Algorithm 3 with PAVA-BC replaced by PAVA generates the optimal bins under the error function \( D \). Accordingly, with the ordered labels \( D \), the optimal bins defined by Definition 3 can be exactly computed under the error function \( D \). 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 \( \{ \hat{P}_b \}_{b=1}^B \). 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 \( \Delta_b \). The optimal bins result in minimisation of a weighted average of the variance of each \( \Delta_b \) 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., \( \{ \Delta_b \}_{b=1}^B = \{ [0, 1] \} \).

**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 optimisation 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 \( D_b^y \) is defined by \( D_b^y := \{ y_i \in D^y \mid P_b(x_i) \in \Delta_b \} \). The input of Algorithm 3 is a set of labels \( D^y = \{ y_i \}_{i=1}^N \) ordered by in ascending order of \( \{ P_b(x_i) \}_{i=1}^N \). This means that each label subset \( D_b^y \) 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 \( D_b^y \) can expressed by

\[
D_b^y = \{ y_i \in D^y \mid P_b(x_i) \in \Delta_b \} = \{ y_i \in D^y \mid i \text{ s.t. } n_b \leq i < n_{b+1} \}.
\]

Accordingly, with the ordered labels \( D^y \), each empirical probability \( \hat{P}_b \) in \( D_b^y \) can be expressed by

\[
\hat{P}_b = \frac{1}{N_b} \sum_{y \in D_b^y} y = \frac{1}{n_{b+1} - n_b} \sum_{j=n_b}^{n_{b+1}-1} y_j.
\]

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

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

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

\[
\sum_{b=1}^B W_b \times D(D_b, \hat{P}_b) = \frac{1}{N} \sum_{b=1}^B \sum_{y \in D_b} (y - \hat{P}_b)^2 = \frac{1}{N} \sum_{b=1}^B \sum_{i=n_b}^{n_{b+1}-1} (y_i - \hat{P}_b)^2 = \frac{1}{N} \sum_{i=1}^N (y_i - g_i)^2.
\]
This formulation translates a problem of choosing bins \( \{ \Delta_b \}_{b=1}^B \) into a problem of finding a monotonically increasing sequence \( \{ y_i \}_{i=1}^N \) that is determined by the choice of indices \( \{ n^*_b \}_{b=1}^B \), so that eq. (11) is minimised. Therefore the optimisation 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 \( \{ y_i \}_{i=1}^N \) is restricted to a form of eq. (10).

Next, consider a standard monotonic regression problem under the square error \( \sum_{i=1}^N (y_i - \hat{y}_i)^2 \) for the ordered set \( \{ y_i \}_{i=1}^N \). PAVA finds a monotonically increasing sequence \( \{ \hat{y}_i \}_{i=1}^N \) that minimises the square error. The solution sequence \( \{ \hat{y}_i \}_{i=1}^N \) 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 \( \{ \hat{y}_i \}_{i=1}^N \) by PAVA is expressed as

\[
\hat{y}_i = \frac{1}{n^*_b - n^*_b+1} \sum_{j=n^*_b}^{n^*_b+1} y_j \quad \text{for each } i \quad \text{s.t. } n^*_b \leq i < n^*_b+1
\]

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

An output of most implementations of PAVA is the solution sequence \( \{ \hat{y}_i \}_{i=1}^N \) 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 \( \{ \hat{y}_i \}_{i=1}^N \) of PAVA by simply finding all indices \( i \) s.t. \( \hat{y}_i \neq \hat{y}_{i+1} \). 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 \( D^y = \{ y_i \}_{i=1}^N \) are ordered in ascending order of \( \{ P_\theta(x_i) \}_{i=1}^N \). If we construct each bin \( \Delta_b \) by

\[
\Delta_b := \left[ \frac{P_\theta(x_{n^*_b-1}) + P_\theta(x_{n^*_b})}{2}, \frac{P_\theta(x_{n^*_b+1}) + P_\theta(x_{n^*_b+1-1})}{2} \right],
\]

it is sufficient to generate each label subset \( D^y_b \) that corresponds to

\[
D^y_b = \{ y_i \in D^y \mid P_\theta(x_i) \in \Delta_b \} = \{ y_i \in D^y \mid i \text{ s.t. } n^*_b \leq i < n^*_b+1 \}.
\]

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 \( \hat{y}_1 \leq \cdots \leq \hat{y}_N \) that minimises some error \( D(\{ \hat{y}_i \}_{i=1}^N, \{ y_i \}_{i=1}^N) \) 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 \( \sum_{i=1}^N (\hat{y}_i - y_i)^2 \). For some partition \( A \) of indices \( I = \{1, \ldots, N\} \) whose element \( A \in A \) is a set of consecutive indices in \( I \), PAVA produces a solution sequence s.t. each element \( \hat{y}_i \) is given by \( \hat{y}_i = (1/|A|) \sum_{i \in A} y_i \) for \( A \) in which \( i \in A \). We refer to each element \( A \) in the partition \( A \) of indices \( I \) as block. PAVA-BC produces a solution sequence that approximates the solution sequence by PAVA under the constraints of the minimum and maximum size of each block. For some partition \( A' \) of indices \( I \), each element \( \hat{y}_i \) of the solution sequence is given by \( \hat{y}_i = (1/|A'|) \sum_{i \in A'} y_i \) 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 \( A' \in A' \) 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 \( \{ P_\theta \}_{b=1}^B \). 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 summarised 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 \( \{ P_\theta \}_{b=1}^B \). Figure 4 was computed using a random forest model trained on the coil2000 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 5 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 represents the size of each bin. The red line in the right panel represents 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 represents the size of each bin. The red line in the right panel represents 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 demonstrate sensitivity of TCE to its hyperparameters, an impact of different dataset size and prevalence, and sensitivity to a small perturbation 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 calibrated 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 $N_{\text{min}}$ in PAVA-BC from $N_{\text{min}} = 1$ to $N_{\text{min}} = 3000$;
- Sensitivity to the maximum bin size $N_{\text{max}}$ 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 perturbation of predictions by a logit-normal noise with scale $\sigma$ from $\sigma = 0.0$ to $\sigma = 1.0$;
- Sensitivity to a choice of significance level $\alpha$ 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 constant 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 constraint $N_{\text{min}} = 3000$ for the total datasize 6000 was 2 at maximum, which was too small to estimate the empirical probabilities $\{\hat{P}_b\}_{b=1}^B$ accurately.

- Table 7 The performance of TCE(P) to evidence the miscalibrated model was consistently reasonable for any maximum binsize constant 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 constraint $N_{\text{max}} = 60$ for the total datasize 6000 was 10 at minimum, which is too large to estimate the empirical probabilities $\{\hat{P}_b\}_{b=1}^B$ 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 heuristic to use such $(N_{\text{min}}, N_{\text{max}})$ for other experiments.

- Table 9 At each model prediction $P_0(x)$, we sample a new prediction from a logit-normal distribution centred at $P_0(x)$ with scale $\sigma$ 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 approximation.

Table 12: The performance of TCE(P) to evidence both the well-calibrated and miscalibrated models was consistently reasonable for any dataset size between $N_{test} = 3000$ and $N_{test} = 60000$, while there was a breakdown point between $N_{test} = 600$ and $N_{test} = 3000$ where TCE(P) was no longer able to do so for the well-calibrated model. This is likely because the dataset size $N_{test} = 600$ was not big enough to estimate the empirical probabilities $\{\hat{P}_b\}_{b=1}^B$ accurately. This result may be improved by using different settings of the minimum and maximum binsize constraints.

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 $N_{min} = 1, 6, 30, 300, 600, 3000$ 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.

<table>
<thead>
<tr>
<th>Test Prevalence</th>
<th>Min Binsize</th>
<th>TCE(P)</th>
<th>TCE(Q)</th>
<th>TCE(V)</th>
<th>ECE</th>
<th>ACE</th>
<th>MCE</th>
<th>MCE(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% (Calibrated)</td>
<td>1</td>
<td>3.4500</td>
<td>5.1000</td>
<td>3.4500</td>
<td>0.1143</td>
<td>0.1651</td>
<td>0.8767</td>
<td>0.6392</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3.3833</td>
<td>4.2000</td>
<td>3.4500</td>
<td>0.0839</td>
<td>0.1142</td>
<td>0.8767</td>
<td>0.5016</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>2.3500</td>
<td>4.3000</td>
<td>3.4500</td>
<td>0.0382</td>
<td>0.0457</td>
<td>0.8767</td>
<td>0.1705</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>2.6333</td>
<td>3.5667</td>
<td>3.4500</td>
<td>0.0271</td>
<td>0.0370</td>
<td>0.2533</td>
<td>0.1189</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>7.2833</td>
<td>10.8833</td>
<td>3.4500</td>
<td>0.0138</td>
<td>0.0150</td>
<td>0.1020</td>
<td>0.0528</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>13.5667</td>
<td>38.7500</td>
<td>3.4500</td>
<td>0.0116</td>
<td>0.0086</td>
<td>0.1020</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>92.2000</td>
<td>92.2000</td>
<td>3.4500</td>
<td>0.0021</td>
<td>0.0021</td>
<td>0.0021</td>
<td>0.0021</td>
</tr>
<tr>
<td>40% (Miscalibrated)</td>
<td>1</td>
<td>88.0667</td>
<td>6.6667</td>
<td>88.0667</td>
<td>0.1417</td>
<td>0.1847</td>
<td>0.8767</td>
<td>0.6111</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>88.0667</td>
<td>8.7000</td>
<td>88.0667</td>
<td>0.1179</td>
<td>0.1389</td>
<td>0.8767</td>
<td>0.4811</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>88.3333</td>
<td>32.2833</td>
<td>88.0667</td>
<td>0.0993</td>
<td>0.0992</td>
<td>0.8767</td>
<td>0.2264</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>87.8667</td>
<td>56.7667</td>
<td>88.0667</td>
<td>0.0971</td>
<td>0.0964</td>
<td>0.2426</td>
<td>0.1827</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>96.1000</td>
<td>96.4667</td>
<td>88.0667</td>
<td>0.0963</td>
<td>0.0951</td>
<td>0.1466</td>
<td>0.1314</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>96.6000</td>
<td>96.7833</td>
<td>88.0667</td>
<td>0.0963</td>
<td>0.0951</td>
<td>0.1099</td>
<td>0.1092</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>93.9500</td>
<td>93.9500</td>
<td>88.0667</td>
<td>0.0951</td>
<td>0.0951</td>
<td>0.0951</td>
<td>0.0951</td>
</tr>
</tbody>
</table>
Table 7: Sensitivity to the maximum binsize $N_{\text{max}} = 6, 30, 300, 600, 3000, 6000$ 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 $N_{\text{max}}$. Note that TCE(V) is a constant across all the row because PAVA does not involve any binsize constraint.

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

Table 8: Sensitivity to the pairs $(N_{\text{max}}, N_{\text{min}})$ 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 $(N_{\text{max}}, N_{\text{min}})$. Note that TCE(V) is a constant across all the row because PAVA does not involve any binsize constraint.

<table>
<thead>
<tr>
<th>Test Prevalence</th>
<th>Binsize Range</th>
<th>TCE(P)</th>
<th>TCE(Q)</th>
<th>TCE(V)</th>
<th>ECE</th>
<th>ACE</th>
<th>MCE</th>
<th>MCE(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[250, 1000]</td>
<td>3.8000</td>
<td>4.2000</td>
<td>3.4500</td>
<td>0.0839</td>
<td>0.1142</td>
<td>0.8767</td>
<td>0.5016</td>
</tr>
<tr>
<td>50% (Calibrated)</td>
<td>[50, 200]</td>
<td>1.8333</td>
<td>4.3000</td>
<td>3.4500</td>
<td>0.0382</td>
<td>0.0457</td>
<td>0.8767</td>
<td>0.1705</td>
</tr>
<tr>
<td></td>
<td>[25, 100]</td>
<td>0.2833</td>
<td>3.5667</td>
<td>3.4500</td>
<td>0.0271</td>
<td>0.0370</td>
<td>0.2533</td>
<td>0.1189</td>
</tr>
<tr>
<td></td>
<td>[5, 20]</td>
<td>7.2833</td>
<td>10.8833</td>
<td>3.4500</td>
<td>0.0138</td>
<td>0.0150</td>
<td>0.1020</td>
<td>0.0528</td>
</tr>
<tr>
<td></td>
<td>[3, 10]</td>
<td>13.5667</td>
<td>38.7500</td>
<td>3.4500</td>
<td>0.0116</td>
<td>0.0086</td>
<td>0.1020</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>[250, 1000]</td>
<td>7.7333</td>
<td>8.7000</td>
<td>88.0667</td>
<td>0.1179</td>
<td>0.1389</td>
<td>0.8767</td>
<td>0.4811</td>
</tr>
<tr>
<td></td>
<td>[50, 200]</td>
<td>45.7667</td>
<td>32.2833</td>
<td>88.0667</td>
<td>0.0993</td>
<td>0.0992</td>
<td>0.8767</td>
<td>0.2264</td>
</tr>
<tr>
<td></td>
<td>[25, 100]</td>
<td>66.1833</td>
<td>56.7667</td>
<td>88.0667</td>
<td>0.0971</td>
<td>0.0964</td>
<td>0.2426</td>
<td>0.1827</td>
</tr>
<tr>
<td></td>
<td>[10, 20]</td>
<td>96.1000</td>
<td>96.4667</td>
<td>88.0667</td>
<td>0.0963</td>
<td>0.0951</td>
<td>0.1466</td>
<td>0.1314</td>
</tr>
<tr>
<td></td>
<td>[3, 10]</td>
<td>96.6000</td>
<td>96.7833</td>
<td>88.0667</td>
<td>0.0951</td>
<td>0.0951</td>
<td>0.0951</td>
<td>0.0951</td>
</tr>
</tbody>
</table>
Table 9: Sensitivity to a small perturbation 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.

<table>
<thead>
<tr>
<th>Test Prevalence</th>
<th>Noise Level</th>
<th>TCE(P)</th>
<th>TCE(Q)</th>
<th>TCE(V)</th>
<th>ECE</th>
<th>ACE</th>
<th>MCE</th>
<th>MCE(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% (Calibrated)</td>
<td>0.00</td>
<td>7.2833</td>
<td>10.8833</td>
<td>3.4500</td>
<td>0.0138</td>
<td>0.0150</td>
<td>0.1020</td>
<td>0.0528</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>8.7167</td>
<td>9.6167</td>
<td>4.8000</td>
<td>0.0113</td>
<td>0.0125</td>
<td>0.0923</td>
<td>0.0527</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>12.8833</td>
<td>11.9000</td>
<td>7.7667</td>
<td>0.0136</td>
<td>0.0156</td>
<td>0.1198</td>
<td>0.0589</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>8.3500</td>
<td>13.0500</td>
<td>3.5500</td>
<td>0.0109</td>
<td>0.0164</td>
<td>0.1143</td>
<td>0.0587</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>61.9500</td>
<td>65.0500</td>
<td>56.1000</td>
<td>0.0615</td>
<td>0.0618</td>
<td>0.3601</td>
<td>0.1498</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>86.1833</td>
<td>84.1000</td>
<td>88.3833</td>
<td>0.1470</td>
<td>0.1478</td>
<td>0.3364</td>
<td>0.2621</td>
</tr>
</tbody>
</table>

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.

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

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.

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

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

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.

<table>
<thead>
<tr>
<th>Train - Test Prevalence</th>
<th>TCE(P)</th>
<th>TCE(Q)</th>
<th>TCE(V)</th>
<th>ECE</th>
<th>ACE</th>
<th>MCE</th>
<th>MCE(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% - 50%</td>
<td>7.2833</td>
<td>10.8833</td>
<td>3.4500</td>
<td>0.0138</td>
<td>0.0150</td>
<td>0.1020</td>
<td>0.0528</td>
</tr>
<tr>
<td>40% - 40%</td>
<td>7.5500</td>
<td>16.2167</td>
<td>8.5667</td>
<td>0.0137</td>
<td>0.0191</td>
<td>0.1632</td>
<td>0.0365</td>
</tr>
<tr>
<td>30% - 30%</td>
<td>8.1667</td>
<td>12.8833</td>
<td>2.8167</td>
<td>0.0125</td>
<td>0.0134</td>
<td>0.1042</td>
<td>0.0313</td>
</tr>
<tr>
<td>20% - 20%</td>
<td>15.9500</td>
<td>22.2167</td>
<td>15.9167</td>
<td>0.0173</td>
<td>0.0153</td>
<td>0.6238</td>
<td>0.0370</td>
</tr>
<tr>
<td>10% - 10%</td>
<td>11.9833</td>
<td>16.7833</td>
<td>15.2333</td>
<td>0.0096</td>
<td>0.0114</td>
<td>0.4361</td>
<td>0.0218</td>
</tr>
<tr>
<td>8% - 8%</td>
<td>15.7000</td>
<td>18.5167</td>
<td>23.1500</td>
<td>0.0087</td>
<td>0.0107</td>
<td>0.0700</td>
<td>0.0234</td>
</tr>
<tr>
<td>6% - 6%</td>
<td>15.5333</td>
<td>17.5500</td>
<td>13.9833</td>
<td>0.0035</td>
<td>0.0109</td>
<td>0.3064</td>
<td>0.0195</td>
</tr>
<tr>
<td>4% - 4%</td>
<td>18.5000</td>
<td>15.6667</td>
<td>20.5833</td>
<td>0.0046</td>
<td>0.0074</td>
<td>0.2240</td>
<td>0.0177</td>
</tr>
<tr>
<td>2% - 2%</td>
<td>13.1167</td>
<td>11.5500</td>
<td>20.7667</td>
<td>0.0052</td>
<td>0.0059</td>
<td>0.0052</td>
<td>0.0131</td>
</tr>
</tbody>
</table>

| 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_digits, pen_digits, 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—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 visual 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


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.


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.

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>TCE</th>
<th>TCE(Q)</th>
<th>ECE</th>
<th>ACE</th>
<th>MCE</th>
<th>MCE(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>coil_2000</td>
<td>LR</td>
<td>8.6189</td>
<td>11.7408</td>
<td>0.0047</td>
<td>0.0111</td>
<td>0.8558</td>
<td>0.0326</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>17.0003</td>
<td>28.9447</td>
<td>0.0071</td>
<td>0.0216</td>
<td>0.4860</td>
<td>0.0381</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>22.2260</td>
<td>6.1758</td>
<td>0.0027</td>
<td>0.0125</td>
<td>0.2465</td>
<td>0.0439</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>20.8687</td>
<td>12.6569</td>
<td>0.0052</td>
<td>0.0098</td>
<td>0.3738</td>
<td>0.0259</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>98.7445</td>
<td>98.7784</td>
<td>0.0652</td>
<td>0.0578</td>
<td>0.7900</td>
<td>0.1649</td>
</tr>
</tbody>
</table>

| 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 |
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.
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.
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.