not common among metrics. The expected signed calibration error (ESCE) [100] or miscalibration score (MCS) [2] is defined as:

|  |  |
| --- | --- |
|  | (35) |

The standard definition of this metric uses equal width bins. To reduce the well-known high local variance of binned estimators, Verhaeghe et al. [100] use the mean of this metric over uniformly distributed bin sizes in the range 0.005 to 0.05. This averaging process is also used by the authors for computing standard ECE, a process with some similarities to SICE. The range of ESCE is . A classifier is over-confident if ESCE<0 and under-confident if ESCE>0. The magnitude of ESCE is equal to or less than ECE. If this indicates the classifier is over and under-confident in different parts of the confidence space.

For multi-class problems, a variant of ESCE exists, called weighted subset MCS (WSMCS) [2]. This computes the MCS individually for each class (i.e. class-wise). All under- or over-confident classes according to this metric are grouped into one of two groups. For each under- or over-confident group, a group-wise MCS is computed as a weighted sum of the class MCS values, with the weights in proportion to the class size (a micro-average). The overall WSMCS is the weighted sum of the two group-MCS values, with weights in proportion to the number of classes in each group (a macro average). This metric takes into account the imbalance between the numbers of under/over-confident classes [2] and is in the range . However, under/over-confident classes can cancel each other out, so a low absolute value may mask high levels of miscalibration in individual classes, which reduces the utility of the metric.

Cabitza et al. [14] define a local (LICI), global (GICI), and directional (DICI) interpretable calibration index based on geometric considerations of the bin-based reliability curve. The LICI of a bin is the normalised diagonal distance between its position on the reliability curve (with confidence and accuracy ) and the nearest point on the line of perfect calibration. The GICI is the sum of LICI values weighted by the number of datapoints in each bin. A separate GICI can be computed for the two sets overall of under or over-confident bins. DICI is the difference between the two GICI values. A single GICI can also be computed over all bins. The distance normalization process in the computation of GICI addresses the criticism of ECE that it depends on the scale of probabilities. DICI has a range , with the sign indicating overall under or overconfidence on average.

## Soft bin metrics

Soft-binning ECE (SBECE), as devised by Karandikar et al. [43], uses soft binning to obtain a metric that is differentiable, allowing it to be used as a loss function to encourage a calibrated model while training the model using gradient descent. The first step is to define a model to classify a confidence value as belonging to a particular bin. If is the vector of bin centres, then the vector of bin probabilities, or bin membership function, is:

|  |  |
| --- | --- |
|  | (36) |

In (36), is a temperature parameter that controls the softness of the binning. As the output tends to a one-hot vector, which is equivalent to standard hard binning. Bohdal et al. [12] show that SBECE is relatively insensitive to temperature parameter values in the range 0.0001 to 0.01.

The soft-binned size, confidence, and accuracy and of bin are:

|  |  |
| --- | --- |
|  | (37) |
|  | (38) |
|  | (39) |

In a similar manner to the mean calibration error in (21), the SBECE is then computed as:

|  |  |
| --- | --- |
|  | (40) |

Differentiable ECE (DECE) is similar to SBECE but uses a different bin membership function [12]. If the upper edges of bin are , weights are defined as , and biases are defined as , then the probability of a confidence belonging to bin is:

|  |  |
| --- | --- |
|  | (41) |

Like SBECE, DECE is relatively insensitive to temperature parameter values in the range 0.0001 to 0.01. However, DECE better approximates ECE than SBECE and produces better calibrated models, as measured by ECE, when used during training. Both soft-bin metrics are reminiscent of kernel-based metrics (see Section 5), with the membership functions over bins taking the role of kernel functions over datapoints. An alternative differentiable ECE metric uses the LogSumExp function to soften the choice of class with the maximum confidence value [102].

Gupta et al. [33] propose a metric called robust expected calibration error (RECE) that incorporates into the metric the uncertainty in predicted top-label confidences. It is not strictly a soft-binned metric as datapoints are still hard-binned, but the influence of points near bin edges is down-weighted. Two variants are proposed. The first, RECE-G, assumes that uncertainty can be represented by a Gaussian distribution. A weight is defined for each point in fixed-width bins that are defined by lower and upper bounds and :

|  |  |
| --- | --- |
|  | (42) |

Function in (42) represents the probability of the bin interval for a Gaussian distribution with mean and standard deviation , which is a hyperparameter fixed to be the same value for all datapoints (0.1 used in experiments). RECE-G is then defined as:

|  |  |
| --- | --- |
|  | (43) |

The second variant, RECE-M, is based on a Gaussian mixture model trained on confidences produced by multiple random test-time augmentations of features. In a comparison of bin-based metrics with 15 bins and using several datasets, the typical ordering of bias from least to most is: RECE-G, RECE-M, class-wise ECE-EW, top-label ECE-EW, then top-label ECE-EM. In addition to low bias, both variants of RECE had lower variance of estimates than the ECE variants [33]. Given the similarity in performance between RECE-G and RECE-M, and considering the greater complexity and randomness of RECE-M, RECE-G should be the preferred metric of the two.

## Overlapping-bin metrics

One of the criticisms of bin-based metrics is the discontinuity at bin boundaries. The CalBin metric by Bella et al. [8] addresses this issue by using overlapping bins of equal mass. The first bin is defined has the first datapoints. The second bin has datapoints . This sliding bin definition is repeated to the end of the dataset, and the errors are averaged over the bins. CalBin is defined as:

|  |  |
| --- | --- |
|  | (44) |

The value for is arbitrary but Bella et al. use .

Another overlapping bin metric is the k-nearest neighbours (KNN) ECE, or ECE-KNN by Peng et al. [75]. This is defined in the same way as standard ECE, except that there is one bin per datapoint, and the other points in each bin are the k-nearest neighbours to the point in question, in terms of confidence value. A partially manual process for selecting k is given. ECE-KNN has a lower bias than ECE-EW, ECE-SWEEP, and ECE-DEBIAS when assessing uncalibrated models [75].

## De-biased metrics

Kumar et al. [49] note that the so-called “plugin estimator” (PL) of the calibration error in (21) is biased. It is possible to construct a de-biased (DB) estimator by subtracting an approximate error term. For the error, the de-biased square estimator (CE2-DB) is:

|  |  |
| --- | --- |
|  | (45) |

The square root of CE2-DB is known as the debiased root mean square calibration error (DRMSCE) in [77]. Like ECE, including a class-wise extension, the computational complexity of DRMSCE for classification is , with only a simple additional debias term required. Therefore, the metric scales well as the number of classes grows. However, full multi-class calibration is not available due to requiring an unspecified be-bias term.

For the error, the de-biased estimator (ECE-DB) is approximated by:

|  |  |
| --- | --- |
|  | (46) |

In (46), is a random sample from the normal approximation of the label distribution in bin , and is the number of Monte Carlo samples [49]. The double summation computes an approximation of the expected value of the plugin ECE estimator, and the large parentheses contain the overall bias. The additional summation makes the computational complexity is . The stochastic nature of ECE-DB limits its reproducibility.

In the limit of infinite size datasets, both CE2-DB and ECE-DB take on values between zero and unity. However, for finite datasets, both metrics can take on small negative values, due to the de-biasing term. Bootstrap methods can be used for hypothesis testing. Using the biased plugin calibration error estimator to test for calibration leads to rejecting well-calibrated models too often [49]. That is, there are too many false alarms when attempting to detect miscalibrated models [104]. Therefore, the de-biased estimator should be used for more refined hypothesis testing. Kumar et al. [49] use equal mass as their binning strategy.

Petersen et al. [77] define an adaptive bin count search (BCS) method to decide the number of bins to use with any binned metric. Like ECE-SWEEP, the method chooses the largest number of bins for which mean confidence values in bins, as computed by the base metric, are monotonic. The implementation of BCS is more efficient than SWEEP as it is based on an interval search method. The method is described in Section A.15, the pure version of which has no adjustable parameters. Petersen et al. [77] compare the performance of ECE-EW, ECE-EM, and DRMSCE, each using BCS or a fixed , while using synthetic data with known zero or non-zero miscalibration. The BCS versions of all metrics generally have a lower bias than non-BCS versions and DRMSCE-BCS outperforms all other metrics tested. This comparison positions DRMSCE-BCS as a good candidate for recommendation.

Xiong et al. [107] show that the confidence scores for data in sparse areas of the feature space tend to be over-confident, and scores in dense areas tend to be under-confident. If the scores for dense and sparse data lie in the same confidence bin, then these can cancel out making a model seem more calibrated according to ECE than it really is. To mitigate this proximity bias or cancellation effect, the proximity-informed ECE (PIECE) metric is proposed. The metric bins datapoints by both confidence value and “proximity” value, where proximity is based on the mean distance of a datapoint to its ten nearest neighbours according to their feature values. PIECE is defined as:

|  |  |
| --- | --- |
|  | (47) |

The bins are equal mass, with the number of confidence bins and number of proximity bins . For the same dataset and model, PIECE is always at least as large as standard ECE, making it a stricter score. The metric can be used to assess the ability of recalibration algorithms to address the proximity bias problem [107]. However, it has the disadvantage that it needs access to feature vectors, which may not be available in all settings. Since bins are two-dimensional, each bin has fewer datapoints than ECE making the variance in bins higher. A similar metric to PIECE, that also analyses feature values, is the kernel-based kernel local calibration error described in Section 5.5.

Yang et al. [108] introduce the partitioned calibration error (PCE). This operates in a similar way to PIECE but is more general as it allows any grouping of datapoints based on confidence or feature values and includes the possibility of averaging over different partitions (ways of grouping) of the same dataset. If there are partitions of the data, each partition has groups, and a general loss function is defined, PCE is defined as:

|  |  |
| --- | --- |
|  | (48) |

PCE takes on many other metrics as special cases. For example, if there is one partition of the dataset into bins of confidence values, and the loss function is , then this is standard ECE. If there is one partition of the dataset putting each datapoint into its own group and then this is the Brier score. Other point-, bin-, and kernel-based metrics may similarly be obtained [108]. The general PCE metric also encompasses the cumulative multi-calibration metric discussed in Section 6.2, which analyses possibly overlapping sub-populations of data. For notational convenience, (48) shows the mean accuracy and confidence , but other statistics may also be used. PCE allows the cancellation effect to be mitigated by grouping the data appropriately. However, as with PIECE, it has the disadvantage that grouping needs access to feature vectors when used in this way. Issues to do with bias and loss relating to grouping effects are discussed more in [76]. PCE may also be classified as a multi-partition metric (see Section 4.8).

Huang et al. [41] define the multi-view calibration error (MVCE) in a similar way to PCE, using the error as the loss function. In their paper, the term “partition calibration error” is defined more simply to be the loss term inside the error. The partition strategy is for each partition to be a random shuffle of the datapoints, followed by an equal-mass scheme used to create the bins. The number of partitions is set to 100 in experiments. Experiments conducted with for all metrics show that MVCE has the lowest bias, followed by ECE-SWEEP then ECE. MVCE and ECE used 32 bins, while ECE-SWEEP automatically sets the number of bins as part of its calculation [41]. The random nature of partitions is reminiscent of RICE, but unlike RICE the confidence values in each partition are not sorted by value. Despite their low bias, the random nature of RICE and MVCE limits their reproducibility.

Pan et al. [73] introduce the field-level expected calibration error (FECE), also known as Field-ECE, which measures calibration bias in sensitive input fields of interest to the decision-maker (e.g. protected characteristics of a person). The mathematical definition of FECE is the same as PIECE in (47) but with the proximity bins replaced by non-overlapping partitions of the data based on feature values and a single bin for confidence values. A large FECE value indicates that predictions are biased in some part of the data, and examining the contributions from induvial bins reveals the location of bias. Experiments in [73] use discrete variables as the sensitive field, setting to the number of possible values. Kelly et al. [44] define a metric called variable-based expected calibration error (VECE) that is the same as FECE, and their experiments examine binned continuous feature values one variable at a time. VECE is shown to reveal location-based bias that is masked by ECE. Ranking features by decreasing VECE highlights variables that are worth investigating for bias.

## Bias analysis

A framework known as bias by construction (BBC) is used by Roelofs et al. [84] to analyse the bias of various metrics. In this framework, the probability estimates produced by classifiers for real data are used to fit a parametric model for the density of confidence values and calibration curves. The fitted models are then used as ground truth to generate large amounts of synthetic data. The framework has been used to compare ECE, ECE-DEBIAS, ECE-SWEEP and kernel density estimation (KDE) metrics. KDE is described in Section 5.6 below. The analysis also compares equal mass and equal width versions of the ECE metrics. In all cases, equal mass versions of ECE show a lower level of bias than the equal width versions as the number of samples increases. The ranking of bias, for realistic calibration curves, from least to most biased is: ECE-SWEEP, ECE-DEBIAS, ECE, KDE. Poor performance for KDE could be due to the standard parameters of the metric, which may only be optimised for the simple synthetic datasets used in [109]. For perfectly calibrated classifiers, ECE-DEBIAS is better than ECE-SWEEP. Therefore, there is little to choose between ECE-DEBIAS and ECE-SWEEP. For either method, at least 500 samples are required to reliably detect a classifier with a 10% calibration error, and 10,000 samples are required to detect one with a 2% error.

Whether equal-width bins or equal data size bins are used, standard binned metrics like ECE increase their value as the number of bins increase, for a fixed number of datapoints. The expected values of binned metrics stay bounded away from zero if the number of datapoints per bin remains bounded on some fixed range of scores, as the maximum bin width becomes arbitrarily small [3].

## Hypothesis-test bin-based metrics

Hypothesis tests can be developed to assess the significance of differences of a measured calibration metric from zero, under the assumption that a model is well calibrated. These tests are usually based on a test statistic, which can on its own be a metric for calibration. Hypothesis tests can be developed for most metrics through bootstrapping schemes. Caution should be noted when interpreting p-values as most tests are based on some kind of approximation. When the approximation does not hold well, a non-significant result for small samples could lead to undeserved claims of good calibration and a statistically significant result for large samples might unjustly represent trivial miscalibration [96]. This section includes bin-based metrics specifically designed with hypothesis testing in mind.

The Hosmer-Lemeshow (HL) statistic has been a popular calibration measure of binary classifiers, especially in the medical field [40]. It was originally used with logistic regression classifiers [53] but has wider applicability. The “c-statistic” version is equivalent to an equal mass binning scheme with bins. The “h-statistic” version is equivalent to an equal width binning scheme with the same number of bins. For either scheme, the test statistic is usually defined as:

|  |  |
| --- | --- |
|  | (49) |

This can be re-written as:

|  |  |
| --- | --- |
|  | (50) |

The HL statistic takes on any nonnegative value and follows a chi-squared distribution with degrees of freedom. From this, a standard chi-squared hypothesis test can be undertaken. A version of the HL statistic for measuring the class-wise performance of multiclass problems has also been defined [87]. The HL test has satisfactory power when used with at least 400 samples [87]. However, one disadvantage of the statistic is that it is highly sensitive to small amounts of miscalibration, especially when the number of datapoints is large. This makes it a more difficult metric to work with when comparing classifiers that, in real situations, are unlikely to be perfectly calibrated. Furthermore, non-significant results can be produced for small samples. Consequently, the HL test is considered by some authors to be an outdated approach [96].

Test for calibration (T-Cal) [53] is a hypothesis test based on a de-biased plugin estimator (DPE) of -ECE.

|  |  |
| --- | --- |
|  | (51) |

DPE is similar in form to the de-biased ECE in (45) but with the de-bias term computed in a different way. In the limit of infinite size datasets, DPE takes on values between zero and unity. However, for finite datasets, the metric can take on small negative values, due to the de-biasing term. For multi-class problems, the binning scheme used for DPE is an equal-volume partition of the simplex. However, it is noted that other binning schemes could be used. To select the optimum number of bins, the basic T-cal test requires knowledge of the smoothness of the calibration map, which is not generally known in practice. Therefore, an adaptive scheme performs hypothesis tests for a range of numbers of bins and rejects the null hypothesis of calibration if any test is rejected. In tests on synthetic data with known properties, T-cal is shown to outperform Cox’s method (see Section 5.12) and -ECE, in terms of missed detections of miscalibration.

Sun et al. [89] show that the distribution of -ECE may be approximated by a normal distribution and thus can be used to compute confidence intervals and perform hypothesis testing. Hypothesis tests based on this normal approximation or T-Cal have similar power, but the normal approximation is computationally more efficient as it is based on analytic computation, whereas T-cal is based on Monte Carlo simulations.

Matsubara et al. [61] define a general calibration error (GCE) metric system. This has ECE and ACE as special cases. To make a general scheme specific, one must choose a loss function, binning scheme, and norm metric. The test-based calibration error (TCE) metric is a specialization of GCE and is good for situations with imbalanced classes [61]. For TCE, the loss is the rejection percentage under a statistical hypothesis test at an alpha confidence level. TCE examines whether each model prediction in a bin can be regarded as an outlier relative to the empirical probability for that bin. The choice of binning scheme and norm for TCE is arbitrary. However, the authors choose the norm and a generic method for deciding on how to arrange the bins called “Near-Optimal Bins Based on PAVA-BC” (NOBB-PAVA-BC). To achieve this, the pooled adjacent violators algorithm (PAVA) used in isotonic regression is modified with block constraints (BC) to have minimum and maximum bin sizes and combined with other steps to form the algorithm. The authors demonstrate the value of TCE by testing it with simulated data, and under-, well-, and over-calibrated models. ECE and ACE are shown to have inconsistent behaviour whereas TCE is more consistent. Although the computation of TCE is , a disadvantage of its practical implementation is that it is 100 times slower to compute than the other metrics. However, the total computation time for 50,000 datapoints is only one minute, which is reasonable for offline scenarios for testing the calibration of models [61]. Nevertheless, the complicated definition of TCE may limit its use.

## Discussion of other binning issues

A general problem with bin-based schemes is that the true calibration is unmeasurable with a finite number of bins. As the number of bins is increased, the measured calibration error increases [49]. It is also the case that any binned version of calibration error underestimates the true calibration error in the limit of infinite data. However, for finite datasets ECE can under or overestimate the calibration error [84].

# Kernel-based and fitted-curve metrics

## Introduction to kernels and fitted-curve metrics

Among the disadvantages of bin-based metrics are the arbitrary grouping of datapoints into bins, and discrete jumps between bins. Kernel or curve-based metrics fit a smooth calibration curve to the data and compare this to the perfect calibration line. In Fig. 2, the calibration error is based on the area between the fitted calibration curve and the perfect line. With kernel or curve-based metrics, small changes in confidence values of individual datapoints result in small changes in metric values. Metrics in this family employ models of varying complexity to fit the data.

## Kernel basics

Kernel-based metrics seek to obtain a smooth calibration curve based on a locally weighted sum of the datapoints. The weighting is defined by the kernel function , which integrates to unity. Kernel estimates generally perform better than bin-based metrics on various criteria, such as continuity of output with respect to input and rate of decline of mean-squared error with sample size [24].

The general difference-kernel estimation of the smooth calibration curve is:

|  |  |
| --- | --- |
|  | (52) |

The parameter is known as the bandwidth, or simply the width. Large widths smooth out the estimate but are unable to model sharp changes in gradient; small widths make the estimate depend more on the local data structure but may be prone to over-fitting. It is typical for the width to be the same for all datapoints. Three common fixed-width methods for choosing are:

* **Normal rule of thumb**. This sets , where is the standard deviation of the samples [109]. This is an optimum choice if the underlying distribution is Gaussian and Gaussian kernels are used.
* **Median heuristic**. This is the median of the distance between all pairs of samples [26].
* **Cross validation**. This divides the data into cross-validation folds and computes the performance of the system on a held-out folds while also varying values of . The overall performance curve for a particular value of is computed as the mean of the performance from held-out folds. The optimum of the performance curve is used to select [24].

Adaptive widths can be used to improve on fixed-width estimates, which can be poor in sparse parts of the distribution [82]. The adaptation makes the widths larger in regions with fewer datapoints.

Detecting miscalibration is only possible with a finite dataset when the conditional probabilities of the classes are sufficiently smooth functions of the predicted confidences, such as kernel-based ones [53]. Smoothness is implied in many calibration measurement schemes but is not usually directly addressed.

Kernel-based calibration metrics are based on the differences between the estimated calibration map and the perfect calibration line (see Fig. 2). Unless otherwise stated, kernel-based metrics take on values between zero and unity and have computational complexity .

## Binary classifier kernel metrics

The mean squared calibration error (MSCE) is defined in [24] as:

|  |  |
| --- | --- |
|  | (53) |

The Gaussian kernel function is used to compute with a width of 0.08, which was optimised through cross-validation, although analysis showed that results are not sensitive to the choice of bandwidth.

Smooth ECE (SECE) is defined in [103] as:

|  |  |
| --- | --- |
|  | (54) |

The Gaussian kernel function is used to compute with a width of 0.01, which was selected via grid search. SECE should not be confused with a different kernel metric of identical name, but denoted SMECE, described as follows, or the similarly-named smooth calibration error, described in Section 5.7.

Smooth ECE (SMECE) [11] is defined in a similar manner to SECE. However, it performs smoothing on the residual rather than the label , which, in addition to a number of other design choices, gives it better mathematical properties. The metric uses a reflected Gaussian kernel to deal with edge effects in the interval , which alleviates bias from the standard Gaussian kernel. SMECE is a consistent calibration measure and is efficient with respect to both sample complexity and runtime. Crucially, it is hyperparameter-free, as a specific scheme is used to choose an optimal kernel bandwidth parameter. A grid-based computation of the metric over grid points is:

|  |  |
| --- | --- |
|  | (55) |

The kernel bandwidth parameter is the standard deviation of the un-reflected Gaussian kernel. Its optimum value is set so that , which can be determined efficiently via binary search. Uncertainty quantification of SMECE can be determined via bootstrapping. The method behind the metric can be used to produce an associated reliability diagram, including a confidence interval for accuracy values at each input confidence value.

Popordanoska et al. [78] define a general kernel density estimator (KDE) and call it ECE-KDE. For the binary classification problem, a partially de-biased beta kernel density estimator (BKDE) specialization is defined as:

|  |  |
| --- | --- |
|  | (56) |

The beta kernel is defined as:

|  |  |
| --- | --- |
|  | (57) |

In (57), and and is the bandwidth parameter. BKDE is a proper scoring rule [80]. Further discussion about BKDE is given below in conjunction with its multi-class extension, the Dirichlet kernel density estimator.

## Dirichlet kernel density estimator

Popordanoska et al. [78] introduce the Dirichlet kernel density estimator (DKDE) to measure strong calibration error in multiclass problems as a different specialization of ECE-KDE. DKDE is known as ECEKDE in [28]. DKDE is computed using the p-norm of a vector based on the label and confidence vectors, and is defined as:

|  |  |
| --- | --- |
|  | (58) |

The Dirichlet kernel is:

|  |  |
| --- | --- |
|  | (59) |

In (59), and is the bandwidth parameter, as before. In experiments, the bandwidth is chosen from a list of possible values by maximizing the leave-one-out likelihood, resulting in typical optimal values of between 0.0001 and 0.001 for . The sensitivity of DKDE with respect to bandwidth is not reported. A de-biasing scheme is available for or . With , DKDE is a proper scoring rule [80].

The advantages of BKDE and DKDE are that they are consistent estimators (unlike ECE or MMCE), scalable with respect to number of classes (unlike ECE or Mix-n-Match), de-biased (unlike Mix-n-Match or MMCE), and differentiable (unlike ECE). Computation of BKDE and DKDE takes time. Scalability of kernel-based metrics with respect to the number of classes is not in general guaranteed [80]; DKDE achieves this by using an appropriate kernel. The property of differentiability allows BKDE and DKDE to be used as a target in gradient-based training algorithms [78]. An empirical test of DKDE in [29] with and found it took in the order of 20 seconds to compute the metric. This may be an acceptable wait time for offline assessments of models.

## Pairwise comparison metrics

Pairwise comparison metrics compare pairs of individual point calibration errors, using a kernel to weight contribution of each summation term. The metrics vary according to the kernel used, which terms to include in the summation, and whether the metric assesses class-wise calibration or strong multiclass calibration.

The maximum mean calibration error (MMCE) is a kernel-based error introduced by Kumar et al. [48]. The motivation behind this metric is to use it as a supplementary target during classifier training. The claim is that other train-time calibration methods based on entropy penalties or temperature smoothing usefully reduce aggregate calibration error but undesirably suppress legitimately confident individual predictions. MMCE can be computed from (60).

|  |  |
| --- | --- |
|  | (60) |

Choice of kernel function is arbitrary, but the implementation selected in [48] is the Laplacian kernel with a width of :

|  |  |
| --- | --- |
|  | (61) |

A more complicated weighted version of MMCE equalises the effect of correct and incorrect examples for multi-class problems, which result in imbalanced datasets when decomposed into binary problems. This is found to improve calibration results relative to using the unweighted version in (60). Although (60) is quadratic in the number of datapoints , the training time for algorithms based on MMCE is only 10% longer than other linear metrics, like NLL [48].

Widmann et al. [104] discuss more general kernel models and suggest that MMCE can only be used for binary classification problems. However, as identified by Kumar et al. [48], the construction of the metric allows it to be applied when assessing the top-label (highest confidence) performance of multi-class classifiers. Furthermore, it can be used to measure class-wise calibration using a one-vs-all strategy.

The Laplace kernel calibration error (LKCE) can be computed using the general kernel formula in (60), using the Laplace Kernel with fixed width . The Laplace kernel results in a consistent calibration measure, which is not true for the widely-used Gaussian kernel. Naïve computation of (60) takes time. However, an approximation described by Błasiok et al. [10] allows this to be computed in time. Performance curves for LKCE are similar to the smooth calibration error (see Section 5.7) but not quite so smooth. The LKCE curves are beneficially less variable than ICE.

The squared kernel calibration error (SKCE) is computed in a similar manner to MMCE [104]. However, whereas MMCE applies only to binary classifiers or top-label confidence, SKCE quantifies strong multiclass calibration and hence is more generally applicable. Several versions of SKCE have been defined, all based on a pairwise error term. This term is defined as:

|  |  |
| --- | --- |
|  | (62) |

Equation (62) has a similar form to the MMCE summand in (60). However, in (62) the and terms are length vectors to account for full multiclass analysis and is a matrix-valued kernel. The matrix-valued kernel for SKCE is chosen as the product of the scalar Laplace kernel (61) and the identity matrix. The width parameter of the kernel is chosen using the median heuristic.

The biased (B) SKCE-B is defined as:

|  |  |
| --- | --- |
|  | (63) |

This is the multiclass extension of MMCE. However, the metric is biased and takes time to compute. Nevertheless, it is a strictly proper metric [28].

The unbiased quadratic (UQ) SKCE-UQ is defined as:

|  |  |
| --- | --- |
|  | (64) |

This metric is unbiased but still takes time to compute. A hypothesis test exists for the SCKE-UQ metric. However, this is based on forming a bootstrap estimate that takes time to compute for Monte Carlo bootstrap samples. Computation of this statistic may be prohibitive for large datasets.

The unbiased linear (UL) SKCE-UL is defined as:

|  |  |
| --- | --- |
|  | (65) |

This metric is unbiased and only takes time to compute. A hypothesis test exists for the SCKE-UL metric, based on an asymptotic approximation. If is the cumulative distribution of the standard normal distribution, and is the standard deviation of the terms in the summand of (65), then the p-value of the metric under the null hypothesis of perfect calibration is:

|  |  |
| --- | --- |
|  | (66) |

There are two issues with SKCE-UL. The first is that the value of the metric relies on the order in which the datapoints are presented to the algorithm. This may be noted from equation (65), where only adjacent pairs of datapoints contribute to the sum. This means that if the same inputs are shuffled, the computed metric may be different. This is an undesirable property. The second issue is that the metric effectively assumes the datapoints are randomly distributed with respect to their characteristics. However, certain data processing pipelines may sort datapoints by confidence value or true label. If that is the case, higher weightings will be encountered than expected on average, potentially resulting in overly high metric values.

The SKCE-UL and SCKE-UQ metrics theoretically lie in the range [0, 1]. However, for nearly perfectly calibrated models with true SKCE≈0, the metric can be slightly negative for certain arrangements of datapoints. This is part of the normal variance associated with computing an unbiased metric. Nevertheless, this property may harm interpretability.

A comparison is made between standard ECE and the three SKCE metrics using 10,000 synthetic datasets each containing 250 samples with known ground truth and calibrated and un-calibrated models. The ECE exhibits both negative and positive bias, whereas SKCE-B is theoretically guaranteed to be biased upward. Hypothesis testing using ECE and consistency resampling (see Section B.2) is found to be unreliable and this gets worse with more classes. The asymptotic approximations for the two unbiased SKCE metrics are good for moderate numbers of classes. However, for 100 classes, SKCE-UL exhibits some mild multimodality in its distribution of values over the datasets, and for 1000 classes it is strongly bi-modal (see Figure 27 in Appendix J.2.3 of [104]). SKCE-UQ appears to have good properties for all tests performed up to 1000 classes.

In conclusion for SKCE, compared to ECE and SKCE-B, SKCE-UL may be preferred as it is unbiased, quick to compute, hypothesis tests are quick to compute, and it has reasonable performance. However, the dependency of SKCE-UL on the order of datapoints is a major disadvantage and it performs poorly for problems with more than 100 classes. The SKCE-UQ is preferable to SKCE-UL, as it is more stable. However, this is at the expense of potentially high computation time for very large datasets, and the Monte Carlo nature of its hypothesis test, which may be undesirable for assurance applications.

Vashistha et al. [98] define the I-trustworthy framework and the associated kernel local calibration error (KLCE) metric, which considers variations of calibration in different parts of the feature space. The metric is defined in the same way as SCKE-UQ but in (62) is replaced by . The I-trustworthy framework analyses one class at a time so the metric only need be applied to binary classifiers. If the kernel is chosen to be diagonal, computational complexity scales linearly with the feature dimensionality. Hypothesis testing is performed via bootstrapping. The framework and metric enable calibration analysis of sup-populations, e.g. by protected characteristics of a person. KLCE is sensitive to the choice of kernel and how its hyperparameters are optimized. In experiments, a radial basis function kernel is used [98]. The disadvantage of the metric is that it requires access to feature values, which may not always be possible. KLCE is a kernel version of the bin-based proximity-informed ECE (see Section 4.12).

Chatterjee et al. [15] introduce the expected conditional maximum mean discrepancy (ECMMD) as a kernel-based metric that can be estimated efficiently from data for both classification and regression problems. Computing the metric for classification proceeds via a series of steps. First, a nearest neighbour graph is computed from the dataset feature values. Second, a possible label for each datapoint is sampled based on the confidence . A kernel-based function is defined between pairs of datapoints, each consisting of pairs of values. For each datapoint, a summand is defined as the mean of computed with the datapoint’s nearest neighbours. The value of the metric is then mean of the summands. Due to the use of nearest neighbours, the computational complexity of the metric is rather than the usual . ECMMD is the kernel equivalent of ECE-KNN. A hypothesis test based on ECMMD is defined. The ECMMD and SKCE hypothesis tests have comparable statistical performance, but ECMMD is computationally much more efficient than the SKCE test. A disadvantage of ECMMD is the stochastic nature of its computation due to the use of sampling. It also has multiple parameters, including and the choice of kernel, which may lead to ambiguity in its definition.

## Smoothed kernel density estimator

Zhang et al. [109] define a smoothed kernel density estimation (SKDE) based ECE estimator. This also referred to as Mix-n-Match in [78], due to its use with the Mix-n-Match recalibration method. The metric performs kernel smoothing for both the confidence estimates and the true labels. The KDE estimate of the confidences is:

|  |  |
| --- | --- |
|  | (67) |

Computation of the full SKDE does not scale well with the number of classes due to the curse of dimensionality, so it is recommended to be used with the top-label or class-wise decomposition of multi-class classifiers into binary classifiers [109]. In the binary case, SKDE can be computed based on a grid approximation to an integral, with points, as:

|  |  |
| --- | --- |
|  | (68) |

The triweight kernel with a fixed width based on the normal rule of thumb is used, since that kernel has been recommended for problems with limited support [109]. Analysis of the metric’s construction shows that its computational complexity is .

Synthetic data with known ground truth is used to compare the SKDE with standard binned ECE in [109]. SKDE is shown always to outperform binning with the difference especially large for small sample sizes.

## Smooth calibration error

The smooth calibration error (SCE) aims to provide a metric that formally varies smoothly with respect to changes in its inputs [10]. The empirical SCE can be computed as:

|  |  |
| --- | --- |
|  | (69) |

The constraints in the definition of SCE produce an implicit weighting function that is 1-Lipschitz smooth, that is, the magnitude of the local gradient is never larger than unity. The Lipschitz condition smooths out the contribution from each neighbourhood of . Unlike standard ECE, this metric is a consistent calibration measure. It provides smoother performance curves than LKCE and is less variable than ICE [10]. The SCE is unusual among other metrics in that it takes on values in the range . The only other metrics that have this property are the bin-based ESCE, WSMCS, and DICI (see Section 4.9). This property allows these metrics to measure under- or over-confidence in addition to the degree of miscalibration.

## Logit Smoothed ECE

Logit smoothed ECE (LSECE) is designed to avoid the problems of discontinuities in ECE [17]. The metric randomly samples datapoints with replacement. The confidence for each sample is converted to logit space via . A normally distributed zero-mean random variable with standard deviation is added to the logit. A kernel estimate of the calibration curve is formed in confidence logit space using a Gaussian kernel of width .

|  |  |
| --- | --- |
|  | (70) |

The randomised logits are mapped back to linear space by the inverse logit function . The LSECE is then computed as the mean absolute difference between the curve and the identity function.

|  |  |
| --- | --- |
|  | (71) |

The randomization process makes the LSECE metric continuous from a theoretical standpoint and in practice the metric is relatively insensitive to the arbitrary parameter , which is analogous to the bin width in ECE. However, the Monte Carlo nature of the metric may be undesirable for assurance applications.

## Integrated calibration index and error percentile

The integrated calibration index (ICI) is a fitted-curve-based metric, and as such, it has some similarities with kernel-based metrics. It is similar to Cox’s intercept and slope method (see Section 5.12) in the sense that it fits a calibration curve and then analyses the curve [40].

The development of the ICI was motivated by Harrell's Emax index, which is the maximum absolute error (MAE) between a smooth calibration curve and the diagonal line of perfect calibration [4]. The smoothed curve is obtained via the locally estimated scatterplot smoothing (LOESS) algorithm. This is a non-parametric regression algorithm that is also known as the Savitzky-Golay filter when independent variables are a fixed width apart. The algorithm fits a low-degree polynomial to datapoints near to the point of interest. Austin et al. [4] utilise a two-degree polynomial, 75% of the full dataset to contribute to each estimate, and a tri-cubic weighting to down-weight datapoints far from the estimation point.

The ICI is the weighted difference between observed and predicted probabilities, in which observations are weighted by the empirical density function of the predicted probabilities. From a theoretical perspective, ICI is given by (72), where is the smoothed estimate of the label proportion for a particular confidence (i.e. the calibration map) and is the density function of the distribution of predicted probabilities.

|  |  |
| --- | --- |
|  | (72) |

The empirical ICI is:

|  |  |
| --- | --- |
|  | (73) |

A metric related to ICI is the error percentile (EP) metric [4], also known as EX [36]. This is the percentile of the absolute difference between observed and predicted probabilities. Common values of are 90 and 50 to give the E90 and E50 metrics. E50 is the median absolute difference and E90 is the 90th percentile of the difference. Confidence intervals for ICI and EP can be estimated using bootstrap methods [4].

ICI, EP, and Emax have advantages over other ways of measuring calibration: they have a simple interpretation and assign a greater weight to dense data areas, which supresses poor estimates from sparse areas. The three metrics were compared by Austin et al. [4] using simulated data with known ground truth while examining the performance of correctly and incorrectly specified models. ICI tends to demonstrate more consistent behaviour in tests than EP or Emax.

## Estimated Calibration Index

The estimated calibration index (ECI) [97], also known as the expected calibration index in [59], is similar in concept to ICI. The definition is:

|  |  |
| --- | --- |
|  | (74) |

In (74), the calibration map is estimated as a multinomial logistic regression model using a cubic spline. This equation is reminiscent of the Brier score, where is replaced by the actual data labels. The advantage of ECI over BS is that is measures calibration only, rather than calibration and discrimination combined [97]. ICI and ECI are curve versions of the label-binned calibration error (see Section 4.6).

## Fit on the test

Kängsepp et al. [42] define a general fit-on-the-test (FOTT) paradigm, where parameters of a calibration function from a family of functions are fit to the data by minimizing the ECE-FOTT loss in (32) through cross validation. Section 4.5 describes how bin-based schemes are a particular family of functions, leading to the tilted-roof reliability diagram. Curve-based function families have also been assessed using this paradigm, as described below.

The piecewise linear (PL) method for evaluating calibration fits a PL calibration map , where parameters of the function are the number of segments, segment boundaries, and the value of the function at the boundaries. A separate calibration map is estimated for each cross-validation fold, and an ensemble average is used to determine the final . The ECE-PL metric for calibration is then given by (32). The optimum number of segments for a calibration task containing 5000 datapoints was three [42].

The piecewise linear in logit-logit space (PL3) method fits a continuous piecewise linear function to logit functions . The independent variable is and the dependent variable is . When one piece (segment) is used, this is equivalent to temperature scaling, a recalibration (calibration improvement) method [31]. When two pieces are used, this is approximately equivalent to beta scaling [47]. As before, the ECE-PL3 metric for calibration is then given by (32).

Other families tested under the FOTT paradigm include ECE-EM, Platt scaling, beta scaling, isotonic regression, spline fitting, and intra-order preserving functions [42].

FOTT metrics have been assessed both with synthetic data, where the true calibration map is known, and with real data, where the true calibration map can be estimated accurately when there exist magnitudes more data than used in computing the calibration metric under test. The dataset CIFAR-5m contains 5 million synthetic images created so that models trained on CIFAR-10 have similar performance and vice versa. Metrics are assessed based on three objectives: (1) the quality of the reliability diagrams; (2) the quality of the calibration error estimates; and (3) Spearman’s rank correlation between the metric ranks and the (approximately) true calibration error ranks, when assessing recalibration methods. For objective (1), PL is the best metric on average, followed by PL3 and beta scaling. However, the beta scaling rank gets worse as the number of datapoints increases due to the small number of parameters in the model. For objective (2), the 15-bin ECE-EM or beta metrics are the best, depending on the task. PL and PL3 also perform well. For objective (3), isotonic regression is best followed by 15-bin ECE-EM. PL is better than average and outperforms PL3 [42].

The above assessment does not provide a clear ranking of calibration metrics as the ranking is different under different objectives. However, PL generally ranks well, and 15-bin ECE-EM also surprisingly ranks well, especially for the important task of ranking re-calibration methods. Since ECE-EM is now considered a “classical” method and PL is relatively easy to compute, these should be considered for use generally. In passing, it is noted that the extensive experiments by Kängsepp et al. [42] required over 10,000 hours of computer time to complete. Thus, it would be costly to recreate similar bespoke experiments for new projects.

## Cox intercept and slope

The Cox intercept and slope (CIS) method measures calibration by performing a regression of the observed outcome against the log odds of the predicted confidence [40]:

|  |  |
| --- | --- |
|  | (75) |

Perfect calibration has intercept (measuring calibration-in-the-large) and calibration slope parameter . If this indicates over-confidence on average, and if this indicates under-confidence on average. If this indicates under-confidence for high probabilities and over-confidence for low probabilities. The converse is true for . Both parameters take on any value on the real line.

The CIS method has some similarities with the PL3 method, with the fitted function for CIS being equivalent to the fit for PL3 with a single segment. However, ECE-PL3 computes an ECE-style measure of the area between the fitted curve and the line of perfect calibration, whereas CIS analysis looks at the values of the intercept and slope to understand calibration. Although CIS provides more information than ECE-PL3, the need to analyse two variables makes CIS harder to use when assessing multiple recalibration algorithms automatically, like the above/below area between curves metrics.

## Hypothesis-test kernel or curve-based metrics

The statistical beta calibration test (SBCT) fits a beta calibration curve to the calibration dataset, computes the area between this curve and the perfect calibration identity line, and uses the result to compute a p-value in a hypothesis test of whether the fitted curve is statistically significantly different from the identity [47]. The test statistic is defined as:

|  |  |
| --- | --- |
|  | (76) |

The p-value is approximated by the quantile of within the distribution . The statistical beta calibration test process has some similarities to the fit-on-the-test paradigm (see Section 5.11), with being similar to ECE-FOTT. However, ECE-FOTT is computed on a sample basis, which puts more emphasis on the regions of the calibration map with more samples, whereas the integral for is computed on a uniform grid of values. The uniform grid enables the possibility of using the above gamma distribution approximation, which is good when at least 300 calibration samples are used. The disadvantage of the SBCT is that it assumes there is a parameterization of the beta distribution that is a good fit for the data. This may not be the case for complex calibration curves with more than one inflection point.

The parabolic Wald statistic (PWS) is defined by Galbraith et al. [24]. A hypothesis test based on this statistic fits a parabola to the data and determines whether the fitted coefficients are significantly different from those of the linear identity function . If is a consistent estimator of the parameter covariance matrix, the PWS is defined as:

|  |  |
| --- | --- |
|  | (77) |

The PWS metric takes on non-negative numbers, is asymptotically distributed according to the chi-squared distribution, and standard significance tests can be constructed based on this. This statistic assumes the calibration curve may be well approximated by a parabola. This may be the case for some datasets but is not in general true if the calibration curve has a more complex character.

Gweon et al. [35] describe a Pearson chi-squared reliability statistic based on k-nearest-neighbours in the confidence prediction space and a Bayesian approach for estimating the expected power of the reliability test for different sample sizes. The use of nearest neighbour makes this metric reminiscent of kernel methods and ECE-KNN.

# Cumulative metrics

## Empirical cumulative calibration error

Reliability diagrams are usually based on binned estimates or sometimes kernel estimates, but the selection of width parameter for either type of representation can be arbitrary. Cumulative metrics sort datapoints by confidence and examine the difference between the cumulative accuracy and the perfect calibration line. The advantage of such metrics is that there is no need to set or estimate arbitrary parameters, such as bin width or parameters of a curve model. Avoiding those arbitrary choices has advantages if calibration metrics are to be used to measure compliance to regulations. The cumulative difference plot (CDP) is defined as for and for as:

|  |  |
| --- | --- |
|  | (78) |

For a perfectly calibrated classifier, the CDP approximates a horizontal flat line at zero.

Two types of empirical cumulative calibration error (ECCE) are defined by Arrieta-Ibarra et al. [3]. The first is the maximum absolute deviation (MAD) of the CDP from zero. This is known as ECCE-MAD or the Kolmogorov-Smirnoff (KS) statistic.

|  |  |
| --- | --- |
|  | (79) |

The second error type is the range of the CDP. This is known as ECCE-R or the Kuiper statistic.

|  |  |
| --- | --- |
|  | (80) |

The variance of is:

|  |  |
| --- | --- |
|  | (81) |

The statistic is equivalent in distribution to the maximum of the absolute value of standard Brownian motion over the unit interval [0, 1]. The statistic is equivalent in distribution to the range of such Brownian motion. This equivalence to Brownian motion can be used to develop a hypothesis test for each statistic and compute p-values (significance levels) [3].

In [3], the ECCE metrics are assessed and compared with standard ECE using synthetic data, where it is possible to compute ground truth statistics. It is shown that the ECCE metrics can distinguish calibrated and miscalibrated classifiers as the number of samples grows large, but this is not the case for ECE. Analysis of classifiers applied to real datasets with large numbers of datapoints produces very low-p-values for ECCE showing that such classifiers are statistically significantly miscalibrated. However, the effect size is small, as seen by the un-normalised ECCE statistics.

Gupta et al. [32] define the Kolmogorov-Smirnoff calibration metric and describe it as a “binning-free calibration measure”. The final form of this metric is identical to ECCE-MAD, but it is derived in a slightly different way. A generalization of the metric also allows the assessment of whether the th most likely class is correct or the top classes are correct [32].

The KS error is zero for perfect calibration and unity for completely un-calibrated systems. The metric can be construed as a percentage difference between two distributions, which aids interpretation of the metric. KS error can be shown to be a special case of kernel-based measures. The relationship between KS error and the cumulative distribution function is the same as the relationship between MCE and the binned probability density function – they are both based on the maximum difference [84].

KS error is compared to RBS, ECE, and CWCE by Gruber et al. [28]. Only KS and RBS are consistent in value with respect to data size. KS error is used, along with ECE, KDE-ECE, MCE, and Brier Score in [32] to assess various recalibration methods on an ImageNet recognition challenge. All metrics give similar rankings for the best recalibration methods. However, since the ground truth for this challenge is not known, it is not possible to determine which metric is best.

## Cumulative multi-calibration metric

Guy et al. [34] discuss the concept of multi-calibration, which is to make sure that all specified sub-populations of the full dataset are calibrated. The , possibly overlapping, sub-populations could be bins based on confidence, groups of datapoints that have similar feature values (if it is desired that all parts of the feature space are calibrated), or a combination of both. The authors propose a metric to measure the multi-calibration error, which is computed as follows. Let be the weight of datapoint and let be the first datapoints of sub-population , sorted by ascending confidence value. The weighted cumulative difference plot (WCDP) of sub-population is then:

|  |  |
| --- | --- |
|  | (82) |

Equation (82) is a weighted version of (78), but only for a single sub-population. A Kuiper statistic for the sub-population is then computed as:

|  |  |
| --- | --- |
|  | (83) |

The multi-calibration metric (MCM) is defined as the worst-case Kuiper statistic weighted by its signal-to-noise ratio under the hypothesis of perfect calibration, assessed over all specified sub-populations:

|  |  |
| --- | --- |
|  | (84) |

The general MCM metric can be used with any specification of sub-population. One specific use case could be to analyse the treatment of people according to protected characteristics to ensure that any automated decision systems are well calibrated for all protected characteristics. However, Guy et al. [34] describe an optional automated method for generating sub-populations based on using a binary tree to partition the data based on feature values. The method retains all internal nodes of the tree, resulting in overlapping sub-populations.

The range of MCM is . The metric has similar advantages and disadvantages to ECCE-MAD and ECCE-R, with the additional advantage of sub-population analysis. MCM is the cumulative equivalent of the partitioned calibration error discussed in Section 4.12.

## Brownian bridge test

Sadatsafavi et al. [86] specify a hypothesis test that makes use of the equivalence of cumulative difference plots with Brownian motion. The test is generally more powerful than that based on ECCE-MAD. The test is based on a Brownian bridge, as follows. Let . Define time and location as:

|  |  |
| --- | --- |
|  | (85) |
|  | (86) |

The maximum absolute value of the bridged random walk is:

|  |  |
| --- | --- |
|  | (87) |

The Brownian bridge test (BBT) p-value is then:

|  |  |
| --- | --- |
|  | (88) |

In (88) is the cumulative distribution function (CDF) of the standard normal distribution, is the CDF of the chi-squared distribution with four degrees of freedom, and is the CDF of the Kolmogorov distribution, which is widely implemented in statistical analysis software [86]. Since BBT is a p-value, its range is .

BBT was compared to the p-value produced by the ECCE-MAD hypothesis test of [3], which is referred to as the Brownian motion (BM) test in [86]. In small test datasets with an effective sample size of less than 30, the p-values of both methods are slightly biased upwards, so the tests are conservative (i.e. reject fewer tests than expected). BBT was more powerful than BM in all cases other than with pure mean shifts in calibration, and in those cases the power was only slightly lower [86]. Based on these results, BBT should be preferred to BM.

# Conclusion

## Summary

This paper analyses a wide range of metrics used to assess the calibration of probabilities produced by machine learning models and has organised these metrics according to families identified as: point, bin, kernel, curve, or cumulative metrics. For each family, Table 1 shows the number of metrics identified, how many have different nominal ranges of value, how many are proper, how many have a built-in associated hypothesis test, and the number that distinguish under-confidence from over-confidence. A list of general pros and cons for each family are shown in Table 2.

| **Type** | **Number** | **Range** | **Range** | **Range** | **Range** | **Proper** | **Hypothesis Test** | **Assesses under / over confidence** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Point | 20 | 12 | 6 | 0 | 2 | 10 | 1 | 4 |
| Bin | 44 | 40 | 1 | 3 | 0 | 0 | 7 | 4 |
| Kernel | 17 | 16 | 0 | 1 | 0 | 0 | 2 | 0 |
| Curve | 9 | 7 | 1 | 0 | 1 | 0 | 1 | 1 |
| Cumulative | 4 | 4 | 0 | 0 | 0 | 0 | 3 | 0 |

**Table 1** Summary of calibration metric families.

| **Type** | **Pros** | **Cons** |
| --- | --- | --- |
| Point | * Includes widely used metrics * Most are easy to understand * Most are easy to compute | * Mixes concepts of calibration and accuracy * Some are very imprecise * Some have adjustable parameters |
| Bin | * Includes ECE (the most widely used metric) and its variants * Many are easy to understand * Most are easy to compute | * Most metrics are biased * Many have adjustable parameters |
| Kernel | * Smoother than bin metrics * Generally unbiased or lower bias than bin metrics | * More complicated definition than other metrics * Can be slow to run for large datasets * Most have adjustable parameters |
| Curve | * Relatively simple definitions * Most have no adjustable parameters | * Some are slow to compute * Make assumptions about form of curve |
| Cumulative | * Assumption free * No adjustable parameters * Better than ECE | * Can (correctly) produce highly significant p-values for large datasets even when effect is small |

**Table 2** General pros and cons for each calibration metric family.

Table 4 in Appendix D summarizes all the individual metrics discussed in this paper. For each metric, the table lists: alternative names; family; value range; whether the metric is proper; whether a hypothesis test is available; whether they distinguish under-confidence from over-confidence; pros and cons; and an external reference where the metric is defined or described in more detail.

## Analysis and recommendations

With so many metrics to choose from, it is necessary to recognise good metric properties to aid their selection. Hagopian et al. [36] define various characteristics that good calibration metrics should have, including reproducibility, representativeness, interpretability, parsimony, and computational efficiency. Three further potentially useful metric properties are catalogued in Table 4. The first two characteristics of being proper or having a hypothesis test are advantages of a metric but are not necessary for its use. The third property of being able to distinguish under-confidence from over-confidence is a major advantage, but lack of this characteristic does not exclude the use of a metric, depending on the use case. Further properties that are only applicable to certain metrics appear in the pros and cons column of Table 4.

No single metric is better than all others, as each one has advantages under certain conditions. Summaries of various numerical comparisons are included in individual metric descriptions above. However, each comparison only covers a small subset of all metrics described so it can be difficult to assess the complete portfolio of possible metrics. Nevertheless, recommendations can be made for a series of use cases.

The first use case is to have a metric that is generally applicable for classifier probability calibration and is used to assess the class-wise calibration performance of models or calibration improvement of recalibration algorithms. To avoid conflating the assessment of accuracy and calibration a pure calibration metric should be selected. Such a metric should at least have Hagopian’s properties. One possible candidate is the standard equal-width ECE. This meets most of the requirements and is widely used and understood. However, complete representativeness is not achieved since the metric is a highly discontinuous function of confidence values – a small change in value could move a datapoint from one bin to another, affecting the contribution from both bins and the resulting metric. The metric has several other drawbacks as outlined in Section 4.2. A better metric is the equal-mass DRMSCE-BCS (see Section 4.12). DRMSCE-BCS is easy to compute, easy to interpret, debiased, has a hypothesis test based on bootstrapping, has no adjustable parameters, outperforms ECE and other metrics in experiments, and has good theoretical properties. The only major disadvantage of DRMSCE-BCS is that its standard definition is for binary rather than multiclass classification. Measurement of class-wise calibration can be achieved through a simple one-vs-rest strategy (see Section 2.3). However, full multi-class measurement would require definition of a multiclass debias term and a modification to the automated binning strategy.

Binned metrics can be “gamed” by models that set all confidence values to a single value or vector equal to the class proportions, producing good outputs according to the metric but not necessarily useful. Several authors (e.g. [18], [87]) therefore recommend that a proper score should be reported alongside any binned metric. The top two choices would be the Brier score or negative log likelihood due to their widespread used and understood properties. As discussed in Section 3.4, log metrics are highly sensitive to the presence of confidence values near zero, violating the representativeness criterion. Therefore, we recommend that the Brier score should be reported along with DRMSCE-BCS.

The second use case is full multiclass calibration. DRMSCE-BCS cannot be used for this purpose – no bin-based metrics have a reasonable extension to full multiclass calibration with . Out of the metrics that do not conflate accuracy and confidence, only the kernel-based DKDE (also known as ECE-KDE) and SKCE metrics measure full multiclass calibration. Within the SKCE variants, SKCE-UQ is preferred (see Section 5.5). Both DKDE and SKCE-UQ are unbiased and can be computed in time. Both metrics require setting the kernel bandwidth parameter, which is done via the median heuristic for SKCE-UQ and through cross-validation for DKDE. DKDE, as used for evaluation in the experiments in [78], is based on the 1-norm whereas SKCE-UQ is based on the 2-norm. A bootstrap method for hypothesis testing is defined an analysed in [104] for SKCE-UQ. A similar bootstrap method could be defined for DKDE but has not been tested in [78]. There is little to choose between DKDE and SKCE-UQ and there appear to be no direct comparisons of the two techniques in the literature. SKCE-UQ appears to have a slight advantage as it uses the 2-norm (for consistency with DRMSCE-BCS) and results of hypothesis tests are available.

The third use case is determining whether miscalibrated classifiers are under or overconfident. Several metrics meet this high-level requirement. Of the point-based metrics, EO is the only one not to have sensitivity to particular confidence values like ECD, NSES, or the Spiegelhalter z statistic. However, as a global metric EO lacks refinement. From the bin-based metrics, ESCE is preferred as it has a simple definition and there is no clear benefit to the more complex DICI and WSMCS. The variant of ESCE by Verhaeghe et al. [100] that averages over different bin widths should be used to reduce bias. Consideration should be given to an equal-mass extension to further reduce bias. If it is required to determine under and overconfidence separately then the bin-based ABC should be used. The kernel-based smooth calibration error (SCE) has a slightly complex definition and thus lacks interpretability. The curve-based CIS has a simple definition, but requires the analysis of two parameters, which complicates analysis. Based on the above analysis, we recommend ESCE as a single metric to measure under and overconfidence.

The fourth uses case is to measure the stronger notion of local calibration, where feature values are used in the assessment. Several metrics meet this high-level requirement: the bin-based PIECE, PCE, MVCE, FECE, and VECE; the kernel-based KLCE and ECMMD; and the cumulative MCM. The metrics ECMMD and MVCE contain a random sampling element, harming their reproducibility. FECE and VECE only use one confidence bin, resulting in a global confidence metric for each feature-partition of the data, and an ensuing lack of refinement in confidence analysis. KLCE is the only metric that allows assessment of full multi-class calibration, and it has a defined hypothesis test based on boot strapping. KLCE is unbiased and is smoother than the remaining bin metrics of PIECE and PCE. The major disadvantage of KLCE is its sensitivity to the choice of kernel and how its hyperparameters are optimized. PIECE bins data in feature space by density. While this allows miscalibration detection in the presence of cancellation between different parts of the feature space, the location of miscalibrated areas cannot easily be identified. Based on the analysis above, we recommend a specialization of the general PCE metric that uses as the loss term (for consistency with DRMSCE-BCS and SKCE-UQ) and that bins both confidence and feature values. Since the number of bins grows exponentially with the number of feature dimensions, resulting in sparsely populated bins, single feature dimensions of interest should be analysed one at a time.

The fifth and final use case is for developers and researchers to perform deeper analysis beyond the use of single numerical metrics for automated comparison and assurance. A model can be under or overconfident for different ranges of confidence. These variations in calibration can cancel each other out to different extents depending on the metric, resulting in ostensibly good metric values that mask a more subtle issue [96]. Therefore, for refined analyses it is recommended that calibration curves with associated confidence intervals should be plotted on the reliability plot, like in Fig. 3, and examined.

## Further work

This paper has outlined the definition and properties of many metrics. Existing comparisons between metrics typically only assess a small number of metrics at a time, and these comparisons are often carried out by metric designers who perform analysis from an angle that may be beneficial to their metric. A large-scale independent study should compare the more promising metrics in this survey using data and models of known levels of miscalibration.

Certain specific lines of research would be useful. DRMSCE-BCS is a recommended metric for top-label and class-wise calibration but is not applicable to full multiclass calibration. A debiasing term and automated binning strategy should be developed for the multiclass setting, including an analysis of scalability as the number of classes grows. Sun et al. [89] show that the distribution of -ECE may be approximated by a normal distribution and thus can be used to compute confidence intervals and perform hypothesis testing. It would be useful to determine whether this analytic process can be applied to DRMSCE-BCS so that bootstrapping may be avoided in hypothesis testing. DKDE and SKCE-UQ are promising metrics for multiclass calibration. A detailed comparison of the two metrics should be carried out, including their ability to perform hypothesis testing. Metrics for measuring local calibration are less analysed than more general calibration metrics, so further work should be conducted in this area to benchmark performance under controlled conditions.

This survey focuses on the probability calibration of classifiers that produce confidence scores for a fixed set of unordered classes. Similar surveys should be carried out to catalogue probability calibration metrics for other problems, including object detection, conformal classification with variable numbers of output classes, ordinal classification for ranked classes, and regression.

Finally, several open problems in probability calibration assessment are discussed in [87]. These include modelling the epistemic uncertainty in models and assessing out-of-distribution (OOD) inputs. These topics should be investigated further from a calibration point of view.

# Acknowledgments

The author thanks Emma Tattershall, Sarah Pengelly, and William Wood for productive discussions about probability calibration issues and for reviews of early versions of this survey.

# Statements and Declarations

**Ethical considerations:** Not applicable.

**Consent to participate**: Not applicable.

**Consent for publication**: Not applicable.

**Declaration of conflicting interest**: The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding statement**: Funding for initial literature survey and write-up activities was provided by the Defence Science Technology Laboratory under ASTRID Task 267 of contract DSTL/AGR/01142/01. Subsequent survey activities and the production of this paper were funded by the QinetiQ Fellow scheme.

**Data Availability**: Since this paper is a review article, no datasets are analysed or generated other than for illustrative purposes.

# References

1. A.N. Angelopoulos, R.F. Barber, S. Bates. Theoretical foundations of conformal prediction. arXiv:2411.11824v3. June 2025.
2. S. Ao, S. Rueger, A. Siddharthan. Two sides of miscalibration: identifying over and under-confidence prediction for network calibration. Conference on Uncertainty in Artificial Intelligence (UAI), PMLR 216:77-87. July 2023.
3. I. Arrieta-Ibarra, P. Gujral, J. Tannen, M. Tygert, C. Xu. Metrics of calibration for probabilistic predictions. Journal of Machine Learning Research (JMLR) 23(1):15886-15940. 2022.
4. P.C. Austin, E.W. Steyerberg. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine 38(21):4051-4065. September 2019.
5. Y. Bai, S. Mei, H. Wang, C. Xiong. Don’t just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. International Conference on Machine Learning (ICML). July 2021.
6. C. Barbera, L. Perini, G. De Toni, A. Passerini, A. Pugnana. Multiclass Local Calibration with the Jensen-Shannon Distance. arXiv:2510.26566. October 2025.
7. J.T. Barron. A general and adaptive robust loss function. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2019.
8. A. Bella, C. Ferri, J. Hernández-Orallo, M.J. Ramírez-Quintana. On the effect of calibration in classifier combination. Applied Intelligence 38:566-585. 2013.
9. G. Bihani, J.T. Rayz. Calibration error estimation using fuzzy binning. North American Fuzzy Information Processing Society Annual Conference. May 2023.
10. J. Błasiok, P. Gopalan, L. Hu, P. Nakkiran. A unifying theory of distance from calibration. Annual ACM Symposium on Theory of Computing. June 2023.
11. J. Błasiok, P. Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. International Conference on Learning Representations (ICLR). May 2024.
12. O. Bohdal, Y. Yang, T. Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. Transactions on Machine Learning Research. August 2023.
13. J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography 135(643):1512-1519. July 2009.
14. F. Cabitza, A. Campagner, L. Famiglini. Global interpretable calibration index, a new metric to estimate machine learning models’ calibration. International Cross-Domain Conference for Machine Learning and Knowledge Extraction. August 2022.
15. A. Chatterjee, Z. Niu, B.B. Bhattacharya. A kernel-based conditional two-sample test using nearest neighbors (with applications to calibration, regression curves, and simulation-based inference). arXiv:2407.16550. July 2024.
16. Z. Chen, C. Heckman, S. Julier, N. Ahmed. Weak in the NEES?: Auto-tuning Kalman filters with Bayesian optimization. International Conference on Information Fusion (FUSION) July 2018.
17. M. Chidambaram, H. Lee, C. McSwiggen, S. Rezchikov. How flawed is ECE? An analysis via logit smoothing. International Conference on Machine Learning (ICML), PMLR 235:8417-8435. July 2024.
18. M. Chidambaram, R. Ge. Reassessing how to compare and improve the calibration of machine learning models. International Conference on Learning Representations (ICLR). April 2025.
19. P. Conde, R.L. Lopes, C. Premebida. A theoretical and practical framework for evaluating uncertainty calibration in object detection. arXiv:2309.00464. September 2023.
20. H. Dawkins, I. Nejadgholi. Region-dependent temperature scaling for certainty calibration and application to class-imbalanced token classification. Annual Meeting of the Association for Computational Linguistics 2:538-544. May 2022.
21. H. Du. Beyond strictly proper scoring rules: The importance of being local. Weather and Forecasting 36(2):457-468. April 2021.
22. P. Flach. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. AAAI Conference on Artificial Intelligence 33(01):9808-9814. July 2019.
23. P. Flach, H. Song. How to assess and improve classifier confidence and uncertainty. A tutorial at ECML-PKDD 2020. Advanced Topics and Conclusion. September 2020. Accessed from <https://classifier-calibration.github.io>.
24. J.W. Galbraith, S. van Norden. Kernel-based calibration diagnostics for recession and inflation probability forecasts. International Journal of Forecasting 27(4):1041-1057. October 2011.
25. M. Gebel. Multivariate calibration of classifier scores into the probability space. PhD thesis, University of Dortmund. April 2009.
26. P. Glaser, D. Widmann, F. Lindsten, A. Gretton. Fast and scalable score-based kernel calibration tests. Conf. on Uncertainty in Artificial Intelligence, PMLR 216:691-700. July 2023.
27. T. Gneiting, A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477):359-378. March 2007.
28. S. Gruber, F. Buettner. Better uncertainty calibration via proper scores for classification and beyond. Advances in Neural Information Processing Systems 35:8618-8632. December 2022.
29. S.G. Gruber, F. Bach. Optimizing estimators of squared calibration errors in classification. Transactions on Machine Learning Research (TMLR). March 2025.
30. T. Guilbert, O. Caelen, A. Chirita, M. Saerens. Calibration methods in imbalanced binary classification. Annals of Mathematics and Artificial Intelligence 92:1319-1352. July 2024.
31. C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger. On calibration of modern neural networks. International Conference on Machine Learning (ICML). August 2017.
32. K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, R. Hartley. Calibration of neural networks using splines. International Conference on Learning Representations (ICLR). May 2021.
33. M. Gupta, S. Basu, C. Arora. How reliable are the metrics used for assessing reliability in medical imaging? International Conference on Medical Image Computing and Computer-Assisted Intervention. October 2023.
34. I. Guy, D. Haimovich, F. Linder, N. Okati, L. Perini, N. Tax, M. Tygert. Measuring multi-calibration. arXiv:2506.11251. June 2025.
35. H. Gweon. A power-controlled reliability assessment for multi-class probabilistic classifiers. Advances in Data Analysis and Classification 17(4):927-949. December 2023.
36. K. Hagopian, N.L. Plackowski, A. Todd, J.A. Richards. Confidence Calibration Metrics. Sandia Report SAND2023-14752, December 2023.
37. R. Hebbalaguppe, J. Prakash, N. Madan, C. Arora. A stitch in time saves nine: a train-time regularizing loss for improved neural network calibration. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2022.
38. D. Hendrycks, M. Mazeika, T. Dietterich. Deep anomaly detection with outlier exposure. International Conference on Learning Representations (ICLR). May 2019.
39. J. Hernández-Orallo, P.A. Flach, C.F. Ramirez. Brier curves: a new cost-based visualisation of classifier performance. International Conference on Machine Learning ICML. June 2011.
40. Y. Huang, W. Li, F. Macheret, R.A. Gabriel, L. Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association 27(4):621-633. April 2020.
41. S. Huang, Y. Wang, L. Mou, H. Zhang, H. Zhu, C. Yu, B. Zheng. MBCT: tree-based feature-aware binning for individual uncertainty calibration. ACM Web Conference. April 2022.
42. M. Kängsepp, K. Valk, M. Kull. On the usefulness of the fit-on-the-test view on evaluating calibration of classifiers. Machine Learning 114(4):105. April 2025.
43. A. Karandikar, N. Cain, D. Tran, B. Lakshminarayanan, J. Shlens, M.C. Mozer, B. Roelofs. Soft calibration objectives for neural networks. Advances in Neural Information Processing Systems 34:29768-29779. December 2021.
44. M. Kelly, P. Smyth. Variable-based calibration for machine learning classifiers. AAAI Conference on Artificial Intelligence 37(7):8211-8219. June 2023.
45. V. Kuleshov, S. Ermon. Reliable confidence estimation via online learning. arXiv:1607.03594. July 2016.
46. M. Kull, P.A. Flach. Reliability maps: a tool to enhance probability estimates and improve classification accuracy. European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Nancy, France. September 2014.
47. M. Kull, T.M. Silva Filho, P. Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics 11(2):5052-5080. June 2017.
48. A. Kumar, S. Sarawagi, U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. International Conference on Machine Learning (ICML). July 2018.
49. A. Kumar, P.S. Liang, T. Ma. Verified uncertainty calibration. Advances in Neural Information Processing Systems 32. December 2019.
50. F. Küppers, J. Kronenberger, A. Shantia, A. Haselhoff. Multivariate confidence calibration for object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops. June 2020.
51. S. Kuzucu, K. Oksuz, J. Sadeghi, P.K. Dokania. On calibration of object detectors: Pitfalls, evaluation and baselines. European Conference on Computer Vision (ECCV). January 2025.
52. R.O. Lane, W.J. Holmes, T. Lamont-Smith. Target recognition in SAR images with low-SWAP processing hardware. International Conference on Synthetic Aperture in Sonar and Radar. September 2023.
53. D. Lee, X. Huang, H. Hassani, E. Dobriban. T-cal: An optimal test for the calibration of predictive models. Journal of Machine Learning Research (JMLR) 24(335):1-72, 2023.
54. J. de Leeuw, K. Hornik, P. Mair. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. Journal of Statistical Software 32(5):1-24. August 2009.
55. W. Liang, C. Dong, L. Zheng, Z. Li, W. Zhang, W. Chen. Calibrating deep neural network using Euclidean distance. arXiv:2410.18321. October 2024.
56. T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. Focal loss for dense object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 2017.
57. J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, B. Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. Advances in Neural Information Processing Systems 33:7498-7512. December 2020.
58. R. Luo, A. Bhatnagar, Y. Bai, S. Zhao, H. Wang, C. Xiong, S. Savarese, S. Ermon, E. Schmerling, M. Pavone. Local calibration: Metrics and recalibration. Conference on Uncertainty in Artificial Intelligence, PMLR 180:1286–1295. August 2022.
59. L. Maier-Hein, A. Reinke, P. Godau, et al. Metrics reloaded: recommendations for image analysis validation. Supplementary Note 2.6. Nature Methods 21:195-212. February 2024.
60. C. Marx, S. Zalouk, S. Ermon, S. Calibration by distribution matching: Trainable kernel calibration metrics. Advances in Neural Information Processing Systems 36:25910-25928. December 2023.
61. T. Matsubara, N. Tax, R. Mudd, I. Guy. TCE: A test-based approach to measuring calibration error. Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1390-1400. August 2023.
62. A.K. Menon, X.J. Jiang, S. Vembu, C. Elkan, L. Ohno-Machado. Predicting accurate probabilities with a ranking loss. International Conference on Machine Learning (ICML) 2012:703-710. June 2012.
63. M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, M. Lucic. Revisiting the calibration of modern neural networks. Advances in neural information processing systems 34:15682-15694. December 2021.
64. R.E. Moore, C. Rosato, S. Maskell. Refining epidemiological forecasts with simple scoring rules. Philosophical Transactions of the Royal Society A 380(2233):20210305. October 2022.
65. J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania. Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems 33:15288-15299. December 2020.
66. M.A. Munir, M.H. Khan, S. Sarfraz, M. Ali. Towards improving calibration in object detection under domain shift. Advances in Neural Information Processing Systems 35. December 2022.
67. A.H. Murphy. A new vector partition of the probability score. J. Applied Meteorology 12(4):595-600. June 1973.
68. L. Neumann, A. Zisserman, A. Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems (MLITS). December 2018.
69. J. Nixon, M.W. Dusenberry, L. Zhang, G. Jerfel, D. Tran. Measuring calibration in deep learning. arXiv:1904.01685v1. April 2019.
70. S. Obadinma, H. Guo, X. Zhu. Class-wise calibration: A case study on COVID-19 hate speech. Canadian Conference on Artificial Intelligence. June 2021.
71. K. Oksuz, T. Joy, P.K. Dokania. Towards building self-aware object detectors via reliable uncertainty quantification and calibration. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2023.
72. R. Padilla, S.L. Netto, E.A.B. da Silva. A survey on performance metrics for object-detection algorithms. International Conference on Systems, Signals and Image Processing (IWSSIP). July 2020.
73. F. Pan. X. Ao, P. Tang, M. Lu, D. Liu, L. Xiao, Q. He. Field-aware calibration: a simple and empirically strong method for reliable probabilistic predictions. ACM Web Conference. April 2020.
74. B. Pathiraja, M. Gunawardhana, M.H. Khan. Multiclass confidence and localization calibration for object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2023.
75. S. Peng, P. Alamzadeh, M. Ester. Better calibration error estimation for reliable uncertainty quantification. International Conference on Machine Learning (ICML). July 2023.
76. A. Perez-Lebel, M. Le Morvan, G. Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. International Conference on Learning Representations (ICLR). May 2023.
77. E. Petersen, M. Ganz, S. Holm, A. Feragen. On (assessing) the fairness of risk score models. ACM Conference on Fairness, Accountability, and Transparency. June 2023.
78. T. Popordanoska, R. Sayer, M. Blaschko. A consistent and differentiable lp canonical calibration error estimator. Advances in Neural Information Processing Systems 35:7933-7946. November 2022.
79. T. Popordanoska, A. Tiulpin, M.B. Blaschko. Beyond classification: Definition and density-based estimation of calibration in object detection. IEEE/CVF Winter Conference on Applications of Computer Vision. January 2024.
80. T. Popordanoska. Advancing Calibration in Deep Learning: Theory, Methods, and Applications. Doctor of Engineering Science (PhD) Thesis, KU Leuven, Belgium. January 2025.
81. R.D. Riley, J. Ensor, K.I. Snell, T.P. Debray, D.G. Altman, K.G. Moons, G.S. Collins. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 353:i3140. June 2016.
82. B. Ristic, B. La Scala, M. Morelande, N. Gordon. Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. International Conference on Information Fusion. June 2008.
83. P. Röchner, H.O. Marques, R.J. Campello, A. Zimek. Evaluating outlier probabilities: assessing sharpness, refinement, and calibration using stratified and weighted measures. Data Mining and Knowledge Discovery 38:3719-3757. July 2024.
84. R. Roelofs, N. Cain, J. Shlens, M.C. Mozer. Mitigating bias in calibration error estimation. International Conference on Artificial Intelligence and Statistics (AISTATS), Valencia, Spain. May 2022.
85. R. Rossellini, J.A. Soloff, R.F. Barber, Z. Ren, R. Willett. Can a calibration metric be both testable and actionable? arXiv:2502.19851. February 2025.
86. M. Sadatsafavi, J. Petkau. Non‐parametric inference on calibration of predicted risks. Statistics in Medicine 43(18):3524-3538. August 2024.
87. T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, P. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. Machine Learning 112(9):3211-3260. September 2023.
88. D.J. Sumler, L. Devlin, S. Maskell, R.O. Lane. An entropic metric for measuring calibration of machine learning models. TRUST-AI: The European Workshop on Trustworthy AI, Bologna, Italy. October 2025.
89. Y. Sun, P. Chaudhari, I.J. Barnett, E. Dobriban. A confidence interval for the l2 expected calibration error. arXiv:2408.08998. August 2024.
90. L. Tao, M. Dong, C. Xu. Dual focal loss for calibration. International Conference on Machine Learning (ICML). July 2023.
91. L. Tao, Y. Zhu, H. Guo, M. Dong, C. Xu. A benchmark study on calibration. International Conference on Learning Representations (ICLR). May 2024.
92. L. Tao, H. Guo, M. Dong, C. Xu. Consistency Calibration: Improving Uncertainty Calibration via Consistency among Perturbed Neighbors. arXiv:2410.12295. October 2024.
93. J. Tödter, B. Ahrens. Generalization of the ignorance score: Continuous ranked version and its decomposition. Monthly Weather Review 140(6):2005-2017. June 2012.
94. A. Torabian, R. Urner. Investigating Calibrated Classification Scores Through the Lens of Interpretability. World Conference on Explainable Artificial Intelligence. July 2024.
95. J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, T. Schön. Evaluating model calibration in classification. International Conference on Artificial Intelligence and Statistics. April 2019.
96. B. Van Calster, G.S. Collins, A.J. Vickers, L. Wynants, K.F. Kerr, L. Barreñada, G. Varoquaux, K. Singh, K.G. Moons, T. Hernandez-Boussard, D. Timmerman. Performance evaluation of predictive AI models to support medical decisions: Overview and guidance. arXiv:2412.10288. December 2024.
97. K. Van Hoorde, S. Van Huffel, D. Timmerman, T. Bourne, B. Van Calster. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. Journal of Biomedical Informatics 54:283-293. April 2015.
98. R. Vashistha, A. Farahi. I-trustworthy Models. A framework for trustworthiness evaluation of probabilistic classifiers. International Conference on Artificial Intelligence and Statistics, PMLR 258:4726-4734. May 2025.
99. R. Vasilev, A. D'yakonov. Calibration of Neural Networks. arXiv:2303.10761. March 2023.
100. J. Verhaeghe, T. De Corte, C.M. Sauer, T. Hendriks, O.W.M. Thijssens, F. Ongenae, P. Elbers, J. De Waele, S. Van Hoecke. Generalizable calibrated machine learning models for real-time atrial fibrillation risk prediction in ICU patients. International Journal of Medical Informatics 175:105086. July 2023.
101. B.C. Wallace, I.J. Dahabreh. Improving class probability estimates for imbalanced data. Knowledge and Information Systems 41(1):33-52. October 2014.
102. Z. Wang, Y. Luo, Z. Chen, S. Wang, Z. Huang. Cal-SFDA: Source-free domain-adaptive semantic segmentation with differentiable expected calibration error. ACM International Conference on Multimedia. October 2023.
103. C. Wang, J. Golebiowski. Towards unbiased calibration using meta-regularization. Transactions on Machine Learning Research (TMLR). September 2024.
104. D. Widmann, F. Lindsten, D. Zachariah. Calibration tests in multi-class classification: A unifying framework. Advances in Neural Information Processing Systems 32. December 2019.
105. D.S. Wilks. Regularized Dawid-Sebastiani score for multivariate ensemble forecasts. Quarterly J. Royal Meteorological Soc. 146:2421-2431, May 2020.
106. L. Wu, B. Lei, D. Xu, D. Zhou. Towards reliable rare category analysis on graphs via individual calibration. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2629-2638. August 2023.
107. M. Xiong, A. Deng, P.W.W. Koh, J. Wu, S. Li, J. Xu, B. Hooi. Proximity-informed calibration for deep neural networks. Advances in Neural Information Processing Systems 36:68511-68538. December 2023.
108. J.Q. Yang, D.C. Zhan, L. Gan. Beyond probability partitions: Calibrating neural networks with semantic aware grouping. Advances in Neural Information Processing Systems 36. February 2024.
109. J. Zhang, B. Kailkhura, T. Y.-J. Han. Mix-n-match: ensemble and compositional methods for uncertainty calibration in deep learning. International Conference on Machine Learning (ICML). July 2020.
110. Remaining classifier probability metrics

Introduction to remaining metrics

This appendix contains metrics that either have not seen widespread use for probability calibration estimation despite existing for a while, have a clearly better alternative, have not been compared to other metrics, or have some aspect that limits their general applicability. The metrics are included here and in Table 1 and Table 4 as a reference for researchers who encounter these metrics elsewhere in the literature.

Mean absolute error

The mean absolute error (MAE) is a point-based metric defined as:

|  |  |
| --- | --- |
|  | (89) |

The MAE is always greater than the MSE (Brier Score), except for unrealistic edge cases where the classifier is always correct and 100% confident, incorrect and 0% confident, or correct/incorrect and 0%/100% confident respectively. In these edge cases, MAE is the same as MSE. The metric takes on values between zero and unity. Reference [22] categorises the MAE metric as “bad” and recommends it is not worth reporting over MSE, as MAE measures the expected loss of a classifier that randomly chooses a class according to its confidence, without taking into account the cost of different wrong decisions, whereas MSE can be interpreted as a cost-conscious metric.

Spiegelhalter z statistic

The Spiegelhalter z statistic is designed to perform statistical hypothesis tests. It defined by [40] as:

|  |  |
| --- | --- |
|  | (90) |

This statistic, under the null hypothesis of perfect calibration, is approximately distributed according to the standard normal distribution and can take on any value on the real line. A disadvantage of the z statistic is that it is undefined when all [36]. If all quantities are close to one of those values without being exactly equal to them, the statistic may take on very high values, harming its stability and interpretability. Since classifiers often produce confidences near these values this could be a practical limitation of the metric. In tests where an already calibrated classifier was de-calibrated, the z statistic produced p-values that were less significant than those produced in the Hosmer-Lemeshow (HL) statistic [40] (see Section 4.14), which suggests the z statistic is less powerful.

Success rate

The success rate (SR), also known as zero-one score or misclassification loss, rewards a probabilistic prediction if the most likely class materializes [27]. In case of ties, the reward is reduced proportionally to the number of most likely classes. If denotes the number of elements (cardinality) of a set, then the SR is:

|  |  |
| --- | --- |
|  | (91) |

The success rate takes on values between zero and unity. The metric is proper but not strictly proper. An asymmetric version of the metric can be defined that assigns different costs to the misclassification of different classes.

Hinge loss

The hinge loss (HL) for a binary classifier where the score is not necessarily a probability and can take on any real number is defined in [62] as:

|  |  |
| --- | --- |
|  | (92) |

The hinge loss is normally used in conjunction with support vector machines (SVMs), where is a score with target values of for positive and negative classes. However, when used for general probabilistic classifiers whose confidence score is in the range , the max function is unnecessary. Therefore, the mean hinge loss is simply:

|  |  |
| --- | --- |
|  | (93) |

This loss function is the same as the loss and therefore takes on values between zero and unity. The hinge loss is not a proper loss [62].

Point metrics for isotonic regression

The authors de Leeuw et al. [54] define a number of loss functions that can be used for isotonic regression as an alternative to the standard square error (Brier score) and other point-based metrics. These include:

* **Huber loss**. This is a piecewise metric that is quadratic for small errors and linear for large ones.
* **General least-squares**. This metric is in the form for some matrix .
* **Asymmetric least-squares**. This uses different weights for negative or positive values of .
* **Soft insensitive loss function** (SILF). This is zero for small errors, linear for medium errors, and quadratic for large errors.

Further analysis of these metrics is beyond the scope of this review as they are not widely used for measuring classifier calibration.

Bin Count Search

Section 4.12 describes a bin count search method for deciding the optimum number of bins via an efficient interval search method. From analysis of the published code[[1]](#footnote-1), the search method maintains a lower bound (initialised to one bin), an upper bound (initialised to the lower of bins or the number of bins required to maintain samples per bin), and a current value (initialised to the lower of or the upper bound). If the current value results in an unacceptable non-monotonic function, the upper bound is set to one below the current value and the new current value is set to halfway between the lower and upper bounds. If the current value results in an acceptable monotonic function, the lower bound is set to the current value and the new current value is set to the lower of double the old current value or halfway between the lower and upper bounds. If, after the above procedure, the current value is equal to the lower bound and less than the upper bound it is incremented by unity. The procedure repeats until the lower bound, current value, and upper bound are identical. In the implementation, , , and . The value of does not affect the value of the computed metric – only its speed of computation. The other two parameters would not affect the value of the metric unless the constraints are violated. A “pure” version of the binning strategy would set and allow the upper bound to grow if the lower bound reaches the upper bound and the function is still monotonic. The pure version of the strategy has no adjustable parameters.

Imbalanced calibration error

The bin-based imbalanced calibration error (ICE) is defined in [30] as:

|  |  |
| --- | --- |
|  | (94) |

In (94), is the normalization constant and is a parameter between zero and unity that depends on the class size proportions . When the classes have equal sizes, and this is equivalent to ECE. The aim of the metric is to deal with imbalanced datasets by increasing the relative weight of bins with a small number of datapoints.

Contraharmonic expected calibration error

The contraharmonic expected calibration error (CECE) is based on the contraharmonic mean of the individual class subset ECEs [70]. If the ECE for a single class is , then CECE is:

|  |  |
| --- | --- |
|  | (95) |

The squared terms in the numerator mean that the CECE is more biased towards the ECE of classes with high values and thus more severely “punishes” models that may have a few badly calibrated classes than the class-wise ECE metric, which may be considered a desirable property [70].

Expected individual calibration error

Wu et al. [106] define a metric called expected individual calibration error (EICE). The theoretical version of this metric is the same as MAE. However, to deal with the issue that individual datapoints are either correct or incorrect, and even for a calibrated inaccurate classifier, they replace by a leave-one-out jackknife estimate of that can in principle be equal to , producing a proper scoring rule. According to [106], ECE measures calibration from a global perspective, does not guarantee calibration at the individual level, and can ignore rare categories. In contrast, being able to analyze calibration at the datapoint level with EICE is beneficial for rare categories. However, a disadvantage of EICE is that this metric requires computation of the Hessian of the model loss with respect to the model parameters, which may be computationally infeasible for large models. If access is not granted to the model’s internal structure, then it is not possible to compute the metric. This limits its applicability.

Max-variance mean-split

Max-variance mean-split (MVMS) is a binning strategy to be used as a variant of standard binning in ECE, and is described in the supplementary material of [95]. MVMS is a recursive partitioning scheme that splits confidence predictions in a potentially multi-class setting along the mean of the class dimension with highest variance in confidence scores. An additional regularization procedure sets a minimum number of datapoints per bin, which is set to 1000 in the experiments. MVMS is not compared to other binning strategies in [95] as it is used as standard for all experiments.

Local calibration error

Luo et al. [58] define the local calibration error (LCE) metric, which has some similarities with PIECE due to its use of feature values. Datapoints are binned according to confidence value. Within each bin, a points-based score is computed as the sum of signed calibration errors weighted by a kernel that operates on the feature space. The multi-class extended definition of LCE used by Barbera et al. [6] is:

|  |  |
| --- | --- |
|  | (96) |

The kernel weighting smooths the point-wise error over its nearest neighbours in a soft manner. In [58], , the Laplace kernel is used, and it is set to operate on a reduced-dimensionality feature set computed using principal components analysis rather than raw features. The kernel bandwidth is chosen by testing multiple values and using the inflection point on a performance curve. In [6], and any bins with fewer than 20 points are discarded to prevent unstable estimates. LCE has similar advantages to PIECE. The disadvantage of LCE is the high number of adjustable parameters (number of bins, kernel shape choice, kernel bandwidth) and the requirement to access feature values.

Fuzzy calibration error

Bihani et al. [9] define the soft-bin fuzzy calibration error metric (FCE) using trapezoidal bin membership functions where maximum bin membership is attained for the middle half of equal-width bins. In experiments, FCE showed better calibration error estimation than ECE especially in multi-class settings, that may have some bins with few points. Unlike ECE, FCE is not sensitive to the binning strategy [9]. A disadvantage of FCE compared to SBECE and DECE is that it only uses information from adjacent bins. If several contiguous bins are sparsely populated, the variance of the estimate for those bins would be high.

Harmonic Calibration Score

The harmonic calibration score (HCS) is a bin-based metric that measures both calibration (via ECE) and accuracy (A) simultaneously [91]. It is defined as:

|  |  |
| --- | --- |
|  | (97) |

The parameter controls the balance between accuracy and calibration, with larger emphasizing calibration more. If then this is the harmonic mean, giving equal emphasis. The aim of the metric is to have a single performance measure to compare models [91]. It shares this property with more standard metrics like Brier score and NLL. However, where it is necessary to measure accuracy and calibration separately, this metric cannot be used.

Well-calibration ratio

The well-calibration ratio (WCR) is a bin-based measure of calibration for multi-class classifiers [25]. Computation of WCR requires a few steps. First the data is partitioned into groups , each of which contains datapoints with whose maximum confidence is associated with one particular class . It is implicitly assumed that there is a mechanism for breaking ties in the maximum confidence. The mean confidence , for is given by:

|  |  |
| --- | --- |
|  | (98) |

The mean correctness is given by:

|  |  |
| --- | --- |
|  | (99) |

For a well-calibrated classifier, the and the should converge to each other. Therefore, the WCR is defined as:

|  |  |
| --- | --- |
|  | (100) |

For a perfectly calibrated classifier, , and this decreases as the calibration worsens, with a minimum value of zero. For a binary classifier, WCR is equivalent to a two-bin bin-based metric. Therefore, it does not provide as fine-grain a measure as other bin-based metrics that have more bins.

The well-calibration ratio is maximised by having all confidence values in a partition equal to the mean correctness for that partition. This property rewards confidences being pushed to the centre. In contrast, the Brier score rewards probabilities near zero or unity. To overcome these drawbacks Gebel [25] defines a “calibration measure” that combines the WCR and the BS. This is defined as:

|  |  |
| --- | --- |
|  | (101) |

Although [25] is a well-cited thesis due to its introduction of the Dirichlet calibration recalibration algorithm, neither the WCR nor the combined calibration measure appear to have seen use outside of [25]. Due to their lack of wide use or analysis, it is recommended these metrics are not used.

Maximum mean discrepancy and calibration estimation risk

Marx et al. [60] describe how the pre-existing kernel-based maximum mean discrepancy (MMD) metric for measuring the distance between general probability distributions can be specialised for both regression and classification. Their formula for MMD in classification is the same as used in SKCE-UQ (see (62) and (64) above), but the implementation of MMD uses a radial basis function as the kernel, in contrast to the Laplace kernel used for SKCE-UQ. MMCE is a special case of MMD applied to top-label calibration [60]. The primary aim of Marx et al. in using MMD is as an auxiliary loss function to train neural net models rather than a metric to measure performance. Using MMD in this way results in higher accuracy and better calibrated models (as measured by ECE) than using ECE-KDE, MMCE, or no auxiliary loss function. In experiments, the kernel bandwidth was set through a validation process, thus not requiring manual assignment of this parameter. However, MMD is sensitive to bandwidth value [60], which limits its value as a calibration performance metric.

Gruber et al. [29] define a metric called calibration estimation risk (CER) in the same way as SKCE-UQ and MMD. An RBF kernel is used, and the kernel bandwidth is set through a validation process

Reliability map

Error correcting output codes (ECOC) are an approach to combining the opinions of many different “experts”, where each expert is a binary classifier trained on a sub-problem of the full multiclass problem (see Section 2.3). In an ideal case, this requires each expert to produce not only probabilities but also the uncertainty in those probabilities. When combining information from the different experts, this allows the opinion of uncertain experts to have a lower weight. Least squares ECOC (LS-ECOC) is a standard method for combining probabilities but assumes all experts are equally certain among themselves and for all datapoints.

Kull et al. [46] propose a “reliability map” to estimate the reliability of experts for each datapoint, rather than their overall reliability. This instance-wise representation enables the learning of nonlinear boundaries even when using linear base models, reduces bias, and produces better models. The reliability is used to develop a modified probability combination algorithm called LS-ECOC-Reliability (LS-ECOC-R), which outperforms LS-ECOC on all synthetic and real datasets tested.

The theoretical local reliability for a confidence is based on the ratio between the variance of estimated true probability at that point and , which is the variance of the Bernoulli distribution associated with the calibration map. The theoretical reliability is defined as

|  |  |
| --- | --- |
|  | (102) |

A reliability of zero means that the estimated true probability is zero or unity. A reliability of unity means that the estimated true probability is the same as the reported confidence .

For practical implementation, datapoints are placed into clusters of size that have similar values of and . An unbiased estimator of the reliability map in one cluster with mean confidence is:

|  |  |
| --- | --- |
|  | (103) |

The overall reliability map is learnt using an Epanechnikov kernel with a fixed bandwidth and local linear regression using as an input. The calibration map is learnt in a similar manner. The bandwidth for calibration map learning was set to 0.01, based on optimization from cross-validation. The bandwidth for reliability map learning was set to times this value, at 0.1. Since local linear regression can produce estimates outside the range [0, 1], and extreme values near 0 or 1 are undesired, the estimates are clipped to 0.001 or 0.999.

To test the reliability map approach, synthetic data with a Gaussian distribution was used so that the true calibration and reliability maps are known. The reliability map approach consistently outperformed a constant reliability estimator across all dataset sizes, with 400 being the smallest size tested.

The clustering approach used in reliability map estimation reduces the effective sample size by a factor of making reliability maps harder to estimate than calibration maps. Typically, 2000 datapoints are required for reliability map estimation when used as part of a wider processing chain, but sometimes only as few as 400 are needed.

Kernel calibration conditional Stein discrepancy

The kernel calibration conditional Stein discrepancy (KCCSD) test statistic is introduced by Glaser et al. [26]. The KCCSD test is a hypothesis test for general probability models that are not necessarily normalised. Normalization is not an issue for classifier models, as it is trivial to normalise a set of un-normalised probabilities by dividing by the sum of those probabilities. However, this can be an issue for some regression models. KCCSD is a special case of the squared kernel calibration error (SKCE) and the maximum mean discrepancy (MMD) metrics. The test is based on a bootstrap approach to estimating quantiles and determining a significance level. Computing the KCCSD statistic involves intricate nested equations and definitions, making its implementation nontrivial. Glaser et al. [26] analyse probabilistic models in general without analysing the classification case in detail. Due to the complexity of this metric, and lack of analysis specifically for classification it is not considered further here.

1. Further discussion
   1. Skill scores

Metrics are only directly comparable when used to compare different classifier models applied to the same dataset. Data from a different scenario may suffer some form of variation, making the prediction task inherently harder. If a scenario is understood, then a reference (i.e. default) prediction and optimum prediction may be calculated. A skill score is the achieved value of a metric in relation to the reference and optimum predictions. The score takes on a value of zero when matching the reference and unity when matching the optimum [27]. Since skill scores can in principle be computed for any of the metrics catalogued in this survey, they are not discussed further.

* 1. Bootstrapping and consistency sampling

All the metrics in this review have some inherent randomness. While the calculation of the metric is usually deterministic given a fixed dataset, the dataset itself is random. Furthermore, some metrics, such as the surrogate interval calibration error, are random, even given a fixed dataset. Therefore, it is of importance to know the uncertainty in the metric to determine the significance of different values.

Some metrics have known distributions, or at least good approximate known distributions. In those cases, p-values can be computed directly. Where the distribution is unknown, Monte Carlo methods can be used to estimate the uncertainty. Two methods for doing this are bootstrapping and consistency resampling [95].

Bootstrapping creates several new datasets from the original one by sampling with replacement. The metric is computed for each bootstrapped dataset and the variability in computed values is used to understand the uncertainty in this metric.

Consistency resampling, a modified version of bootstrapping, assumes that there is a known calibration map , which in practice can be estimated from the data. Bootstrapping is used to generate samples just for . The calibration map is used for each bootstrapped sample to produce . This is then used to randomly select a label with probability . The final bootstrapped dataset is composed of pairs of values . In contrast to standard bootstrapping, consistency resampling produces labels that were not necessarily in the original dataset but are consistent with the calibration map. This method allows smaller datasets to be used than standard bootstrapping, at the expense of requiring the calibration map to be estimated.

* 1. Consistency calibration

Point-based metrics primarily focus on reliability and are measures of mismatch between confidence and accuracy. Tao et al. [92] propose an alternative view of calibration based on the concept of consistency. Under this view, models that are highly confident should be consistent in assigning the same top-class decision when there are minor perturbations to the data. If is the input data, is a perturbed version of it, is some distance measure, is a small value, is the number of Monte Carlo runs, and is the class index, the definition of consistency for a single datapoint is:

|  |  |
| --- | --- |
|  | (104) |

A classifier is said to be consistent if the consistency matches the confidence, i.e. . A probability calibration metric for consistency is not directly defined in [92]. However, such a metric could easily be constructed by replacing or in any of the other point-based metrics in this section with or as appropriate.

1. List of symbols

| **Symbol** | **Definition** |
| --- | --- |
|  | Soft-binned accuracy |
|  | Number of bins |
|  | Bin index |
|  | Confidence of datapoint |
|  | Mean confidence in bin |
|  | Kernel density estimate of confidence |
|  | Number of grid points |
|  | Grid point index |
|  | Number of proximity bins |
|  | Kernel bandwidth or proximity bin index, according to context |
|  | Logit function |
|  | Pairwise error term |
|  | Indicator function |
|  | Datapoint index |
|  | Second class index or datapoint index, according to context |
|  | Number of classes |
|  | Class index or other index, according to context |
|  | Kernel function |
|  | Max parameter in surrogate interval calibration error |
|  | Loss function |
|  | Number of Monte Carlo runs |
|  | Reliability map cluster size or Monte Carlo run index |
|  | Total number of datapoints |
|  | Number of datapoints in bin |
|  | Parameter of the p-norm |
|  | Proportion of datapoints in bin |
|  | Number of partitions of data |
|  | Partition index |
|  | Random sample from label distribution |
|  | Reliability map estimate |
|  | Number of datapoints in a sliding bin |
|  | Sampled value |
|  | Soft-binned size |
|  | Scoring function |
|  | Transpose operator |
|  | Bin membership function |
|  | Bin width |
|  | Biases in a model |
|  | Weights in a model |
|  | Class importance |
|  | | Feature vector of datapoint |
|  | Class label of datapoint |
|  | True proportions of each class in the dataset |
|  | Calibration map |
|  | Mean accuracy in bin |
|  | Normalization constant |
|  | Exponent in alpha scores or other exponent, according to context |
|  | Focus parameter |
|  | Kronecker delta function |
|  | Small value |
|  | Set of eligible bins |
|  | Mean |
|  | Inverse logit function |
|  | Standard deviation |
|  | Temperature parameter |
|  | Number of elements (cardinality) of a set |

**Table 3** Notation for probability calibration used in this paper.

1. Summary Table of Calibration Metrics

Table 4 below summarizes all the individual metrics discussed. For each metric, the table lists: alternative names; family; value range (zero indicates perfect calibration unless otherwise stated); whether the metric is proper; whether a hypothesis test (HT) is available; whether they distinguish under-confidence from over-confidence (UO); pros and cons; and an external reference where the metric is defined or described in more detail.

| **Metric name** | **Family** | **Range** | **Proper** | **HT** | **UO** | **Pros** | **Cons** | **Ref** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Brier score (BS) | Point |  | Yes | No | No | * Widely used * Easy to understand * Easy to compute * Robust to test set size | * Mixes calibration and accuracy * Cannot achieve a perfect score without perfect accuracy | [27] [87] |
| Root Brier score (RBS) | Point |  | Yes | No | No | * Easy to understand * Easy to compute * Robust to test set size | * Mixes calibration and accuracy * Cannot achieve a perfect score without perfect accuracy * Less used than Brier Score | [28] |
| Negative log likelihood (NLL), binary cross-entropy, ignorance, logarithmic score, predictive deviance, logistic loss | Point |  | Yes | No | No | * Widely used * Easy to compute | * Mixes calibration and accuracy * Cannot achieve a perfect score without perfect accuracy * Unstable wrt outliers or label errors | [31] |
| Focal loss (FL) | Point |  | No | No | No | * Emphasises hard, misclassified examples * Improves calibration more than NLL when used as a loss function. | * Adjustable focussing parameter * Primarily used as a loss function rather than a metric * Only computed for the top-confidence class rather than the full class computation | [56] |
| Entropic calibration difference (ECD) | Point |  | Yes | No | Yes | * Easy to compute * Can assess under/over confidence (0=perfect) * Theory relates to non-classifier metrics | * Dominated by points with very low or high confidence | [88] |
| Global squared bias (GSB), reliability-in-the-large | Point |  | No | No | No | * Very easy to understand * Easy to compute | * Imprecise: Can be zero if different parts of the reliability diagram are un-calibrated but cancel out | [24] |
| Multi-class difference of confidence and accuracy (MDCA) | Point |  | No | No | No | * Very easy to understand * Easy to compute * Differentiable: can be used for optimization | * Imprecise: Can be zero if different parts of the reliability diagram are un-calibrated but cancel out * Primarily used as a loss function rather than a metric | [37] |
| Expected to observed (EO) ratio | Point |  | No | No | Yes | * Very easy to understand * Easy to compute * Can assess under/over confidence (1=perfect) | * Imprecise: Can be zero if different parts of the reliability diagram are un-calibrated but cancel out | [81] |
| Success rate, zero-one score, mis-classification loss | Point |  | Yes | No | No | * Very easy to understand * Easy to compute | * Imprecise: only check whether most likely class is correct but doesn’t depend on probability | [27] |
| Dawid-Sebastiani score (DSS) | Point |  | Yes | No | No | * Easy to compute * Takes uncertainty in estimate into account | * Not strictly proper * Approximates distribution by two moments only | [64] |
| Normalised squared error score (NSES) | Point |  | No | No | Yes | * Easy to compute * Takes uncertainty in estimate into account * Can assess under/over confidence (1=perfect) | * Not widely used for classifiers | [64] |
| Mean absolute error (MAE), hinge loss | Point |  | No | No | No | * Easy to compute | * Worse than Brier score * Almost never worth reporting | [22] [45] |
| Expected individual calibration error (EICE) | Point |  | Yes | No | No | * Enables analysis of rare classes | * Requires access to model structure and parameters * Computationally intensive for large models * Difficult to compute | [106] |
| Pointwise error | Point |  | No | No | No | * Easy to compute | * Adjustable parameter * Values other than rarely used * Chebyshev norm is not robust (depends on one datapoint) | [54] |
| L1eps error | Point |  | Yes | No | No | * Easy to compute * Robust to outliers * Continuous derivatives for all degrees | * Adjustable parameter * Non-standard range | [54] |
| Spiegelhalter z statistic | Point |  | No | Yes | Yes | * Easy to compute * Can assess under/over confidence (0=perfect) | * Not as powerful as HL statistic * Undefined if all confidences are in {0, 0.5, 1} * Unstable if all confidences are near {0, 0.5, 1} | [40] |
| Pseudo-spherical score (PSS) | Point |  | Yes | No | No | * Easy to compute | * Adjustable parameter * Hard to interpret meaning | [13] |
| Power score | Point |  | Yes | No | No | * Easy to compute | * Adjustable parameter * More widely used PLS is a special case | [21] |
| Proper linear score (PLS) | Point |  | Yes | No | No | * Easy to compute | * Same as Brier score for binary classifiers | [13] |
| Soft F1 score (SF1) | Point |  | No | No | No | * Puts emphasis on mistakes when these are small in number | * Hard to interpret meaning | [38] |
| Expected Calibration Error (ECE), estimated CE, empirical CE, -ECE | Bin |  | No | No | No | * Easy to understand * Easy to compute * Widely used | * Adjustable parameter (#bins) * Few datapoints in fixed bins: high variance * Assumes perfect calibration in empty bins * Only checks most likely class * Conflates calibration and sharpness * Depends on the scale of probabilities * Highly discontinuous function of confidence values * The most criticised metric | [31] [87] |
| Root mean-square calibration error (RMSCE), reliability-in-the-small, reliability, -ECE | Bin |  | No | No | No | * Easy to understand * Easy to compute * Commonly used * Better than ECE * Good theoretical properties | * Adjustable parameter (#bins) * For fixed-width bin variant: few datapoints in fixed bins: high variance * Discontinuous function of confidence values, highly discontinuous for fixed width bins | [18] [49] [53] |
| Marginal calibration error (MCE) | Bin |  | No | No | No | * Easy to compute * Explicitly addresses class-wise multi-class calibration * General form of other metrics | * Several adjustable parameters * Few datapoints in fixed bins: high variance * Can be dominated by small probabilities * Does not address full multi-class calibration * Not widely used | [49] |
| Class-wise ECE (CWCE), macro subset ECE (MSECE), static calibration error (SCE) | Bin |  | No | No | No | * Easy to compute * Explicitly addresses class-wise multi-class calibration * Deals with imbalanced datasets * Special case of MCE | * Adjustable parameter (#bins) * Equal weights for classes mean small classes have a proportionately high influence * Few datapoints in fixed bins: high variance * Can be dominated by small probabilities * Does not address full multi-class calibration * Not widely used | [69] [70] [87] |
| Weighted subset expected calibration error (WSECE). | Bin |  | No | No | No | * Easy to compute * Explicitly addresses class-wise multi-class calibration * Special case of MCE | * Adjustable parameter (#bins) * Few datapoints in fixed bins: high variance * Not widely used * Can be dominated by small probabilities | [70] |
| Adaptive calibration error (ACE), equal-mass ECE (ECE-EM) | Bin |  | No | No | No | * Easy to understand * Easy to compute * Commonly used * Better than ECE | * Adjustable parameter (#bins) * Conflates calibration and sharpness * Discontinuous function of confidence values * Can be dominated by small probabilities | [69] |
| Thresholded Adaptive Calibration Error (TACE), ECE@k | Bin |  | No | No | No | * Easy to compute * Better than ECE * Prevents many small probabilities washing out averages | * Multiple adjustable parameters * Discontinuous function of confidence values | [69] [30] |
| Above/below area between curves (ABC), integrated calibration error | Bin |  | No | No | Yes | * Easy to understand * Easy to compute | * Multiple variables make it hard to assess multiple recalibration algorithms automatically * All the cons of ECE | [36] |
| Distance from calibration error (DCE) | Bin |  | No | No | No | * Has some good theoretical properties | * Practical implementation not well studied * Adjustable parameter (#bins) * Few datapoints in fixed bins: high variance * Highly discontinuous function of confidence values | [1] [10] |
| Imbalanced calibration error (ICE) | Bin |  | No | No | No | * Easy to compute * Deals with imbalanced datasets * Prevents many small probabilities washing out averages | * Adjustable parameter (#bins) * Discontinuous function of confidence values | [30] |
| Sanders Modified Brier Score (SMBS), Murphy’s Brier Score | Bin |  | No | No | No | * Easy to compute * Can be decomposed into uncertainty, reliability, and resolution | * Adjustable parameter (#bins) * Discontinuous function of confidence values * Mixes calibration and accuracy * Cannot achieve perfect score without perfect accuracy * Difference between SMBS and the better-known Brier Score is small | [36] |
| Region-balanced ECE (RBECE) | Bin |  | No | No | No | * Easy to compute * Deals with imbalanced datasets * Prevents many small probabilities washing out averages | * Ignores some datapoints * Multiple adjustable parameters * Discontinuous function of confidence values | [20] |
| Label-binned calibration error (ECE-LB), Probability deviation error (PDE) | Bin |  | No | No | No | * Easy to compute * Lower bias than ECE | * Multiple adjustable parameters * Discontinuous function of confidence values | [84] |
| Sweep calibration error (ECE-SWEEP) | Bin |  | No | No | No | * Better than ECE * No adjustable parameters (unlike most bin metrics) * Low levels of bias | * Discontinuous function of confidence values | [84] |
| Contraharmonic expected calibration error (CECE) | Bin |  | No | No | No | * Easy to compute * Penalises high ECE in individual classes | * Same cons as ECE | [70] |
| Maximum calibration error (MCE), -ECE | Bin |  | No | No | No | * Gives worse case calibration: may be useful in safety applications | * Adjustable parameter (#bins) * Few datapoints in fixed bins: high variance * Highly sensitive to placement of bins * Highly discontinuous function of confidence values | [31] |
| Max-variance mean-split (MVMS) | Bin |  | No | No | No | * Has multiclass considerations * Better than ECE | * Multiple adjustable parameters * Discontinuous function of confidence values | [95] |
| Fit-on-the-test ECE (ECE-FOTT) | Bin |  | No | No | No | * Better than ECE * No adjustable parameters (unlike most bin metrics) | * Cross-validation slow and more complex to run * Discontinuous function of confidence values | [42] |
| Surrogate interval calibration error (SICE) | Bin |  | No | No | No | * Better than ECE * Number of bins is optimally chosen | * Monte Carlo slow and more complex to run * Monte Carlo techniques are not good for assurance * Minor adjustable parameter * Discontinuous function of confidence values * Not as good as Laplace kernel error | [10] |
| Cutoff calibration error (CCE) | Bin |  | No | No | No | * Better than ECE * No adjustable parameters (unlike most bin metrics) * Continuous function of confidence values. | * Slightly complex to implement interval searching | [85] |
| Expected signed calibration error (ESCE), miscalibration score (MCS) | Bin |  | No | No | Yes | * Can assess under/over confidence * Averaging over number of bins makes it smoother than other bin metrics | * Discontinuous function of confidence values | [100] |
| Weighted subset miscalibration score (WSMCS) | Bin |  | No | No | Yes | * Can assess under/over confidence * Explicitly addresses class-wise multi-class calibration * Takes into account the imbalance between the numbers of under/over-confident classes | * Under/over-confident classes can cancel each other out, masking miscalibration in individual classes * Discontinuous function of confidence values * Does not address full multi-class calibration | [100] |
| Global interpretable calibration index (GICI) | Bin |  | No | No | No | * Claimed to be more interpretable than other metrics | * Only defined for top-label classification * Discontinuous function of confidence values * In contrast to most metrics, high is good | [14] |
| Directional interpretable calibration index (DICI) | Bin |  | No | No | Yes | * Can assess under/over confidence | * Only defined for top-label classification * Discontinuous function of confidence values | [14] |
| Soft-binning ECE (SBECE) | Bin |  | No | No | No | * Differentiable: can be used for optimization | * Adjustable parameter (temperature) * Worse match to theoretical ECE than DECE | [43] |
| Differentiable ECE (DECE) | Bin |  | No | No | No | * Differentiable: can be used for optimization * Better match to theoretical ECE than SBECE | * Adjustable parameter (temperature) | [12] |
| Fuzzy calibration error metric (FCE) | Bin |  | No | No | No | * Smoother than ECE | * Only uses information from adjacent bins | [9] |
| Robust expected calibration error (RECE), RECE-G | Bin |  | No | No | No | * Lower bias than ECE, SCE, ACE * Less discontinuous function of confidence values than ECE | * Multiple adjustable parameters * Discontinuous function of confidence values * Same cons as ECE | [33] |
| De-biased square calibration error (CE2-DB), De-biased root mean square calibration error (DRMSCE) | Bin |  | No | Yes | No | * Unbiased metric, unlike most other bin metrics | * Same cons as ECE apart from bias | [49] |
| De-biased root mean square calibration error with bin count search (DRMSCE-BCS) | Bin |  | No | Yes | No | * Unbiased metric, unlike most other bin metrics * No adjustable parameters * Lower bias than standard DRMSCE | * Only checks most likely class * Discontinuous function of confidence values | [49] [77] |
| De-biased ECE (ECE-DB) | Bin |  | No | Yes | No | * Unbiased metric, unlike most other bin metrics | * Monte Carlo slow and more complex to run * Monte Carlo techniques are not good for assurance * Same cons as ECE | [49] |
| Proximity-informed ECE (PIECE) | Bin |  | No | No | No | * Considers bias from under/over-confidences in dense/sparse regions of data space * Mitigates under/over-confidence cancellation effect * A stricter score than ECE | * Needs access to feature vectors, not always available * Two dimensional bins mean each bin has fewer datapoints than ECE * Discontinuous function of features due to reliance on a fixed number of neighbours | [107] |
| Local calibration error (LCE) | Bin |  | No | No | No | * Considers bias from under/over-confidences in dense/sparse regions of data space * Mitigates under/over-confidence cancellation effect * Continuous function of features due to kernel | * Needs access to feature vectors, not always available * Several adjustable parameters * Not commonly used | [6] |
| Partitioned calibration error (PCE) | Bin |  | No | No | No | * More general than PIECE: allows any grouping of datapoints * Allows averaging over different groupings of data * Mitigates under/over-confidence cancellation effect * A stricter score than ECE | * Needs access to feature vectors, not always available * Multi-dimensional bins mean each bin has fewer datapoints than ECE or possibly PIECE | [108] |
| Multi-view calibration error (MVCE) | Bin |  | No | No | No | * Allows averaging over different groupings of data * Mitigates under/over-confidence cancellation effect * Lower bias than ECE and ECE-SWEEP | * Adjustable parameter (#partitions) * Monte Carlo techniques are not good for assurance | [41] |
| Field-level expected calibration error (FECE), Field-ECE, variable-based expected calibration error (VECE) | Bin |  | No | No | No | * Considers bias from under/over-confidences in dense/sparse regions of data space * Mitigates under/over-confidence cancellation effect * Can be used to find areas of bias | * Needs access to feature vectors, not always available * Adjustable parameter (#bins) | [73] |
| Hosmer-Lemeshow (HL) statistic | Bin |  | No | Yes | No | * Widely used | * Adjustable parameter (#bins) * Less intuitive than other bin metrics * Highly sensitive to small amounts of miscalibration, for large datasets * Produces non-significant results in small samples * Discontinuous function of confidence values | [40] |
| Test for calibration (T-cal) | Bin |  | No | Yes | No | * Unbiased metric, unlike most other bin metrics * Use of number of bins makes it smoother than other bin metrics * Better than ECE | * Complicated definition * Multiple adjustable parameters * Unusual use of squaring compared to other bin metrics – less interpretable * Slower to compute than ECE * Could be overly sensitive as null hypothesis is rejected if any number of bins rejects the hypothesis | [53] |
| Test-based calibration error (TCE) | Bin |  | No | Yes | No | * Good for situations with imbalanced classes * Includes method for arranging the bins * Better than ECE | * Complicated definition * Slower to compute than ECE * Discontinuous function of confidence values | [61] |
| CalBin | Bin |  | No | No | No | * Averaging over overlapping bins makes it smoother than other bin metrics * Better than ECE | * Adjustable parameter (#points in a bin) | [8] |
| K-nearest neighbours ECE (ECE-KNN) | Bin |  | No | No | No | * Averaging over overlapping bins makes it smoother than other bin metrics * Lower bias than ECE-EW, ECE-SWEEP, and ECE-DEBIAS | * Requires a partially manual process for selecting k | [75] |
| Harmonic calibration score (HCS) | Bin |  | No | No | No | * Measures both calibration and accuracy simultaneously | * Adjustable parameter (balance between accuracy and calibration) * Measures both calibration and accuracy simultaneously | [91] |
| Well-calibration ratio (WCR) | Bin |  | No | No | No | * Works with full multiclass calibration | * Complicated definition * Unusual compared to other bin metrics * Imprecise: for a binary classifier, equivalent to a two-bin bin-based metric * Not mentioned outside of original reference | [25] |
| Mean squared calibration error (MSCE) | Kernel |  | No | No | No | * Similar definition to Brier score, a popular metric * Simple definition – good interpretability * Smoother than bin metrics | * Adjustable parameter (kernel width) * Gaussian kernel introduces bias from edge effects * Slow to compute for large datasets: | [24] |
| Smooth ECE (SECE) | Kernel |  | No | No | No | * Simple definition – good interpretability * Smoother than bin metrics | * Adjustable parameter (kernel width) * Gaussian kernel introduces bias from edge effects * Slow to compute for large datasets: * Has same unabbreviated name as SMECE, which has better properties | [103] |
| Smooth ECE (SMECE) | Kernel |  | No | No | No | * Smoothing on residual gives better mathematical properties * Reflected Gaussian kernel alleviates bias from the standard Gaussian kernel * Consistent calibration measure and is efficient with respect to both sample complexity and runtime * Hyperparameter-free * Smoother than bin metrics * Fast to compute for large datasets: | * Reflected Gaussian kernel makes computation harder * Slightly more complicated definition than SECE reduces interpretability * Has same unabbreviated name as SECE | [11] |
| Dirichlet kernel density estimator (DKDE), , ECE-KDE, ECEKDE, | Kernel |  | No | No | No | * Full multiclass calibration * Consistent and de-biased * Differentiable: can be used for optimization * Smoother than bin metrics | * Slightly complex definition * Slow to compute for large datasets: | [78] |
| Beta kernel density estimator (BKDE) | Kernel |  | No | No | No | * Consistent and de-biased * Differentiable: can be used for optimization * Smoother than bin metrics | * Binary or class-wise calibration only * Slow to compute for large datasets: | [78] |
| Maximum mean calibration error (MMCE) | Kernel |  | No | No | No | * Has a multiclass version * Better properties than NLL – more stable * Smoother than bin metrics | * Multiclass version has complex definition * Biased and inconsistent metric * Slow to compute for large datasets: | [48] |
| Laplace kernel calibration error (LKCE) | Kernel |  | No | No | No | * Laplace kernel is consistent whereas Gaussian is not * Calibration map is less up-and-down than SICE * Smoother than bin metrics | * Slow to compute for large datasets: [a linear approximation exists] * Not as smooth as smooth calibration error (SCE) | [48] [10] |
| Unbiased linear squared kernel calibration error (SKCE-UL) | Kernel |  | No | Yes | No | * Unbiased * Works with full multiclass calibration * Fast to compute for large datasets: * Smoother than bin metrics | * Performance degrades with number of classes starting between 100 and 1000 classes | [104] |
| Unbiased quadratic Squared kernel calibration error (SKCE-UQ) | Kernel |  | No | Yes | No | * Unbiased * Works with full multiclass calibration * Works well with any number of classes * Smoother than bin metrics | * Monte Carlo HT slow and more complex to run * Monte Carlo techniques are not good for assurance * Slow to compute for large datasets: | [104] |
| Kernel local calibration error (KLCE) | Kernel |  | No | Yes | No | * Unbiased * Can work with full multiclass calibration * Smoother than bin metrics | * Several adjustable parameters * Only tested with top-label classification. * Requires access to feature values * Monte Carlo HT slow and more complex to run * Slow to compute for large datasets: | [98] |
| Maximum mean discrepancy (MMD) | Kernel |  | No | No | No | * Unbiased * Works with full multiclass calibration * Smoother than bin metrics | * Several adjustable parameters * Slow to compute for large datasets: * Sensitive to choice of kernel bandwidth | [60] |
| Calibration estimation risk (CER) | Kernel |  | No | No | No | * Unbiased * Smoother than bin metrics | * Only used with top-label classification * Slow to compute for large datasets: * Sensitivity to kernel bandwidth not assessed | [29] |
| Expected Conditional Maximum Mean Discrepancy (ECMMD) | Kernel |  | No | Yes | No | * Smoother than bin metrics * More computationally efficient at than some other kernel metrics | * Monte Carlo techniques are not good for assurance * Slightly complex definition | [15] |
| Smoothed kernel density estimator (SKDE) | Kernel |  | No | No | No | * Better than all bin metrics at assessing calibration errors * Smoother than other kernel metrics and all bin metrics * Fast to compute for large datasets: | * Slightly complex definition * Computation scales badly with number of classes, recommended as a top-label or class-wise metric * Biased metric | [109] |
| Least squares error correcting output codes reliability (LS-ECOC-R) | Kernel |  | No | No | No | * Low bias * Produces better results than some other metrics * Smoother than bin metrics | * Multiple adjustable parameters * Complex definition * Needs more datapoints than other kernel methods | [46] |
| Smooth calibration error (SCE) | Kernel |  | No | No | Yes | * Consistent calibration measure * Smoother performance curves than LKCE * Less up-and-down than SICE * Smoother than bin metrics | * Non-trivial definition: hard to interpret meaning * Slow to compute for large datasets: | [10] |
| Logit smoothed ECE (LSECE) | Kernel |  | No | No | No | * Smoother than bin metrics | * Adjustable parameter (sigma) * Monte Carlo based metrics are not good for assurance * Slow to compute for large datasets: | [17] |
| Kernel calibration conditional Stein discrepancy (KCCSD) | Kernel |  | No | Yes | No | * Works with full multiclass calibration * Smoother than bin metrics * Part of a wider theory on probabilistic algorithms | * Implementation is highly complex * Monte Carlo HT slow and complex to run * Analysis specific to classifiers is not available | [26] |
| Integrated calibration index (ICI) | Curve |  | No | No | No | * Relatively simple definition * Similar definition to ECE – helps gain acceptability * Easy to compute * Assigns a greater weight to dense data areas than sparse ones – more stable * Better than the similar EP | * Multiple adjustable parameters * Slow to compute for large datasets: | [4] [40] |
| Error percentile (EP), EX | Curve |  | No | No | No | * Relatively simple definition * Easy to compute * Easy to interpret * Assigns a greater weight to dense data areas than sparse ones – more stable | * Multiple adjustable parameters * Not as good as the similar ICI * Slow to compute for large datasets: | [4] |
| Estimated calibration index (ECI), expected calibration index | Curve |  | No | No | No | * Similar definition to BS – helps gain acceptability * Unlike BS, purely measures calibration * Smoother than bin metrics | * Non-trivial definition | [97] |
| Piecewise linear ECE (ECE-PL) | Curve |  | No | No | No | * Better than all metrics for reliability diagram quality * Better than average and outperforms PL3 when ranking recalibration methods * Similar definition to ECE – helps gain acceptability * No adjustable parameters * Easier to explain than PL3 * Fast to compute for large datasets: | * Cross-validation: slow to compute and slightly complex to implement | [42] |
| Piecewise linear in logit-logit space ECE (ECE-PL3) | Curve |  | No | No | No | * Has the popular temperature scaling as a special case * Better than all metrics except ECE-PL for reliability diagrams quality * Similar definition to ECE – helps gain acceptability * No adjustable parameters * Fast to compute for large datasets: | * Cross-validation: slow to compute and slightly complex to implement * Harder to explain than ECE-PL | [42] |
| FOTT beta scaling | Curve |  | No | No | No | * Better than all metrics except ECE-PL and ECE-PL3 for reliability diagram quality * Better than all metrics for quality of the calibration error estimates * No adjustable parameters * Fast to compute for large datasets: | * Cross-validation: slow to compute and slightly complex to implement * Rank of performance gets worse as the number of datapoints increases | [42] |
| Cox intercept and slope (CIS) | Curve |  | No | No | Yes | * Can assess under/over confidence * Can assess more than one aspect of calibration * No adjustable parameters * Classic method * Fast to compute for large datasets: | * Multiple variables make it hard to assess multiple recalibration algorithms automatically. * Not as flexible as the more general ECE-PL3 | [40] |
| Statistical beta calibration test (SBCT) | Curve |  | No | No | No | * Uniform emphasis across confidence values rather than data samples * Relatively easy to compute * Similar definition to ECE – helps gain acceptability * Fast to compute for large datasets: | * The metric is only as good as the beta fit to the data | [47] |
| Parabolic Wald statistic | Curve |  | No | Yes | No | * Relatively easy to compute * No adjustable parameters * Fast to compute for large datasets: | * The metric lacks interpretability | [24] |
| Empirical cumulative calibration error maximum absolute deviation (ECCE-MAD), Kolmogorov-Smirnoff (KS) | Cumulative |  | No | Yes | No | * Assumption free * No adjustable parameters * Better than ECE * Generalization available for top r classes form of multiclass problems * Easily interpretable | * Can (correctly) produce highly significant p-values for large datasets even when effect is small | [3] |
| Empirical cumulative calibration error range (ECCE-R), Kuiper | Cumulative |  | No | Yes | No | * Assumption free * No adjustable parameters * Better than ECE | * Can (correctly) produce highly significant p-values for large datasets even when effect is small * Slightly biased so that miscalibration false alarms are more likely than they should be | [3] |
| Multi-calibration metric (MCM) | Cumulative |  | No | No | No | * Assumption free * Assessment of sub-population calibration | * Requires access to feature values * Automated sub-population creation requires specification of adjustable parameters | [34] |
| Brownian bridge test (BBT) | Cumulative |  | No | Yes | No | * Assumption free * No adjustable parameters * Better than ECE * Easily interpretable | * Slightly biased so that miscalibration false alarms are more likely than they should be | [86] |

**Table 4** Summary of calibration metrics. HT indicates whether a hypothesis test is available. UO indicates the metric can distinguish under and overconfidence.

1. <https://github.com/e-pet/risk-score-fairness>, published 16th June 2025. [↑](#footnote-ref-1)