TCE: A Test-Based Approach to Measuring Calibration Error (Supplementary Material)

Takuo Matsubara[1,2] **Niek Tax**[3] **Richard Mudd**[3] **Ido Guy**[3]

[1]The Alan Turing Institute
[2]Newcastle University
[3]Meta Platforms, Inc.

This supplement contains all the additional results referred to in the main text. Appendix A contains the proof that the optimisation criterion of eq. (8) is indeed minimised using PAVA. Appendix B shows an example of bins obtained using PAVA-BC that caused mild violation of the monotonic constraint of the empirical probabilities $\{\widehat{P}_b\}_{b=1}^B$. Finally, additional experimental results are presented in Appendix C.

A OPTIMAL BINS BASED ON PAVA

The optimal bins defined by Definition 3 can be exactly computed under the error function D specified by eq. (9) which corresponds to the variance of each $\mathcal{D}_b^y$. The optimal bins result in minimisation of a weighted average of the variance of each $\mathcal{D}_b^y$ over all b, where the weights are proportional to the size of each bin. The following proposition shows that Algorithm 3 with PAVA-BC replaced by PAVA generates the optimal bins under the error function D. In what follows, we assume a standard setting where the solution of eq. (8) is at least not a set of only one single bin, i.e., $\{\Delta_b\}_{b=1}^1 = \{[0,1]\}$.

Proposition 1. *The minimum of eq. (8) in Definition 3 under the error function* D *in eq. (9) is attained at bins computed by Algorithm 3 with PAVA-BC replaced by PAVA.*

Proof. First, we show that the optimasation problem of eq. (8) in Definition 3 under the loss function D in eq. (9) is equivalent to the monotonic regression problem under the squared error. Recall that, given a choice of bins $\{\Delta_b\}_{b=1}^B$, each label subset $\mathcal{D}_b^y$ is defined by $\mathcal{D}_b^y := \{y_i \in \mathcal{D}^y \mid P_\theta(x_i) \in \Delta_b\}$. The input of Algorithm 3 is a set of labels $\mathcal{D}^y = \{y_i\}_{i=1}^N$ ordered by in ascending order of $\{P_\theta(x_i)\}_{i=1}^N$. This means that each label subset $\mathcal{D}_b^y$ is a set of consecutive elements in the ordered set $\{y_i\}_{i=1}^N$. Therefore, there exist corresponding indices n_b and n_{b+1} s.t. each label subset $\mathcal{D}_b^y$ can expressed by

$$\mathcal{D}_b^y = \{y_i \in \mathcal{D}^y \mid P_\theta(x_i) \in \Delta_b\} = \{y_i \in \mathcal{D}^y \mid i \text{ s.t. } n_b \leq i < n_{b+1}\}.$$

Accordingly, with the ordered labels $\mathcal{D}^y$, each empirical probability $\widehat{P}_b$ in $\mathcal{D}_b^y$ can be expressed by

$$\widehat{P}_b = \frac{1}{N_b} \sum_{y \in \mathcal{D}_b^y} y = \frac{1}{n_{b+1} - n_b} \sum_{j=n_b}^{n_{b+1}-1} y_j.$$

Define a set of scalars $\{g_i\}_{i=1}^N$ whose element $g_i \in [0,1]$ corresponds to the empirical probability $\widehat{P}_b$ of the bin index b if $n_b \leq i < n_{b+1}$. Namely,

$$g_i := \widehat{P}_b = \frac{1}{n_{b+1} - n_b} \sum_{j=n_b}^{n_{b+1}-1} y_j \quad \text{for each} \quad i \quad \text{s.t.} \quad n_b \leq i < n_{b+1}. \tag{1}$$

Under these notations, the optimisation criterion in eq. (8) can be rewritten as

$$\sum_{b=1}^B W_b \times \mathrm{D}(\mathcal{D}_b, \widehat{P}_b) = \frac{1}{N} \sum_{b=1}^B \sum_{y \in \mathcal{D}_b} \left(y - \widehat{P}_b\right)^2 = \frac{1}{N} \sum_{b=1}^B \sum_{i=n_b}^{n_{b+1}-1} \left(y_i - \widehat{P}_b\right)^2 = \frac{1}{N} \sum_{i=1}^N (y_i - g_i)^2. \tag{2}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

This formulation translates a problem of choosing bins $\{\Delta_b\}_{b=1}^B$ into a problem of finding a monotonically increasing sequence $\{g_i\}_{i=1}^N$ that is determined by the choice of indices $\{n_b\}_{b=1}^B$, so that eq. (11) is minimised. Therefore the optimasation problem of eq. (8) in Definition 3 under the loss function D in eq. (9) is equivalent to the monotonic regression problem under the squared error whose solution sequence $\{g_i\}_{i=1}^N$ is restriced to a form of eq. (10).

Next, consider a standard monotonic regression problem under the square error $\sum_{i=1}^N (y_i - \widehat{y}_i)^2$ for the ordered set $\{y_i\}_{i=1}^N$. PAVA finds a monotonically increasing sequence $\{\widehat{y}_i\}_{i=1}^N$ that minimises the square error. The solution sequence $\{\widehat{y}_i\}_{i=1}^N$ by PAVA is given in a form of eq. (10); see e.g. [de Leeuw et al., 2009, Henzi et al., 2022]. This means that there exists a set of indices $\{n_b^*\}_{b=1}^B$ s.t. the solution sequence $\{\widehat{y}_i\}_{i=1}^N$ by PAVA is expressed as

$$\widehat{y}_i = \frac{1}{n_{b+1}^* - n_b^*} \sum_{j=n_b^*}^{n_{b+1}^*} y_j \quad \text{for each} \quad i \quad \text{s.t.} \quad n_b^* \leq i < n_{b+1}^*$$

and the sequence $\{\widehat{y}_i\}_{i=1}^N$ satisfies the monotonic constraint $\widehat{y}_1 \leq \cdots \leq \widehat{y}_N$ holds. We can obtain such a solution sequence $\{\widehat{y}_i\}_{i=1}^N$ by applying any standard implementation of PAVA.

An output of most implementations of PAVA is the solution sequence $\{\widehat{y}_i\}_{i=1}^N$ rather than the associated indices $\{n_b^*\}_{b=1}^B$. However, the indices $\{n_b^*\}_{b=1}^B$ can be easily recovered from a given solution sequence $\{\widehat{y}_i\}_{i=1}^N$ of PAVA by simply finding all indeces i s.t. $\widehat{y}_i \neq \widehat{y}_{i+1}$. Finally, we consider constructing bins $\{\Delta_b\}_{b=1}^B$ based on the recovered indices $\{n_b^*\}_{b=1}^B$. Recall that the set of labels $\mathcal{D}^y = \{y_i\}_{i=1}^N$ are ordered in ascending order of $\{P_\theta(x_i)\}_{i=1}^N$. If we construct each bin Δ_b by

$$\Delta_b := \left[\frac{P_\theta(x_{n_b^*-1}) + P_\theta(x_{n_b^*})}{2}, \frac{P_\theta(x_{n_{b+1}^*-1}) + P_\theta(x_{n_{b+1}^*})}{2} \right],$$

it is sufficient to generate each label subset $\mathcal{D}_b^y$ that corresponds to

$$\mathcal{D}_b^y = \{y_i \in \mathcal{D}^y \mid P_\theta(x_i) \in \Delta_b\} = \{y_i \in \mathcal{D}^y \mid i \text{ s.t. } n_b^* \leq i < n_{b+1}^*\}.$$

Then the optimisation criterion in eq. (8), which is translated to the error of the monotonic regression problem of PAVA, is minimised by the choice of bins produced in this procedure. Observing that Algorithm 3 with PAVA-BC replaced by PAVA performs this procedure concludes the proof. □

B MILD VIOLATION OF MONOTONICITY BY PAVA-BC

A monotonic regression algorithm finds a monotonically increasing sequence $\widehat{y}_1 \leq \cdots \leq \widehat{y}_N$ that minimises some error $\mathrm{D}(\{\widehat{y}_i\}_{i=1}^N, \{y_i\}_{i=1}^N)$ for a given ordered set $\{y_i\}_{i=1}^N$. PAVA is one of the most common monotonic regression algorithms that uses the square error $\sum_{i=1}^N (\widehat{y}_i - y_i)^2$. For some partition $\mathcal{A}$ of indices $I = \{1, \ldots, N\}$ whose element $A \in \mathcal{A}$ is a set of consequentive indices in I, PAVA produces a solution sequence s.t. each element $\widehat{y}_i$ is given by $\widehat{y}_i = (1/|A|) \sum_{i \in A} y_i$ for A in which $i \in A$. We refer to each element A in the partition $\mathcal{A}$ of indices I as *block*. PAVA-BC produces a solution sequence that approximates the solution sequence by PAVA under the contraints of the minimum and maximum size of each block. For some partition $\mathcal{A}'$ of indices I, each element $\widehat{y}_i$ of the solution sequence is given by $\widehat{y}_i = (1/|A'|) \sum_{i \in A'} y_i$ for A' in which $i \in A'$ in the same manner as PAVA. PAVA-BC meets the minimum and maximum size constraints of each block $A' \in \mathcal{A}'$ at the cost of the possibility of mild violation of the monotonic constraint. It depends on the minimum and maximum size constraints, data, and models whether violation of the monotonic constraint occurs by PAVA-BC. Figure 3 shows an example where bins based on PAVA-BC did not violate the monotonicity of the empirical probabilities $\{\widehat{P}_b\}_{b=1}^B$. Figure 3 was computed using a random forest model trained on the *satimage* dataset used in Section 4.2, and corresponds to Figure 2 presented in Section 3. The total estimation error in eq. (8) and an average of the estimation error within each bin in eq. (9) for each set of the bins in Figure 3 were summerised in Table 1 presented in Section 3. Figure 4 shows an example where bins based on PAVA-BC violated the monotonic constraint of the empirical probabilities $\{\widehat{P}_b\}_{b=1}^B$. Figure 4 was computed using a random forest model trained on the *coil_2000* dataset used in Section 4.2. The total estimation error in eq. (8) for each set of the bins in Figure 4 was 0.0509, 0.0517, and 0.0521 for PAVA, PAVA-BC, binning based on 10-quantiles, respectively. An average of the estimation error within each bin in eq. (9) for each set of the bins in Figure 3 was 0.0834, 0.0627, and 0.0520 for PAVA, PAVA-BC, binning based on 10-quantiles, respectively.

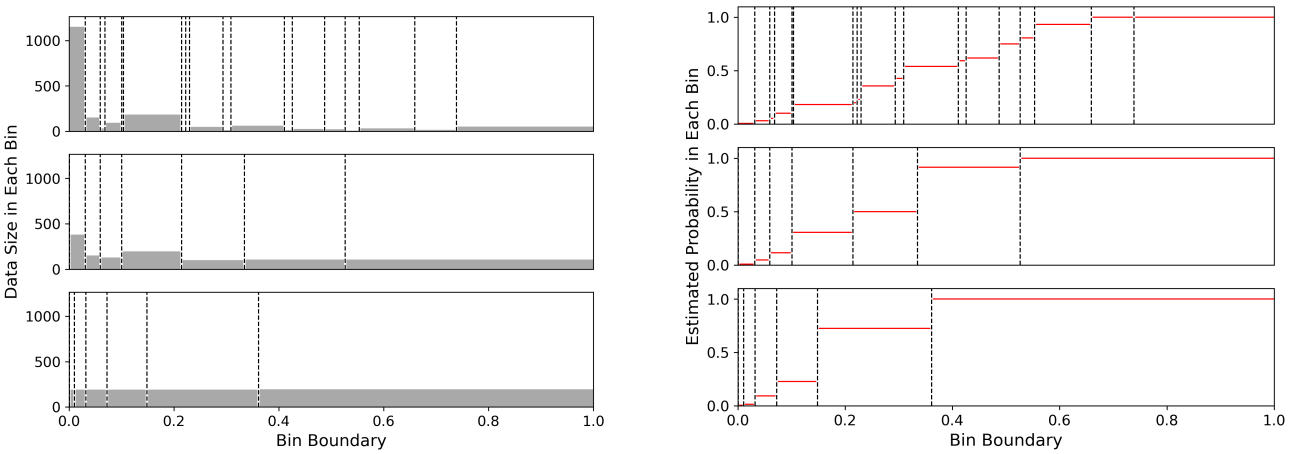


Figure 3: Comparison of bins based on three different approaches for a random forest model on the satimage dataset: (top) PAVA, (middle) PAVA-BC, (bottom) binning based on 10-quantiles. The dotted line in the left and right panels represents the boundary of each bin. The grey bar in the left panel repsents the size of each bin. The red line in the right panel repsents the empirical probability of each bin.

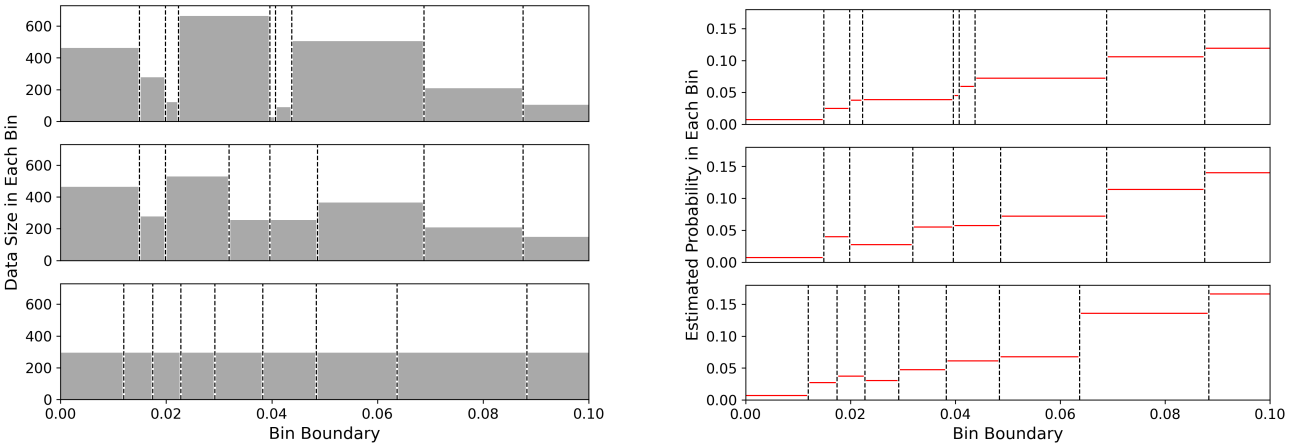


Figure 4: Comparison of bins based on three different approaches for a random forest model on the satimage dataset: (top) PAVA, (middle) PAVA-BC, (bottom) binning based on 10-quantiles. Each xaxis is restricted to a range $[0.0, 0.1]$ as the majority of bins were contained in the range in this example. The dotted line in the left and right panels represents the boundary of each bin. The grey bar in the left panel repsents the size of each bin. The red line in the right panel repsents the empirical probability of each bin. A random forest model trained on the coil_2000 dataset was used.

C ADDITIONAL EXPERIMENTS

We present additional experiments in each section that complement the experiments illustrated in the main text. We use the same settings as the main text for the minimum and maximum size for bins based on PAVA-BC as well as the bin number B for equi-spaced and quantile-based bins.

C.1 SIMULATION STUDY OF TCE

We perform detailed simulation studies of TCE in the same simplified setting as Section 4.1. We demonstarate sensitivity of TCE to its hyperparameters, an impact of different dataset size and prevalence, and sensitivity to a small purtabation to model predictions. In all experiments, we generated training and test data from the Gaussian discriminant analysis in Section 4.1, each with the prevalence $P_{\text{training}}(y)$ and $P_{\text{test}}(y)$, and compute TCE of a logistic model fitted to the training data. In all experiments except ones on an impact of different dataset size and prevalence, we set the training data size to 14000 and set the test data size to 6000. We then examine two cases where the model is calibated and miscalibrated synthetically, setting $P_{\text{training}}(y) = 0.5$ and $P_{\text{test}}(y) = 0.5$ for the first case and setting $P_{\text{training}}(y) = 0.5$ and $P_{\text{test}}(y) = 0.4$ for the second case. In summary, we present the following experimental analyses:

- Sensitivity to the minimum bin size N_{min} in PAVA-BC from $N_{\text{min}} = 1$ to $N_{\text{min}} = 3000$;
- Sensitivity to the maximum bin size N_{min} in PAVA-BC from $N_{\text{max}} = 6$ to $N_{\text{max}} = 6000$;
- Sensitivity to a pair of $(N_{\text{min}}, N_{\text{max}})$ in PAVA-BC chosen so that each binsize fall into selected ranges;
- Sensitivity to a small purtabation of predictions by a logit-normal noise with scale σ from $\sigma = 0.0$ to $\sigma = 1.0$;
- Sensitivity to a choice of significance level α in the Binomial test from $\alpha = 0.0001$ to $\alpha = 0.1$;
- Comparison of TCE by different choices of test, binomial test and t-test;
- Comparison of TCE by different total sizes N of test dataset from $N = 30$ to $N = 60000$;
- Comparison of TCE by different prevalences P of dataset from $P = 0.5$ to $P = 0.02$.

Tables 6 to 13 presents the result of each experiment above in order. In each table, TCE(P) denotes TCE based on PAVA-BC, TCE(Q) denotes TCE based on quantile-binning, and TCE(V) denotes TCE based on PAVA. For reference, we include values of ECE, ACE, MCE, and MCE(Q), where MCE(Q) denotes MCE based on quantile-binning. Observations from each result in are summarised as follows.

- Table 6: The performance of TCE(P) to evidence the well-calibrated model was consistently reasonable for any minimum binsize constaint between $N_{\text{max}} = 1$ and $N_{\text{max}} = 600$, while there was a breakdown point between $N_{\text{min}} = 600$ and $N_{\text{min}} = 3000$ where TCE(P) was no longer able to do so. This is likely because the number of bins produced under the contraint $N_{\text{min}} = 3000$ for the total datasize 6000 was 2 at maximum, which was too small to estimate the empirical probabilities $\{\widehat{P}_b\}_{b=1}^{B}$ accurately.
- Table 7: The performance of TCE(P) to evidence the miscalibrated model was consistently reasonable for any maximum binsize constaint between $N_{\text{max}} = 300$ and $N_{\text{max}} = 6000$, while there was a breakdown point between $N_{\text{min}} = 60$ and $N_{\text{min}} = 300$ where TCE(P) was no longer able to do so. This is likely because the number of bins produced under the contraint $N_{\text{max}} = 60$ for the total datasize 6000 was 100 at minimum, which is too large to estimate the empirical probabilities $\{\widehat{P}_b\}_{b=1}^{B}$ accurately.
- Table 8: The performance of TCE(P) to evidence both the well-calibrated and miscalibrated models was arguably the most reasonable when $(N_{\text{min}}, N_{\text{max}})$ was chosen so that the number of bins produced falls into the range $[5, 20]$. This suggests a huristic to use such $(N_{\text{min}}, N_{\text{max}})$ for other experiments.
- Table 9: At each model prediction $P_\theta(x)$, we sample a new prediction from a logit-normal distribution centred at $P_\theta(x)$ with scale σ to generate a perturbed prediction by a small noise. All calibration error metrics were shown to have similar sensitivities to the noise. The scale between $\sigma = 0.10$ and $\sigma = 0.50$ was the breakdown point where each metric started to produce an unreasonable score for the well-calibrated model.
- Table 10: The performance of TCE(P) to evidence both the well-calibrated and miscalibrated models was consistently reasonable for any significant level between $\alpha = 0.001$ and $\alpha = 0.1$, while there was a breakdown point between $\alpha = 0.1$ and $\alpha = 0.5$ where TCE(P) was no longer able to do so for the well-calibrated model.

- Table 11: TCE based on the Binomial test outperformed one based on the t-test in the majority of the settings. It is possible that the Binomial test produces more accurate outcomes than the t-test, given that it is an exact test whose test statistics does not involve any apporoximation.
- Table 12: The performance of TCE(P) to evidence both the well-calibrated and miscalibrated models was consistently reasonable for any dataset size between $N_{\text{test}} = 3000$ and $N_{\text{test}} = 60000$, while there was a breakdown point between $N_{\text{test}} = 600$ and $N_{\text{test}} = 3000$ where TCE(P) was no longer able to do so for the well-calibrated model. This is likely because the dataset size $N_{\text{test}} = 600$ was not big enough to estimate the empirical probabilities $\{\widehat{P}_b\}_{b=1}^{B}$ accurately. This result may be improved by using different settings of the minimnum and maximum binsize constaints.
- Table 13: The performance of TCE(P) on both the well-calibrated and miscalibrated models was reasonable for any prevalence. While there was a fluctuation in values of TCE(P) for different values of prevalence, TCE(P) overall produced better values than TCE(Q and TCE(V).

Table 6: Sensitivity to the minimum binsize $N_{\text{min}} = 1, 6, 30, 300, 600, 3000$ in PAVA-BC. For comparison purpose, the number of bins B of quantile-binning and equispaced-binning was varied as $B = 1000, 500, 100, 50, 10, 5, 1$ along with N_{min}. Note that TCE(V) is a constant across all the row because PAVA does not involve any binsize constraint.

Test Prevalence	**Min Binsize**	**TCE(P)**	**TCE(Q)**	**TCE(V)**	**ECE**	**ACE**	**MCE**	**MCE(Q)**
50% (Calibrated)	1	3.4500	5.1000	3.4500	0.1143	0.1651	0.8767	0.6392
	6	3.3833	4.2000	3.4500	0.0839	0.1142	0.8767	0.5016
	30	2.3500	4.3000	3.4500	0.0382	0.0457	0.8767	0.1705
	60	2.6333	3.5667	3.4500	0.0271	0.0370	0.2533	0.1189
	300	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	600	13.5667	38.7500	3.4500	0.0116	0.0086	0.1020	0.0236
	3000	92.2000	92.2000	3.4500	0.0021	0.0021	0.0021	0.0021
40% (Miscalibrated)	1	88.0667	6.6667	88.0667	0.1417	0.1847	0.8767	0.6111
	6	88.0667	8.7000	88.0667	0.1179	0.1389	0.8767	0.4811
	30	88.3333	32.2833	88.0667	0.0993	0.0992	0.8767	0.2264
	60	87.8667	56.7667	88.0667	0.0971	0.0964	0.2426	0.1827
	300	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	600	96.6000	96.7833	88.0667	0.0963	0.0951	0.1099	0.1092
	3000	93.9500	93.9500	88.0667	0.0951	0.0951	0.0951	0.0951

Table 7: Sensitivity to the maximum binsize $N_{\max} = 6, 30, 300, 600, 3000, 6000$ in PAVA-BC. For comparison purpose, the number of bins B of quantile-binning and equispaced-binning was varied as $B = 1000, 500, 100, 50, 10, 5, 1$ along with $N_{\max}$. Note that TCE(V) is a constant across all the row because PAVA does not involve any binsize constraint.

Test Prevalence	Max Binsize	TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
50% (Calibrated)	6	5.8500	5.1000	3.4500	0.1143	0.1651	0.8767	0.6392
	30	3.0000	4.2000	3.4500	0.0839	0.1142	0.8767	0.5016
	60	2.3667	4.3000	3.4500	0.0382	0.0457	0.8767	0.1705
	300	3.7667	3.5667	3.4500	0.0271	0.0370	0.2533	0.1189
	600	3.3833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	3000	3.4500	38.7500	3.4500	0.0116	0.0086	0.1020	0.0236
	6000	3.4500	92.2000	3.4500	0.0021	0.0021	0.0021	0.0021
40% (Miscalibrated)	6	5.5000	6.6667	88.0667	0.1417	0.1847	0.8767	0.6111
	30	9.1000	8.7000	88.0667	0.1179	0.1389	0.8767	0.4811
	60	14.3833	32.2833	88.0667	0.0993	0.0992	0.8767	0.2264
	300	79.6667	56.7667	88.0667	0.0971	0.0964	0.2426	0.1827
	600	85.6500	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	3000	88.0667	96.7833	88.0667	0.0963	0.0951	0.1099	0.1092
	6000	88.0667	93.9500	88.0667	0.0951	0.0951	0.0951	0.0951

Table 8: Sensitivity to the pairs $(N_{\max}, N_{\min})$ in PAVA-BC selected so that the number of bins produced falls into ranges $[250, 1000], [50, 200], [25, 100], [10, 20], [3, 10]$. For comparison purpose, the number of bins B of quantile-binning and equispaced-binning was varied as $B = 1000, 500, 100, 50, 10, 5, 1$ along with $(N_{\max}, N_{\min})$. Note that TCE(V) is a constant across all the row because PAVA does not involve any binsize constraint.

Test Prevalence	Binsize Range	TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
50% (Calibrated)	[250, 1000]	3.8000	4.2000	3.4500	0.0839	0.1142	0.8767	0.5016
	[50, 200]	1.8333	4.3000	3.4500	0.0382	0.0457	0.8767	0.1705
	[25, 100]	0.2833	3.5667	3.4500	0.0271	0.0370	0.2533	0.1189
	[5, 20]	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	[3, 10]	13.5667	38.7500	3.4500	0.0116	0.0086	0.1020	0.0236
40% (Miscalibrated)	[250, 1000]	7.7333	8.7000	88.0667	0.1179	0.1389	0.8767	0.4811
	[50, 200]	45.7667	32.2833	88.0667	0.0993	0.0992	0.8767	0.2264
	[25, 100]	66.1833	56.7667	88.0667	0.0971	0.0964	0.2426	0.1827
	[10, 20]	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	[3, 10]	96.6000	96.7833	88.0667	0.0963	0.0951	0.1099	0.1092

Table 9: Sensitivity to a small purtabation to model predictions by a logit-normal noise with scale $\sigma = 0.01, 0.05, 0.10, 0.50, 1.00$. The maximum and minimum binsize of PAVA-BC were set to 1200 and 300. The number of bins of quantile-binning and equispaced-binning was set 10.

Test Prevalence	Noise Level	TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
50% (Calibrated)	0.00	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	0.01	8.7167	9.6167	4.8000	0.0113	0.0125	0.0923	0.0527
	0.05	12.8833	11.9000	7.7667	0.0136	0.0156	0.1198	0.0589
	0.10	8.3500	13.0500	3.5500	0.0109	0.0164	0.1143	0.0587
	0.50	61.9500	65.0500	56.1000	0.0615	0.0618	0.3601	0.1498
	1.00	86.1833	84.1000	88.3833	0.1470	0.1478	0.3364	0.2621
40% (Miscalibrated)	0.00	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	0.01	96.4000	96.4000	89.6167	0.0962	0.0951	0.1511	0.1332
	0.05	94.7667	95.5333	89.1667	0.0962	0.0951	0.1496	0.1420
	0.10	93.8500	95.9667	86.5833	0.0967	0.0951	0.1852	0.1412
	0.50	86.6667	83.9000	81.2667	0.1071	0.1055	0.2513	0.2203
	1.00	90.3167	88.8500	91.2167	0.1713	0.1698	0.4577	0.3648

Table 10: Sensitivity to a choice of significance level $\alpha = 0.001, 0.005, 0.01, 0.05, 0.1, 0.5$. The maximum and minimum binsize of PAVA-BC were set to 1200 and 300. The number of bins of quantile-binning and equispaced-binning was set 10.

Test Prevalence	Significant Level	TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
50% (Calibrated)	0.001	1.4500	4.6833	0.1833	0.0138	0.0150	0.1020	0.0528
	0.005	2.4667	5.5667	1.1333	0.0138	0.0150	0.1020	0.0528
	0.010	3.0500	6.2000	1.6500	0.0138	0.0150	0.1020	0.0528
	0.050	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	0.100	12.8500	15.2000	6.8667	0.0138	0.0150	0.1020	0.0528
	0.500	53.1000	55.3333	46.5667	0.0138	0.0150	0.1020	0.0528
40% (Miscalibrated)	0.001	77.8000	83.3833	76.1000	0.0963	0.0951	0.1466	0.1314
	0.005	86.3000	92.8000	80.0833	0.0963	0.0951	0.1466	0.1314
	0.010	90.1833	95.2167	83.1167	0.0963	0.0951	0.1466	0.1314
	0.050	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	0.100	97.2167	96.9167	90.1667	0.0963	0.0951	0.1466	0.1314
	0.500	99.3000	98.7167	97.7500	0.0963	0.0951	0.1466	0.1314

Table 11: Comparison of TCE based on the Binomial test and the t-test. TCE(Q)-B denotes TCE(Q) based on the Binomial test and TCE(Q)-T denotes TCE(Q) based on the t-test; the same applies for the other columns. The maximum and minimum binsize of PAVA-BC and the number of bins of quantile-binning and equispaced-binning were varied as in Table 8.

Test Prevalence	Binsize Range	TCE(P)-B	TCE(P)-T	TCE(Q)-B	TCE(Q)-T	TCE(V)-B	TCE(V)-T
50% (Calibrated)	[250, 1000]	3.8000	33.6667	4.2000	31.9167	3.4500	34.2167
	[50, 200]	1.8333	36.0000	4.3000	31.4333	3.4500	34.2167
	[25, 100]	0.2833	31.3667	3.5667	40.4333	3.4500	34.2167
	[5, 20]	7.2833	37.8000	10.8833	41.8500	3.4500	34.2167
	[3, 10]	13.5667	46.5000	38.7500	68.8167	3.4500	34.2167
40% (Miscalibrated)	[250, 1000]	7.7333	50.2833	8.7000	45.1833	88.0667	97.7333
	[50, 200]	45.7667	73.2667	32.2833	71.1667	88.0667	97.7333
	[25, 100]	66.1833	96.5333	56.7667	85.3833	88.0667	97.7333
	[5, 20]	96.1000	99.2667	96.4667	98.4833	88.0667	97.7333
	[3, 10]	96.6000	98.6333	96.7833	98.4833	88.0667	97.7333

Table 12: Comparison of TCE by different total sizes $N_{\text{test}} = 30, 60, 300, 600, 3000, 6000, 30000, 60000$ of test dataset. The training prevalence was $P_{\text{training}}(y) = 0.5$ for all datasets. The maximum and minimum binsize of PAVA-BC were set by $N_{\text{max}} = N_{\text{test}}/20$ and $N_{\text{min}} = N_{\text{test}}/5$. The number of bins of quantile-binning and equispaced-binning was set 10.

Test Prevalence	Data Size	TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
50% (Calibrated)	30	0.0000	0.0000	0.0000	0.2293	0.2631	0.4164	0.5660
	60	0.0000	3.3333	0.0000	0.0923	0.2158	0.7148	0.4208
	300	5.3333	11.0000	6.3333	0.0774	0.0867	0.1971	0.2057
	600	1.0000	4.5000	1.6667	0.0368	0.0445	0.3404	0.1270
	3000	8.0667	4.6333	4.7667	0.0190	0.0182	0.1209	0.0304
	6000	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	30000	16.1633	31.7167	0.7833	0.0036	0.0061	0.9045	0.0164
	60000	19.1483	45.7600	4.4417	0.0035	0.0043	0.0949	0.0100
40% (Miscalibrated)	30	13.3333	6.6667	36.6667	0.3164	0.3377	0.6569	0.6338
	60	0.0000	3.3333	0.0000	0.1072	0.1611	0.7148	0.4208
	300	27.3333	37.3333	48.3333	0.1240	0.1368	0.1971	0.2665
	600	14.1667	8.0000	26.5000	0.0694	0.0685	0.5824	0.1350
	3000	92.2333	91.7667	76.7667	0.0964	0.0958	0.1495	0.1358
	6000	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	30000	99.4700	99.2300	97.4433	0.0907	0.0906	0.9045	0.1064
	60000	99.7783	99.6600	98.9000	0.0923	0.0923	0.0972	0.1065

Table 13: Comparison of TCE by different prevalences P of training and test dataset. The training data size was 14000 and the test data size was 6000. The maximum and minimum binsize of PAVA-BC were set to 1200 and 300. The number of bins of quantile-binning and equispaced-binning was set 10.

Train - Test Prevalence		TCE(P)	TCE(Q)	TCE(V)	ECE	ACE	MCE	MCE(Q)
Calibrated	50% - 50%	7.2833	10.8833	3.4500	0.0138	0.0150	0.1020	0.0528
	40% - 40%	7.5500	16.2167	8.5667	0.0137	0.0191	0.1632	0.0365
	30% - 30%	8.1667	12.8833	2.8167	0.0125	0.0134	0.1042	0.0313
	20% - 20%	15.9500	22.2167	15.9167	0.0173	0.0153	0.6238	0.0370
	10% - 10%	11.9833	16.7833	15.2333	0.0096	0.0114	0.4361	0.0218
	8% - 8%	15.7000	18.5167	23.1500	0.0087	0.0107	0.0700	0.0234
	6% - 6%	11.5333	17.5500	13.9833	0.0035	0.0109	0.3064	0.0195
	4% - 4%	18.5000	15.6667	20.5833	0.0046	0.0074	0.2240	0.0177
	2% - 2%	13.1167	11.5500	20.7667	0.0052	0.0059	0.0052	0.0131
Miscalibrated	50% - 40%	96.1000	96.4667	88.0667	0.0963	0.0951	0.1466	0.1314
	40% - 30%	96.5667	96.1833	82.7500	0.0872	0.0869	0.1485	0.1262
	30% - 20%	94.9500	94.6667	88.5833	0.0846	0.0846	0.2146	0.1247
	20% - 10%	95.8833	95.5833	96.4333	0.0868	0.0868	0.6238	0.1500
	10% - 8%	32.3500	26.7000	42.5667	0.0151	0.0173	0.4361	0.0502
	8% - 6%	42.3167	38.7833	45.8333	0.0164	0.0186	0.3259	0.0477
	6% - 4%	47.0833	39.9500	65.9500	0.0167	0.0188	0.3064	0.0440
	4% - 2%	56.5833	42.4333	72.4500	0.0142	0.0142	0.2240	0.0337
	2% - 0%	99.9167	96.9000	100.0000	0.0181	0.0181	0.0181	0.0382

C.2 RESULTS ON OTHER UCI DATASETS

Algorithms in Section 4.2 are all trained with the default hyperparameters in the scikit-learn package, except that the maximum depth in the random forest is set to 10 and the number of hidden layers in the multiple perceptron is set to 1 with 1000 units. For better comparison, we add TCE based on quantile bins, denoted TCE(Q) in each table, to five metrics presented in the main text. The following Table 14 compares six different calibration error metrics computed for eight UCI datasets that were not presented in the main text: coil_2000, isolet, letter_img, mammography, optimal_degits, pen_degits, satimage, spambase [Dua and Graff, 2017, van der Putten and van Someren, 2000-2009, Elter et al., 2007]. The prevalence of the spambase dataset is well-balanced and that of the rest is imbalanced. The following Figures 5 and 6 shows the visual representations of TCE, ECE, and ACE—the test-based reliability diagram and the standard reliability diagram—each for the logistic regression and the gradient boosting algorithm. We selected four datasets, abalone, coil_2000, isolet, and webpage, to produce the visual representations in Figures 5 and 6.

C.3 RELIABILITY DIAGRAMS OF RESULTS ON IMAGENET1000

The following Figure 7 shows the viaual representations of TCE, ECE, and ACE—the test-based reliability diagram and the standard reliability diagram—for four different deep learning models presented in the main text, where we omit the model ResNet50 whose result sufficiently resembles that of ResNet18.

References

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

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

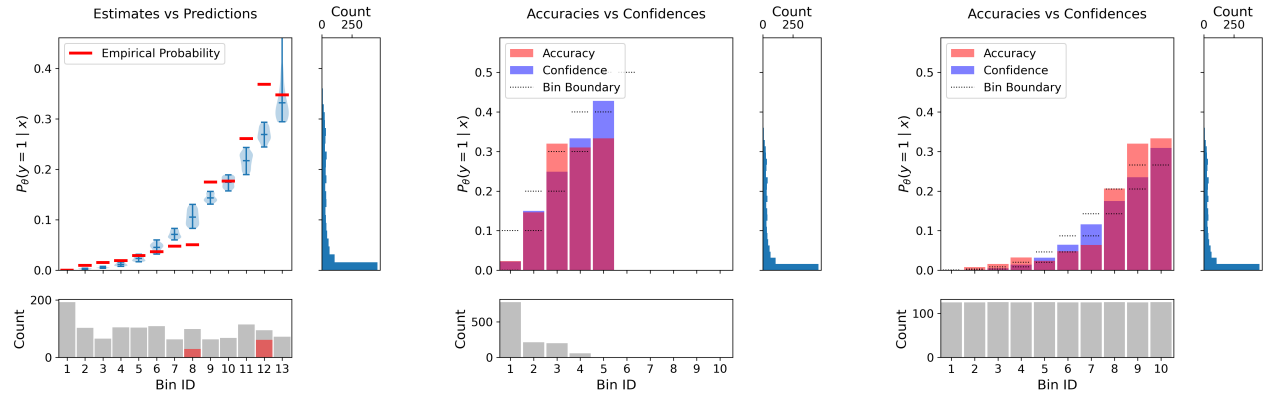
M Elter, R Schulz-Wendtland, and T Wittenberg. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Medical Physics*, 34(11), 2007.

Alexander Henzi, Alexandre Mösching, and Lutz Dümbgen. Accelerating the Pool-Adjacent-Violators Algorithm for Isotonic Distributional Regression. *Methodology and Computing in Applied Probability*, 24(4):2633–2645, 2022.

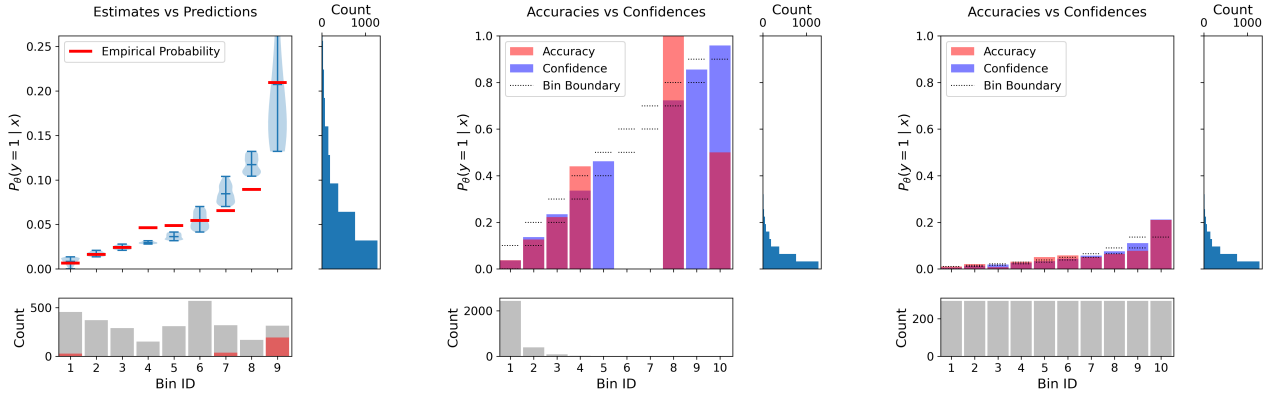
Peter van der Putten and Maarten van Someren. Coil challenge 2000: The insurance company case. Technical report, Sentient Machine Research, Amsterdam and Leiden Institute of Advanced Computer Science, 2000-2009.

Table 14: Comparison of six calibration error metrics for five algorithms trained on eight UCI datasets. The same setting of TCE presented in Section 4 is used. TCE(Q) and MCE(Q) denotes TCE and MCE each based on quantile bins where the number of bins is set to 10.

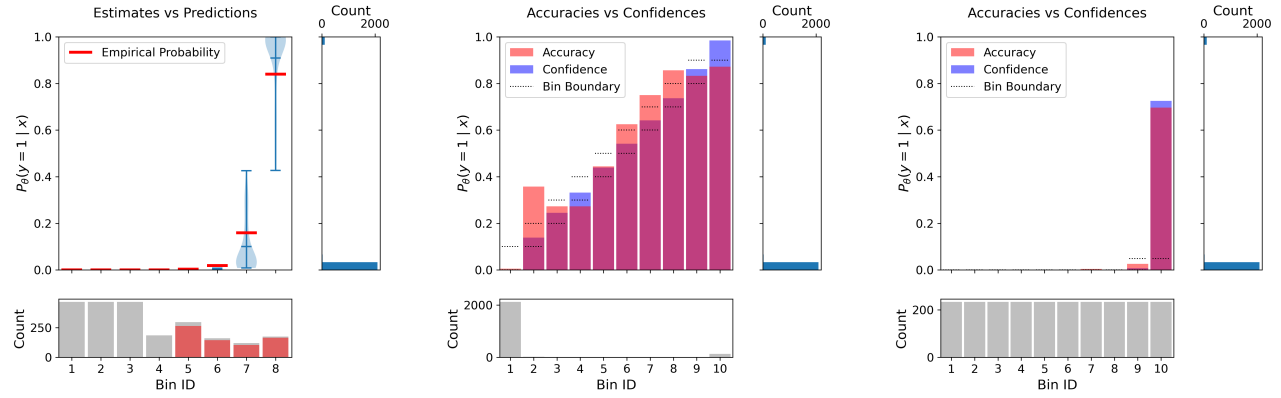
Data	Algorithm	TCE	TCE(Q)	ECE	ACE	MCE	MCE(Q)
coil_2000	LR	8.6189	11.7408	0.0047	0.0111	0.8558	0.0326
	SVM	17.0003	28.9447	0.0071	0.0216	0.4860	0.0381
	RF	22.2260	6.1758	0.0027	0.0125	0.2465	0.0439
	GB	20.8687	12.6569	0.0052	0.0098	0.3738	0.0259
	MLP	98.7445	98.7784	0.0652	0.0578	0.7900	0.1649
isolet	LR	28.8462	27.3932	0.0131	0.0051	0.2183	0.0286
	SVM	11.5812	13.2479	0.0064	0.0028	0.1969	0.0194
	RF	66.4530	52.5214	0.0524	0.0507	0.3635	0.2137
	GB	25.2991	16.6667	0.0198	0.0174	0.4463	0.1123
	MLP	9.8291	17.5641	0.0049	0.0031	0.4173	0.0232
letter_img	LR	10.5167	12.0500	0.0025	0.0008	0.1617	0.0042
	SVM	11.8667	14.8167	0.0019	0.0017	0.6257	0.0146
	RF	26.5000	20.2500	0.0097	0.0033	0.5179	0.0131
	GB	25.9500	18.7333	0.0067	0.0029	0.3653	0.0109
	MLP	19.9833	9.9833	0.0010	0.0001	0.4550	0.0007
mammography	LR	25.0671	26.7660	0.0027	0.0065	0.3594	0.0208
	SVM	20.2683	20.1490	0.0067	0.0088	0.6741	0.0353
	RF	19.4039	9.2996	0.0047	0.0016	0.4465	0.0043
	GB	14.5156	15.4098	0.0061	0.0034	0.5355	0.0124
	MLP	20.5663	26.9747	0.0042	0.0027	0.4351	0.0113
optical_digits	LR	11.6251	27.1649	0.0098	0.0037	0.2251	0.0135
	SVM	4.8043	10.6762	0.0042	0.0028	0.6608	0.0157
	RF	49.6441	38.3155	0.0451	0.0433	0.5432	0.2271
	GB	13.0486	11.2693	0.0181	0.0168	0.5639	0.1122
	MLP	4.6856	12.1590	0.0037	0.0034	0.5992	0.0306
pen_digits	LR	20.4063	23.1049	0.0121	0.0060	0.1652	0.0252
	SVM	9.7635	10.2790	0.0017	0.0010	0.4735	0.0068
	RF	29.6240	22.8623	0.0152	0.0132	0.4535	0.0592
	GB	9.9151	13.0988	0.0077	0.0058	0.6543	0.0303
	MLP	9.4603	10.0061	0.0014	0.0004	0.6457	0.0037
satimage	LR	23.6665	23.0968	0.0215	0.0223	0.7312	0.0767
	SVM	10.2020	21.8540	0.0229	0.0163	0.1666	0.0870
	RF	29.1041	20.1450	0.0265	0.0214	0.2084	0.1328
	GB	23.2004	19.8861	0.0154	0.0235	0.2101	0.0902
	MLP	58.0528	58.4671	0.0352	0.0328	0.5049	0.1384
spambase	LR	33.6713	56.1188	0.0256	0.0267	0.1539	0.0895
	SVM	12.8168	34.5402	0.0177	0.0227	0.2207	0.0465
	RF	66.0391	49.4569	0.0635	0.0601	0.2056	0.1616
	GB	20.2028	20.4200	0.0295	0.0277	0.1409	0.0891
	MLP	60.9703	67.1253	0.0413	0.0397	0.2931	0.1076



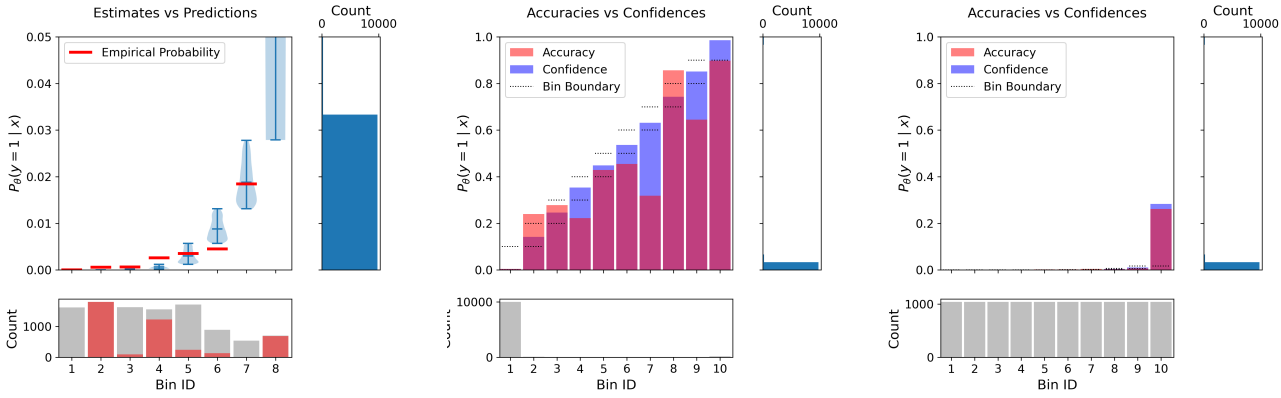
(a) abalone



(b) coil_2000

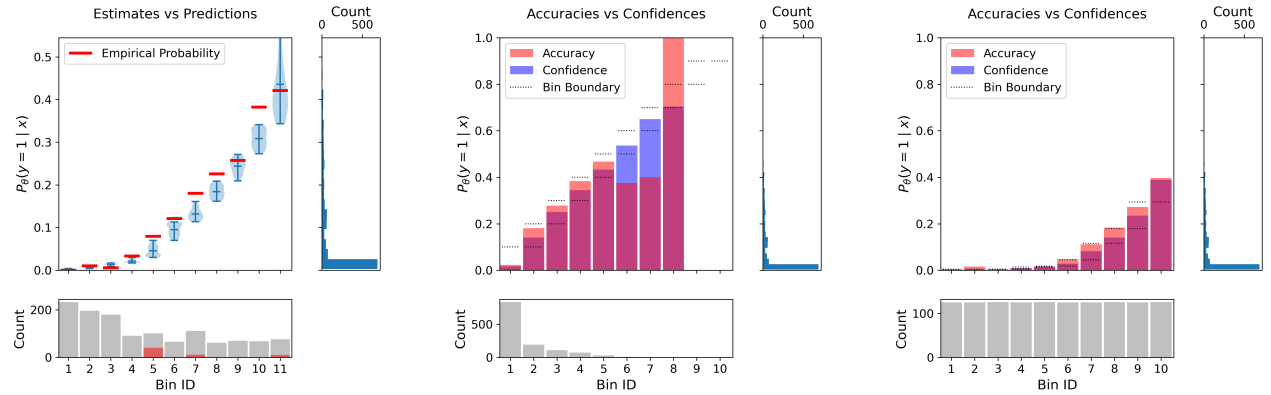


(c) isolet

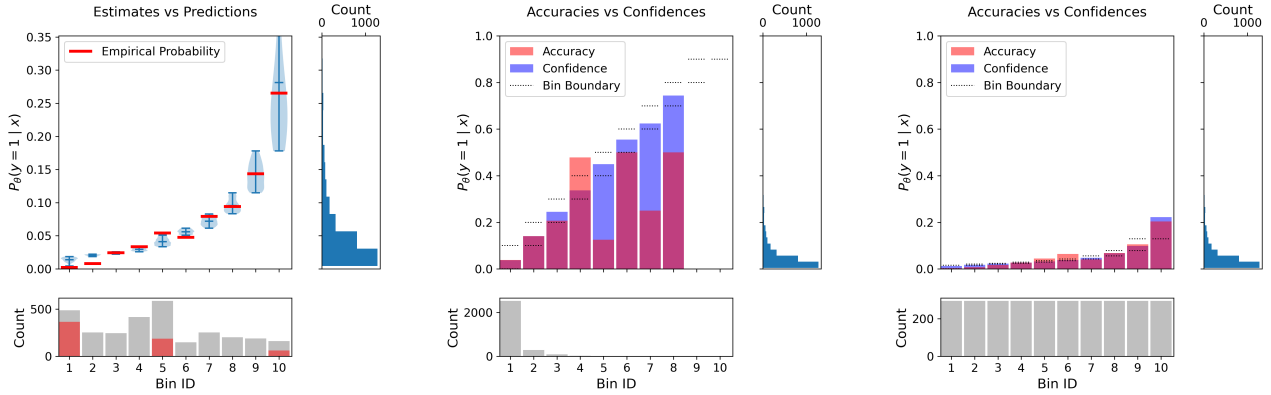


(d) webpage

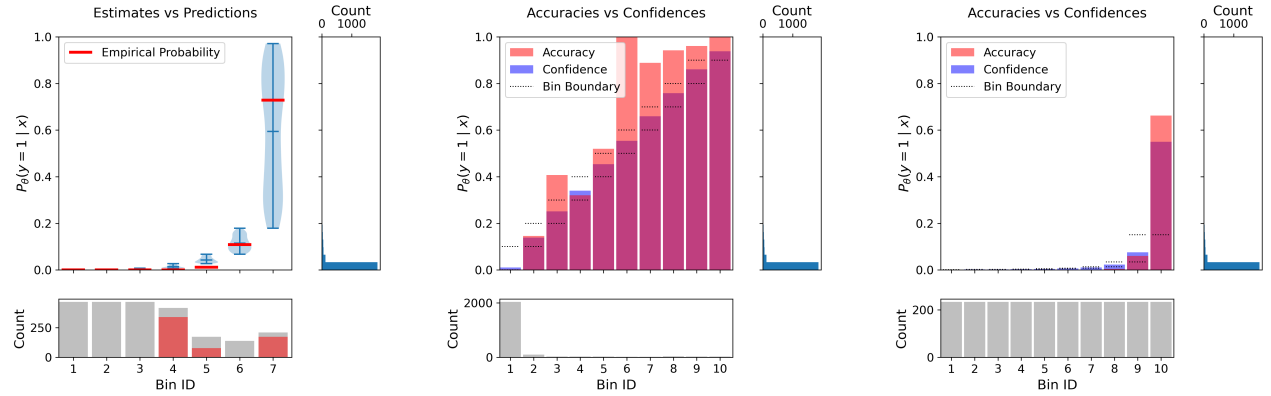
Figure 5: Comparison of visual representations of TCE, ECE and ACE for the logistic regression algorithm. (Left) The test-based reliability diagram of TCE. (Middle) The reliability diagram of ECE. (Right) The reliability diagram of ACE. Each row corresponds to a result on the dataset: (a) abalone, (b) coil_2000, (c) isolet, and (d) webpage.



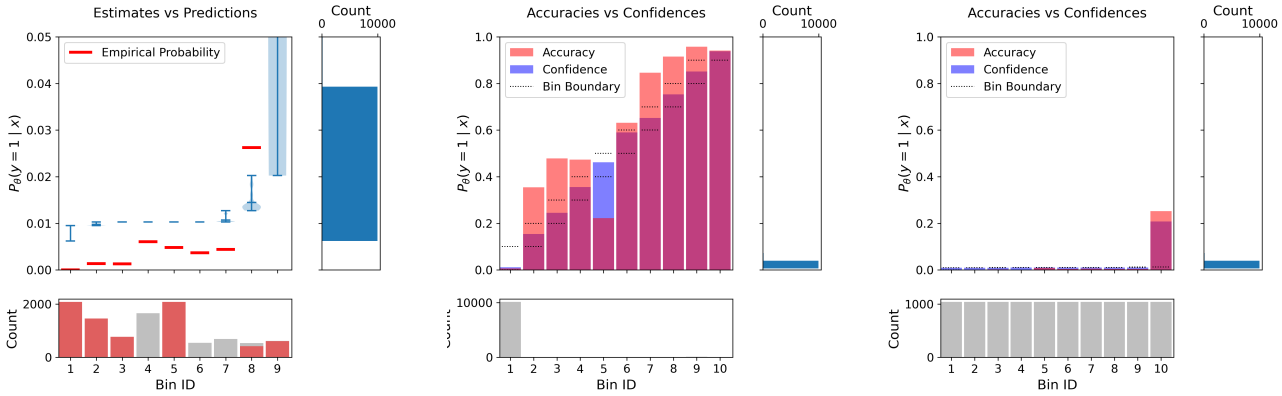
(a) abalone



(b) coil_2000

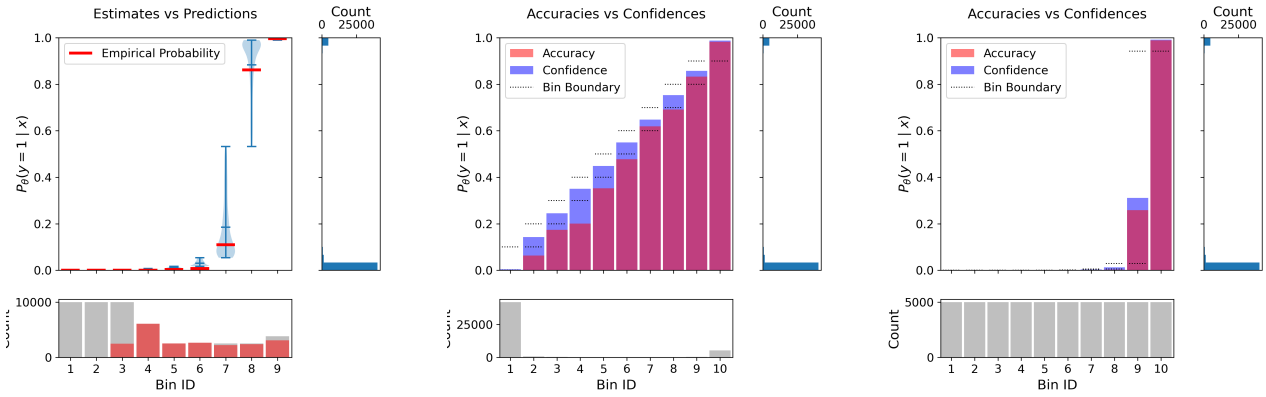


(c) isolet

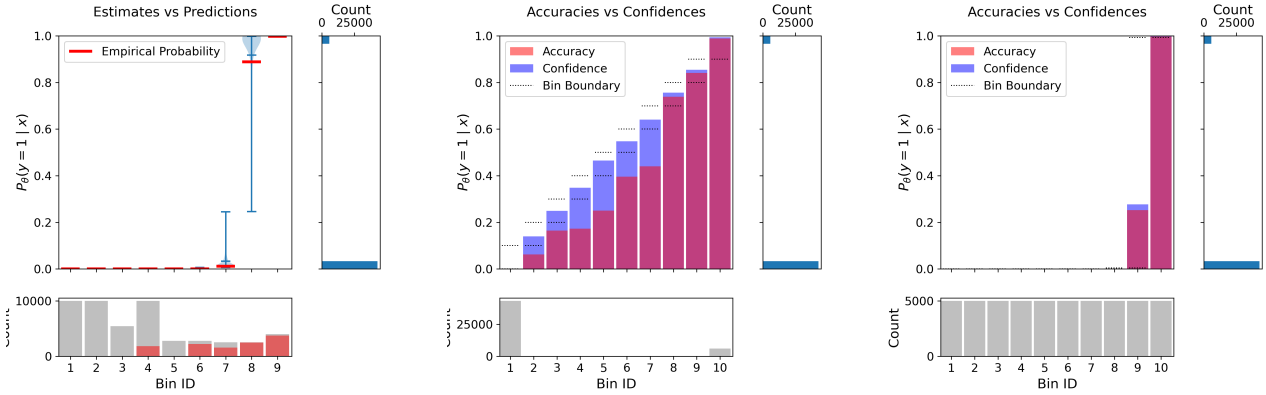


(d) isolet

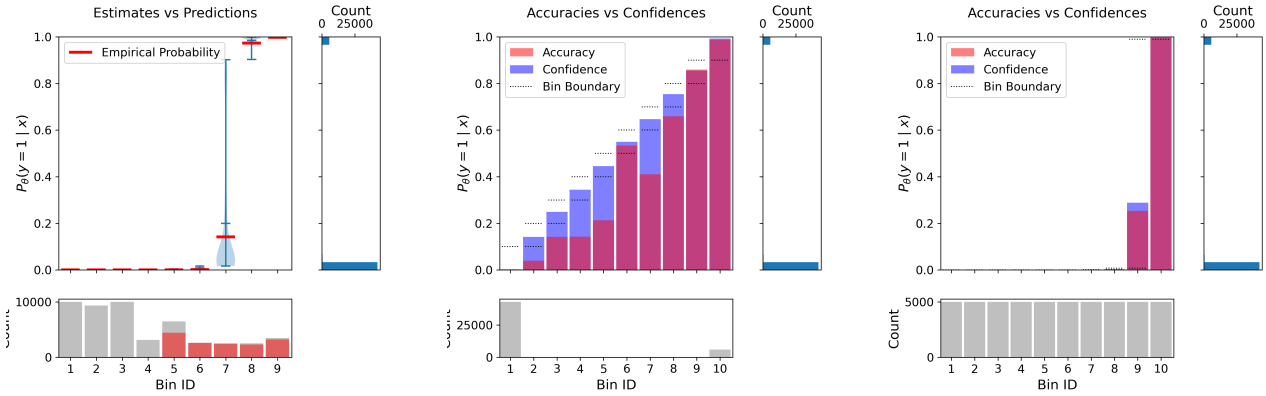
Figure 6: Comparison of visual representations of TCE, ECE and ACE for the logistic regression algorithm. (Left) The test-based reliability diagram of TCE. (Middle) The reliability diagram of ECE. (Right) The reliability diagram of ACE. Each row corresponds to a result on the dataset: (a) abalone, (b) coil_2000, (c) isolet, and (d) webpage.



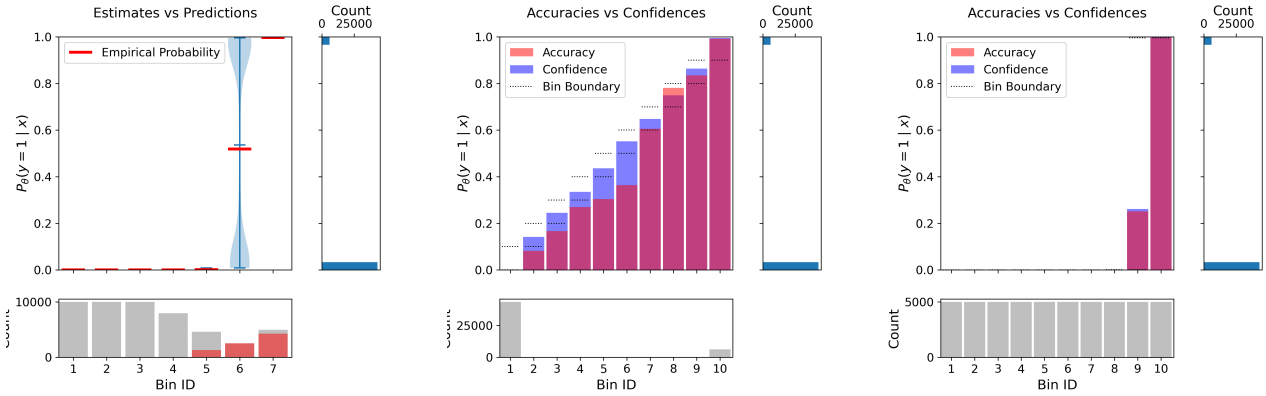
(a) AlexNet



(b) VGG19



(c) ResNet 18



(d) ResNet 152

Figure 7: Comparison of visual representations of TCE, ECE, and ACE on the ImageNet 1000 dataset. (Left) The test-based reliability diagram of TCE, (Middle) The reliability diagram of ECE (Right) The reliability diagram of ACE. Each row corresponds to a result for the model: (a) AlexNet, (b) VGG19, (c) ResNet 18, and (d) ResNet 152.