13th International Workshop on Innovative Simulation for Healthcare, 009
21st International Multidisciplinary Modeling & Simulation Multiconference

2724-0371 © 2024 The Authors.
doi: 10.46354/i3m.2024.iwish.009

Calibration error as loss function in binary classification

Stephan Dreiseitl[1,*]

[1]Dept. of Software Engineering, University of Applied Sciences Upper Austria, A-4232 Hagenberg, Austria

[*]Corresponding author. Email address: stephan.dreiseitl@fh-hagenberg.at

Abstract

Calibration, as an evaluation criterion for classification models, is often not given as much consideration as discrimination, possibly because it is harder to measure. Nevertheless, it is a crucial measure in biomedical data analysis, because a probability estimate by a model should reflect the relative frequency with which it occurs. Poor calibration can lead to suboptimal decisions; by ensuring a close correspondence between predicted probabilities and actual chance of outcome, medical professionals can make these decisions in the most informed manner possible.

We consider the special case of binary classification, and calibration measures for this task. We show how the multi-class calibration measure *expected calibration error* can be adapted to the two-class case, and used directly as a loss function in training neural network models. We also consider the alternative two-class calibration measure of the Hosmer-Lemeshow test statistic, and demonstrate empirically how calibration measures, both stand-alone and in combination with cross-entropy error, can serve as loss functions for classifying two sample data sets from the biomedical domain. Our experiments demonstrate that explicitly optimizing for calibration loss results in models that are well calibrated without losing their ability to discriminate between two classes.

Keywords: Calibration error; neural network loss functions; neural network calibration.

1. Introduction

In the biomedical domain, the performance of binary classifiers is often assessed by two distinct metrics that measures their discriminatory power and their calibration. When a binary classifier calculates an estimate $\hat{p}_i$ of the class 1 membership probability $P(\text{class} = 1 \mid x_i)$ for a given feature vector x_i, its discriminatory power is most often assessed by the area under the ROC curve (AUC) (Hanley and McNeil, 1982; Lasko et al., 2005; Zou et al., 2007), although this is not without controversy (Cook, 2007; Hand, 2009; Flach et al., 2011; Janssens and Martens, 2020).

It is far harder to evaluate the calibration of a binary classifier, because — in contrast to discrimination — there is no gold standard against which a classifier output in the form of a class-membership probability can be measured (Van Calster et al., 2019; Silva Filho et al., 2023). For a long time, the de-facto standard for measuring classifier calibration was the Hosmer-Lemeshow variant of a chi-squared goodness-of-fit test (Hosmer and Lemeshow, 1980, 2000); several improvements were subsequently published in the literature (Pigeon and Heyse, 1999a,b) after minor flaws in the methodology were exposed (Kuss, 2002; Hosmer et al., 1997). Alternatively, calibration is also commonly assessed via graphical means (Copas, 1983; Finazzi et al., 2011; Austin and Steyerberg, 2014; Nattino et al., 2017). If a binary classifier is found to be insufficiently calibrated, it can be re-calibrated after training by a variety of methods (Zadrozny and Elkan, 2002; Naeini et al., 2015).

Regarding the calibration of neural network classifiers, and irregardless of how this is assessed, there are conflicting reports on how well such models are calibrated: In an early reference, Niculescu-Mizil and Caruana (2005) report that neural networks are well classified, but that seems to have changed with deep learning. Guo et al. (2017) observe that such deep neural networks, usually trained by minimizing a negative log-likelihood loss function, are

not well calibrated. However, the authors note that a simple variant of Platt scaling (Platt, 1999) that they call *temperature scaling* can improve the calibration of classifiers in a post-processing step.

The uncertain stature of the calibration of deep learning models has motivated research into loss functions that directly incorporate terms that penalize models that are not well-calibrated. An overview of the literature on this topic is given in Section 2.

In this work, we investigate how calibration measures can be incorporated specifically in binary classifiers. We want to empirically assess the performance of such classifiers on much smaller data sets than usually analyzed in deep learning settings, such as the data sets available for most biomedical binary classification tasks. In Section 3, we argue that there is a subtle but important difference between calibration measures in multi-class vs. binary classification settings. We modify calibration measures used in deep learning for multi-class problems to take advantage of these differences.

The results of applying such alternative loss functions to biomedical classification problems are presented in Section 4. We give concluding remarks in Section 5.

2. State of the Art

Interest in the calibration of deep learning models has significantly increased in recent years, mostly spurred by the publication of Guo et al. (2017) that deep-learning models trained on maximum likelihood-derived loss functions can exhibit poor calibration. One must note that calibration in the context of deep learning models is usually measured by *expected calibration error* (Naeini et al., 2015), which — similarly to the groupings originally introducted by Hosmer and Lemeshow (1980) — uses bins B_k to group probability estimates. Recall that for binary classification, $\hat{p}_i$ is the class 1 membership probability $P(\text{class} = 1 \mid x_i)$ for feature vector x_i. For more than two classes, (Naeini et al., 2015) define $\hat{p}'_i$ to be the highest probability estimate over all classes. Note that these are not the same, and we use two different symbols $\hat{p}'_i$ and $\hat{p}_i$ to distinguish them. Writing y_i for the correct (gold standard) class for case i and $\hat{y}_i$ for the predicted class (i.e., the class for which $\hat{p}_i$ is highest), they further define the *accuracy* acc and *confidence* conf of bins B_k (of case indices) as

$$\text{acc}(B_k) := \frac{1}{|B_k|} \sum_{i \in B_k} \mathbf{1}(\hat{y}_i = y_i) \tag{1}$$

and

$$\text{conf}(B_k) := \frac{1}{|B_k|} \sum_{i \in B_k} \hat{p}'_i \,. \tag{2}$$

Thus, acc is the same as the traditional accuracy metric in machine learning, and conf is how high (close to 1) the average highest probability assessment is in a bin.

With these terms, K denoting the number of bins, and n being the total number of cases, expected calibration error (ECE) is defined as

$$\text{ECE} := \sum_{k=1}^{K} \frac{|B_k|}{n} |\text{acc}(B_k) - \text{conf}(B_k)|, \tag{3}$$

the bin-size-weighted average difference between the accuracy and confidence terms. Due to the binning and the discrete counting operation via the indicator function $\mathbf{1}$ in the definition of accuracy in Equation (1), ECE is not directly suitable as a (component of a) loss function in neural network training. Recently, Bohdal et al. (2023) introduced the notion of *differentiable expected calibration error* DECE to remedy this situation. DECE uses *soft binning*, which learns a simple model of the probability $o_{k,i}$ that the highest probability estimate $\hat{p}_i$ of case i is placed into bin B_k. With this, and n now denoting the number of cases in a minibatch, DECE is defined as

$$\text{DECE} := \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} o_{ki}}{n} |\text{acc}(B_k) - \text{conf}(B_k)|. \tag{4}$$

Comparing this expression with the definition of ECE in Equation 3, one can observe that the set size term $|B_k|$ is now replaced by the sum term $\sum_{i=1}^{n} o_{ki}$ that estimates how many elements of a minibatch belong to bin B_k. In Equation 4, the terms for accuracy and confidence also have to be modified slightly to

$$\text{acc}(B_k) := \frac{1}{\sum_{i=1}^{n} o_{ki}} \sum_{i=1}^{n} o_{ki} \mathbf{1}(\hat{y}_i = y_i) \tag{5}$$

and

$$\text{conf}(B_k) := \frac{1}{\sum_{i=1}^{n} o_{ki}} \sum_{i=1}^{n} o_{ki} \hat{p}'_i \,, \tag{6}$$

where the sum terms $\sum_{i=1}^{n} o_{ki}$ again play a similar role.

Bohdal et al. (2023) report that using DECE as part of a meta-calibration setup improves model calibration specifically also on test sets. Other authors employ soft-calibration variants of ECE directly in loss functions, combined with standard negative log-likelihood terms (Kumar et al., 2018; Karandikar et al., 2021).

The literature also contains publications that do not employ ECE or DECE, such as the work of Mukhoti et al. (2020), Einbinder et al. (2022) and Tao et al. (2023). A literature overview on calibration in the context of neural networks is given by Wang (2024); Minderer et al. (2021) report on a thorough investigation of various factors in neural network design that influence the calibration of models.

3. Materials and Methods

The problem domain of binary classification and its associated measures of calibration is subtly different from the more prevalent multi-class classification tasks encountered in deep learning settings. In particular, calibration measures such as those of Hosmer and Lemeshow (2000) and its variants measure the agreement between the following two binned entities:

- Average true class 1 prevalence $\bar{y}_k = \frac{1}{|B_k|} \sum_{i \in B_k} y_i$, and
- average class 1 membership probability estimates $\bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} \hat{p}_i$,

with the notation introduced in Section 2. The original goodness-of-fit test statistic for the calibration of a logistic regression model (Hosmer and Lemeshow, 1980) can then be written as

$$\sum_{k=1}^{K} |B_k| \left[\frac{(\bar{p}_k - \bar{y}_k)^2}{\bar{p}_k} + \frac{\big((1-\bar{p}_k) - (1-\bar{y}_k)\big)^2}{(1-\bar{p}_k)} \right]. \quad (7)$$

This test statistic was empirically observed to follow a chi-squared distribution with $K-2$ degree of freedoms when evaluated on the training set, and K degrees of freedom when evaluated on the test set.

Note that the terms $\bar{y}_k$ and $\bar{p}_k$ differ from the terms used in the definitions of accuracy and confidence in Equations (1) and (2) in the following way:

- Accuracy measures the average agreement between predicted and true class, and is thus model-dependent. Average true class 1 prevalence $\bar{y}_k$ is model-independent.
- Confidence employs $\hat{p}'_i$, the highest probability estimate for a given case (the probability estimate of the most likely class), whereas estimated class-membership probability $\hat{p}_i$ for binary classification is always for class 1.

We can now modify ECE, as given in Equation (3), to use the terms particular to binary classification. We denote this variant by BECE, defined as

$$\text{BECE} := \sum_{k=1}^{K} \frac{|B_k|}{n} |\bar{y}_k - \bar{p}_k|, \quad (8)$$

and use it as a loss function, or as a component of a loss function, much the same way as ECE is used in deep learning research.

In summary, we set out to investigate how well the following loss functions perform, both in terms of calibration and discrimination, on two small sample data sets from the biomedical domain. As above, let $\hat{p}$ and y denote the vectors of estimated class 1 membership probabilities and the true class labels (0 or 1), respectively, for a data set of size n. The binary cross entropy loss function is included here both as a baseline reference, and as a component of other loss functions:

- *Binary cross entropy* $L_{\text{BCE}}(\hat{p}, y) := -\frac{1}{n} \sum_{i=1}^{n} \big[y_i \log(\hat{p}_i) + (1-y_i) \log(1-\hat{p}_i) \big]$
- *Binary expected calibration error* $L_{\text{BECE}}(\hat{p}, y) := \sum_{k=1}^{K} \frac{|B_k|}{n} |\bar{y}_k - \bar{p}_k|$,
- *Hosmer-Lemeshow calibration error* $L_{\text{HL}}(\hat{p}, y) := \sum_{k=1}^{K} |B_k| \left[\frac{(\bar{p}_k - \bar{y}_k)^2}{\bar{p}_k} + \frac{\big((1-\bar{p}_k) - (1-\bar{y}_k)\big)^2}{(1-\bar{p}_k)} \right]$.
- *Combined cross-entropy/expected calibration error* $L_{\text{BCE-BECE}}(\hat{p}, y) := \alpha L_{\text{BCE}}(\hat{p}, y) + (1-\alpha) L_{\text{BECE}}(\hat{p}, y)$.
- *Combined cross-entropy/Hosmer-Lemeshow error* $L_{\text{BCE-HL}}(\hat{p}, y) := \beta L_{\text{BCE}}(\hat{p}, y) + (1-\beta) L_{\text{HL}}(\hat{p}, y)$.

In the last two of these, the hyperparameters α and β weigh the relative contributions of the two error terms. These hyperparameters have to be estimated empirically, e.g. by cross-validation. In contrast, the hyperparameter K of numbers of bins is usually set to a fixed size of 10.

Note that there are two possible approaches to implementing the binning process necessary for calculating the calibration terms in the loss functions: Either divide the interval $[0, 1]$ into K equal-width parts, or divide the sorted probability estimate vector $\hat{p}$ into K parts of (nearly) equal size. We decided on the first alternative to avoid the non-differentiable sorting operation that would prevent gradient flow during the network training process.

All experiments were implemented in Python 3.11 and Pytorch 2.2.1. Final model assessment was performed with the `pycaleva` package (Weigl, 2022) that computes the area under the ROC curve, the Hosmer-Lemeshow and Pigeon-Heyse variants of a goodness-of-fit test, and the calibration belt figure of Finazzi et al. (2011).

We performed the experiments summarized in Section 4 on the following two data sets:

Acute Myocardial Infarction Data Set

This data set comprises information on 1253 patients with symptoms of acute myocardial infarction collected at the Edinburgh Royal Infirmary in Scotland in 1993–1994; the modeling task is to predict whether these patients have a heart attack. Patient data consists of 3 numerical and 29 binary features, with a class distribution of 274 positives, and 979 negatives. The numerical features were standardized to mean zero and standard devtion 1. The gold standard diagnosis of presence or absence of AMI was obtained by majority voting of three independent experts. More details on the data set can be found in the original publication (Kennedy et al., 1996).

Coronary Artery Disease Data Set

This second data set from the domain of cardiovascular disease prediction was collected by four study centers in the United States, Hungary and Switzerland between 1981 and 1987 (Detrano et al., 1989). It consists of 918 cases (410 negative, 508 positive) with 11 features, 5 of which are numerical, and 6 categorical. We removed two features

related to ST elevation, because these are highly correlated with the gold standard; including these makes the modeling task much easier. Again, the numerical features were standardized, and the categorical features one-hot encoded. For this data set, the task is to predict the presence or absence of coronary artery disease.

4. Results and Discussion

This section summarizes our experiments of all combinations of the five loss functions and the two data sets. Due to the small size of the data sets, the neural network model architecture was kept very simple, at just one hidden layer with 20 neurons. We employed the Adam optimizer with a constant learning rate of $\eta = 0.0001$, and did not do any L_1 or L_2 regularization. Since the goal of our experiments is not to find the optimal model, but to check what influence various loss functions have on performance metrics, we found this choice to be justified. All numbers below were obtained on a test set that consisted of a random subset containing 30% of the original data set; the remaining 70% made up the training set. Note that the calibration curves in all the figures in this section employ the *deciles of risk* data grouping strategy, where the bins are made up of an equal number of $\hat{p}_i$ estimates. Therefore, the points on the calibration curves are not equally spaced on the x-axis.

The `pycaleva` package reports on a number of different metrics, which are also included in the figures generated by this package. These metrics are:

- The *Brier score* $\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2$, the average sum-of-squares difference between model outputs and ground truth (Brier, 1950);
- the *adaptive calibration error* ACE which, for binary classification, corresponds to BECE with binnings according to sorted probabilities, and not fixed cutoffs (Nixon et al., 2019);
- the *maximum calibration error* MCE := $max_{1\leq k\leq n}|\bar{p}_k - \bar{y}_k|$, i.e., the largest difference between average estimated class 1 membership and true class 1 membership over all bins, and
- the *area within LOWESS curve*, i.e., the area between a nonparametrically smoothed curve of the $(\hat{p}_i, y_i)$ scatter plot and the diagonal line indicating perfect calibration (Weigl, 2022).

4.1. Results on Acute Myocardial Infarction Data Set

We first trained our model using the binary cross-entropy loss as a baseline. In terms of calibration, one can observe that overtraining only has a small effect on the area under ROC curve, but does greatly effect the calibration of the model. This can be seen in Figure 1, where the left plot shows good calibration, and the right plot poor calibration due to overtraining of the model. In contrast to this, training on the expected calibration error loss showed entirely different time evolutions, with the results reversed compared to binary cross-entropy loss (see Figure 2, with the left plot displaying undertraining, and the right plot good model fit) . The same characteristics can be seen in the plots for the Hosmer-Lemeshow loss (in Figure 3), where models are still undertrained after 4000 epochs (left plot), but do not exhibit overtraining even after a much larger number of epochs (right plot). As a last series of experiments, we combined binary cross-entropy error with the calibration errors for two new loss functions, as described above. The relative weightings of the error and calibration terms were set to $\alpha = 0.5$ and $\beta = 0.5$ in these experiments. Figure 4 shows the results of combining cross-entropy error with the binary expected calibration error; one can observe sufficient calibration after 4000 epochs (left plot), but overtraining starts already after 8000 epochs (right plot). The situation is reversed when combining cross-entropy error with the Hosmer-Lemeshow calibration term, as can be seen in Figure 5: After 4000 epochs, the network is still undertrained (left plot), whereas it shows decent calibration and no overtraining even after 20000 epochs (right plot).

The metrics for all these training runs are summarized in Table 1. Note that some error functions tend to overtraining (L_{BCE}, $L_{\text{BCE-BECE}}$), while others are more robust and can also be trained with a higher number of epochs (L_{BECE}, L_{HL}, $L_{\text{BCE-HL}}$). However, these error functions need more training time, as they exhibit poor performance with a lower number of training epochs.

4.2. Results on Coronary Artery Disease Data Set

We ran the same series of experiments also on our second data set. Here, we do not show the same calibration curves an in Section 4.1, but rather highlight how the calibration belts introduced by Finazzi et al. (2011) can provide an alternative visual assessment of a model's calibration. A total of four such plots are given in Figures 6 and Figures 7; in the left plots, one can observe visually that the models are not well calibrated.

All the performance metrics are summarized in Table 2. It can be seen that compared to the results of Table 1, underfitting the models requires training with far fewer epochs, i.e., there is a wider range of possible epoch values in which the models have adequate calibration. What is similar to the results on the acute mycardial infarction data set is the discrepency between discrimination and calibration measures: while some models are well calibrated and others are not, the discriminatory ability of the models, as measured by AUC curve, is in most cases on the same level, with the poorly calibrated models only a few percentages below the well-calibrated models.

4.3. Discussion

In the last years, there have been a number of publications investigating the calibration of (mostly deep) neural network models (Guo et al., 2017; Kumar et al., 2018; Mukhoti et al., 2020; Karandikar et al., 2021; Bohdal et al., 2023) on

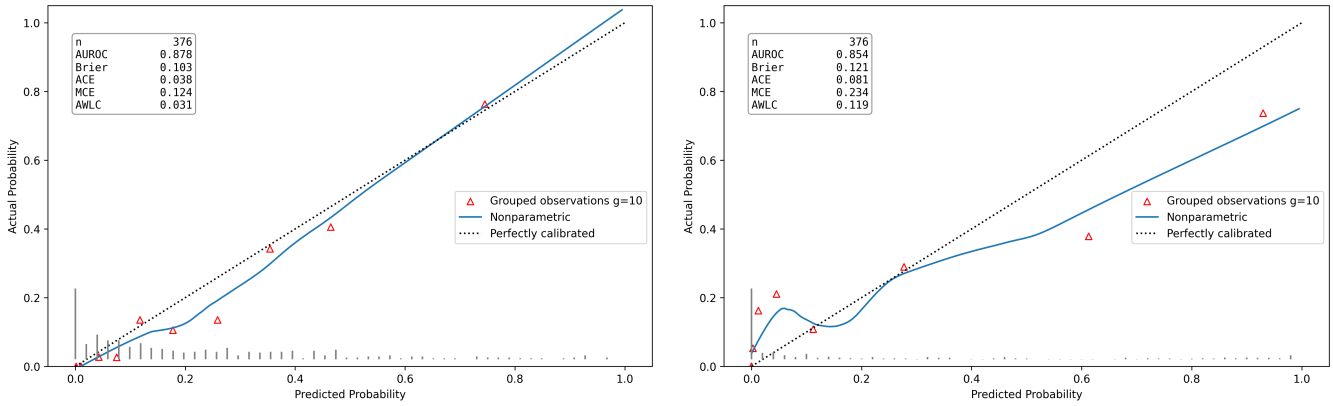

Figure 1. Two calibration curves of our neural network trained on the acute myocardial infarction data set using the binary cross-entropy error function. The left plot shows calibration after 4000 epochs, and the right after 8000 epochs. One can clearly observe calibration getting worse, which is also evidenced by the metrics given in the plots. The red triangles are the points $(\hat{p}_k, \hat{y}_k)$ for $K = 10$ bins; the tick marks along the x-axis represent the distribution of probability estimates $\hat{p}_i$ for $n = 376$ data points.

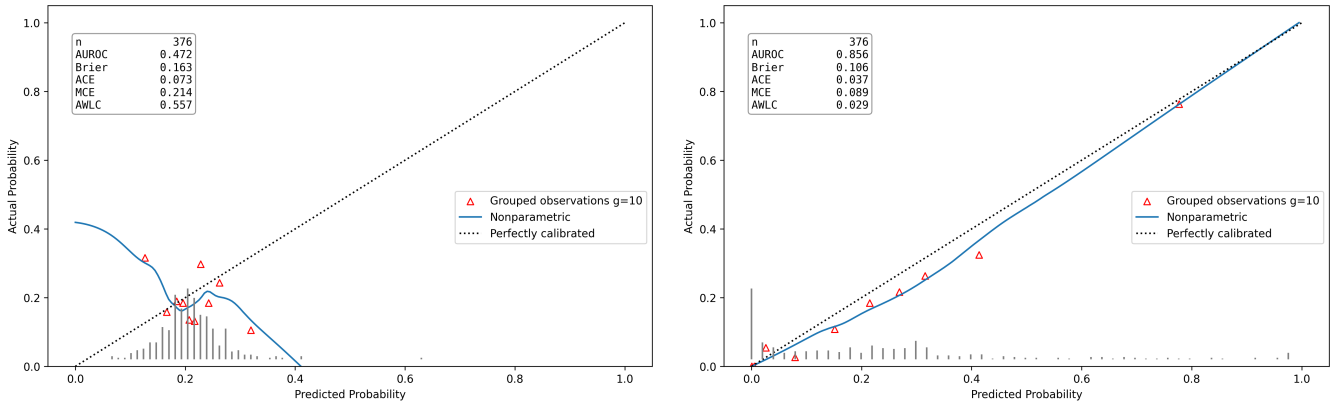

Figure 2. Two calibration curves of our neural network trained on the acute myocardial infarction data set, but now using the expected calibration error loss function. The left plot shows calibration after 4000 epochs, the right after 20000 epochs.

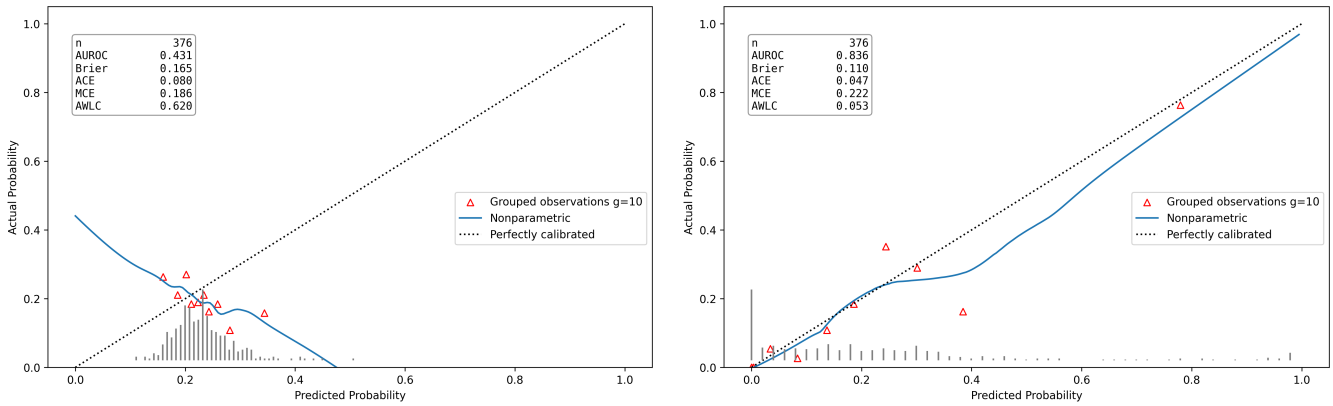

Figure 3. Two calibration curves of our neural network trained on the acute myocardial infarction data set with the Hosmer-Lemeshow loss function. The left plot shows calibration after 4000 epochs, the right after 20000 epochs.

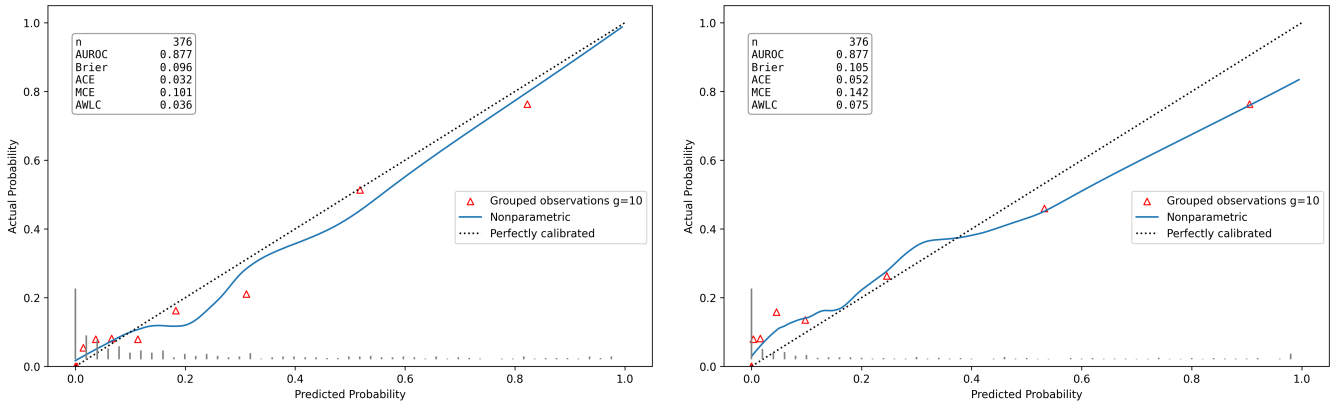

Figure 4. Two calibration curves of our neural network trained on the acute myocardial infarction data set with a combination of the binary cross-entropy and the binary expected calibration error loss loss functions. The left plot shows calibration after 4000 epochs, the right after 8000 epochs.

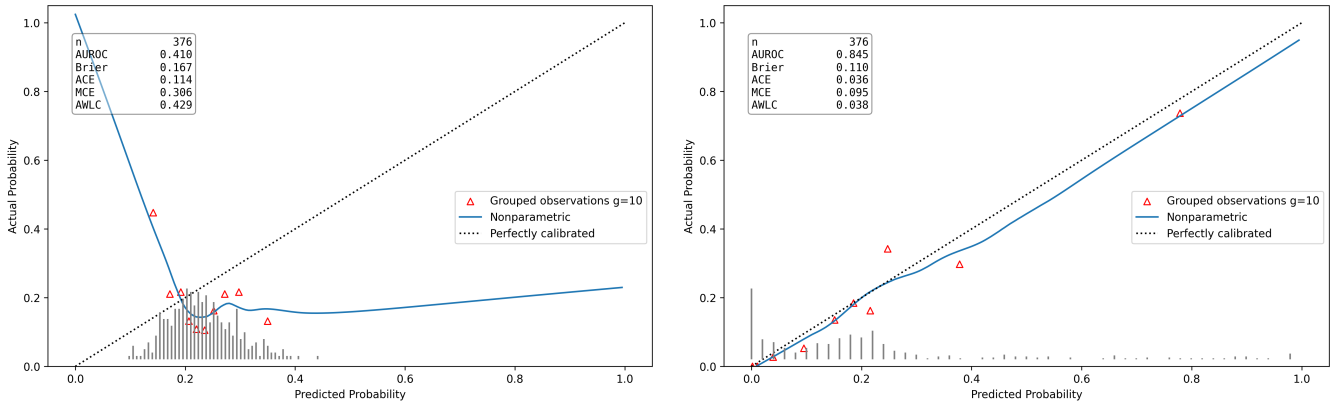

Figure 5. Two calibration curves of our neural network trained on the acute myocardial infarction data set with a combination of the binary cross-entropy and the Hosmer-Lemeshow calibration error loss loss functions. The left plot shows calibration after 4000 epochs, the right after 20000 epochs.

Table 1. Performance metrics of neural network models trained using different loss functions on the acute myocardial infarction data set. In addition to the number of epochs for training the network, the metrics listed are AUC (area under the ROC curve), ACE (expected calibration error for $K = 10$ groups, same as ECE for binary classifiers), p-value (of the Pigeon-Heyse improvement to the Hosmer-Lemeshow goodness-of-fit test), and AWLC (area within LOWESS curve). The definitions of the loss functions are given in Section 3.

	good calibration					poor calibration				
Metric	L_{BCE}	L_{BECE}	L_{HL}	$L_{\text{BCE-BECE}}$	$L_{\text{BCE-HL}}$	L_{BCE}	L_{BECE}	L_{HL}	$L_{\text{BCE-BECE}}$	$L_{\text{BCE-HL}}$
epochs	4000	20000	20000	4000	20000	8000	4000	4000	8000	4000
AUC	0.878	0.856	0.836	0.877	0.845	0.854	0.472	0.431	0.877	0.410
ACE	0.038	0.037	0.047	0.032	0.036	0.081	0.073	0.080	0.052	0.114
p-value	0.720	0.265	0.242	0.480	0.856	< 0.0001	0.005	0.046	< 0.0001	< 0.0001
AWLC	0.031	0.029	0.053	0.036	0.038	0.119	0.557	0.620	0.075	0.429

Table 2. Performance metrics of neural network models trained using different loss functions on the coronary artery disease data set. All notations are the same as in Table 1.

	good calibration					poor calibration				
Metric	L_{BCE}	L_{BECE}	L_{HL}	$L_{\text{BCE-BECE}}$	$L_{\text{BCE-HL}}$	L_{BCE}	L_{BECE}	L_{HL}	$L_{\text{BCE-BECE}}$	$L_{\text{BCE-HL}}$
epochs	4000	20000	20000	8000	8000	10000	1000	1000	1000	1000
AUC	0.914	0.906	0.912	0.903	0.897	0.884	0.882	0.615	0.862	0.531
ACE	0.046	0.068	0.042	0.063	0.073	0.087	0.102	0.082	0.105	0.082
p-value	0.777	0.219	0.784	0.167	0.349	< 0.0001	0.049	0.494	0.023	0.159
AWLC	0.039	0.027	0.045	0.044	0.038	0.060	0.091	0.160	0.105	0.146

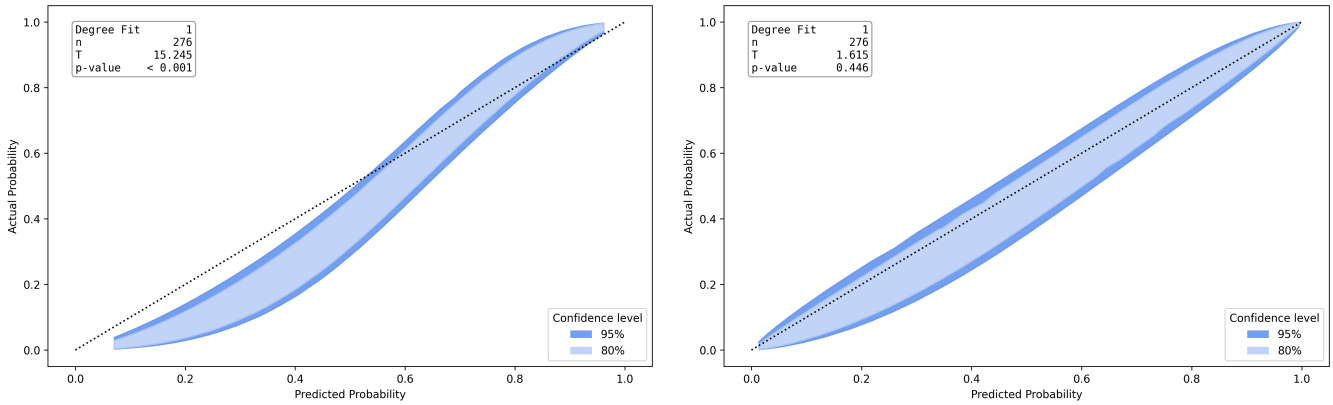

Figure 6. Two calibration belts of our neural network trained on the coronary artery disease data set with the binary expected calibration error loss functions. The left plot shows calibration after 1000 epochs, the right after 20000 epochs.

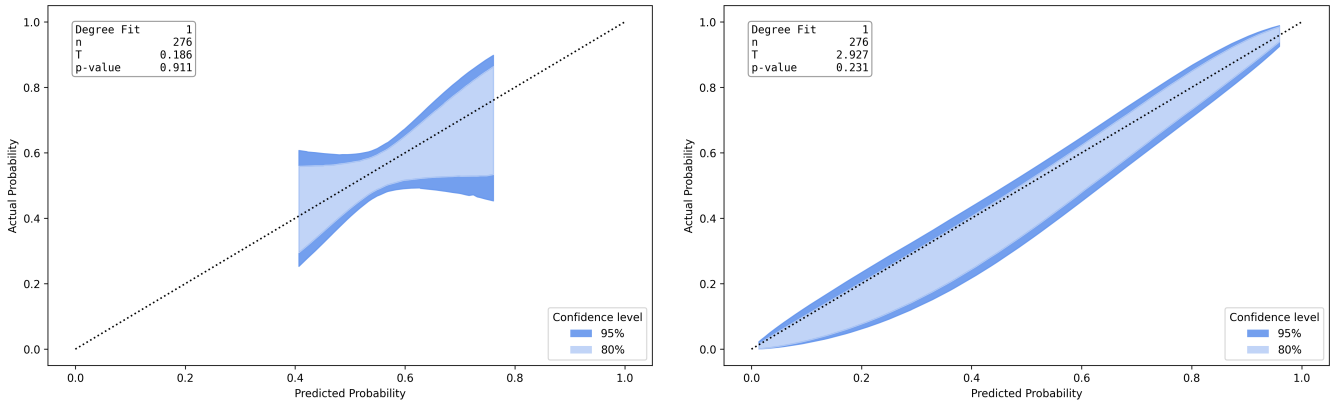

Figure 7. Two calibration belts of our neural network trained on the coronary artery disease data set with the Hosmer-Lemeshow loss function. The left plot shows calibration after 1000 epochs, the right after 20000 epochs.

multi-class classification tasks. In this work, we focused on binary classification and demonstrated that in this case, expected calibration error ECE can directly be used as a loss function for neural network training. We saw that models trained on these loss functions exhibit characteristics similar to the ones trained on binary cross-entropy loss, the loss function more traditionally used for these tasks.

In particular, what can be observed from a synopsis of the results on both data sets is the following:

- Models trained with binary cross-entropy loss function can be overtrained more easily compared to those trained on calibration losses.
- Visual inspection of calibration curves and belts is more informative than blindly trusting the p-values of calibration tests. This can be seen in the left plot of Figure 3, where the model is clearly not calibrated well, but the p-value of the Pigeon-Heyse test is barely significant at 0.046; and in the left plot of Figure 7, with a p-value of 0.494, but very poor calibration due to a narrow range of probability estimate $\hat{p}_i$.
- Models trained on calibration losses nevertheless exhibit good AUC values.
- It is easily possible to combine binary cross-entropy loss with calibration loss functions, and the performance of models trained with these losses is comparable to the performance of models trained only on either cross-entropy or calibration loss.

5. Conclusion

This work focuses on loss functions that include calibration terms in order to emphasize the importance of calibration in properly trained predictive models. This approach is particularly important in the domain of biomedical data analysis, where diagnosis and treatment decisions are based in no small part on the probabilities of putative findings. While a binary threshold of 0.5 may treat all probability estimates larger than that as positive, it is clear that there is a large difference between estimates of 0.51 and 0.99. Physicians may act differently when presented with these different values; acting on this difference is sensible only when these values are an accurate representation of the ground truth probabilities.

We demonstrated empirically that models trained on loss functions that optimize for calibration performance are able to achieve results comparable to models trained on binary cross-entropy error. The novelty of our approach lies in making calibration error a central component of our loss function; this is not traditionally done when training predictive models, even in fields where well-calibrated models are of primary importance.

We note that these investigations will still need to be expanded to more diverse and larger data sets, possibly also from other domains, before the findings reported here can be seen as conclusive evidence that calibration losses are viable alternatives to more established loss functions. Future work can expand and improve upon the results presented here by developing other variants of incorporating calibration terms into loss functions, and evaluating these variants in a wider range of application areas.

References

Austin, P. and Steyerberg, E. (2014). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33(3):517–535.

Bohdal, O., Yang, Y., and Hospedales, T. (2023). Meta-calibration: Learning of model calibration using differentiable expected calibration error. *Transactions on Machine Learning Research*, 08/2023.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78:1–3.

Cook, N. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115:928–935.

Copas, J. (1983). Plotting *p* against *x*. *Applied Statistics*, 32(1):25–31.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, W., Schmid, J., Sandhu, S., Guppy, K., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5):304–310.

Einbinder, B.-S., Romano, Y., Sesia, M., and Zhou, Y. (2022). Training uncertainty-aware classifiers with conformalized deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 22380–22395.

Finazzi, S., Poole, D., Luciani, D., Cogo, P., and Bertolini, G. (2011). Calibration belt for quality-of-care assessment based on dichotomous outcomes. *PLoS One*, 6(2):e16110.

Flach, P., Hernández-Orallo, J., and Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning*, pages 657–664, Bellevue, WA, USA.

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

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123.

Hanley, J. and McNeil, B. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hosmer, D., Hosmer, T., Cessie, S. L., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression, 2nd Edition*. Wiley-Interscience Publication.

Janssens, A. and Martens, F. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*, 49(4):1397–1403.

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., and Roelofs, B. (2021). Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 29768–29779.

Kennedy, R., Burton, A., Fraser, H., McStay, L., and Harrison, R. (1996). Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *European Heart Journal*, 17(8):1181–1191.

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2805–2814.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, 21:3789–3801.

Lasko, T., Bhagwat, J., Zhou, K., and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38(5):404–415.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. (2021). Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. (2020). Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2901–2907.

Nattino, G., Lemeshow, S., Phillips, G., Finazzi, S., and Bertolini, G. (2017). Assessing the calibration of dichotomous outcome models with the calibration belt. *The Stata Journal*, 17(4):1003–1014.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–41.

Pigeon, J. G. and Heyse, J. F. (1999a). A cautionary note about assessing the fit of logistic regression models. *Journal of Applied Statistics*, 26(7):847–853.

Pigeon, J. G. and Heyse, J. F. (1999b). An improved goodness of fit statistic for probability prediction models. *Biometrical Journal*, 41(1):71–82.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Silva Filho, T. M., Song, H., Nieto, M. P., Santos-Rodriguez, R., Kull, M., and Flach, P. A. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260.

Tao, L., Dong, M., and Xu, C. (2023). Dual focal loss for calibration. In *Proceedings of the 40th International Conference on Machine Learning*, page 33833–33849.

Van Calster, B., McLernon, D., van Smeden, M., Wynants, L., and Steyerberg, E. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Wang, C. (2024). Calibration in deep learning: A survey of the state-of-the-art.

Weigl, M. (2022). pycaleva: A framework for calibration evaluation of binary classification models. https://github.com/MartinWeigl/pycaleva.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

Zou, K., O'Malley, A., and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115:654–657.