Proceedings of Machine Learning Research 230:1–20, 2024 Conformal and Probabilistic Prediction with Applications

Calibrated Explanations for Multi-Class

Tuwe Löfström TUWE.LOFSTROM@JU.SE
Department of Computing
Jönköping University
Sweden

Helena Löfström HELENA.LOFSTROM@JU.SE
Department of Computing
Jönköping University
Sweden

Ulf Johansson ULF.JOHANSSON@JU.SE
Department of Computing
Jönköping University
Sweden

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Calibrated Explanations is a recently proposed feature importance explanation method providing uncertainty quantification. It utilises Venn-Abers to generate well-calibrated factual and counterfactual explanations for binary classification. In this paper, we extend the method to support multi-class classification. The paper includes an evaluation illustrating the calibration quality of the selected multi-class calibration approach, as well as a demonstration of how the explanations can help determine which explanations to trust.

Keywords: Calibrated Explanations, Multi-Class, Venn-Abers, XAI

1. Introduction

Explaining the process of machine learning models generating their predictions is crucial for increasing users' appropriate trust in high-stakes applications, such as defence (Gunning et al., 2021). A user who trusts the model too much accepts both correct and incorrect predictions with a risk of making low-quality decisions. With a measure of confidence attached to the predictions and the features most relevant to the predictions, users are expected to get an understanding of the internal reasoning of the model, and the possibilities for high-quality decisions increase. However, both the confidence of the predictions and the weights attached to the features in the explanations could become misleading due to the reality of often poorly calibrated underlying models (Löfström et al., 2023).

External calibration methods can help convert internal confidences from poorly calibrated machine learning models into well-calibrated probabilities which are useful when guiding users' decisions. Two of the most well-known calibration techniques are Platt scaling (Platt, 1999) and isotonic regression (Zadrozny and Elkan, 2001), which fit a logistic or isotonic function to the confidence measures and the true targets on a calibration set. This approach requires that the underlying model is a *scoring classifier*. For scoring classifiers, a larger belief in the positive class is indicated by higher scores, making them applicable to two-class problems only. Another calibration technique applied to scoring classifiers is

Venn-Abers predictors (Vovk and Petej, 2012). The output is multi-probabilistic predictors with unique validity properties. Although these multi-probabilistic predictors are highly informative, they only apply to two-class problems.

Multi-class problems are, in contrast to a two-class situation, when more than two classes exist, although each instance should be given precisely one class label. The standard approach to multi-class problems has been to use either a one-vs-all or an all-vs-all schema, requiring training multiple models, before calibrating each class and then aggregating the results into probability estimates. Additionally, due to this drawback, the possibility to analyse and explain a single model is no longer available (Johansson et al., 2021a).

In a recently proposed method (Johansson et al., 2021b), multi-class calibration is performed on inherently multi-class models capable of predicting not only a label but also a confidence measure for each class label. The calibration is done by first calibrating each class label, one at a turn, in a one-vs-all calibration. The class label with the highest calibrated probability estimate is used as the predicted class and the multi-probabilistic prediction from Venn-Abers for the predicted class against all other classes is also returned.

In this study, we integrate the previously proposed method for multi-class calibration (Johansson et al., 2021b) into *Calibrated Explanations*, a recently proposed feature importance explanation method (Löfström et al., 2024) providing well-calibrated explanations with uncertainty quantification of both prediction and feature weights.

In the next section, we describe the theoretical foundation necessary for the proposed solution. Section 3 outlines the proposed solution and describes how multi-class support is integrated into Calibrated Explanations. In Section 4, we introduce the suggested approach and describe the experimental setup, including the data sets used. In Section 5, we first demonstrate the approach and then present and analyse the results obtained when using each of the three types of underlying models. Finally, in Section 6, we give the main conclusions and outline some directions for future work.

2. Background

2.1. Probabilistic Prediction

A probabilistic predictor outputs a predicted class label and a probability distribution over the labels. Validity refers to the extent to which the predicted probability distributions align with the statistical tests, as evidenced by subsequent observations of the labels. Gammerman et al. (1998) showed that validity can not, in a general sense, be achieved for probabilistic prediction. *Calibration*, on the other hand, is when the following holds:

$$p(c_j \mid p^{c_j}) = p^{c_j}, \tag{1}$$

where p^{c_j} is the probability estimate for class label c_j. This implies that predictions with a probability estimate of 0.95 should be accurate approximately 95% of the time. In essence, the accuracy of observed outcomes should align with the predicted probabilities. While most predictive models produce probability estimates, these estimates are often poorly calibrated. Poor calibration can be handled by applying an external calibration method. Platt scaling (Platt, 1999) and isotonic regression (Zadrozny and Elkan, 2001) are the two most frequently used. The standard procedure for these external methods is to perform the actual calibration on a separate part of the available labelled data called the *calibration set*.

2.2. Venn-Abers Predictors

Venn predictors are probabilistic predictors that restrict the statistical tests for validity to calibration and output multiple probabilities for each label, with one of them being the valid one (Vovk et al., 2004). In Venn prediction, instances are divided into a number of *categories*, based on a so-called *Venn taxonomy*. The relative frequency for each class label is calculated as the calibrated probability within the category of the test instance. To achieve validity, the test instance is also included in the calculation. As the true label is unknown for test instances, every possible label is assigned in turn, resulting in C multi-probability distribution, where C is the number of possible labels. In order to calibrate an underlying pre-trained multi-class model, inductive Venn prediction (Lambrou et al., 2015) is used, setting aside a calibration set that is divided into categories according to the Venn taxonomy. As outlined below, these multi-probabilistic predictions can be transformed into probability intervals for each label. The size of these intervals provides a basic measure of the confidence in the estimation.

Deciding on the taxonomy to use is vital for Venn predictors. Venn-Abers prediction (Vovk and Petej, 2012) offers automated taxonomy optimisation using isotonic regression, providing optimised probability intervals for binary classification. One isotonic regressor is trained for each of the negative and the positive classes, defining the lower and upper bound of the multi-probability distribution. Since the instance must be one or the other, it follows that the interval must contain the true probability. Even though the true label is not known, the width and location of the interval can provide rich support for decision-making. A smaller interval indicates higher certainty about the prediction, while a larger interval indicates more uncertainty.

Let us assume a test object x_{n+1} for which a prediction is sought, let $Z = \{z_1, \ldots, z_n\}$, where $n = l + q$, be a training set. Each instance $z_i = (x_i, y_i)$ consists of two parts, an object x_i and a target y_i. Normally, calibration requires a separate calibration set, motivating a split of the training set into a proper training set Z_l with l instances, and a calibration set $Z_q = \{z_1, \ldots, z_q\}$[1]. Use a scoring classifier, i.e., a classifier restricted to two-class problems that, when predicting a test object x_i, outputs a *prediction score* s_i, where a higher value indicates a larger belief in label 1. To predict a class label from a scoring classifier, the prediction is 1 if $s > t$, and otherwise 0. A Venn-Abers predictor requires a scoring classifier as the underlying model. Instead of using a fixed threshold, an increasing function g is fitted using a number of prediction scores with known true targets. This function, $g(s)$, can then be interpreted as the probability that the label for x is 1, i.e., it is a calibrator. Venn-Abers predictors use isotonic regression (Zadrozny and Elkan, 2001) for the fitting. Normally, the score s is defined as the probability estimate for the positive class from a classifier h. Inductive Venn-Abers prediction for binary classification follows these steps:

1. To derive the isotonic calibrator g_0, use $\{(s_1, y_1), \ldots, (s_q, y_q), (s_{n+1}, 0)\}$ and to derive the isotonic calibrator g_1, use $\{(s_1, y_1), \ldots, (s_q, y_q), (s_{n+1}, 1)\}$.

1. As we assume random ordering, the calibration set is indexed $1, \ldots, q$ rather than $l+1, \ldots, n$, for indexing convenience.

2. The probability interval for $y_{n+1} = 1$ is defined as $[g_0(s_{n+1}), g_1(s_{n+1})]$ (henceforth referred to as $[p_{low}, p_{high}]$, representing the lower and upper bounds of the interval).

3. The regularised probability estimate for $y_{n+1} = 1$, minimising the log loss (Vovk and Petej, 2012), can be defined as:

$$p = \frac{p_{high}}{1 - p_{low} + p_{high}}$$

In summary, Venn-Abers produces a calibrated (regularised) probability estimate p together with a probability interval with a lower and upper bound $[p_{low}, p_{high}]$.

2.3. Multi-Class Calibration

Zadrozny and Elkan (2002) suggested a one-vs-all approach for multi-class calibration where a binary classifier is trained for each class label before calibrating. This approach is applicable to any underlying model and any binary calibration technique. How to combine the calibrated estimates is not altogether obvious.

A pair-wise (all-vs-all) approach was suggested by Manokhin (2017). This approach was applied using standard and cross Venn-Abers for multi-class calibration on models built by logistic regression, support vector machines and neural networks. They showed that probabilistic models calibrated with Venn-Abers were generally better calibrated than the uncalibrated predictors.

Wenger et al. (2020) points out that most modern classifiers are inherently multi-class and are thus an alternative to the traditional approach with one-vs-all or all-vs-all schemes. This allows for the predictions from the inherently multi-class model on a calibration set can be used directly for calibration, avoiding training a number of models before calibrating.

In two recent papers, two additional approaches were suggested (Johansson et al., 2021a,b). Both approaches operate on inherently multi-class models, providing a set of confidence measures for all labels. In the first approach (Johansson et al., 2021a), a probability estimate for a test instance is produced using the label predicted by the underlying model as the positive class, and all other labels are regarded as belonging to the negative class. After this, the actual calibration is performed in the standard way for Venn-Abers, resulting in a probability estimate for the positive class, i.e., the label predicted by the underlying model. A drawback of this approach is that the calibrated probability for the predicted class is not guaranteed to be the highest calibrated probability among all classes. In the second approach (Johansson et al., 2021b), a one-vs-all calibration is done, calibrating once for each label, using that label as the positive class and all the remaining classes as the negative. The label with the highest calibrated probability is used as the predicted class. It will often be the same class as the class predicted by the underlying model, but there is no guarantee that the calibrated prediction and the prediction from the underlying model will be the same.

2.4. Calibrated Explanations

Löfström et al. (2024) introduced a local explanation method providing feature importance with uncertainty quantification[2]. The initial release supported binary classification and support for regression has been proposed (Löfström et al., 2023). Calibrated Explanations produce instance based explanations, which can be either factual or counterfactual. A *factual explanation* is composed of a *calibrated prediction* from the underlying model accompanied by an *uncertainty interval* and a collection of *factual feature rules*, each composed of a *feature weight with an uncertainty interval* and a *factual condition*, covering that feature's instance value. *Counterfactual explanations* contain a collection of *counterfactual feature rules*, each composed of a *prediction estimate with an uncertainty interval* and a *counterfactual condition*, covering alternative instance values for the feature. For binary classification, the explanation explains the calibrated probability estimate (and its level of uncertainty) for the positive class.

The core of the algorithm is defined based on a numeric estimate and a lower and an upper bound, defining an uncertainty interval for the numeric estimate. For binary classification, the probability estimate of the positive class is calibrated using a Venn-Abers calibrator (Vovk and Petej, 2012), producing a lower and an upper bound for the calibrated probability estimate (using a regularised mean of these bounds as the numeric estimate). Slightly more formally, Calibrated Explanations creates its explanations for binary classification in the following manner:

1. Use a scoring classifier h, trained using the proper training set Z_l, producing score s when predicting an object $h(x)$.

2. To get a calibrated prediction for a test object[3] x, apply a Venn-Abers calibrator to get a calibrated prediction p and uncertainty interval $[p_{low}, p_{high}]$ for the positive class.

3. For each feature f:

   (a) Change the value of feature f in a systematic way (use each category for categorical features and a sample of values for numerical values[4]), producing slightly perturbed versions of object x. Use the calibrator from step 2 to estimate the (averaged) prediction $p.f$ and uncertainty intervals $[p_{low}.f, p_{high}.f]$.

   (b) The feature importance for feature f is defined as the difference between the calibrated prediction p, achieved on the original object x, and the estimated (averaged) calibrated prediction $p.f$, achieved on the perturbed versions of x.

   (c) The uncertainty intervals for the feature importance is defined analogously by calculating the difference between p and the uncertainty intervals $[p_{low}.f, p_{high}.f]$ for the perturbed versions of x.

   (d) A factual condition is formed, with `feature = categorical instance value` for categorical features and `feature` $\leq$ `threshold` or `feature` $>$ `threshold`

2. Calibrated Explanations can be installed using, e.g., `pip install calibrated-explanations` or accessed at github.com/Moffran/calibrated_explanations.
3. The index $n + 1$ is omitted to reduce clutter.
4. For details on how the sampling is done, see Löfström et al. (2024).

for numerical features. The `threshold` is defined so that the factual condition incorporates the numerical instance value for that feature. Since the factual condition must always include the feature value, only one factual condition is formed for each feature.

As opposed to the calibration performed on the original object x in step 2, where exchangeability is expected between calibration and test instances, no guarantees can be given for the calibration performed on the perturbed instances. Even though the feature values are defined to match the feature values in the calibration set, the feature values used to exchange the original feature values may in some cases result in perturbed instances that are not exchangeable with the calibration set.

2.5. Counterfactual Explanations

The main difference when using Counterfactual Calibrated Explanations is that the conditions are counterfactual rather than factual. This means that the categorical counterfactual condition is using the $\neq$ condition. Numerical counterfactual conditions are defined as $\leq$-conditions and $>$-conditions but with conditions that exclude the instance value. One counterfactual condition is formed for each alternative categorical feature value, and numerical counterfactual conditions can allow both counterfactual $\leq$-conditions and $>$-conditions excluding the feature value to be formed.

3. Contribution

In order for Calibrated Explanations to explain multi-class problems, the multi-class calibration approach suggested in Johansson et al. (2021b) is applied to step 2 in the description above. In other words, for each class label c, a Venn-Abers calibrator is initialised using that label as the positive class and the probability estimate for that label as the score s^c from the underlying model. More formally, the multi-class approach is as follows:

1. For each class label $c \in C$, where C is the set of possible class labels:
   (a) Use $\{(s_1^c, y_1 = c), \ldots, (s_q^c, y_q = c), (s_{n+1}^c, 1)\}$, where $y_i = c$ is 1 when true and 0 otherwise, to derive the isotonic calibrator g_c and use $\{(s_1^c, y_1 = c), \ldots, (s_q^c, y_q = c), (s_{n+1}^c, 0)\}$ to derive the isotonic calibrator $g_{\neg c}$.
   (b) The probability interval for $y_{n+1} = c$ is defined as $[g_{\neg c}(s_{n+1}^c), g_c(s_{n+1}^c)]$ (henceforth referred to as $[p_{low}^c, p_{high}^c]$, representing the lower and upper bounds of the interval).
   (c) The regularised probability estimate for $y_{n+1} = c$, minimising the log loss (Vovk and Petej, 2012), can be defined as:

$$p^c = \frac{p_{high}^c}{1 - p_{low}^c + p_{high}^c}$$

2. Output the class label $c = \texttt{argmax}_{c \in C}\ p^c$ together with the probability interval $[p_{low}^c, p_{high}^c]$ and the regularised probability p^c. Also output the Venn-Abers calibrator defined using g_c and $g_{\neg c}$ to be used in step 3 in the description of Calibrated Explanations.

```python
from calibrated_explanations import WrapCalibratedExplainer
# Load and pre-process your data
# Divide it into proper training, calibration, and test sets

# Initialize the WrapCalibratedExplainer with your model
model = WrapCalibratedExplainer(ModelOfYourChoice())

# Train your model using the proper training set
model.fit(X_proper_training, y_proper_training)

# Calibrate your model using the calibration set
model.calibrate(X_calibration, y_calibration)

# Create and plot factual explanations
factual_explanations = model.explain_factual(X_test)
factual_explanations.plot_all()
factual_explanations.plot_all(uncertainty=True)

# Create and plot counterfactual explanations
counterfactual_explanations = model.explain_counterfactual(X_test)
counterfactual_explanations.plot_all()
```

Figure 1: Code example on using *calibrated-explanations*

The proposed solution necessitates the initiation of one Venn-Abers for each class, resulting in a complexity for the initialisation of $O(|C|\ q \log q)$, compared to $O(q \log q)$ for binary classification. For each instance, one Venn-Abers prediction is also made for each class once. Once the class predicted by Venn-Abers is determined, i.e., the class with the highest regularised probability, only the Venn-Abers calibrator for that class is used to construct the explanations, making the complexity of Calibrated Explanations for multi-class only slightly worse than Calibrated Explanations for binary classification. See Löfström et al. (2024) for further details on the complexity of Calibrated Explanations.

The proposed improvement was introduced into the existing implementation of *calibrated-explanations* in version v0.3.2. Calibrated Explanations with multi-class is using identical function calls as for binary classification. Figure 1 show an example on how to initialise a `CalibratedExplainer` and create factual and counterfactual explanations from a trained model.

4. Method

In the suggested approach, an underlying model is first trained using a proper training set before a `CalibratedExplainer` is initiated using the model and the calibration set. During the initialisation, one Venn-Abers calibrator is initialised for each class. When explaining a test object, the one-vs-all calibration procedure described in Section 3 determines the

predicted class c and returns the Venn-Abers calibrator for that class which is used to generate the factual or counterfactual feature rules.

The approach is evaluated in two ways. In Section 5.1, the multi-class approach suggested in Johansson et al. (2021b) is evaluated on a number of data sets to illustrate its ability to calibrate. In Section 5.2, the *calibrated-explanations* integration of multi-class calibration is demonstrated on a number of instances.

4.1. Experimental setup

To illustrate the multi-class calibration, two setups were compared, as described below:

- No external calibration (NoCal): Uses the output from the underlying model as probability estimates. Since this setup requires no calibration set, all available labelled data were used for inducing the models.
- Venn-Abers (VA): The model is induced using the proper training set, setting aside a calibration set used for calibration. For each test instance, each class label is calibrated in a one-vs-all calibration as described above, and the label c with the highest probability p^c is outputted together with the Venn-Abers defined using g_c and $g_{\neg c}$. When comparing Venn-Abers calibrations to the uncalibrated model, the regularised value p^c is used.

The main purpose of the comparison is to provide an insight into how the interval of the Venn-Abers calibration work. Consequently, the evaluation is limited to the two setups above, as additional calibration techniques would obfuscate the purpose.

In the experimentation, the `RandomForestClassifier` from *scikit-learn* was used and all parameter values were left at the default settings. The `VennAbers` from *calibrated-explanations* is used as a Venn-Abers calibrator. For the evaluation, standard 10x10-fold stratified cross-validation was used. For Venn-Abers, the proper training set consisted of 2/3 of all the training instances and the calibration set of 1/3. As mentioned, all training data was used for generating the non-calibrated models. In the experiments, 19 publicly available multi-class data sets from the UCI repository (Dua and Graff, 2017) were used. The data sets characteristics are presented in Table 1, where *#class* is the number of classes, *#inst.* is the number of instances and *#attrib.* is the number of input attributes.

For the evaluation, accuracy and area under the ROC-curve (AUC) are used to measure the predictive performance. Log losses and the expected calibration error (ECE) are used to evaluate the calibration.

The log loss is calculated using

$$\lambda_{log} = \begin{cases} -\log p^c & \text{if correct} \\ -\log(1 - p^c) & \text{if incorrect} \end{cases} \tag{2}$$

where log is the binary logarithm and p^c the estimate for the label predicted by Venn-Abers. The log loss function used (from scikit-learn) make sure that the probabilities are never exactly 0 or 1 to avoid infinite results.

ECE is calculated by dividing the probability estimates for the predicted class into M (here $M = 10$) equally sized bins, before taking a weighted average of the absolute differences

Table 1: Data sets

Data set	#class	#inst.	#attrib.
balance	3	625	4
cars	4	1728	6
cmc	3	1473	9
cool	3	768	8
glass	6	214	9
heat	3	768	8
image	7	2310	19
iris	3	150	4
steel	7	1941	27
tae	3	151	5
user	5	403	5
wave	3	5000	40
vehicle	4	846	18
whole	3	440	7
wine	3	178	13
wineR	6	1599	11
wineW	7	4898	11
vowel	11	990	11
yeast	10	1484	8

between the mean of the prediction probabilities (mop) and the fraction of correct (foc) predictions, as described in Equation (3):

$$ECE = \sum_{i=1}^{M} \frac{|B_i|}{n} \Big| mop(B_i) - foc(B_i) \Big| \tag{3}$$

where n is the size of the data set and B_i represents bin i.

All demonstrations are done using the *glass* data set. A similar setup, using a random forest and dividing data into proper training and calibration sets are the same as above. The test set is composed of four instances from each class, to ensure coverage of all classes. A confusion matrix is used to guide the use and interpretation of the explanations provided by *calibrated-explanations*.

Code for running the experiments in Section 5.1 can be accessed at the GitHub repository for *calibrated-explanations* in the evaluation sub-folder. All plots used for demonstrations in Section 5.2 can be found in `calibrated-explanations/notebooks/plots` and code for generating the plots in the notebook `demo_multiclass_glass.ipynb` in the notebooks folder. Explanation plots for all data sets used in Section 5.1 can be found in `calibrated-explanations/evaluation/multiclass/plots`.

5. Results

5.1. Evaluation of Venn-Abers for Multi-Class

Looking at the predictive performance in Table 2, not surprisingly, it can be seen that access to more training data is beneficial for the underlying models, as seen for both accuracy and

AUC. The [Low, High] columns show the average lower and upper bounds from the Venn-Abers calibrators. The expectation is that the accuracy for Venn-Abers should lie within the interval, which is the case for all data sets but the CMC data set, which is off by 0.1 percentage point from the lower bound. Considering the calibration results, we see that Venn-Abers calibration does not affect log loss much but is able to significantly improve the ECE.

Table 2: Predictive performance and calibration results

	Accuracy				**AUC**		**ECE**		**Log loss**	
	NoCal	**VA**	**[Low,**	**High]**	**NoCal**	**VA**	**NoCal**	**VA**	**NoCal**	**VA**
balance	0.831	0.862	[0.797,	0.874]	0.938	0.884	0.064	0.043	0.246	0.280
cars	0.983	0.974	[0.942,	0.982]	0.982	0.962	0.068	0.023	0.104	0.078
cmc	0.521	**0.521**	[0.522,	0.561]	0.642	0.637	0.131	0.022	0.708	0.648
cool	0.947	0.943	[0.912,	0.963]	0.964	0.937	0.011	0.021	0.099	0.137
glass	0.795	0.736	[0.612,	0.857]	0.763	0.721	0.094	0.039	0.461	0.523
heat	0.986	0.985	[0.933,	0.993]	0.933	0.916	0.023	0.041	0.058	0.079
image	0.981	0.975	[0.932,	0.984]	0.979	0.962	0.040	0.032	0.075	0.083
iris	0.954	0.950	[0.810,	0.985]	0.926	0.917	0.019	0.095	0.122	0.198
steel	0.785	0.770	[0.728,	0.797]	0.820	0.804	0.076	0.022	0.451	0.428
tae	0.666	0.589	[0.518,	0.701]	0.679	0.651	0.055	0.012	0.616	0.645
user	0.905	0.893	[0.802,	0.937]	0.835	0.821	0.074	0.055	0.274	0.290
vehicle	0.754	0.752	[0.712,	0.795]	0.853	0.830	0.056	0.027	0.410	0.413
vowel	0.967	0.931	[0.800,	0.950]	0.933	0.902	0.258	0.091	0.364	0.225
wave	0.852	0.848	[0.836,	0.857]	0.818	0.813	0.113	0.011	0.389	0.340
whole	0.699	0.710	[0.695,	0.729]	0.508	0.498	0.090	0.040	0.653	0.610
wine	0.980	0.970	[0.805,	0.996]	0.972	0.939	0.082	0.114	0.127	0.178
wineR	0.706	0.669	[0.642,	0.689]	0.747	0.707	0.045	0.013	0.528	0.569
wineW	0.701	0.656	[0.643,	0.668]	0.763	0.724	0.054	0.009	0.519	0.554
yeast	0.616	0.599	[0.570,	0.651]	0.679	0.666	0.022	0.015	0.616	0.632
Mean	**0.823**	**0.807**	**[0.748,**	**0.841]**	**0.828**	**0.805**	**0.072**	**0.038**	**0.359**	**0.364**

For some data sets, like Vowel in Figure 2, the uncalibrated random forest is extremely underconfident. Luckily, calibration using Venn-Abers is able to reduce the ECE substantially. In the calibration plot, the low (green) and high (red) bounds are included to indicate how the uncertainty intervals are distributed. It is evident that Vowel provides fairly wide uncertainty intervals on average, with the lines for the low and high bounds far apart.

Another similar example is the Wave data set, where the random forest is again clearly underconfident, see Figure 3. Here, calibration create a number of very confident predictions, leading to almost perfect calibration. The intervals are very narrow for this data set.

The random forests are often underconfident, especially for the higher estimates. The one very different example is the CMC data set, see Figure 4, where the random forest is actually overconfident. It is, of course, reassuring to see that Venn-Abers is able to calibrate these models too, producing generally lower estimates, thus resulting in significantly lower ECE:s.

One of the data sets where the calibrated models are worse than the uncalibrated is WINE, see Figure 5. Similar to the Vowel data set in Figure 2, the intervals are wide[5]. As

5. As the probabilities are binned into ten equally wide bins, probabilities for Venn-Abers, low, and high will not be binned identically, explaining why the high bounds only occur in the top four bins. Since

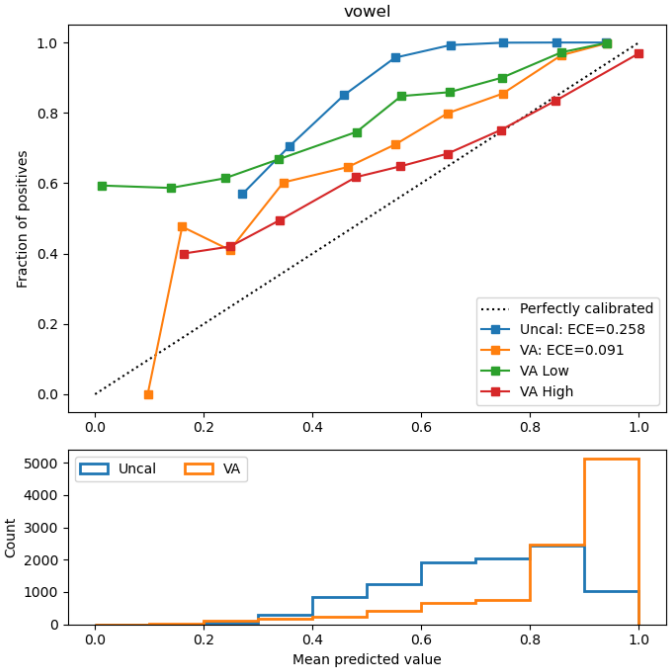


Figure 2: Vowel data set

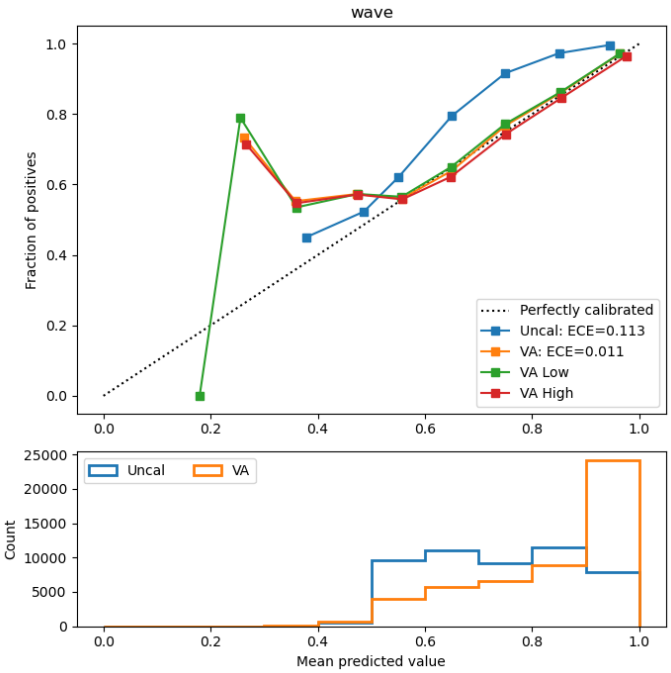


Figure 3: Wave data set

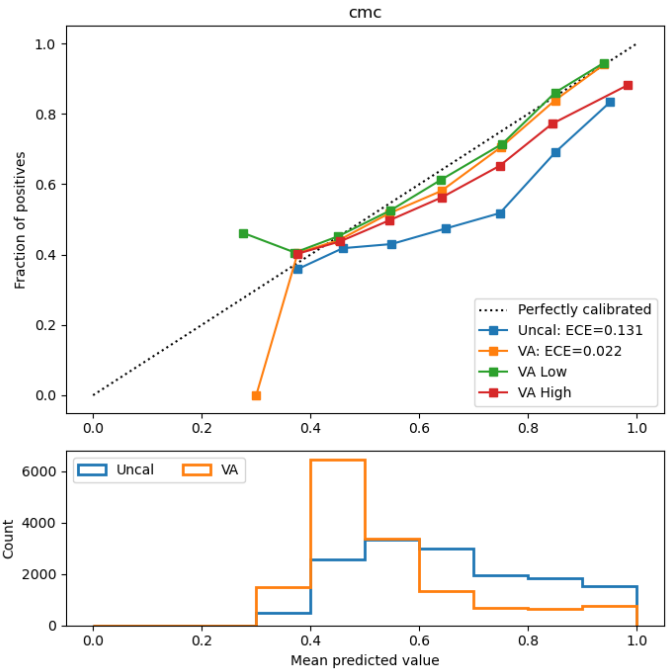


Figure 4: CMC data set

can be seen, the upper bounds of the interval is aligned with the diagonal, revealing that the confidence intervals are well-calibrated. The regularization will reduce higher probabilities more, which is clearly detrimental for the ECE in this case.

Summarising the experiment, the main result is that the calibration is able to successfully calibrate the underlying models in terms of ECE. The price paid is a small loss in predictive performance.

5.2. Demonstrating Multi-Class Explanations using Calibrated Explanations

All demonstration included in this paper are taken from the glass data set, with a test set of four instances being explained per class. Knowing the class distribution and which classes are easiest and which are hardest provides important cues for interpreting the explanations generated, as only the predicted class is shown in the explanation. The confusion matrix with the calibrated predictions for the calibration instances from each class is shown in Table 3. Here, a leave-one-out approach is used, so that the calibration set used for instance z_i is $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_q\}$.

As can be seen in Table 3, *tableware* (t), with only two instances in the calibration set, is entirely correctly predicted. The least correct class is the *vehic wind float* (vwf), with none of the four instances being correctly predicted. When looking at the precision, vwf has no predictions whatsoever, while both instances predicted as containers are in fact containers. For the other classes, both recall and precision varies in the range $0.667 - 0.875$.

Venn-Abers is the regularized probability derived from low and high, it is always between these values for every instance. When aggregating these plots, one instance may fall into different bins for Venn-Abers, low and high bounds.

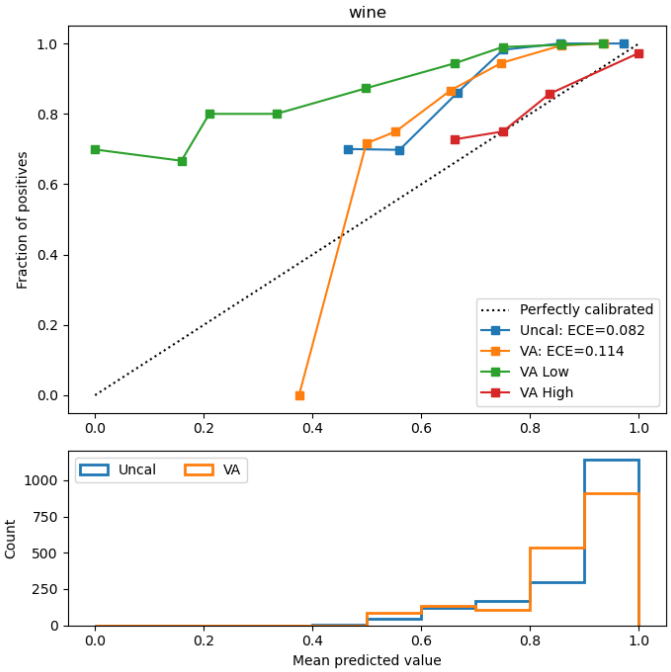


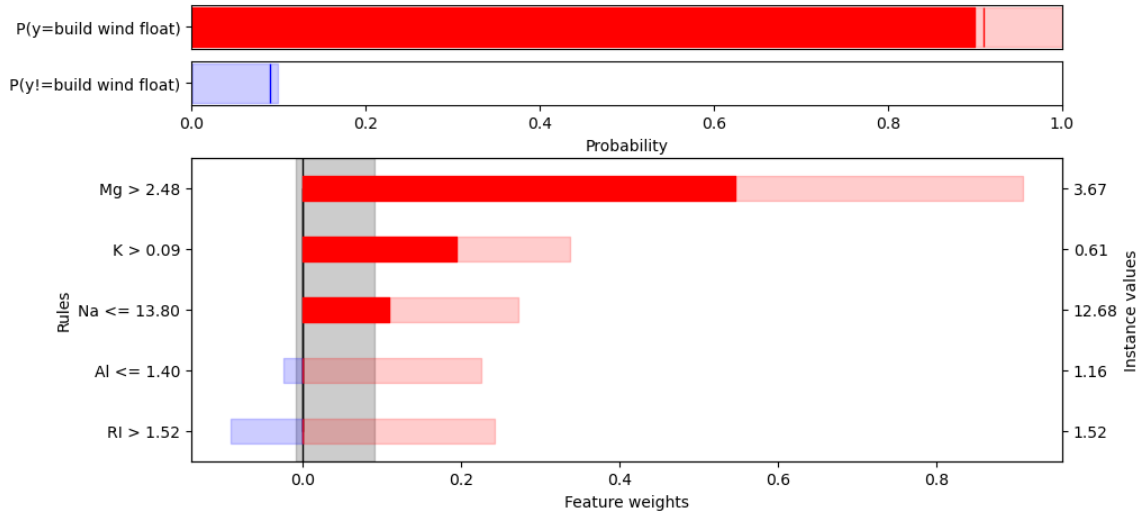
Figure 5: WINE data set

Table 3: Confusion Matrix for the glass data set

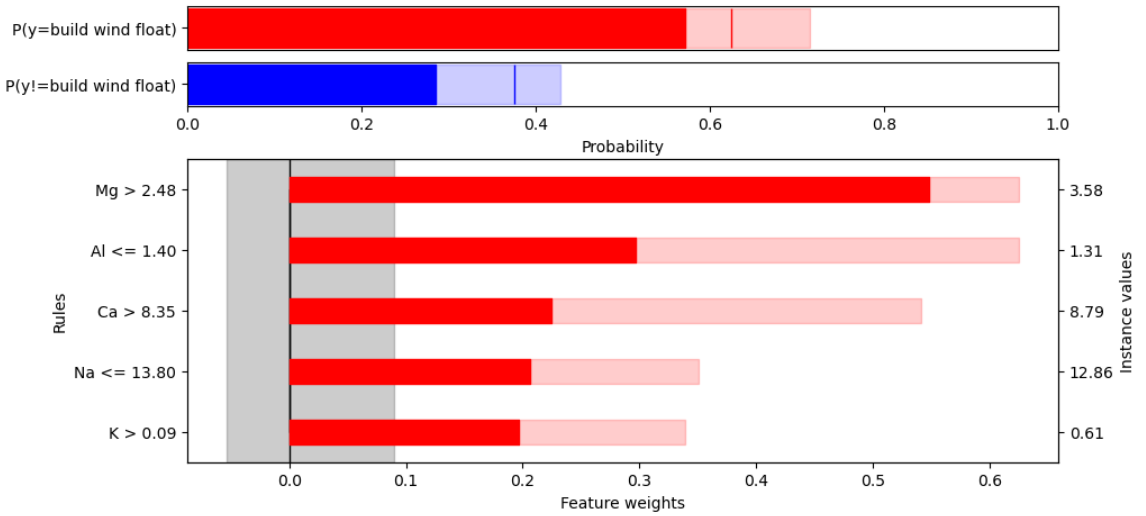
		Predicted						
Full class names	**Actual**	**bwf**	**bwnf**	**c**	**h**	**t**	**vwf**	**Recall**
build wind float	**bwf**	**18**	4	0	0	0	0	0.818
build wind non-float	**bwnf**	4	**18**	0	1	1	0	0.750
containers	**c**	0	1	**2**	0	0	0	0.667
headlamps	**h**	1	0	0	**7**	0	0	0.875
tableware	**t**	0	0	0	0	**2**	0	1.000
vehic wind float	**vwf**	2	2	0	0	0	**0**	0.000
	Precision	0.720	0.720	1.000	0.875	0.667	NaN	

The most common pattern to expect is an instance predicted as *build wind float* (*bwf*) or *build wind non-float* (*bwnf*), as these are the most commonly predicted classes (as well as the most common classes). Figure 6 show two factual explanations for instances predicting class *bwf*[6].

6. Instance numbers are indicated in the repository at calibrated-explanations/notebooks/plots.



(*a*) A factual explanation with uncertainty for instance 2



(*b*) Factual explanation with uncertainty for instance 22

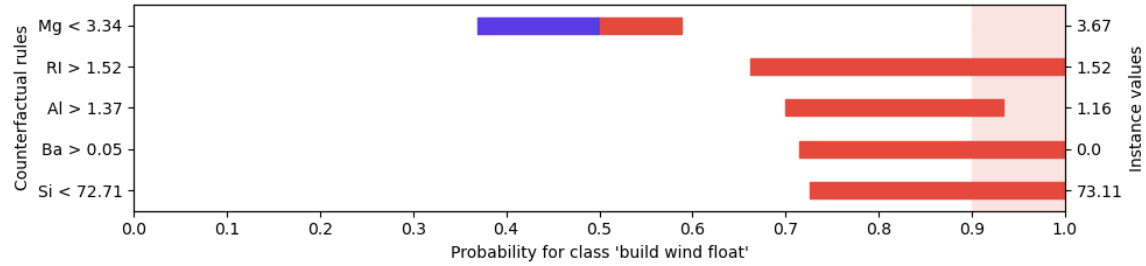
Figure 6: Factual explanation with uncertainty for *Build wind float* (instance 22).

The top red bar in Figure 6(*a*) shows the calibrated probability of the class predicted by Venn-Abers. As can be seen, the probability for the predicted class is very high, with the uncertainty interval indicated by the lighter red part. The lower box contains the most important factual conditions with feature importance in bright colour and uncertainty in lighter colour. The actual feature value of the instance can be seen to the right of each bar and the factual condition for which the feature importance applies can be seen to the left of the bar. A red feature importance indicate that a factual condition have a positive impact on the probability for the predicted class whereas a blue feature importance indicates a negative impact on the probability of the predicted class. The uncertainty interval provides

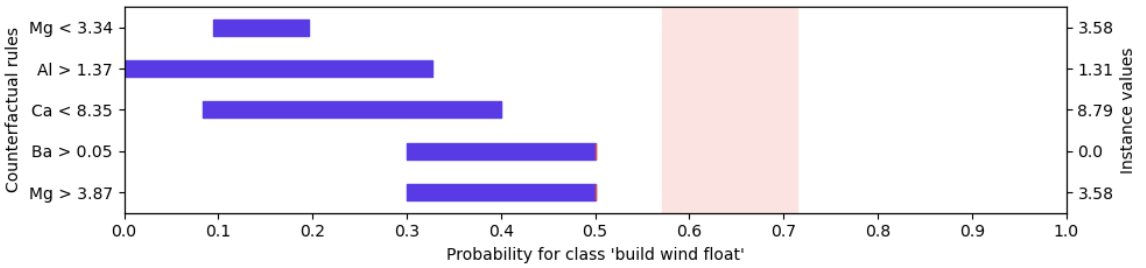
an estimate of the lower and upper bounds of the feature importance. As can be seen for the last two factual conditions, the uncertainty interval can point in different directions, indicated by the colour shifting from blue to red. Only the first factual condition, indicating a positive impact on the probability for class *bwf* due to `Mg > 2.48`, is strongly affecting the prediction.

Another instance also predicted as class *bwf* can be seen in Figure 6(*b*). Here, the calibrated probability of the class predicted by Venn-Abers is much lower, with a wider uncertainty interval. All factual conditions and their feature importance and uncertainty interval indicate support for the positive class, sometimes with a large degree of uncertainty.

Calibrated Explanations include the option of creating counterfactual explanations, providing complementary insights. The counterfactual explanations for instances 2 and 22 can be seen in Figure 7.



(*a*) A counterfactual explanation for instance 2



(*b*) A counterfactual explanation for instance 22

Figure 7: Counterfactual explanations for *Build wind float*.

The explanation in Figure 7(*a*) shows five counterfactual conditions indicating what the probability interval for the class predicted by Venn-Abers would have been if the instance would have had a counterfactual feature value, as indicated by the condition. The lighter coloured area in the background correspond to the prediction interval from Venn-Abers on the actual instance values. The first counterfactual condition, `Mg < 3.34`, indicate that the belief in the predicted class would drop to approximately $[0.37, 0.59]$, i.e., still inconclusive. All other counterfactual conditions would still provide fairly strong support for the predicted class. Looking instead on the counterfactual explanation in Figure 7(*b*), all the counterfactual conditions substantially decrease the already low support for the predicted class.

Considering the knowledge from the confusion matrix in combination with the factual and counterfactual explanations, it is easy to see that we have strong indication that instance 2 is in fact correctly predicted as class *bwf*. This is supported both by the fact that *bwf* is one of the most common classes and by the fact that both the factual and counterfactual explanations strongly support it. For instance 22, on the other hand, the explanations rather indicate against the prediction being accurate, indicated by low initial probability for the class predicted by Venn-Abers which is clearly decreased further by all the counterfactual conditions. As it turns out, the actual class for instance 2 is *bwf* while the actual class for instance 22 is *vwf*.

Considering the confusion matrix in Table 3, with no instances being predicted as *vwf*, the explanation in Figure 8 is particularly interesting.

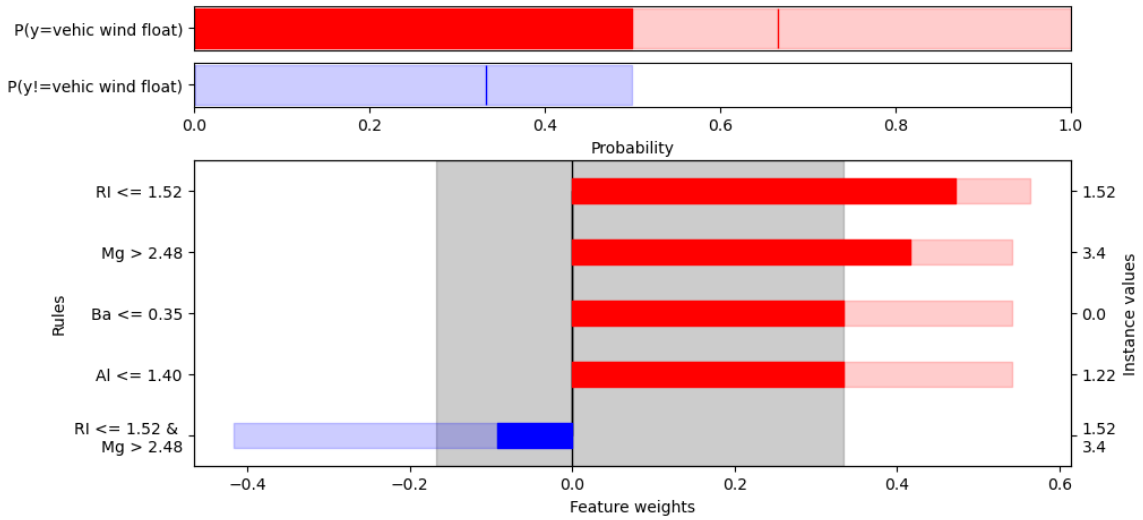
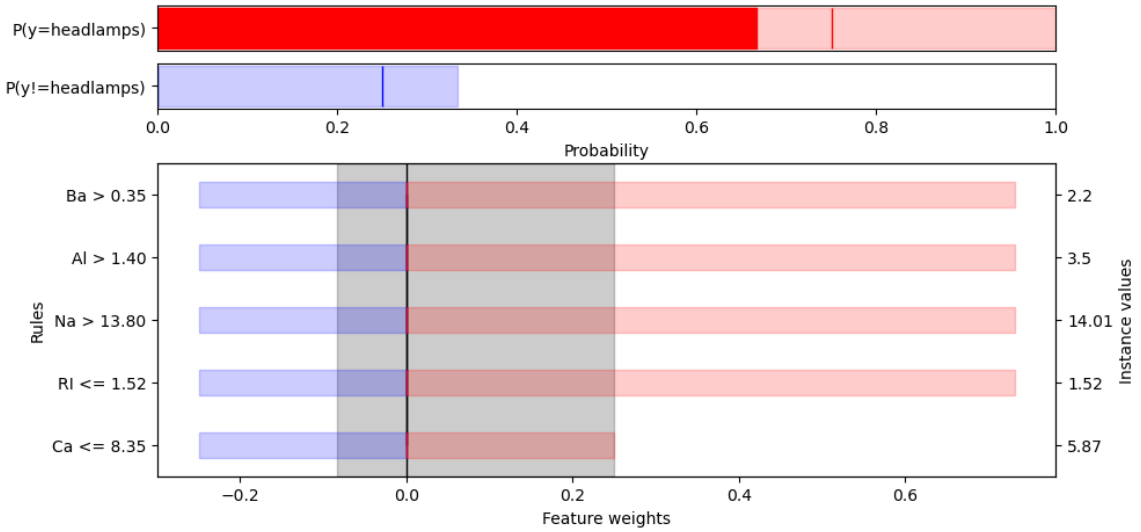


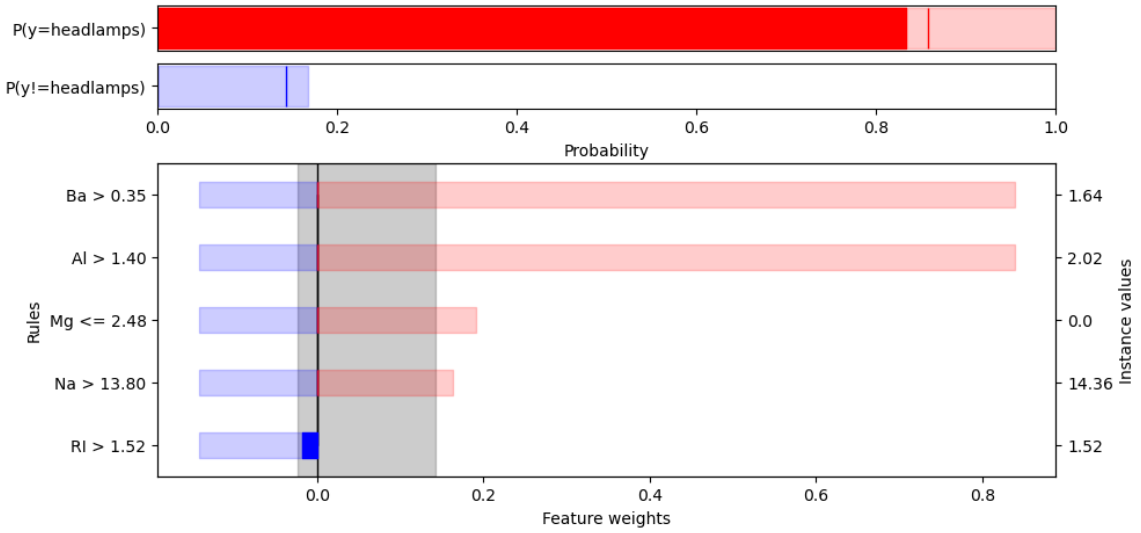
Figure 8: A Calibrated Explanation with uncertainty and conjunctive rules for *Vehic wind float* (instance 23)

This instance is predicted as a class which is clearly difficult, as indicated by the confusion matrix. The prediction has strong support for the predicted class, even though the uncertainty is large. The four first factual conditions clearly support the predicted class. Only the conjunctive condition show some slight support against the predicted class. Considering that the confusion matrix revealed that only four instances belonging to the *vwf* class exists in the calibration set, the degree of uncertainty in the prediction should not provide too much concern. Considering that this is a very difficult class and the explanation show strong support for the predicted class, it is reasonable to trust the prediction. Not surprisingly, this is a correct assumption for this instance.

As can be seen in the confusion matrix in Table 3, the classes *bwf*, *bwnf* and *vwf* are the classes most often mistaken for each other. In the final comparison, focus is instead placed on one of the other classes by showing explanations for *headlamps* (h). Figure 9 show two factual explanations for instances predicted as class h.



(a) A factual explanation for instance 8

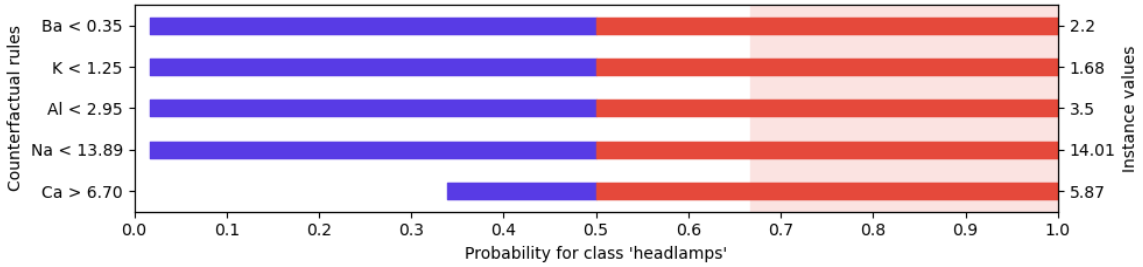


(b) A factual explanation for instance 14

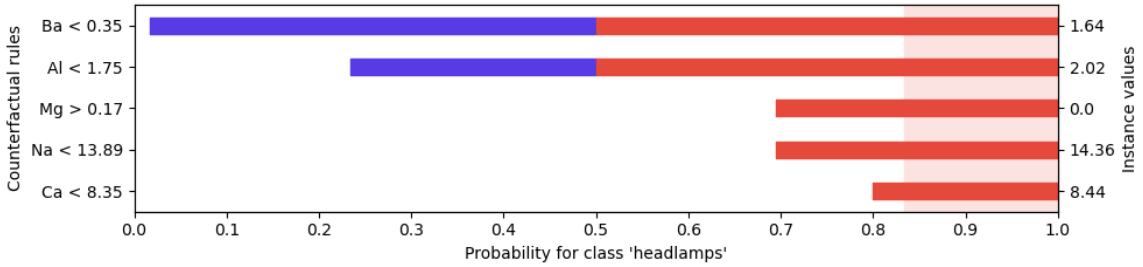
Figure 9: Factual explanations with uncertainty intervals for *Headlamps*.

Both explanations provide strong support for the predicted class, even though instance 8 (in Figure 9(a)) indicate more uncertainty by having a wider uncertainty interval at the top bar. The first four factual conditions for instance 8 provide fairly strong support for the predicted class h, even though the uncertainty is extremely large. For instance 14 (in Figure 9(b)), only two factual conditions provide strong support (with extreme uncertainty) for the predicted class h. The remaining factual conditions are inconclusive or indicate support against the predicted class. Nothing in these explanations provide any strong evidence against the predictions from Venn-Abers for these instances, even though there is a lot of uncertainty.

Looking at the same instances using counterfactual explanations (see Figure 10), some further insights might be drawn. Looking first at instance 8 in Figure 10(*a*), the first four counterfactual conditions all have 100% uncertainty, i.e., indicating less belief and more uncertainty in the predicted class *h*. The final counterfactual condition is also very uncertain, decreasing the belief in the predicted class. In Figure 10(*b*), the first two counterfactual conditions are also very uncertain. The three remaining counterfactual conditions for instance 14 are all favouring the predicted class *h*, even though the uncertainty is increased compared to the prediction made by Venn-Abers (in lighter red).



(*a*) A counterfactual explanation for instance 8



(*b*) A counterfactual explanation for instance 14

Figure 10: Counterfactual explanations for *Headlamps*.

Taking the evidence from both the factual and counterfactual explanations for instances 8 and 14, the already weak evidence in favour of the predicted class *h* is decreased further by the counterfactual explanations for instance 8. The combination of more uncertainty in the probability for the predicted class to begin with, in combination with the extreme uncertainty in both the factual and counterfactual explanations indicate that the prediction of instance 8 should not be trusted. For instance 14, on the other hand, the counterfactual explanation indicates that only two of the suggested changes would potentially change the support against the predicted class (even though the uncertainty is extreme). When looking at the ground truth, it turns out that the class for instance 8 is actually *container* (*c*), while the class of instance 14 is *h*. Looking at other instances also predicted as class *h*, they are all correct and very similar to the explanations achieved for instance 14, further corroborating the reasoning given above.

In summary, a number of examples have been provided where factual and counterfactual explanations have been analysed with the help of the confusion matrix. The analysis has

focused on the predicted classes and given a number of examples on how to use both factual and counterfactual explanations to help determine which instances might be correctly predicted and which might not when predicted with the same class.

6. Concluding Discussion

The paper provides a solution for how to provide calibrated explanations with uncertainty estimates for multi-class problems. The proposed solution combine previous work on multi-class calibration and calibrated explanations, providing a useful tool for analysis and decision support when working with multi-class problems. The suggested approach is available as a python library from github.com/Moffran/calibrated_explanation installable using `pip` or `conda`.

The paper included experimental results indicating the strength of the calibration method adopted for multi-class problems on a number of data sets. The experimental results corroborated previously published results on the soundness of the multi-class calibration approach.

The results also included a demonstration of how the factual and counterfactual explanations could be used in combination with a confusion matrix to provide sufficient insights to enable a decision maker to determine which predictions to trust and which to distrust.

One interesting direction for future work would be to explore possibilities to include multi-probabilistic predictors with well-calibrated probability intervals for each class instead of only considering the predicted class. A functional feature worth considering is to add the possibility to select for which class to extract an explanation.

Acknowledgments

The authors acknowledge the Swedish Knowledge Foundation and industrial partners for financially supporting the research and education environment on Knowledge Intensive Product Realisation SPARK at Jönköping University, Sweden. Projects: PREMACOP grant no. 20220187, AFAIR grant no. 20200223, and ETIAI grant no. 20230040.

References

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

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Uncertainty in artificial intelligence*, pages 148–155. Morgan Kaufmann Publishers Inc., 1998.

David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Authorea Preprints*, 2021.

Ulf Johansson, Tuwe Löfström, and Henrik Boström. Calibrating multi-class models. In *Conformal and Probabilistic Prediction and Applications*, pages 111–130. PMLR, 2021a.

Ulf Johansson, Tuwe Löfström, and Henrik Boström. Well-calibrated and sharp interpretable multi-class models. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 193–204. Springer International Publishing Cham, 2021b.

Antonis Lambrou, Ilia Nouretdinov, and Harris Papadopoulos. Inductive venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74(1):181–201, 2015.

Helena Löfström, Tuwe Löfström, Ulf Johansson, and Cecilia Sönströd. Investigating the impact of calibration on the quality of explanations. *Annals of Mathematics and Artificial Intelligence*, pages 1–18, 2023.

Helena Löfström, Tuwe Löfström, Ulf Johansson, and Cecilia Sönströd. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246:123154, 2024.

Tuwe Löfström, Helena Löfström, Ulf Johansson, Cecilia Sönströd, and Rudy Matela. Calibrated explanations for regression, 2023.

Valery Manokhin. Multi-class probabilistic classification using inductive and cross Venn–Abers predictors. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pages 228–240, Stockholm, Sweden, 2017. PMLR.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. In *Advances in Neural Information Processing Systems*, pages 1133–1140, 2004.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 178–190. PMLR, 26–28 Aug 2020.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proc. 18th International Conference on Machine Learning*, pages 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775151.