Proceedings of Machine Learning Research 204:1–21, 2023 Conformal and Probabilistic Prediction with Applications

Confidence Classifiers with Guaranteed Accuracy or Precision

Ulf Johansson ULF.JOHANSSON@JU.SE
Cecilia Sönströd CECILIA.SONSTROD@JU.SE
Tuwe Löfström TUWE.LOFSTROM@JU.SE
Dept. of Computing, Jönköping University, Sweden

Henrik Boström HENRIK.BOSTROM@KTH.SE
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

In many situations, probabilistic predictors have replaced conformal classifiers. The main reason is arguably that the set predictions of conformal classifiers, with the accompanying significance level, are hard to interpret. In this paper, we demonstrate how conformal classification can be used as a basis for a classifier with reject option. Specifically, we introduce and evaluate two algorithms that are able to perfectly estimate accuracy or precision for a set of test instances, in a classifier with reject scenario. In the empirical investigation, the suggested algorithms are shown to clearly outperform both calibrated and uncalibrated probabilistic predictors.

Keywords: Conformal prediction, Classification, Classification with reject option, Precision

1. Introduction.

Binary classification, where a model approximates the function $f(\mathbf{X}, y) \rightarrow \{0, 1\}$, often provides a natural representation of the problem at hand when using machine learning models for decision-making. Some examples from different domains are loan applications (approve/reject), return prediction (return or not) and spam filters (spam or not). Decision-making can be fully automated, or involve a human, acting with the aid of the model predictions, possibly augmented with explanations and/or estimates of how credible the prediction is. For most model types, the actual class prediction is based on internal probability estimates from the model. In binary classification, a threshold on the probability estimate will determine the predicted class, with the typical value being 0.5. In the standard classification setup, all instances are predicted by the model, including those around this threshold, for which the model is very uncertain.

An alternative is to use *classification with reject option*, where the classifier can refrain from making predictions for certain instances. This option of not predicting all instances can be exploited in several different ways. For human-in-the-loop situations, it is possible to split the workload between human and machine decision-making, so that easy and clear-cut instances are predicted by the classifier, and more difficult cases are referred to the human. When acting upon an erroneous prediction that has an associated cost or risk, having the classifier refrain from making uncertain predictions can actually be beneficial,

see e.g. Hanczar and Sebag (2014). For classification with reject option, there is a clear trade-off between accuracy and rejection rate (Hansen et al., 1997). For some applications, there is also a difference in which type of error is most costly, and consequently a trade-off between precision and recall.

Conformal classification, with its predicted label sets and guaranteed validity for a confidence level provided by the user, may initially appear as a very strong option for the classifier with reject scenario. Intuitively, the conformal classifier would predict the instances with singleton label sets, while rejecting to make predictions when the label sets are empty or contain several labels. However, if used in this way, it must be noted that the accuracy of the singleton predictions would not meet the provided confidence level. In fact, the error rate of the singleton predictions would most likely be significantly higher than indicated by the confidence level supplied by the user. Even more importantly, the empirical accuracy of the singleton predictions could not be estimated in a straightforward way. Furthermore, it is not trivial to set up a conformal set predictor guaranteeing precision, rather than accuracy.

With this in mind, the most obvious choice for classification with reject option would be to use probabilistic predictors. In this paper, however, we introduce an alternative approach that is in fact based on conformal classifiers, but using confidence-credibility scores rather than set predictions. The suggested procedure requires access to a set of test instances, i.e., it can not be used in a streaming scenario, but the result is a classifier with reject option inheriting the validity guarantees from conformal classification.

In the empirical investigation, we will demonstrate the suggested approach, showing that the accuracy estimates of the conformal classifier will be well-calibrated for all reject levels. We will also extend the approach to well-calibrated precision estimates, by using a Mondrian conformal classifier.

2. Background.

2.1. Probabilistic prediction and calibration

Probabilistic predictors output a probability distribution over the possible labels. The probabilistic predictor is *well-calibrated* if the probability of a certain label, given a probability estimate for that label, is equal to the probability estimate, i.e.,

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

where p^{c_j} is the probability estimate for class j. If tested empirically, the (average) probability estimate for the predicted label should correspond to the empirical accuracy. If we, as an example, make a (large) number of predictions, with the (mean) probability estimate 0.95, we would expect the empirical accuracy of these predictions to be 0.95.

While almost all classifiers return such probability estimates, these may not be well-calibrated. Instead, techniques like decision trees (Provost and Domingos, 2003) and naive Bayes (Niculescu-Mizil and Caruana, 2005) often produce classifiers that are very poorly calibrated. More recent studies show that this also applies to both modern (i.e., deep) neural networks (Guo et al., 2017) and traditional neural networks (Johansson and Gabrielsson, 2019), although to a lesser degree. There are, however, many so-called *external* calibration

methods that transform classifier scores into calibrated probability estimates. One example is *Platt scaling* (Platt, 1999), where a sigmoid function is fitted to the probability estimates from the model, using a separate calibration set. The function is

$$\hat{p}(c_j \mid p^{c_j}) = \frac{1}{1 + e^{As+B}}, \tag{2}$$

where $\hat{p}(c_j \mid p^{c_j})$ is the calibrated probability estimate that the instance belongs to class c_j, given the probability estimate p^{c_j} from the model. A and B are found by minimizing a specific loss function, using a gradient descent search, for details see (Platt, 1999).

2.2. Conformal classification

In conformal classification, a test instance is tentatively labeled $(\boldsymbol{x}_{k+1}, \tilde{y})$, and then a p-value statistic is calculated to attempt to reject the hypothesis that $\tilde{y}$ is the true label y_{k+1} at a significance level ϵ. This procedure is repeated for all possible labels, resulting in the label set $\tilde{y} \subseteq Y$ containing all labels that were not rejected. This label set is, under mild conditions, guaranteed to contain the true target y_{k+1} with a probability of $1 - \epsilon$.

When deciding which labels can be rejected, a so-called *nonconformity function* $A : X \times Y \to \mathbb{R}$ is used. The purpose of the nonconformity function is to measure the relative strangeness of the instance $(\boldsymbol{x}, \tilde{y})$, i.e., the tentative label together with the input feature values, compared to a set of instances with known target values. If the p-value for that combination is lower than the threshold ϵ, the label can be rejected.

In practice, the nonconformity function is normally based on the prediction of a machine learning model, called the *underlying model*. In this study, we will employ a frequently used option, the *hinge loss* function,

$$\Delta\left[h\left(\boldsymbol{x}_i\right), \tilde{y}\right] = 1 - \hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}_i\right), \tag{3}$$

where $\hat{P}_h\left(\tilde{y} \mid \boldsymbol{x}\right)$ is the probability estimate provided by the machine learning model h for the instance $\boldsymbol{x_i}$ and the label $\tilde{y}$.

In more details, an *inductive conformal classifier* (ICP) (Papadopoulos et al., 2002; Vovk et al., 2005; Papadopoulos, 2008), is constructed in the following way:

1. Divide the available labeled training data Z into two disjoint subsets: a proper training set Z^t and a calibration set Z^c, where $|Z^c| = q$.
2. Train the underlying machine learning model h using the proper training set Z^t.
3. Apply the nonconformity function, here, Eq. 3, to all calibration examples in Z^c to produce a list of calibration scores $\alpha_1, \ldots, \alpha_q$.

To find the predicted label set of a test instance $\boldsymbol{x}_{k+1}$:

1. Obtain the (probability) prediction $h(\boldsymbol{x}_{k+1})$ from the underlying model.
2. Tentatively assign the label $\tilde{y} \in Y$ for $\boldsymbol{x}_{k+1}$, and measure the resulting nonconformity of $(\boldsymbol{x}_{k+1}, \tilde{y})$.

3. Find the resulting p-value as

$$p_{k+1}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z^c : \alpha_i \geq \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{q+1} + \theta_{k+1}\frac{\left|\left\{z_i \in Z^c : \alpha_i = \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{q+1}, \quad (4)$$

   where $\theta_{k+1} \sim U[0,1]$.

4. Repeat steps 2-3 for each possible label $\tilde{y} \in Y$.

5. Compare the p-values to the significance level ϵ, rejecting all labels $\tilde{y}$ where $p_{k+1}^{\tilde{y}} < \epsilon$.

6. Include all labels that are not rejected in the final predicted label set Γ_{k+1}^{ϵ}.

Using this procedure, the probability for the resulting label set to include the true label is exactly $1-\epsilon$, as long as the calibration set and the test set are exchangeable.

By conditioning the confidence predictions on some characteristic of a test instance, a so-called *Mondrian* conformal classifier (Vovk et al., 2005), can provide confidence measures with similar guarantees, but in a category-wise manner. The most common example is a class-conditional (Mondrian) conformal classifier, where the validity applies for each of the classes separately. The Mondrian framework, however, can use any taxonomy for dividing the instances into categories, and then the validity guarantees will apply to each category κ_j.

To create a Mondrian conformal classifier, (4) simply needs to be redefined as:

$$p_{k+1}^{\tilde{y},\kappa_j} = \frac{\left|\left\{z_i \in Z^{\kappa_j} : \alpha_i \geq \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{r+1} + \theta_{k+1}\frac{\left|\left\{z_i \in Z^{\kappa_j} : \alpha_i = \alpha_{k+1}^{\tilde{y}}\right\}\right| + 1}{r+1}, \quad (5)$$

where $Z^{\kappa_j} \subseteq Z^C$, given by some condition(s), and r is the number of calibration instances in the category κ_j.

While the theory behind conformal classification is solid, and the guarantees strong, it is still rather awkward to utilize conformal classifiers for decision making. Specifically, a decision-maker would most likely be tempted to focus on singleton predictions, i.e., label sets containing only one label. The problem is, however, that when looking at the singleton predictions, it is not straightforward to estimate the probability that these are in fact correct. In particular, the probability of an error is almost guaranteed to be (significantly) higher than ϵ. The explanation is, of course, that the guarantees only hold *a priori*; once a predicted label set has been observed, the likelihood of an error is highly dependent on the size of the set. Specifically, if the label set contains all labels, it cannot be an error; and if it is empty, it must be an error. Looking at two-class problems, and the quite frequent situation where there are no (or very few) empty predictions, *all* (or almost all) errors most come from the singleton predictions. With this in mind, many researchers and practitioners have moved away from conformal classification, instead using probabilistic predictors, including *Venn predictors*. There is, however, a less frequently used option for conformal classifiers, called *confidence-credibility* predictions. These are also based on the p-values, and for each test instance $\boldsymbol{x}_j$ we can calculate the following three values:

- The predicted class label $\hat{y}_j$, i.e., the label with the highest $p_j^{\tilde{y}}$.

- The *confidence*, calculated as one minus the second largest p-value.
- The *credibility*, which is equal to the largest p-value.

The connection between confidence-credibility measures, and the set predictions, is that the confidence represents the highest significance level where we get a singleton prediction, while the credibility is equal to the significance level where all labels are rejected.

In this paper, we will utilize the confidence measure, and a set of test instances, to produce classifiers with reject option having statistical guarantees. While the suggested approach is heavily inspired by Linusson et al. (2018), it is now generalized to obtain precision, instead of just accuracy.

To appreciate the approach, it is vital to understand exactly what a confidence score λ_j for the test instance x_j represents. Instead of providing a probability estimate for just the instance x_j, the correct interpretation is actually that if we look at *all* m predictions with a confidence of at least λ_j these will contain (on average) $n(1-\lambda_j)$ errors, where n is the *total* number of predictions made, i.e., the size of the test set. So, in a classifier with reject scenario, if we reject all instances with a confidence score lower than λ_j, the expected error rate of the predicted instances is:

$$\frac{n \cdot (1-\lambda_j)}{m} \tag{6}$$

If the conformal classifier is a standard ICP, the guarantees will be for overall error rate, i.e., accuracy, but if a Mondrian setup is used, the guarantees will apply to each category individually. In this paper, we will use a Mondrian taxonomy where the categories are determined from the label predicted by the underlying model. Consequently, if we look at the category where the model predicted the positive class (label 1), the guarantees will in fact be for precision.

2.3. Related Work

Related work in classification with reject option, not using the conformal prediction framework, includes Li and Sethi (2006), where a design methodology for confidence-based classifiers was proposed, with SVMs as the underlying models. The objective is to control the error rate of a classifier with reject option, but the approach does not yield any statistical guarantees. In Hanczar and Dougherty (2008), classification with reject option is set up so that the user can specify a desired accuracy level, and the rejection region is found from this. In Johansson et al. (2023), conformal classification is employed to automatically suggest which accuracy levels that can be exactly met, with statistical guarantees, using classification with reject option. In Fisch et al. (2022), conformal classification is adapted to achieve guarantees on the false positive rate in the prediction sets produced.

3. Method.

As described in the introduction, the purpose of the study is to investigate how conformal classifiers can be used as the basis for classifiers with reject option, capable of estimating the accuracy or precision level, with statistical guarantees.

The classifiers used were standard decision trees and random forests (Breiman, 2001), as implemented in scikit learn. Parameters were left at default values, except the decision tree *min_leaf* parameter, which was set to 7.

In the first experiment, the targeted metric is accuracy, i.e., all models should estimate their accuracy, given a certain reject level. In the second experiment, the goal is instead to estimate the precision. In both experiments, and for both model types, three different setups were used:

- **Uncal** - Uncalibrated models. Here, the probabilitiy estimates produced by the decision tree or random forest model are used as the basis for rejecting instances in the subsequent classification. More specifically, the accuracy or precision estimate for a certain subset of the instances is the mean probability estimate for the predicted label of these instances. In Experiment 1, all instances are considered, but in Experiment 2, only the instances predicted as the positive class are considered.

- **Platt** - Platt scaling. Models are calibrated using Platt scaling, with the *CalibratedClassifierCV* method in scikit-learn, and the resulting probabilities are used in the classification with reject option, in an identical way as for the uncalibrated models.

- **Conf** - Conformal classification. A conformal classification model is trained and calibrated, as described above. Confidence scores are calculated for the test instances and used as a basis for the classification with reject option. Here, the accuracy or precision estimate is found using (6). In Experiment 1, a standard conformal classifiers is used, but in Experiment 2, a Mondrian version where the categories are determined by the predicted label of the underlying model is employed. As a consequence, only instances belonging to the positive class are included in the calibration set in Experiment 2.

The evaluation of the classification with reject scenarios was performed using reject proportions of $10\%, 20\%$, etc, up to a 90% rejection rate on test set instances, sorted on either the uncalibrated scores from the underlying model (*Uncal*), calibrated probabilities (*Platt*) or confidence scores (*Conf*).

The testing protocol employed was repeated hold-out with 100 repetitions, for each data set, of a 75/25 split into training and test sets, respectively. For the setups using a calibration set, i.e., *Platt* and *Conf*, a further partition of the training set, into 67% training instances and 33% calibration instances, was used. In the second experiment, the calibration instances belonging to the negative class was, for the conformal approach, just ignored. *Uncal* was, of course, trained on the entire training set.

The 10 publicly available benchmarking data sets used (see Table 1), are all two-class problems, from either the UCI repository (Bache and Lichman, 2013) or the PROMISE Software Engineering Repository (Sayyad Shirabad and Menzies, 2005). We can see that most data sets are rather small, and that some of them are fairly unbalanced. In pc4, and to a lesser degree kc1, kc2, transfusion, diabetes, and wbc, the positive class is in a clear minority.

Table 1: Data set descriptions - showing number of instances, number of input attributes and the proportion of instances that belongs to the positive class.

Data set	#inst	#attrib	prop. pos.	Source
creditA	690	16	0.44	UCI
diabetes	768	9	0.35	UCI
german	1000	21	0.70	UCI
kc1	2109	22	0.26	Promise
kc2	522	22	0.27	Promise
kr-vs-kp	3196	36	0.52	UCI
pc4	1458	38	0.13	Promise
transfusion	748	5	0.26	UCI
ttt	958	10	0.65	UCI
wbc	699	10	0.35	UCI

4. Results.

To establish a baseline for classification with reject option, we first present the predictive performance, measured using accuracy and AUC in Table 2 below. Regarding accuracy, the expectation is that uncalibrated models would outperform calibrated models, based on the advantage of having more training data available for model building. This holds for *Uncal* vs. *Conf*, but models calibrated with Platt scaling actually obtains the the highest accuracy, for both decision trees and random forest. However, differences are small on average and also vary between data sets, despite the large number of repetitions. For AUC, the results are more in line with expectations, i.e., uncalibrated models obtain the best performance, on average, followed by conformal and Platt calibrated models, respectively.

Table 2: Experiment 1: Predictive performance

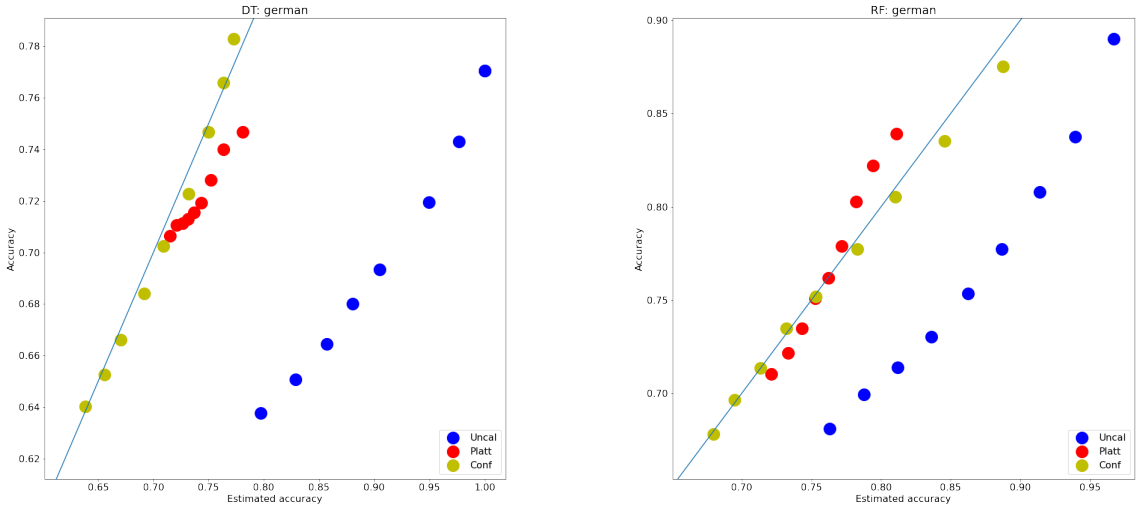
	DT						RF					
	Acc			AUC			Acc			AUC		
Data sets	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf
creditA	.844	.839	.840	.843	.837	.838	.875	.869	.869	.872	.866	.866
diabetes	.725	.721	.720	.698	.665	.689	.764	.759	.759	.725	.715	.719
german	.636	.700	.627	.521	.500	.523	.664	.699	.662	.531	.501	.529
kc1	.681	.734	.682	.580	.509	.574	.751	.748	.743	.612	.567	.597
kc2	.744	.765	.750	.678	.631	.675	.782	.787	.779	.696	.680	.694
kr-vs-kp	.977	.974	.974	.977	.974	.974	.989	.985	.985	.989	.985	.984
pc4	.873	.875	.870	.714	.632	.700	.900	.897	.896	.663	.699	.645
transfusion	.728	.737	.730	.615	.536	.609	.720	.738	.722	.605	.541	.607
ttt	.915	.899	.898	.899	.880	.880	.983	.965	.952	.976	.961	.935
wbc	.948	.944	.944	.945	.938	.939	.970	.967	.968	.968	.964	.966
Mean	**.807**	**.819**	**.803**	**.747**	**.710**	**.740**	**.840**	**.841**	**.833**	**.764**	**.748**	**.754**

Turning to the results for the conformal classifier, Table 3 shows the empirical error rates (*Err.*) and average number of labels in the prediction set (*AvgC*) for significance levels $\epsilon = 0.01, 0.05$, and 0.1. Starting with the error rates, we as expected see that they match the ϵ values, almost perfectly. The results for classification efficiency follow clear and obvious patterns, with fewer labels in the prediction set as ϵ increases. In addition, harder data sets also lead to larger label sets, while the stronger random forest models have smaller prediction sets on average.

Table 3: Experiment 1: Conformal classification

	$\epsilon = 0.01$				$\epsilon = 0.05$				$\epsilon = 0.1$			
	DT		RF		DT		RF		DT		RF	
Data sets	Err.	AvgC	Err.	AvgC	Err.	AvgC	Err.	AvgC	Err.	AvgC	Err.	AvgC
creditA	.010	1.884	.011	1.764	.050	1.455	.051	1.277	.100	1.160	.100	1.071
diabetes	.010	1.942	.009	1.800	.049	1.711	.049	1.548	.097	1.478	.099	1.357
german	.010	1.956	.011	1.909	.050	1.789	.051	1.728	.103	1.623	.102	1.562
kc1	.010	1.956	.012	1.907	.050	1.783	.052	1.666	.100	1.569	.104	1.436
kc2	.011	1.925	.008	1.759	.051	1.640	.052	1.452	.103	1.380	.100	1.285
kr-vs-kp	.010	1.049	.011	1.014	.052	.959	.050	.955	.101	.902	.101	.901
pc4	.010	1.831	.010	1.332	.052	1.253	.050	1.139	.104	1.061	.097	1.015
transfusion	.010	1.936	.009	1.901	.048	1.691	.053	1.661	.097	1.484	.103	1.461
ttt	.009	1.563	.010	1.170	.051	1.144	.053	.995	.101	1.010	.104	.917
wbc	.011	1.529	.012	1.149	.047	1.027	.053	.969	.099	.935	.105	.906
Mean	**.010**	**1.757**	**.010**	**1.570**	**.050**	**1.445**	**.051**	**1.339**	**.100**	**1.260**	**.102**	**1.191**

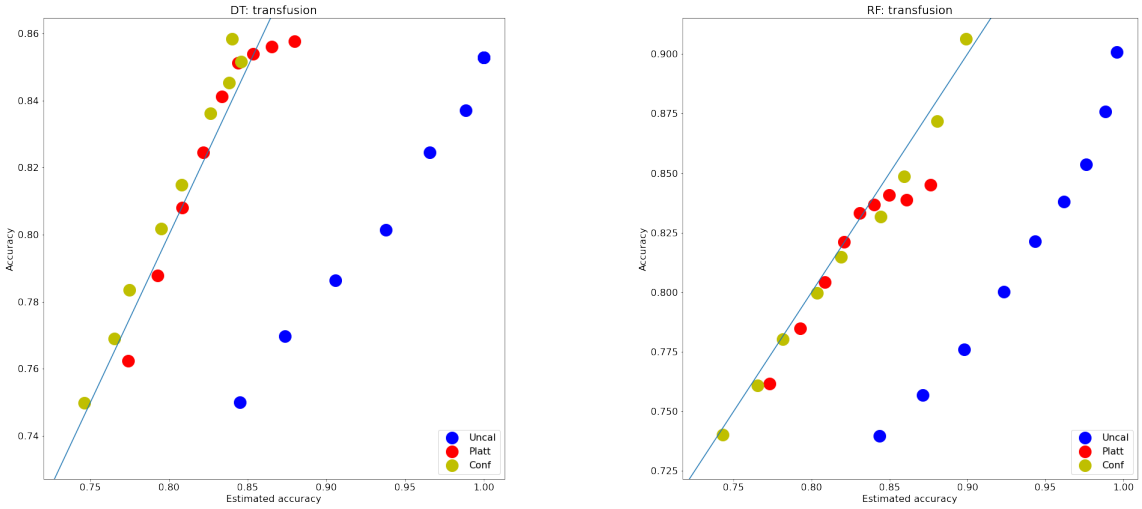
As an introduction to the results for the classification with reject option, Fig. 1-3 show the decision tree and random forest models for classification with reject option, for three different data sets. Each dot represents a rejection level, from 10% up to 90%, with x-axis as the estimated accuracy and the y-axis as the actual accuracy obtained.



(a) DT (b) RF

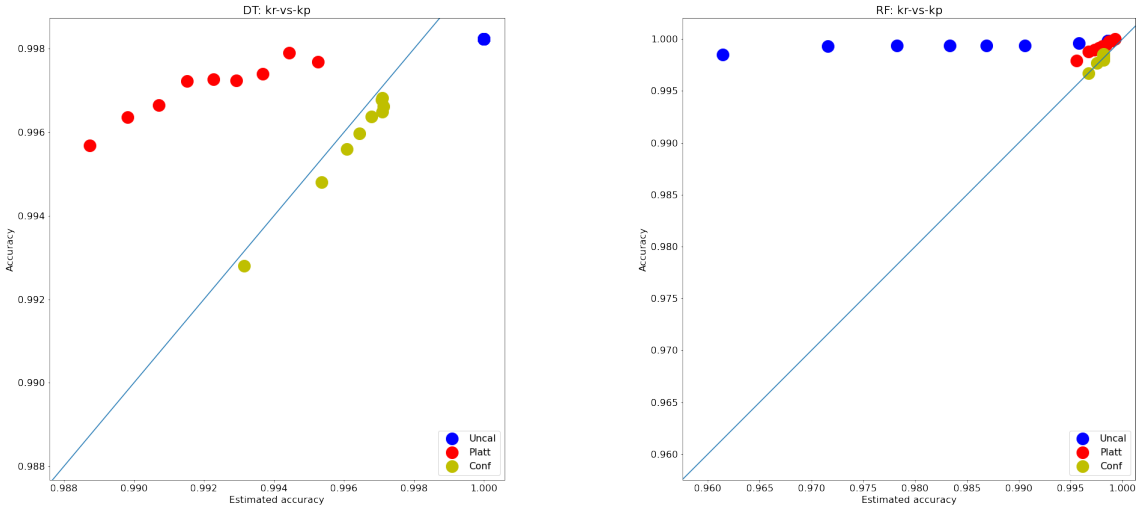
Figure 1: German data set: accuracy estimation

In Figures 1a – 1b, showing the *german* data set, uncalibrated models, both decision trees and random forests, are seen to be overconfident, i.e. over-estimate their accuracy, for all rejection levels. The conformal models exhibit near perfect estimation of actual accuracy, and the Platt scaling models produce reasonable good estimates. Actual accuracies, for the lower rejection levels, are in line with the overall accuracies, presented in Table 2 above, with Platt scaling models obtained the highest accuracies, followed by uncalibrated models and conformal models. However, this advantage in predictive performance is diminished as the rejection level increases, and for 90% rejection rate (represented by the topmost dots in each color, conformal and uncalibrated models obtain higher accuracy on the predictions that they do make. For the *Transfusion* data set, shown in Figures 2a – 2b, the picture is quite similar, i.e. perfectly estimated performance from the conformal models at all rejection levels, with good estimates from the Platt scaling models and consistent over-estimation of predictive performance from uncalibrated models. However, for the higher rejection levels, actual accuracy for Platt scaling models is seen to level off and this is not caught by the accuracy estimates, making the models over-estimate their predictive performance. For the quite easy *kr-vs-kp* data set, in Figures 3a – 3b, models exhibit a highly differing behaviour when performing classification with reject option. Looking first at decision tree models, uncalibrated models have identical values for both estimated and actual accuracies, at all rejection levels, and again over-estimate actual accuracy. Inspection of the estimated accuracy, reveals that, in fact, all instances have a score of 1, which entails that no instances are excluded on any level, and thus that the model will behave identically at all levels. Platt scaling models obtain the highest accuracy, but for this data set, consistently under-estimates its performance. Conformal classifiers provide correct estimations of accuracy on all rejection levels. For random forest models, both Platt scaling and conformal classification have high and largely correctly estimated accuracies at all levels, while uncalibrated random forests under-estimates its similarly high accuracy at all lower rejection levels.



(a) DT (b) RF

Figure 2: Transfusion data set: accuracy estimation



(a) DT (b) RF

Figure 3: kr-vs-kp data set: accuracy estimation

Aggregated accuracy results for the classification with reject option are shown in Tables 4 – 6, for rejection proportions of 50%, 70% and 90%, respectively. For all three rejection levels, uncalibrated models, both decision trees and random forests, consistently and substantially over-estimate accuracy. Conformal models are seen to provide better accuracy estimates than Platt scaling models, for all three levels shown in these tables. Compared to the full accuracy results, in Table 2 above, accuracies for classification with reject are slightly more varied between models and setups. For decision trees, Platt scaling has slightly better performance than uncalibrated trees and the conformal approach, but

differences are generally small. For random forest models, uncalibrated models have the highest average accuracy for all three rejection levels.

Table 4: Experiment 1: Accuracy estimation. Top 50%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.
creditA	1.000	.914	.910	.914	.909	.909	.928	.954	.948	.951	.950	.950
diabetes	1.000	.828	.816	.833	.813	.818	.887	.885	.874	.881	.878	.879
german	.905	.693	.737	.716	.709	.703	.863	.753	.762	.762	.753	.752
kc1	.992	.770	.797	.783	.764	.763	.897	.835	.839	.826	.832	.825
kc2	1.000	.854	.864	.863	.852	.847	.959	.928	.903	.914	.922	.917
kr-vs-kp	1.000	.998	.992	.997	.997	.996	.987	.999	.998	.999	.998	.998
pc4	1.000	.940	.944	.958	.941	.939	.991	1.000	.979	.997	.995	.996
transfusion	.966	.825	.833	.841	.808	.815	.944	.821	.831	.833	.819	.815
ttt	1.000	.982	.963	.978	.976	.978	.908	1.000	.998	1.000	.994	.995
wbc	1.000	.981	.976	.982	.978	.978	1.000	1.000	.987	.995	.992	.992
Mean	**.986**	**.879**	**.883**	**.887**	**.875**	**.875**	**.936**	**.918**	**.912**	**.916**	**.913**	**.912**

Table 5: Experiment 1: Accuracy estimation. Top 30%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.
creditA	1.000	.914	.922	.914	.913	.914	.954	.958	.961	.955	.956	.955
diabetes	1.000	.828	.838	.857	.826	.831	.936	.932	.905	.926	.924	.927
german	.977	.743	.752	.728	.750	.747	.914	.808	.782	.803	.810	.805
kc1	1.000	.776	.809	.787	.768	.771	.940	.858	.862	.844	.856	.849
kc2	1.000	.854	.881	.873	.865	.859	.991	.970	.923	.955	.953	.959
kr-vs-kp	1.000	.998	.994	.997	.997	.997	.996	1.000	.999	1.000	.998	.998
pc4	1.000	.940	.947	.958	.941	.939	.998	1.000	.983	.998	.995	.995
transfusion	1.000	.853	.854	.854	.838	.845	.976	.854	.850	.841	.859	.849
ttt	1.000	.982	.970	.981	.979	.981	.939	1.000	.999	1.000	.994	.996
wbc	1.000	.981	.980	.984	.981	.979	1.000	1.000	.990	.996	.992	.992
Mean	**.998**	**.887**	**.895**	**.893**	**.886**	**.886**	**.964**	**.938**	**.925**	**.932**	**.934**	**.933**

Table 6: Experiment 1: Accuracy estimation. Top 10%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.	Est.	Acc.
creditA	1.000	.914	.939	.912	.914	.915	.982	.964	.974	.964	.959	.955
diabetes	1.000	.828	.861	.864	.828	.835	.982	.977	.933	.966	.960	.968
german	1.000	.770	.781	.747	.773	.783	.967	.890	.811	.839	.887	.875
kc1	1.000	.776	.825	.784	.769	.773	.983	.883	.886	.861	.891	.876
kc2	1.000	.854	.907	.864	.866	.852	1.000	.988	.944	.971	.958	.971
kr-vs-kp	1.000	.998	.995	.998	.997	.997	.999	1.000	.999	1.000	.998	.998
pc4	1.000	.940	.954	.958	.942	.938	1.000	1.000	.987	.997	.995	.995
transfusion	1.000	.853	.880	.858	.846	.852	.996	.901	.877	.845	.899	.906
ttt	1.000	.982	.978	.990	.981	.982	.974	1.000	1.000	1.000	.994	.997
wbc	1.000	.981	.987	.980	.982	.977	1.000	1.000	.993	.996	.992	.992
Mean	**1.000**	**.890**	**.911**	**.895**	**.890**	**.890**	**.988**	**.960**	**.940**	**.944**	**.953**	**.953**

Table 7 shows signed and absolute errors for the accuracy estimations made by the different setups on each rejection level. Starting with decision trees, the uncalibrated models are extremely overconfident, on all rejection levels. Platt scaling and the conformal approach, on the other hand, show no systematic error, with average signed errors very close to zero. A direct comparison of absolute errors shows that the conformal estimations are more accurate, on each rejection level. Turning to the random forests, the uncalibrated models are slightly overconfident, on average. They also have fairly large absolute errors. For Platt scaling and conformal, the results here are quite similar to the decision tree results. In fact, the estimations are generally very good, for both approaches. Again, conformal absolute errors are smaller than Platt absolute errors, on every rejection level.

Table 7: Experiment 1: Accuracy estimation. The table shows signed errors and absolute errors, averaged over all data sets for the different rejection levels.

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Rejected	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.
10%	.093	.093	.002	.006	-.001	.003	-.003	.053	.002	.008	.001	.002
20%	.098	.098	-.002	.007	.000	.002	.003	.053	.002	.007	.001	.002
30%	.104	.104	-.003	.010	.000	.002	.010	.053	.000	.007	.001	.001
40%	.106	.106	-.003	.010	.000	.003	.015	.051	-.003	.005	.001	.002
50%	.108	.108	-.003	.011	.000	.003	.019	.047	-.004	.006	.002	.002
60%	.111	.111	-.002	.011	.000	.004	.024	.044	-.005	.009	.002	.003
70%	.111	.111	.001	.011	.000	.003	.026	.041	-.006	.013	.001	.004
80%	.110	.110	.007	.014	-.002	.004	.029	.038	-.006	.016	.002	.005
90%	.110	.110	.015	.019	-.001	.005	.028	.033	-.004	.017	.000	.006
Mean	**.106**	**.106**	**.001**	**.011**	**.000**	**.003**	**.017**	**.046**	**-.003**	**.010**	**.001**	**.003**

All-in-all, Experiment 1 has demonstrated a clear advantage of using the conformal approach in a classifier with reject scenario targeting accuracy, compared to both uncalibrated and calibrated probabilistic predictors.

Turning to the results from Experiment 2, Table 8 shows the precision and recall of decision tree and random forest models, for the three setups used. Here, it can be seen that uncalibrated models again perform well, but also that Platt models are stronger on precision and conformal models stronger on recall.

Table 8: Experiment 2: Precision and recall

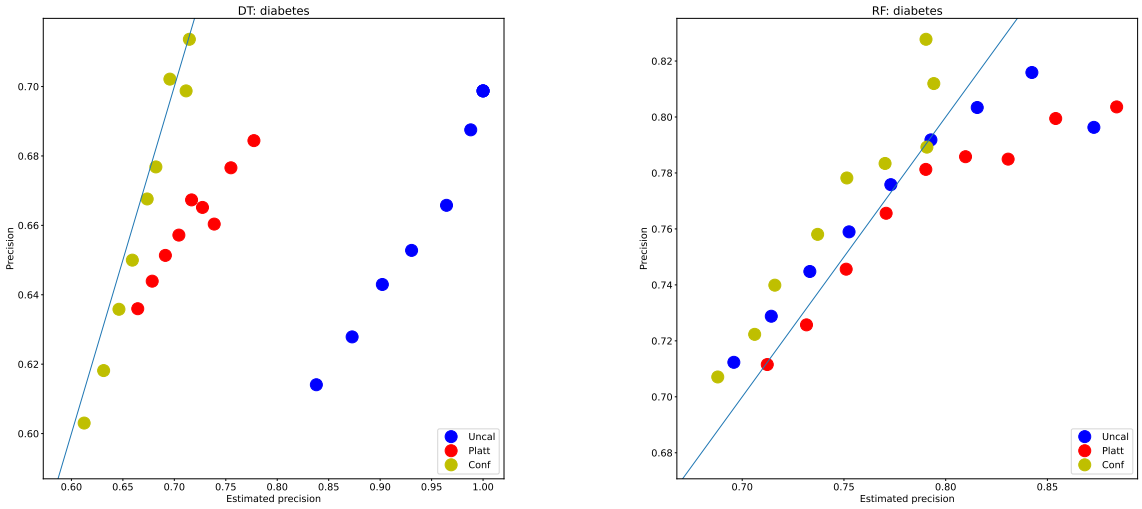
	DT						RF					
	Precision			Recall			Precision			Recall		
Data sets	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf	Uncal	Platt	Conf
creditA	.816	.820	.821	.832	.804	.803	.879	.867	.869	.846	.842	.840
diabetes	.598	.625	.592	.599	.467	.567	.694	.694	.689	.602	.575	.587
german	.709	.700	.714	.804	.998	.764	.720	.706	.721	.858	.988	.857
kc1	.405	.493	.403	.372	.034	.325	.554	.599	.530	.322	.166	.297
kc2	.515	.591	.536	.539	.381	.516	.616	.636	.604	.520	.440	.500
kr-vs-kp	.976	.977	.978	.981	.973	.971	.987	.983	.980	.994	.989	.991
pc4	.522	.557	.514	.509	.312	.474	.751	.665	.736	.338	.422	.310
transfusion	.446	.493	.469	.405	.149	.356	.456	.503	.450	.358	.119	.345
ttt	.919	.908	.912	.952	.940	.931	.978	.972	.944	.999	.982	.993
wbc	.918	.915	.912	.932	.914	.916	.956	.955	.955	.965	.955	.960
Mean	**.682**	**.708**	**.685**	**.692**	**.597**	**.662**	**.759**	**.758**	**.748**	**.680**	**.648**	**.668**

Table 9 below shows the results for the Mondrian conformal classifier used in Experiment 2. Due to space limitations, we only show results for $\epsilon = 0.05$. First of all, it is important to remember that this Mondrian conformal classifier uses a taxonomy where the category investigated is when the underlying model has predicted the label 1. So, the error rates that are found to match the significance level, are measured only on the test instances predicted as label 1, but regardless of whether the true target values are 0 or 1. Errors of course still mean that the correct label is not included in the prediction set, so it becomes interesting to look at the results for when the true target is 0 and 1 separately, which is presented in the columns *err0* and *err1*. What we see is the expected behavior; for most data sets all errors actually come from instances where the underlying model is incorrect, i.e., the true target is 0. Only for the easiest data sets, i.e., *kr-vs-kp*, *ttt*, and *wbc*, some errors are made when the true target is 1. For these instances, the prediction set is actually empty, meaning that the conformal classifier needs empty predictions, i.e., errors, to meet the significance level. Comparing *avgC*, *precision* and *recall*, we see that the stronger underlying RF models lead to smaller prediction sets, as well as higher precision and recall.

Table 9: Experiment 2: Mondrian conformal classifier

	DT						RF					
Data sets	error	err0	err1	avgC	prec.	rec.	error	err0	err1	avgC	prec.	rec.
creditA	.059	.328	.000	1.520	.821	.803	.045	.343	.000	1.317	.869	.840
diabetes	.051	.126	.000	1.826	.592	.567	.046	.147	.000	1.763	.689	.587
german	.053	.184	.000	1.768	.714	.764	.048	.171	.000	1.722	.721	.857
kc1	.047	.079	.000	1.891	.403	.325	.047	.101	.000	1.857	.530	.297
kc2	.045	.098	.000	1.851	.536	.516	.056	.140	.000	1.766	.604	.500
kr-vs-kp	.050	.997	.030	.958	.992	.943	.051	.999	.031	.955	.994	.960
pc4	.055	.114	.000	1.842	.514	.474	.057	.216	.000	1.696	.736	.310
transfusion	.049	.092	.000	1.899	.469	.356	.058	.106	.000	1.885	.450	.345
ttt	.052	.586	.001	1.137	.912	.930	.050	.767	.008	1.001	.952	.985
wbc	.048	.519	.003	1.225	.913	.913	.044	.835	.006	1.077	.956	.954
Mean	**.051**	**.312**	**.003**	**1.592**	**.687**	**.659**	**.050**	**.382**	**.005**	**1.504**	**.750**	**.663**

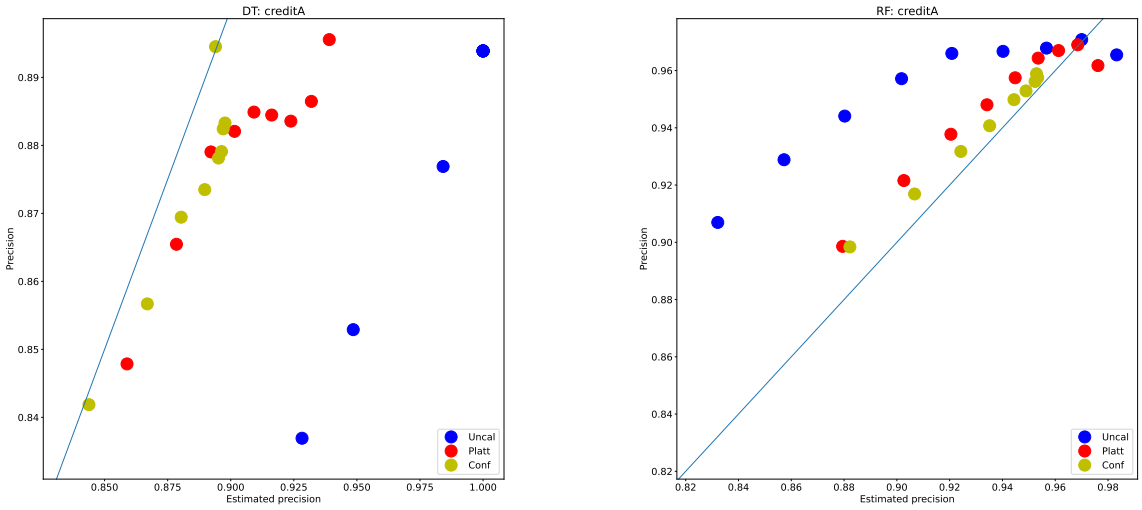
Figures 4 – 7, show plots of estimated precision against empirical precision, for all rejection levels, for four different data sets. For the Diabetes data set, in Figures 4a – 4b, uncalibrated decision trees have the same problems with precision estimation as previously seen with accuracy, i.e. the model substantially over-estimates its precision, for every rejection level. Platt models are slightly over-confident, and also achieve lower precision at high rejection levels. The Mondrian conformal approach is seen to work very well, with almost perfect estimation and the best precision scores for the top rejection levels. For random forest models, all three setups perform similarly for the lower rejection levels, giving reasonably good precision estimates, but this does not hold on top rejection levels, where models are either under-confident (Mondrian conformal) or over-confident (Platt and uncalibrated).



(a) DT (b) RF

Figure 4: Diabetes data set: precision estimation

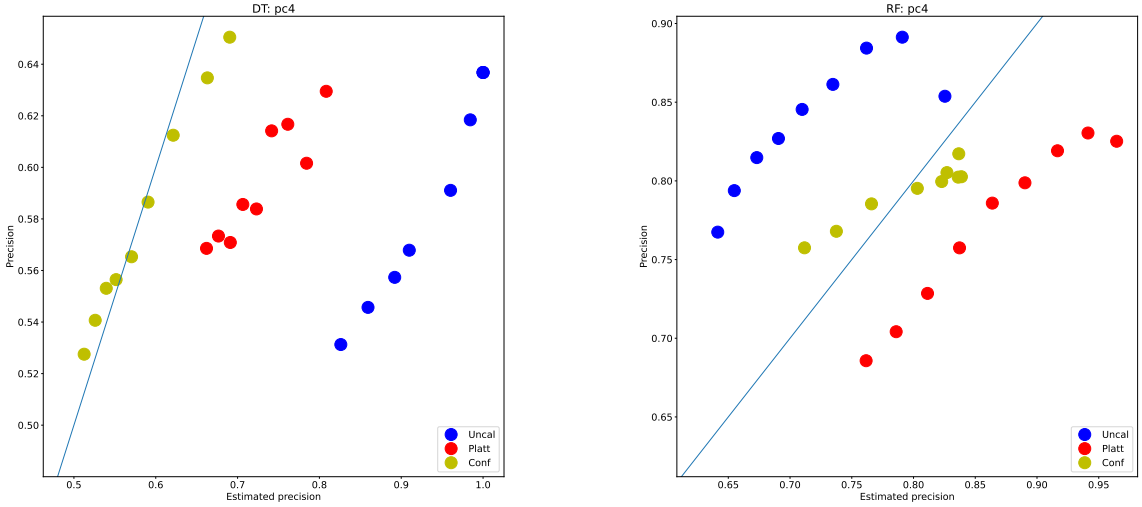
For the CreditA data set, Figures 5a – 5b, uncalibrated decision tree models are unable to differentiate between rejection levels, as seen by several identical estimated and empirical precision scores. Mondrian conformal and Platt scaling perform similarly, but with a pronounced tendency for Platt models to over-estimate precision at the high rejection levels. Uncalibrated random forest models for this data set substantially under-estimate precision on the lower rejection level, whereas both Platt and Mondrian conformal models give reasonably good estimates.



(a) DT (b) RF

Figure 5: CreditA data set: precision estimation

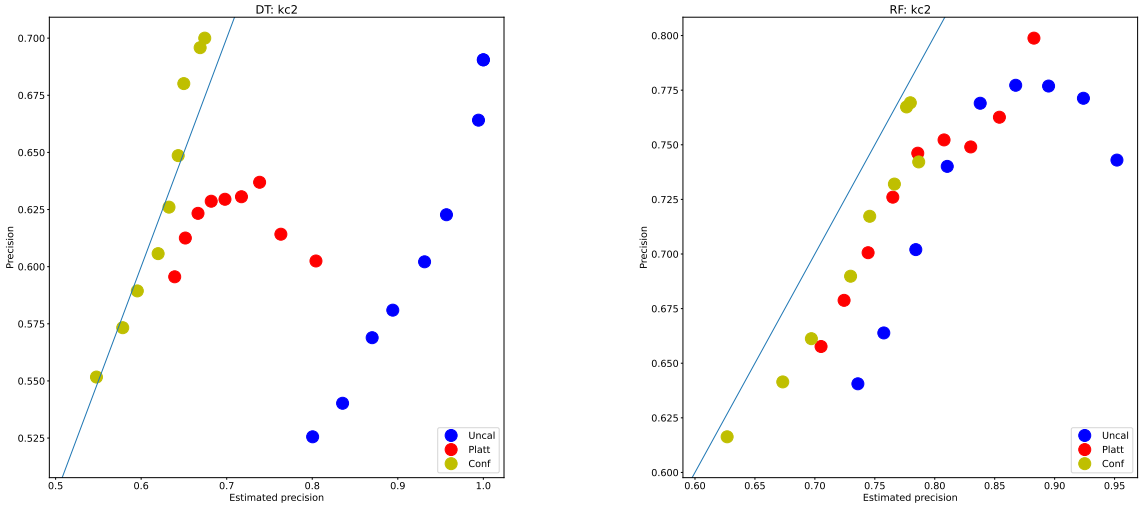
For the PC4 data set, Figures 6a – 6b, the Mondrian conformal approach again provides much better precision estimates than Platt and uncalibrated models, for both decision trees and random forests. Regarding precision performance, however, uncalibrated random forest models outperform both Mondrian conformal and Platt calibration models.



(a) DT (b) RF

Figure 6: PC4 data set: precision estimation

For the KC2 data set, Figures 7a – 7b, arguably the most interesting result is the marked drop in precision on the highest rejection levels, seen for both Platt calibrated decision trees, and uncalibrated random forest models.



(a) DT (b) RF

Figure 7: KC2 data set: precision estimation

Aggregated precision results for the classification with reject option are shown in Tables 10 – 12, for rejection proportions of 50%, 70% and 90%, respectively. For all three rejection levels, Mondrian conformal models achieve very good average precision estimations, compared to actual precision, and this also holds on individual data set level. Uncalibrated decision trees is seen to really struggle on this task, sometimes giving precision estimates

of 100%, while actual precision is under 60%. In fact, the average difference between estimated and actual precision, for uncalibrated decision trees, is over 20 percentage points, for all three rejection levels. Random forest models do better, but estimates degrade as rejection proportion goes up. Calibration using Platt scaling works worse for precision than for accuracy, with larger differences between estimated and actual precision.

Table 10: Experiment 2: Precision estimation. Top 50%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.
creditA	1.000	.894	.909	.885	.895	.878	.921	.966	.945	.957	.944	.950
diabetes	.964	.666	.717	.667	.674	.668	.773	.776	.790	.781	.751	.778
german	.948	.738	.740	.717	.749	.746	.878	.788	.755	.763	.773	.778
kc1	.866	.473	.559	.527	.457	.480	.717	.635	.643	.664	.577	.606
kc2	.931	.602	.698	.629	.633	.626	.838	.769	.786	.746	.746	.717
kr-vs-kp	1.000	.998	.992	.997	.996	.996	.989	.999	.997	.999	.997	.997
pc4	.960	.591	.723	.584	.570	.565	.710	.845	.864	.786	.823	.800
transfusion	.784	.491	.650	.502	.508	.514	.803	.474	.625	.500	.507	.503
ttt	1.000	.977	.964	.975	.971	.967	.929	1.000	.998	1.000	.993	.994
wbc	1.000	.964	.962	.955	.947	.950	.991	.989	.979	.982	.973	.975
Mean	**.945**	**.739**	**.791**	**.744**	**.740**	**.739**	**.855**	**.824**	**.838**	**.818**	**.808**	**.810**

Table 11: Experiment 2: Precision estimation. Top 30%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.
creditA	1.000	.894	.924	.884	.898	.883	.957	.968	.961	.967	.953	.958
diabetes	1.000	.699	.739	.660	.696	.702	.816	.803	.831	.785	.791	.789
german	.997	.770	.755	.731	.782	.771	.923	.833	.773	.792	.814	.824
kc1	.942	.511	.576	.550	.504	.518	.763	.669	.675	.667	.634	.636
kc2	.994	.664	.738	.637	.650	.680	.895	.777	.830	.749	.787	.742
kr-vs-kp	1.000	.998	.993	.997	.996	.996	.995	.999	.998	.999	.997	.997
pc4	1.000	.637	.761	.617	.621	.612	.762	.884	.917	.819	.837	.817
transfusion	.843	.502	.684	.493	.522	.503	.857	.508	.655	.498	.512	.525
ttt	1.000	.977	.969	.979	.976	.974	.954	1.000	.999	1.000	.992	.994
wbc	1.000	.964	.971	.956	.953	.954	.998	.997	.984	.987	.973	.975
Mean	**.978**	**.762**	**.811**	**.750**	**.760**	**.759**	**.892**	**.844**	**.862**	**.826**	**.829**	**.826**

Table 12: Experiment 2: Precision estimation. Top 10%

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Data sets	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.	Est.	Prec.
creditA	1.000	.894	.939	.896	.894	.895	.983	.965	.976	.962	.953	.959
diabetes	1.000	.699	.777	.684	.715	.714	.873	.796	.884	.804	.790	.828
german	1.000	.770	.779	.737	.798	.758	.970	.912	.802	.840	.872	.887
kc1	1.000	.570	.604	.696	.545	.563	.835	.684	.724	.716	.651	.674
kc2	1.000	.691	.804	.602	.674	.700	.952	.743	.883	.799	.776	.767
kr-vs-kp	1.000	.998	.996	.997	.996	.996	1.000	1.000	.999	.999	.997	.996
pc4	1.000	.637	.808	.629	.690	.650	.825	.854	.965	.825	.827	.805
transfusion	.927	.476	.764	.474	.509	.520	.929	.544	.707	.500	.567	.476
ttt	1.000	.977	.979	.977	.977	.969	.981	1.000	1.000	1.000	.992	.996
wbc	1.000	.964	.986	.961	.953	.955	1.000	.998	.990	.993	.975	.970
Mean	**.993**	**.767**	**.844**	**.765**	**.775**	**.772**	**.935**	**.850**	**.893**	**.844**	**.840**	**.836**

Table 13 summarizes Experiment 2 by showing signed and absolute errors for the precision estimations made by the different setups on each rejection level. Similar to the accuracy estimations, uncalibrated decision trees produce very overconfident precision estimations, on all rejection levels. For Platt scaling, however, the decision tree results are quite different, compared to when targeting accuracy. Here, the estimations are systematically overconfident, and often fairly large. The conformal approach, however, makes no systematic mistake, and both signed errors and absolute errors are close to zero. For random forests, both uncalibrated models and Platt scaling are systematically overconfident. While the absolute errors for Platt scaling are much smaller than for the uncalibrated models, only the conformal approach show no systematic tendency. In addition, the conformal absolute errors are close to zero, and smaller than the two competing approaches, on every rejection level.

Interestingly enough, Experiment 2 has shown that using the conformal approach may be even more beneficiary when the classifier with reject scenario targets precision, instead of accuracy. In particular, basing the estimations on a conformal predictor, with the associated guarantees, clearly outperformed the standard procedure of using probabilistic predictors.

Summarizing the experimentation, it is very interesting to see the advantage of using the more informed confidence measure produced by a conformal classifier, compared to the probability estimates of probabilistic predictors. The main difference is, of course, that the confidence measure for every instance represents a property of a set of instances, while a standard probability estimate is restricted to a single instance. The clear and substantial benefit of this, in a classifier with reject option, is a key result of this study. Another important novel contribution is the fact that by using a Mondrian approach, the classifier with reject option could produce well-calibrated precision estimations.

Table 13: Experiment 2: Precision estimation. The table shows signed errors and absolute errors, averaged over all data sets, for the different rejection levels.

	DT						RF					
	Uncal		Platt		Conf		Uncal		Platt		Conf	
Rejected	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.	Sig.	Abs.
10%	.162	.162	.033	.035	-.005	.007	.009	.087	.017	.024	-.013	.015
20%	.175	.175	.033	.036	.001	.006	.015	.088	.021	.027	-.007	.013
30%	.185	.185	.037	.040	.001	.006	.019	.086	.020	.027	-.006	.013
40%	.195	.195	.041	.045	.002	.007	.025	.084	.019	.027	.000	.012
50%	.206	.206	.047	.051	.001	.007	.031	.084	.020	.030	-.001	.012
60%	.212	.212	.054	.056	.001	.006	.039	.083	.027	.036	.004	.011
70%	.216	.216	.061	.064	.000	.011	.048	.085	.036	.042	.003	.010
80%	.220	.220	.075	.076	.001	.015	.060	.087	.045	.052	.004	.013
90%	.225	.225	.078	.097	.003	.015	.085	.095	.049	.058	.004	.021
Mean	**.200**	**.200**	**.051**	**.055**	**.001**	**.009**	**.037**	**.086**	**.028**	**.036**	**-.001**	**.014**

5. Concluding remarks.

We have in this paper evaluated conformal prediction for classification with reject option. Utilizing the strong theoretical properties of conformal prediction, standard and Mondrian conformal classifiers were used as the basis for classifiers with reject option, targeting either accuracy or precision. The empirical evaluation, using ten publicly available data sets, showed that the suggested method produced very exact accuracy and precision estimates, for all rejection levels investigated.

A direct comparison with probabilistic predictors clearly demonstrates the advantage of the conformal approach. Even when calibrating the probabilistic predictors using Platt scaling, the resulting estimations were outperformed by the conformal classifiers, in particular for precision. Specifically, only the conformal models showed no systematic bias when estimating either accuracy or precision for the different rejection levels and using the two underlying models decision trees and random forests.

Acknowledgements

The authors acknowledge the Swedish Knowledge Foundation, Jönköping University, and the industrial partners for financially supporting the research through the AFAIR project with grant number 20200223 and the PREMACOP project with grant number 20220187, as part of the research and education environment SPARK at Jönköping University, Sweden.

References

Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Dr.Regina Barzilay. Conformal prediction sets with limited false positives. In *Proceedings of the 39th International Conference on Machine Learning*, pages 6514–6532. PMLR, 17–23 Jul 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330, 2017.

Blaise Hanczar and Edward R. Dougherty. Classification with reject option in gene expression data. *Bioinform.*, 24(17):1889–1895, 2008.

Blaise Hanczar and Michèle Sebag. Combination of one-class support vector machines for classification with reject option. In *Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer Berlin Heidelberg, 2014.

Lars Kai Hansen, Christian Liisberg, and Peter Salamon. The error-reject tradeoff. *Open Systems & Information Dynamics*, 4(2):159–184, 1997.

U. Johansson and P. Gabrielsson. Are traditional neural networks well-calibrated? In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

Ulf Johansson, Tuwe Löfström, Cecilia Sönströd, and Helena Löfström. Conformal prediction for accuracy guarantees in classification with reject option. *The 20th International Conference on Modeling Decisions for Artificial Intelligence. In press*, 2023.

Mingkun Li and Ishwar K. Sethi. Confidence-based classifier design. *Pattern Recognition*, 39(7):1230–1240, 2006. ISSN 0031-3203.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Classification with reject option using conformal prediction. In *PAKDD*, pages 94–105. Springer, 2018.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632. ACM, 2005.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. *Tools in Artificial Intelligence*, 18:315–330, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356. Springer, 2002.

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.

Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Mach. Learn.*, 52(3):199–215, 2003.

J. Sayyad Shirabad and T.J. Menzies. The PROMISE Repository of Software Engineering Databases. School of IT and Engineering, Univ. of Ottawa, Canada, 2005.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., 2005.