Understanding Systematic Miscalibration in Machine Learning Classifiers

Markelle Kelly* Padhraic Smyth

Department of Computer Science, University of California, Irvine

Abstract

The deployment of machine learning classifiers in high-stakes domains requires well-calibrated confidence scores for model predictions. In this paper we show that standard calibration measurement approaches used in machine learning can obscure significant systematic miscalibration with respect to variables of interest. We demonstrate this phenomenon on multiple well-known datasets, and show that it can persist after the application of widely-used recalibration methods. To mitigate this issue, we propose strategies for detection, visualization, and quantification of systematic miscalibration. We also examine the limitations of score-based recalibration methods and explore potential modifications. Finally, we discuss the implications of these findings, emphasizing that an understanding of calibration beyond simple aggregate measures is crucial for endeavors such as fairness and model interpretability.

Index Terms: Human-centered computing—Visualization—Visualization techniques

1 Introduction

Predictive models built by machine learning algorithms are increasingly informing decisions across high-stakes applications such as medicine [47], employment [11], and criminal justice [23]. There is also broad recent interest in developing collaborative systems where information from both humans and machine learning models is used to make predictions and decisions [5, 16, 29, 51, 57]. An important aspect of machine predictions in such contexts is **calibration**. In particular, for machine learning classifiers, a well-calibrated model is one where the class probabilities produced by the model closely match the empirical frequency of how often the model's predicted class matches the true class label. Calibration error can be measured empirically by the difference between a model's self-perceived accuracy (the probability it assigns to the predicted class, also known as confidence) and the actual accuracy of its predictions as a function of confidence.

In practice, however, it is well-documented that machine learning classifiers such as deep neural networks tend to produce poorly-calibrated class probabilities [19, 39, 53]. As a result, a variety of recalibration techniques have been developed, which aim to ensure that a model's confidence matches its true accuracy. The most widely used approach is the post-hoc calibration method, which uses a separate labeled dataset to learn a mapping from the original model's class probabilities to calibrated probabilities, often using a relatively simple one-dimensional mapping (e.g., [30, 32, 43]). These methods have been shown to generally work well in the sense that the empirical calibration error of the model tends to improve significantly after this post-hoc calibration step. A typical approach is to estimate the expected calibration error empirically on a labeled test set and compare this metric before and after recalibration is performed. A commonly used metric in this context in the machine learning literature is the *expected calibration error* (ECE, [19]), which we define more precisely in Section 3.

In this paper we show that aggregate measures of calibration error such as ECE can hide significant systematic miscalibration, where the calibration error of a model varies significantly as a function of a variable of interest. The variable of interest can be an input variable to the model or some other metadata variable—we focus in particular in this paper on real-valued variables. An example of such a variable is *age*, for prediction problems involving individuals. It has been well-documented that machine learning classifiers often exhibit "age bias," where a model's accuracy systematically varies as a function of age. This bias has been demonstrated across a number of application areas, including in facial image analysis (for gender classification [4], emotion detection [28], and face recognition [34]), in credit-scoring [49], and in prediction of hospital mortality [42]. We show in this paper that such biases can occur not just in accuracy but also in calibration, for instance, a model can be systematically overconfident for some age ranges and underconfident for others.

As an illustrative example, consider a simple neural-network classifier trained to predict the presence of cardiovascular disease using a benchmark medical diagnosis dataset[1]. After applying Platt scaling (a standard post-hoc calibration technique [43]), the model appears to be well-calibrated in the sense that the aggregate calibration error, as measured by ECE, has a relatively low value of about 0.8% (values of 10% to 20% are relatively common in practice for machine learning models). This low ECE is reflected in the reliability diagram shown in Figure 1a. Reliability diagrams [38] plot empirical classification accuracy or error as a function of a model's predicted class probability or confidence. A well-calibrated model will have a curve close to the diagonal on the plot, as in Figure 1a. Thus, based on both the reliability diagram and and the low ECE value of 0.8%, a user of this model might reasonably conclude that the model is generally well-calibrated. However, comparing the model's actual classification error and the model's predicted classification error, with respect to the variable *Patient Age*, as illustrated in Figure 1b, reveals an undesirable pattern of systematic miscalibration. The model is underconfident by upwards of five percentage points for younger patients, and is significantly overconfident for older patients.

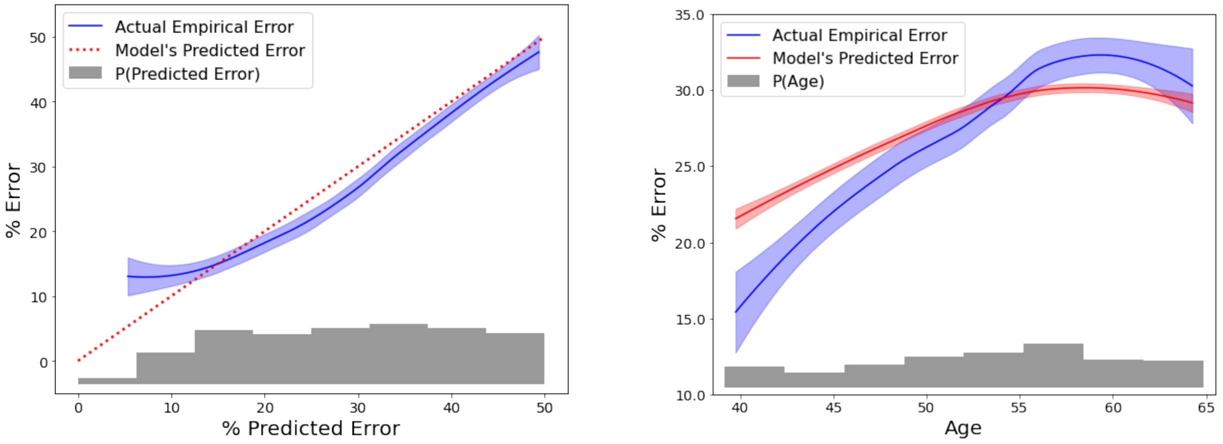
In this paper we demonstrate that this type of systematic miscalibration appears to be relatively common in practice across well-known machine learning models and datasets, and that standard calibration measures such as ECE can hide such miscalibration. In particular, our contributions are as follows: (i) we introduce a new calibration metric (VECE) that assesses calibration on a per-variable basis, (ii) we propose corresponding *variable-wise calibration plots* for visualization of systematic miscalibration; (iii) we perform a case study investigating systematic miscalibration over tabular, text, and image datasets, with a focus on interpretation and visualization; and (iv) we show that a relatively simple variable-wise tree-based calibration method can significantly reduce systematic miscalibration across a variety of classifiers and datasets. Our code is available online at https://github.com/markellekelly/variable-wise-calibration.

2 Related Work

Visualization of Model Performance by Variable: Although there are many visualization techniques that support diagnosis and

*e-mail: kmarke@uci.edu

[1] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset



(a) Reliability diagram

(b) Variable-wise calibration plot

Figure 1: LOESS-smoothed calibration plots, with 95% confidence intervals, for a neural network predicting cardiovascular disease, after recalibration with Platt scaling: (a) reliability diagram, (b) actual and model-predicted error as a function of patient age. This dataset consists of 70,000 records of patient data (49,000 train, 6,000 validation, 15,000 test), with a binary prediction task of determining the presence of cardiovascular disease. Both plots examine the differences between the model's predicted error (1-the confidence score, or probability the model assigns to its predictions) and the model's actual error (1-accuracy). The variable-wise calibration plot compares these with respect to patient age, revealing disparities that are not apparent in the reliability diagram. Note: here we include a smoothed reliability diagram based on error for ease of comparison with the variable-wise calibration plot; we include a more traditional, binned reliability diagram in the Appendix.

understanding of predictive models in general, there is limited prior work on visualization of model performance with respect to a particular variable of interest. One such technique is partial dependence plots [18, 36], which visualize the effect of an input feature of interest on model predictions. Another approach is dashboards such as FairVis [10] and Fairlearn [7] which enable the exploration of model performance (e.g., accuracy, false positive rate) across various data subgroups. However, there is no prior work we are aware for visualization of *calibration properties* of a model as a function of a variable of interest, i.e., how a model's own predictions of accuracy (or error) vary as a function of a particular variable.

Quantification of Model Performance by Variable: Beyond visualizations such as fairness dashboards, mathematical methods for detecting systematic variation in model performance have also been developed, centered on the concept of disaggregated evaluation: computing metrics of interest individually for sensitive sub-populations [35, 46]. For model calibration in particular, several group-wise evaluation techniques have been introduced, which compute calibration error across various categorical subsets of the data [20, 40]. However, none of this prior work focuses on quantifying and characterizing calibration error with respect to continuous variables.

Visualization for Uncertainty and Calibration: The use of different visualization methods for presenting information about uncertainty is well-established in prior work [8, 45]. Given the popularity of machine learning across many different application domains (e.g., for AI-assisted decision making), there has been significant recent interest in applying these general techniques to provide tools to better understand the predictions of machine learning models and their associated uncertainties. For example, pie charts or icon charts can be used to visualize the uncertainty (a single probability) about a binary model prediction [6]. Beyond including model confidence for single predictions (e.g., [59]), more global uncertainty visualizations have also been used, such as plotting continuous (regression) model predictions with 95% confidence intervals against a variable of interest [3, 12].

In the more specific context of understanding calibration properties of machine learning classifiers, visualization methods have generally been limited to reliability diagrams, which consist of bar charts that plot the model's predicted probabilities against true class frequencies [17, 37], with extensions such as resampled "consistency bars" to communicate additional data-driven uncertainty about degrees of miscalibration [9]. Our work expands on reliability diagrams, introducing a new type of variable-dependent calibration plot that explores calibration as a function of a particular variable of interest, drawing from the literature on uncertainty visualization.

3 Quantifying Calibration Error

Consider a classification problem mapping inputs x to predictions for labels $y \in 1, \ldots, K$. Let f be a black-box classifier which outputs label probabilities $f(x) \in [0,1]^K$ for each $x \in X$. Then, for the standard 0-1 loss function, the predicted label is $\hat{y} = \operatorname{argmax}(f(x)) \in 1, \ldots, K$ and the corresponding confidence score is $s = s(x) = P_f(y = \hat{y}|x) = \max(f(x))$. It is of interest to determine whether such a model is *well-calibrated*, that is, whether its confidence matches the true probability that a prediction is correct.

For a given confidence score s, we define $\mathrm{Acc}(s) = P(y = \hat{y}|s) = \mathbb{E}\left[\mathbb{I}[y = \hat{y}|s]\right]$.Then the ℓ_p calibration error (CE), as a function of the confidence score s, is defined as the difference between accuracy and confidence score [32]:

$$\mathrm{CE}(s) = |P(y = \hat{y}|s) - s|^p = |\mathrm{Acc}(s) - s|^p \tag{1}$$

where $p \geq 1$. In this paper, we will focus on the expectation of the ℓ_1 calibration error with $p = 1$, known as the ECE:

$$\begin{aligned} \mathrm{ECE} &= \mathbb{E}_{P(s)}[CE(s)] \\ &= \int_s P(s)|P(y = \hat{y}|s) - s|ds \end{aligned} \tag{2}$$

where an ECE of zero corresponds to "perfect" calibration. In practice, ECE is often estimated empirically on a labeled test dataset by creating B bins over s according to some binning scheme [19]:

$$\widehat{\mathrm{ECE}} = \sum_{b=1}^{B} \frac{n_b}{n} |\mathrm{Acc}_b - \mathrm{Conf}_b| \tag{3}$$

where n_b is the number of datapoints in bin b, n is the total number of datapoints, and Acc_b and Conf_b are the accuracy and average value of confidence s, respectively, in bin $b = 1, \ldots, B$.

Now, consider a real-valued variable V taking values v. To evaluate model calibration with respect to V, we introduce the notion of *variable-wise calibration error* (VCE), defined pointwise as a function of v:

$$\mathrm{VCE}(v) = \left| P(y = \hat{y}|v) - \mathbb{E}[s(v)] \right| \tag{4}$$

where $\mathbb{E}[s(v)]$ is the expected score conditioned on a particular value v:

$$\mathbb{E}[s(v)] = \int_s s \cdot P(s|v) ds \tag{5}$$

In general, conditioning on v will induce a distribution over inputs x, which in turn induces a distribution $P(s|v)$ over scores s. For example, in the context of Figure 1b, at $v = 45$, the model accuracy $P(y = \hat{y}|v)$ is estimated to be $100 - 21 = 79\%$ and the expected score $\mathbb{E}[s(v)]$ is estimated to be 76%, so the $\mathrm{VCE}(v)$ is approximately 3%.

The expected value of $\mathrm{VCE}(v)$, with respect to V, is defined as:

$$\mathrm{VECE} = \mathbb{E}_{P(v)}[\mathrm{VCE}(v)] = \int_v P(v)\mathrm{VCE}(v) dv \tag{6}$$

Note that this differs from the definition of ECE in Equation 2 in that it measures the calibration error with respect to variable V, rather than with respect to the score s—we investigate the differences between the two in more detail in the Appendix.

As with ECE, we can compute an empirical estimate of VECE by binning, where bins b are now defined by some binning scheme (e.g., equal weight) over values v of the variable V (rather than over scores s):

$$\widehat{\mathrm{VECE}} = \sum_{b=1}^{B'} \frac{n_b}{n} |\mathrm{Acc}_b - \mathrm{Conf}_b| \tag{7}$$

where b is some bin corresponding to a sub-range of V, n_b is the number of points within this bin, and Acc_b and Conf_b are empirical estimates of the model's accuracy and the model's average confidence (average score) within bin b.

In the Appendix, we establish several theoretical results regarding the ECE and VECE. In particular, we prove that the ECE and VECE can differ significantly, by a gap of up to $0.5 - \frac{1}{2K}$, where K is the number of classes (e.g., in the binary case, a gap of up to 0.25). We also show that, when a model is consistently over-confident (as defined precisely in the Appendix), the ECE and VECE are equal.

4 Mitigation of Systematic Miscalibration

4.1 Variable-wise Calibration Plots

Classical reliability diagrams, which plot a model's confidence scores against its true accuracy, are widely used in calibration evaluation [19, 37]. As a complement to reliability diagrams, we introduce *variable-wise calibration plots*, building on the concept of variable-wise calibration error (VECE) from Section 3. These plots visualize the differences in accuracy and confidence along the dimension of a variable of interest, allowing for visual interpretation of any variable-wise miscalibration. Figure 1b provides an example of a variable-wise calibration plot for the variable *Age*.

Variable-wise calibration plots display smoothed curves for the model's predicted and actual error rates as a function of the variable of interest V. To put the differences in curves into perspective, these plots include 95% confidence bars for the actual and predicted error. [3] found that including 95% confidence intervals in visualizations improved the confidence of machine learning experts in making model selection decisions. Compared to related uncertainty visualization techniques, error bars are visually simple and easy to interpret, as long as they are explicitly labeled (e.g., as 95% confidence intervals) [6]. One limitation is that error bars can over-emphasize the range within them, although this was shown for a general audience [13] and may be less of a concern for users with statistical backgrounds.

We recommend that variable-wise plots also include a histogram of the data with respect to V. This aggregate distributional information provides important context, and is often a key component of multi-view data visualizations [24, 26, 50]. In particular, the histograms in variable-wise calibration plots function as a non-parametric expression of uncertainty [44, 48, 56].

For ease of interpretation in the results below we use the model's error rate and predicted error, rather than accuracy and confidence, although they are equivalent. Particularly for models with high accuracy, this framing emphasizes important differences in performance (e.g., "doubling the error rate from two to four percent" rather than "reducing the accuracy from 96% to 94%"). We note, however, that the choice to present this "negatively" (e.g., five percent error rate) or "positively" (e.g., 95% accuracy) could alter users' perceptions of the risks involved, potentially affecting their decisions [14, 52].

To generate these plots, we first compute the individual error $\mathbb{I}[y \neq \hat{y}]$ and predicted error $1 - s(x) = 1 - \max(f(x))$ for each observation. We then construct nonparametric error curves with LOESS, with quadratic local fit and an assumed symmetric distribution of the errors, with empirically-chosen smoothing factors between 0.8 and 0.9. (Further details are available in our code.) This approach allows us to obtain 95% confidence bars based on the standard error.

In Section 5 we explore how these plots can be used to characterize miscalibration across different models, datasets, and variables.

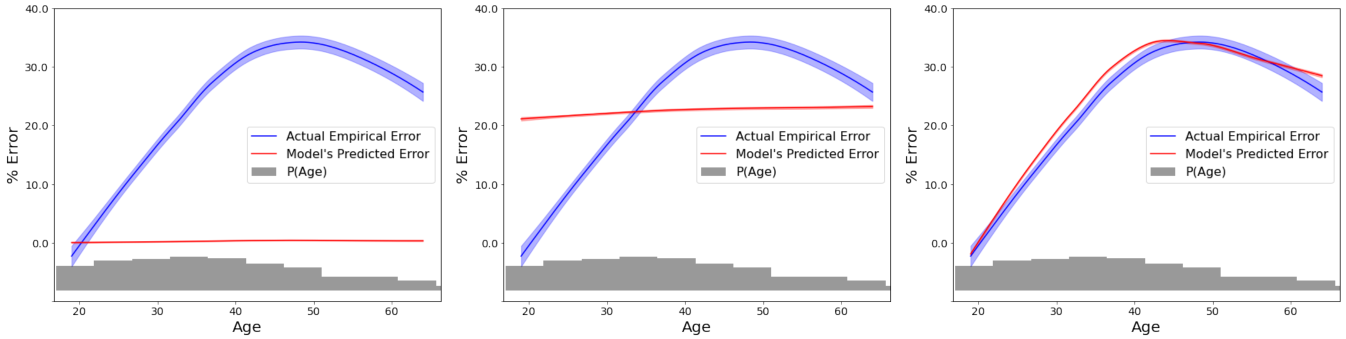
4.2 Recalibration Methods

We find empirically that standard score-based recalibration techniques often reduce ECE while neglecting variable-wise systematic miscalibration. Because calibration error can vary as a function of a feature of interest v, we propose incorporating information about v during recalibration.

We introduce the concept of variable-wise recalibration, a family of recalibration methods that adjust confidence scores with respect to some variable of interest V. As an illustrative example, we perform experiments in Section 5 with a modification of probability calibration trees [33]. This technique involves performing logistic calibration separately for data splits, defined by decision trees trained over the input space. We alter the method to train decision trees for y with only v as input, then perform beta calibration at each leaf [30]. In the multi-class case, we use Dirichlet calibration, an extension of beta calibration for k-class classification [31]. Our use of split-based recalibration using decision trees is intended to provide a straightforward illustration of the potential benefits of variable-wise calibration, rather than to provide a state-of-the-art methodology that can balance ECE and VECE (which we leave to future work).

5 Systematic Miscalibration in Practice

In this section, we explore several examples where a model appears to be well-calibrated according to ECE, but is hiding systematic miscalibration relative to some variable of interest. For each dataset and variable of interest V, we evaluate several score-based calibration methods and our variable-wise recalibration, based on ECE, VECE, and variable-wise calibration plots. In particular, we calibrate with scaling-binning [32], Platt scaling [43], beta calibration [30], and (for the multi-class case) Dirichlet calibration [31]. These examples demonstrate the application and interpretation of our variable-wise calibration plots as well as the potential benefits of variable-wise recalibration. We show that variable-wise calibration plots enable meaningful characterization of the relationships between variables of interest, predicted, and true empirical error, providing more detailed model insight than a single number (i.e., ECE or VECE).

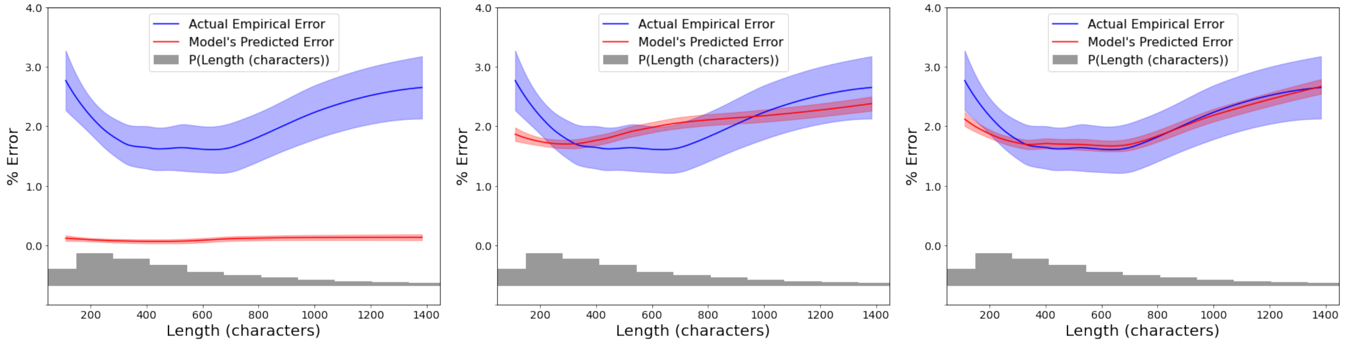


(a) Uncalibrated

(b) Beta-calibrated

(c) Variable-wise calibrated

Figure 2: Variable-wise calibration plots for the Adult Income model for *Age*



(a) Uncalibrated

(b) Beta-calibrated

(c) Variable-wise calibrated

Figure 3: Variable-wise calibration plots for the Yelp model for *Review Length*

For each example, the model is trained on a subset of the full dataset. The remaining data is split into a calibration subset and a test subset. Each calibration method is trained on the same calibration set, and all metrics and figures are produced from the final test set. The ECE and VECE are computed with an equal-support binning scheme, with $B = B' = 10$. Additional details regarding datasets, models, and calibration can be found in the Appendix.

5.1 Adult Census Records: Predicting Income

The UCI Adult Income dataset[2] is comprised of 1994 Census records, where the goal is to predict whether an individual's annual income is greater than $50,000. We model this data with a simple feed-forward neural network and we explore the model's calibration error with respect to age (i.e. let V=age). Uncalibrated, this model has an ECE and VECE of 20.67% (see Table 1). The ECE and VECE are equal precisely because of the model's consistent overconfidence as a function of both the confidence score and the V variable (see Appendix). The overconfidence with respect to age is reflected in the variable-wise calibration plot (Figure 2a). The model's error rate varies significantly as a function of age, with very high error for individuals around age 50, and much lower error for younger and older people. However, its confidence remains nearly constant at close to 100% (i.e., a predicted error close to 0%) across all ages.

After recalibrating, the ECE is dramatically reduced, with beta calibration achieving an ECE of 1.65%. However, the corresponding VECE is still high at over 9%. As shown in Figure 2b, the model's self-predicted error has increased substantially, but this estimate remains near-constant for all ages. Thus, despite a significant improvement in ECE, this recalibrated model still harbors unfairness with respect to age, exhibiting overconfidence in its predictions for individuals in the 35-65 age range, and underconfidence for those outside of it. As the model is no longer consistently overconfident, the ECE and VECE diverge, as predicted theoretically.

	ECE	VECE
Uncalibrated	20.67%	20.67%
Scaling-binning	2.27%	9.25%
Platt scaling	4.57%	10.13%
Beta calibration	1.65%	9.59%
Variable-wise calibration	**1.64%**	**2.11%**

Table 1: Adult Income model calibration error

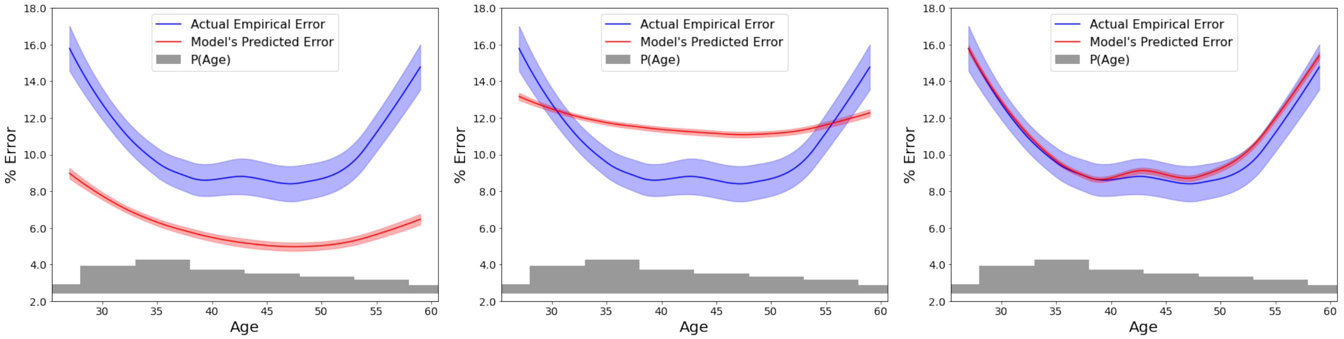
Our simple variable-wise calibration obtains a significantly lower VECE of 2.11%, while simultaneously reducing the ECE. This improvement in VECE is reflected in Figure 2c. The model's predicted error now varies with age to match the true error rate. In this case, variable-wise recalibration improves the age-wise systematic miscalibration of the model, without detriment to the overall calibration error.

5.2 Yelp Reviews: Predicting Sentiment

To explore the phenomenon of systematic calibration in an NLP context, we use a fine-tuned large-scale language model, BERT [27], on the Yelp review dataset[3]. The goal is to predict whether a review has a positive or negative rating based on its text. Note that, in this context, there are no interpretable features being provided directly

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://www.yelp.com/dataset

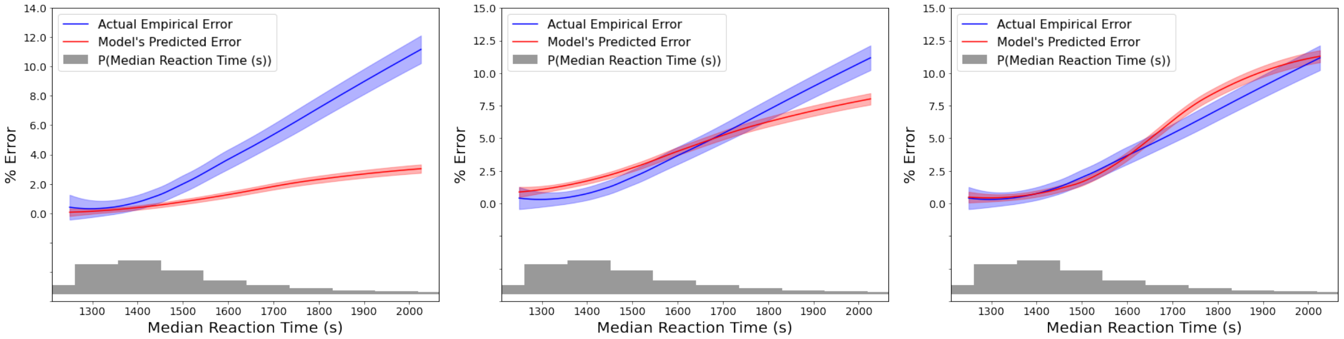


(a) Uncalibrated

(b) Calibrated with Platt scaling

(c) Variable-wise calibrated

Figure 4: Variable-wise calibration plots for the Bank Marketing model for *Age*



(a) Uncalibrated

(b) Dirichlet calibrated

(c) Variable-wise calibrated

Figure 5: Variable-wise calibration plots for the CIFAR-10H model for *Median Reaction Time*

as input to the model. Instead, to better diagnose model behavior, we can analyze real-valued characteristics of the text, such as the length of each review or statistics related to how often certain parts-of-speech occur in the review. Here we focus on review length in characters.

Figure 3a shows the model's error and predicted error with respect to review length. The error rate is lowest for reviews around 300-700 characters, around the median review length. Very short and very long reviews are associated with a higher error rate. Again, this model is consistently overconfident, with an uncalibrated ECE and VECE of 1.93% (see Table 2).

	ECE	**VECE**
Uncalibrated	1.93%	1.93%
Scaling-binning	4.23%	4.23%
Platt scaling	3.04%	0.64%
Beta calibration	1.73%	0.37%
Variable-wise calibration	**1.70%**	**0.23%**

Table 2: Yelp model calibration error

Of the score-based recalibration methods, beta calibration obtains the lowest ECE of 1.73% and a substantially reduced VECE of 0.37%. Figure 3b reflects this; the model's predicted error aligns more closely with its actual error rate, although it is still notably overconfident for very short reviews. Again, we observe that the ECE and VECE are equal for the uncalibrated, consistently overconfident model, and diverge after recalibration; these observations hold true in general for each example dataset.

Variable-wise recalibration reduces the VECE slightly further, while offering a small improvement to the overall ECE. After variable-wise calibration, the predicted error curve matches the true relationship between review length and true error rate more faithfully, reducing overconfidence for short reviews (Figure 3c).

5.3 Bank Marketing: Predicting Customer Subscriptions

We also investigate miscalibration on a simple neural network modeling the UCI Bank Marketing dataset[4]. The model predicts whether a bank customer will subscribe to a bank term deposit as a result of direct marketing.

The uncalibrated model is again overconfident, with ECE and VECE over 4.5% (see Table 3). Consider the calibration error with respect to customer age, both before (Figure 4a) and after (Figure 4b) recalibration. The best-performing recalibration technique, Platt scaling, uniformly increases the predicted error across age, resulting in underconfidence for most ages and overconfidence at the edges of the distribution. The ECE and VECE are both reduced, but the 2.83% VECE can be further improved.

	ECE	**VECE**
Uncalibrated	4.69%	4.69%
Scaling-binning	4.37%	3.39%
Platt scaling	2.38%	2.83%
Beta calibration	2.48%	2.77%
Variable-wise calibration	**2.10%**	**0.52%**

Table 3: Bank Marketing model calibration error

[4] https://archive.ics.uci.edu/ml/datasets/bank+marketing

Variable-wise recalibration achieves the lowest ECE, while reducing VECE to about half of one percent. Figure 4c reflects this improvement. The predicted error after variable-wise recalibration matches the true error rate more closely, reducing the miscalibration with respect to customer age.

5.4 CIFAR-10H: Image Classification

As a multi-class example, we investigate variable-wise miscalibration on CIFAR-10H, a 10-class image dataset including labels and reaction times from human annotators [41]. We use a standard deep learning image classification architecture (a DenseNet model) to predict the image category. Instead of Platt scaling and beta calibration, here we use Dirichlet calibration, to accomodate multiple classes.

Consider the calibration error with respect to median annotator reaction time. Dirichlet calibration achieves the lowest overall ECE; variable-wise calibration obtains the lowest VECE (see Table 4). As shown in Figure 5, variable-wise recalibration reduces underconfidence for examples with low median reaction times (where the majority of data points lie). It is intuitive that taking annotator reaction time into account could improve overall calibration, and in this case it does achieve competitive ECE and VECE.

	ECE	VECE
Uncalibrated	1.90%	1.92%
Scaling-binning	3.83%	3.60%
Dirichlet calibration	**0.80%**	1.12%
Variable-wise calibration	1.31%	**0.60%**

Table 4: CIFAR-10H model calibration error

6 Discussion

Complex uncertainty visualizations may be overwhelming or difficult to interpret for the general public—for a layperson, a lower-precision categorization of uncertainty or calibration might be preferred [6,54]. On the other hand, detailed information about uncertainty information enables knowledgeable users (such as developers of machine learning models) to calibrate their trust in a model [59] and make decisions more confidently [3]. Thus, our visualizations are likely to be most useful for (i) machine learning developers (e.g., in evaluating a model in development), and (ii) users (or potential users) of models (e.g., in determining whether or not a model should be deployed), who have the numeracy and statistical literacy to understand concepts such as probability and confidence intervals.

Medical diagnosis with machine learning models is an example of a domain where variable-wise calibration plots may be particularly useful. In the application of machine learning in clinical contexts, the variation between training datasets and prediction (test) sets (for example due to changes in age distributions of patients) can lead to significant degradation in model performance [58]. In particular, clinical radiology is an area where machine learning is rapidly moving from research labs into clinical use: since 2018 the U.S. Food and Drug Administration (FDA) has doubled the number of machine learning models approved for clinical use in radiology [2]. While there is broad consensus among radiologists that machine learning methods show considerable promise, there is also a general sense that these methods may have "blind spots" and cannot be fully trusted. To quote a recent American College of Radiologist's survey [1]:

> Establishing the safety and efficacy of AI algorithms before clinical use was critically important to the survey respondents; more than 60% indicated they want some form of external validation of AI models across representative data sets, and an equal number indicated they would like to be able to assess the performance of an AI model on their own patient data before deploying it into their clinical workflows.

In this context, methods for systematic characterization of model performance can play an important diagnostic role in the development and deployment of prediction models [2,15]. More concretely, radiologists could create variable-wise calibration plots using a set of their own patient data to evaluate a model. These plots would provide insight into when the model is more or less likely to be correct or overconfident, across variables such as age and weight, helping inform the practitioners' decisions on when to trust the model or whether to use it at all. Similarly, [21] note the importance of such global plots for "fairness-focused debugging," determining whether an issue the radiologist experiences is a "one-off" or part of a wider systematic problem (i.e. miscalibration). However, we note that visualizations such as these capture only part of the full picture and, alone, can lead to over-trust and limited, superficial evaluation [25]; thus, they should be used as part of a more comprehensive model evaluation workflow.

A user study to evaluate the effectiveness of variable-wise calibration plots is an important next step. For example, these visualizations could be tested in the context of decision-making for medical diagnosis. Criteria to evaluate include comprehension (whether users can correctly interpret the plots), performance (whether the plots improve the accuracy of a user's decision-making, e.g. compared to a reliability diagram or single confidence score), and trust (whether the plots affect users' trust in the model or confidence in their own decisions) [22,55].

Limitations In this work we focused on characterization and mitigation of miscalibration for a single variable at a time; analyzing miscalibration as a function of multiple variable is a potentially interesting direction for future work. For example, although we did not observe recalibration with respect to one variable V worsening VECE for another variable, this behavior has not been analyzed theoretically. In addition, the tree-based variable-based calibration technique used in the paper is primarily for illustration; the development of new methods for simultaneously reducing score-based and variable-based miscalibration also merits further investigation.

7 Conclusions

In this paper we demonstrated that the visualization of calibration from the perspective of specific variables of interest can offer new insight into model behavior and can provide actionable avenues for improvement. For example, the detection and mitigation of systematic miscalibration across continuous variables is essential for ensuring the fairness of machine learning models, particularly for demographic variables such as age. We showed that traditional measures of calibration such as ECE and reliability diagrams can hide significant miscalibration with respect to variables of potential importance to a developer or user of a classification model. To better detect and characterize this systematic miscalibration, we introduced the VECE measure and corresponding variable-wise calibration plots. In a case study across several datasets and models, we showed that variable-wise calibration plots, VECE, and variable-wise recalibration are empirically useful for understanding and mitigating systematic miscalibration. Looking forward, we recommend moving beyond purely score-based calibration analysis to mitigate biases in measurement of calibration error.

Acknowledgements

This material is based upon work supported in part by the HPI Research Center in Machine Learning and Data Science at UC Irvine, by the National Science Foundation under grants number 1900644 and 1927245, and by a Qualcomm Faculty Award.

References

[1] B. Allen, S. Agarwal, L. Coombs, C. Wald, and K. Dreyer. 2020 ACR data science institute artificial intelligence survey. *Journal of the American College of Radiology*, 18(8):1153–1159, 2021.

[2] B. Allen, K. Dreyer, R. Stibolt Jr, S. Agarwal, L. Coombs, C. Treml, M. Elkholy, L. Brink, and C. Wald. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: Try it, buy it, check it. *Journal of the American College of Radiology*, 18(11):1489–1496, 2021.

[3] S. Z. Arshad, J. Zhou, C. Bridon, F. Chen, and Y. Wang. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*, pp. 352–360, 2015.

[4] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pp. 327–359. Springer, 2021.

[5] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. Optimizing ai for teamwork. *CoRR*, abs/2004.13102, 2020.

[6] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

[7] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[8] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*, pp. 3–27. Springer, 2014.

[9] J. Bröcker and L. A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007.

[10] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 46–56. IEEE, 2019.

[11] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27, 2016.

[12] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021.

[13] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014.

[14] K. Damnjanovic and V. Gvozdenovic. Influence of the probability level on the framing effect. *Psychological topics*, 25(3):405–429, 2016.

[15] D. Daye, W. F. Wiggins, M. P. Lungren, T. Alkasab, N. Kottler, B. Allen, C. J. Roth, B. C. Bizzo, K. Durniak, J. A. Brink, et al. Implementation of clinical artificial intelligence in radiology: Who decides and how? *Radiology*, p. 212151, 2022.

[16] A. De, N. Okati, A. Zarezade, and M. G. Rodriguez. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 5905–5913, 2021.

[17] T. S. Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, P. Flach, et al. Classifier calibration: How to assess and improve predicted class probabilities: a survey. *arXiv preprint arXiv:2112.10327*, 2021.

[18] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pp. 1189–1232, 2001.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

[20] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

[21] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019.

[22] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics*, 25(1):903–913, 2018.

[23] E. E. Joh. Feeding the machine: Policing, crime data, & algorithms. *Wm. & Mary Bill Rts. J.*, 26:287, 2017.

[24] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[25] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

[26] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, 2013.

[27] J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[28] E. Kim, D. Bryant, D. Srikanth, and A. Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 638–644, 2021.

[29] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

[30] M. Kull, T. S. Filho, and P. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54, pp. 623–631, 2017.

[31] M. Kull, M. Perello-Nieto, M. Kängsepp, T. S. Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.

[32] A. Kumar, P. S. Liang, and T. Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3787–3798, 2019.

[33] T. Leathart, E. Frank, G. Holmes, and B. Pfahringer. Probability calibration trees. In M.-L. Zhang and Y.-K. Noh, eds., *Proceedings of the Ninth Asian Conference on Machine Learning*, vol. 77, pp. 145–160, 2017.

[34] S. Lee, S. Oh, M. Kim, and E. Park. Measuring embedded human-like biases in face recognition models. In *Computer Sciences and Mathematics Forum*, vol. 3, p. 2. MDPI, 2022.

[35] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

[36] C. Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

[37] A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1):41–47, 1977.

[38] A. H. Murphy and R. L. Winkler. Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79(387):489–500, 1984.

[39] Y. Ovadia, E. Fertig, B. Lakshminarayanan, S. Nowozin, D. Sculley, J. Dillon, J. Ren, Z. Nado, and J. Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.

[40] F. Pan, X. Ao, P. Tang, M. Lu, D. Liu, L. Xiao, and Q. He. Field-aware calibration: A simple and empirically strong method for reliable probabilistic predictions. In *Proceedings of The Web Conference: WWW 2020*, pp. 729–739, 2020.

[41] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9617–9626, 2019.

[42] S. R. Pfohl, A. Foryciarz, and N. H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621, 2021.

[43] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[44] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson. Visualizing summary statistics and uncertainty. *Computer Graphics Forum*, 29(3):823–832, 2010.

[45] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *IFIP Working Conference on Uncertainty Quantification*, pp. 226–249. Springer, 2011.

[46] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020.

[47] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[48] E. Sakhaee and A. Entezari. A statistical direct volume rendering framework for visualization of uncertain data. *IEEE transactions on visualization and computer graphics*, 23(12):2509–2520, 2016.

[49] P. Snel and S. van Otterloo. Practical bias correction in neural networks: a credit default prediction case study. *Computers and Society Research Journal*, (3), 2022.

[50] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics*, 25(1):672–681, 2018.

[51] M. Steyvers, H. Tejeda, G. Kerrigan, and P. Smyth. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11), 2022.

[52] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. In *Behavioral decision making*, pp. 25–41. Springer, 1985.

[53] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467, 2019.

[54] A. M. Van der Bles, S. Van Der Linden, A. L. Freeman, J. Mitchell, A. B. Galvao, L. Zaval, and D. J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society open science*, 6(5):181870, 2019.

[55] N. D. Weinstein and P. M. Sandman. Some criteria for evaluating risk messages. *Risk Analysis*, 13(1):103–114, 1993.

[56] D. Weiskopf. Uncertainty visualization: Concepts, methods, and applications in biological data visualization. *Frontiers in Bioinformatics*, p. 10, 2022.

[57] B. Wilder, E. Horvitz, and E. Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 1526–1533, 7 2020.

[58] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 279–290, 2021.

[59] Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.

A Proofs for Section 5

Theorem A.1 (VECE bound)**.** *There exist K-ary classifiers f and variables V such that the classifier f has ECE = 0 and variable-wise VECE $= 0.5 - \frac{1}{2K}$.*

Proof. Let V be a continuous variable with density $P(v)$. Recall that VECE $= \int_v P(v)|P(y=\hat{y}|v) - \mathbb{E}[s|v]|dv$ where $P(y=\hat{y}|v)$ is the accuracy of model f as a function of v, and the score s is the probability that the model assigns to its label prediction $\hat{y}$. The reliability diagram for a K-ary classifier has scores $s \in [\frac{1}{K}, 1]$ where the leftmost value for this interval is a result of the fact that the score is defined as the maximum of K class probabilities. Let $\gamma = 0.5 + \frac{1}{2K}$ be the midpoint of this interval.

Assume that the scores s have a uniform distribution of the form $s \sim U(\gamma - \alpha, \gamma + \alpha)$, where α is some constant and $0 \leq \alpha \leq 0.25$, and that the scores s and the variable V are independent. Further assume that the accuracy of the model f depends on v and s in the following manner

$$P(y=\hat{y}|v \leq v_t, s \leq \gamma) = 1 - \alpha \qquad P(y=\hat{y}|v \leq v_t, s > \gamma) = 1$$
$$P(y=\hat{y}|v > v_t, s \leq \gamma) = \frac{1}{K} \qquad P(y=\hat{y}|v > v_t, s > \gamma) = \frac{1}{K} + \alpha$$

where v_t is defined such that $P(v \leq v_t) = P(v > v_t) = 0.5$.

The marginal accuracy as a function of the score (marginalizing over v) can be written as

$$\begin{aligned} P(y=\hat{y}|s \leq \gamma) &= \gamma - \frac{\alpha}{2} \\ P(y=\hat{y}|s > \gamma) &= \gamma + \frac{\alpha}{2}. \end{aligned}$$

The marginal accuracy as a function of z (marginalizing over s) is

$$\begin{aligned} P(y=\hat{y}|z \leq z_t) &= 1 - \frac{\alpha}{2} \\ P(y=\hat{y}|z > z_t) &= \frac{1}{K} + \frac{\alpha}{2}. \end{aligned}$$

This setup is designed so that the score is close to the accuracy as a function of s (to minimize ECE), but the variable-wise expected scores $\mathbb{E}[s|z] = \gamma$ are relatively far away from accuracy as a function of z.

Under these assumptions we can write the ECE as

$$\begin{aligned} \text{ECE} &= \int_s p(s) \cdot |P(y=\hat{y}|s) - s|ds \\ &= \int_{\gamma-\alpha}^{\gamma} \frac{1}{2\alpha}|\gamma - \frac{\alpha}{2} - s|ds + \int_{\gamma}^{\gamma+\alpha} \frac{1}{2\alpha}|\gamma + \frac{\alpha}{2} - s|ds \\ &= \frac{\alpha}{4}. \end{aligned} \tag{8}$$

We can write the VECE as

$$\begin{aligned} &\int_{-\infty}^{v_t} P(v)|P(y=\hat{y}|v) - \mathbb{E}[s|v]|dv + \int_{v_t}^{\infty} P(v)|P(y=\hat{y}|v) - \mathbb{E}[s|v]|dv \\ &= \int_{-\infty}^{v_t} P(v) \cdot |1 - \frac{\alpha}{2} - \gamma|dv + \int_{v_t}^{\infty} P(v) \cdot |\frac{1}{K} + \frac{\alpha}{2} - \gamma|dv \\ &= (0.5 - \frac{1}{2K} - \frac{\alpha}{2}) \int_v P(v)dv \\ &= 0.5 - \frac{1}{2K} - \frac{\alpha}{2}. \end{aligned} \tag{9}$$

Thus, as $\alpha \to 0$, VECE $\to (0.5 - \frac{1}{2K})$ and ECE $\to 0$. □

Theorem A.2 (ECE bound)**.** *There exist K-ary classifiers f and variables V such that the classifier f has VECE $= 0$ and ECE $= 0.5 - \frac{1}{2K}$.*

Proof. Let V be a continuous variable with density $P(V)$. Recall that a K-ary classifier has scores $s \in [\frac{1}{K}, 1]$, where we let $\gamma = 0.5 + \frac{1}{2K}$ be the midpoint of this interval. Assume that f produces scores from two uniform distributions, with equal probability: $s \sim U(\frac{1}{K}, \frac{1}{K} + \alpha)$ and $s \sim U(1-\alpha, 1)$, where α is some constant $0 \leq \alpha \leq 0.25$, and that the scores s and the variable V are independent. Finally, suppose the accuracy of the model $P(y=\hat{y}) = \gamma$ is independent of s and V.

Under these assumptions we can write the VECE as

$$\begin{aligned} \text{VECE} &= \int_{-\infty}^{\infty} P(v) \cdot |P(y=\hat{y}|v) - \mathbb{E}[s|v]|dv \\ &= \int_{-\infty}^{\infty} P(v) \cdot |\gamma - \gamma|dv \\ &= 0. \end{aligned} \tag{10}$$

We can write the ECE as

$$\begin{aligned} \text{ECE} &= \int_s p(s) \cdot |P(y=\hat{y}|s) - s|ds \\ &= \frac{1}{2}\int_{\frac{1}{K}}^{\frac{1}{K}+\alpha} \frac{1}{\alpha}|\gamma - s|ds + \frac{1}{2}\int_{1-\alpha}^{1} \frac{1}{\alpha}|\gamma - s|ds \\ &= 0.5 - \frac{1}{2K} - \frac{\alpha}{2}. \end{aligned} \tag{11}$$

Thus, as $\alpha \to 0$, ECE $\to (0.5 - \frac{1}{2K})$ and VECE $= 0$. □

Definition A.3 (Consistent overconfidence)**.** Let f be a classifier with scores s. For a variable V taking values v, f is *consistently overconfident* if $\mathbb{E}[s|v] > P(y=\hat{y}|v), \forall v$, i.e., the expected value of the model's scores f as a function of v is always greater than the true accuracy as a function of v.

Consistent underconfidence is defined analogously with $\mathbb{E}_{[}s|v] < P(y=\hat{y}|v), \forall v$. In the special case where the variable V is defined as the score itself, we have $s > P(y=\hat{y}|s), \forall s$, etc.

Theorem A.4 (Equality conditions for ECE and VECE)**.** *Let f be a classifier that is consistently under- or over-confident with respect both to s and any variable V. Then the ECE and VECE of f are equal.*

Proof. Without loss of generality, suppose f is consistently underconfident with respect to its scores s and V.

Then we have, by consistent underconfidence and the law of total probability:

$$\begin{aligned} \text{ECE} &= \int_s p(s) \cdot |P(y=\hat{y}|s) - s|ds \\ &= \int_s p(s) \cdot P(y=\hat{y}|s)ds - \mathbb{E}[s] \\ &= P(y=\hat{y}) - \mathbb{E}[s] \end{aligned} \tag{12}$$

$$\begin{aligned} \text{VECE} &= \int_v p(v) \cdot |P(y=\hat{y}|v) - \mathbb{E}[s|v]|dv \\ &= \int_v p(v) \cdot P(y=\hat{y}|v)dv - \int_v p(v)\mathbb{E}[s|v]dv \cdot \\ &= \int_v p(v) \cdot P(y=\hat{y}|v)dv - \mathbb{E}[s] \\ &= P(y=\hat{y}) - \mathbb{E}[s] = \text{ECE} \end{aligned} \tag{13}$$

□

B Calibration, Model, and Dataset Details

Here, we include additional information for each dataset and model discussed in Section 5. Code for reproducing all tables and plots is available online[5].

On each dataset, we test several existing recalibration techniques: Platt scaling, scaling-binning, beta calibration, and (for the multiclass case) Dirichlet calibration. For scaling-binning, we calibrate over 10 bins, and for Dirichlet calibration, we use a lambda value of 1e-3, values chosen based on the respective authors' provided examples. Here and in Section 7, we present the uncalibrated and variable-wise calibrated output, along with the best-performing score-based calibration method (for the Adult and Yelp datasets, beta calibration; for Bank Marketing, Platt scaling; for CIFAR, Dirichlet calibration).

Our variable-wise recalibration is performed as follows. Given the calibration set, a decision tree classifier is trained to predict the outcome y with input v (the single variable of interest). We use a maximum depth of two and a minimum leaf size of $0.1*$ the size of the calibration set. The calibration set is then split according to the leaf nodes of the trained decision tree, and separately the rest of the dataset is split according to the same rules. Standard beta calibration is then performed separately for each split, using the subset of the original calibration set as the new calibration set, and computing recalibrated probabilities for the subset of the original dataset.

We note the VECE for each numeric variable in each dataset before and after the recalibration described. We find in general empirically that variable-wise calibration with respect to one variable is not detrimental to the VECE of other variables.

B.1 Cardiovascular Disease (Introduction)

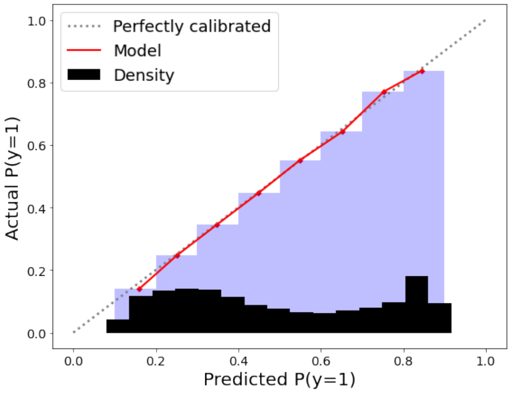


Figure 6: Standard reliability diagram to supplement the smoothed plot in Figure 1a.

B.2 Adult Income

The Adult Income dataset was modeled with a multi-layer perceptron, with two hidden layers of sizes 100 and 75. The model's accuracy is 79%. Of the 48,842 observations, 32,561 were used for training, 2,500 were used for calibration, and 13,781 were used for testing. The dataset includes six continuous variables: age, fnlwgt (the estimated number of people an individual represents), education-num (a number representing the individual's years of education), capital-gain, capital-loss, and hours-per-week (the number of hours per week that an individual works).

Based on the beta-calibrated model, education-num and age rank the highest in VECE, as shown in Section 6. For all six variables, VECE is reduced by performing recalibration with respect to age.

	Uncalibrated	**Beta**	**Variable-wise**
education-num	20.67%	9.95%	**8.53%**
age	20.67%	9.59%	**2.11%**
hours-per-week	20.67%	7.94%	**6.02%**
fnlwgt	20.67%	5.06%	**4.10%**
capital-gain	20.67%	1.50%	**1.39%**
capital-loss	20.67%	1.50%	**1.39%**

Table 5: VECE for numeric variables in the Adult Income dataset: uncalibrated, beta-calibrated, and after variable-wise recalibration **with respect to age**.

B.3 Yelp

The Yelp dataset was modeled with a fine-tuned BERT model. The model's accuracy is 97.7%. 100,000 observations were randomly sampled from the full Yelp dataset. Of these, 70,500 were used for training, 10,000 were used for calibration, and 19,500 were used for testing. Several continuous features were generated from the raw text reviews, including length in characters, number of special characters, and proportions of each part of speech. Based on the beta-calibrated model, review length ranked highest in VECE, followed by proportion of stop words, as shown in Table 6.

	Uncalibrated	**Beta**	**Variable-wise**
Length (characters)	1.93%	0.37%	**0.23%**
Stop-word Proportion	1.93%	0.29%	**0.28%**
Named Entity Count	1.93%	**0.21%**	0.22%

Table 6: VECE for numeric variables in the Yelp dataset: uncalibrated, beta-calibrated, and after variable-wise recalibration **with respect to length in characters**.

B.4 Bank Marketing

The Bank Marketing dataset was modeled with a multi-layer perceptron, with two hidden layers of sizes 100 and 75. The model's accuracy is 88.9%. Of the 45,211 total observations, 31,647 were used for training, 1,000 were used for calibration, and 12,564 were used for testing. Based on the scaling-binning-calibrated model, account balance ranked highest in VECE, followed by age, as shown in Table 7.

	Uncalibrated	**Platt-scaling**	**Variable-wise**
Account balance	5.35%	4.17%	**3.22%**
Age	4.69%	2.83%	**0.52%**

Table 7: VECE for numeric variables in the Bank Marketing dataset: uncalibrated, calibrated with Platt scaling, and after variable-wise recalibration **with respect to age**.

B.5 CIFAR-10H

The CIFAR-10H dataset was modeled with a DenseNet model. The model's accuracy is 96.6%. Of the 10,000 total observations, 4,057 were used for training, 2,000 were used for calibration, and 3,943 were used for testing.

[5] https://github.com/markellekelly/variable-wise-calibration