International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00573-2

REGULAR PAPER

Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests

Stefan Haas[1,2] · **Eyke Hüllermeier**[1,3]

Received: 30 January 2023 / Accepted: 21 May 2024

Abstract
Due to the inherent presence of uncertainty in machine learning (ML) systems, the usage of ML is until now out of scope for many critical (financial) business processes. One such process is goodwill assessment at car manufacturers, where a large part of goodwill cases is still assessed manually by human experts. To increase the degree of automation while still providing an overall reliable assessment service, we propose a selective uncertainty-aware automated decision making approach based on uncertainty quantification through conformal prediction. In our approach, goodwill requests are still shifted to human experts in case the risk of a wrong assessment is too high. Nevertheless, ML can be introduced into the process with reduced and controllable risk. We hereby determine the risk of wrong ML assessments through two hierarchical conformal predictors that make use of the prediction set and interval size as the main criteria for quantifying uncertainty. We also utilize conformal prediction's property to output empty prediction sets if no prediction is significant enough and abstain from an automatic decision in that case. Instead of providing mathematical guarantees for limited risk, we focus on the risk vs. degree of automation trade-off and how a business decision maker can select in an a posteriori fashion a trade-off that best suits the business problem at hand from a set of pareto optimal solutions. We also show empirically on a goodwill data set of a BMW National Sales Company that by only selecting certain requests for automated decision making we can significantly increase the accuracy of automatically processed requests. For instance, from 92 to 98% for labor and from 90 to 98% for parts contributions respectively, while still maintaining a degree of automation of approximately 70%.

Keywords Uncertainty quantification · Conformal prediction · Selective classification · Prescriptive machine learning

1 Introduction

Many business processes in industry are still based on manual human execution steps, checks and assessments. These manual processes are often in place for years, if not decades. Hence, a lot of historical transactional data slumbers in IT systems that could be used to design data driven decision agents using supervised machine learning (SML). Trained machine learning (ML) models can then be used to either fully automate business processes through automated decision making (ADM) or at least to assist during the process in the form of a decision support system (DSS), where unlike in ADM the human expert is still in control over the final decision. Automating business processes is beneficial since it reduces process costs and potentially also increases standardization. Consequently, there is a noticeable shift from the usage of ML for *predictive* modeling towards *prescriptive* modeling, where appropriate actions are supposed to be triggered in real world scenarios. This trend has recently been coined *prescriptive machine learning* [17].

Nevertheless, the usage of data-induced decision agents is not free of risk. The decisions of an ML model cannot be considered correct all the time, for instance, there might be issues related to the (training) data, such as data and concept drift or shift, inadequate or wrong supervision (human decisions cannot always be considered as ground truth) or even inherent non-determinism in the dependency between input and output. This last uncertainty is often referred to as *aleatoric* uncertainty. Uncertainty with regards to the quality and amount of training data is known as *approximation*

✉ Stefan Haas
stefan.sh.haas@bmwgroup.com

Eyke Hüllermeier
eyke@lmu.de

[1] Institute of Informatics, LMU Munich, Munich, Germany

[2] BMW Group, Munich, Germany

[3] Munich Center for Machine Learning, Munich, Germany

Published online: 14 June 2024

uncertainty. Identifying the right type of model for a particular problem is referred to as *model uncertainty*. Both previous uncertainties can be attributed to *epistemic* uncertainty, which is reducible unlike *aleatoric* uncertainty [18].

With the before mentioned uncertainties, it is hardly conceivable that high-stake business domains will immediately go from a purely manual human decision process to a fully ML automated process at once since this would entail a lot of (financial) risk. A practical approach could be to first automate rather clear or certain cases and still leave the more complex or uncertain cases to a human expert. In recent years, the topic of uncertainty quantification in machine learning has gained a lot of attention [12, 22, 33]. The capability of a machine learning model to quantify its uncertainty related to a certain query could be utilized to quantify the risk of a wrong decision. Knowing the potential risk of a wrong decision for a particular query could then serve as a means to distinguish between fully automated decision making and decision support. Roughly speaking, when the risk of a wrong decision is high, the machine learning model is (at most) supposed to be used as a decision support and the final decision must be left to a human expert. In contrast, if the risk of a wrong decision is considered low, the process can be fully automated through automated decision making.

A versatile method for quantifying uncertainty, that is also widely used in practice, is conformal prediction [8, 9, 20, 23, 36]. As a foundation, conformal prediction only requires a model that is capable of outputting heuristic probabilities which makes it almost model agnostic and broadly applicable. Consequently, in this paper we will evaluate how uncertainty quantification with conformal prediction can be used to draw an uncertainty-aware decision boundary between automated decision making and decision support, where the final decision is still up to a human expert. We will do this by means of a case study using a goodwill data set of a BMW National Sales Company (NSC) containing customer goodwill requests and manual contribution decisions made by human experts.

2 Machine learning for automated decision making

In many business domains there is a demand for automating repetitive tasks through machine learning with the main goal to free work force and thereby save costs. One such exemplary business process is goodwill assessment, where a (car) manufacturer compensates customers in cases of product related queries outside of the warranty window (usually after 3–5 years). The aim of granting goodwill is to keep customers satisfied and loyal to the brand. To a large extent, these goodwill assessments are still carried out manually at BMW. Business experts check the goodwill requests, which contain extensive information regarding the vehicle and the present problem, and subsequently grant a certain repair cost contribution percentage (binned to ten percent steps, i.e., elements of $\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$) separately for labor and parts.

Since this manual process is in place for years, there is plenty of data that can be used for machine learning. This data comes in the form

$$\mathcal{D} = \left\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\right\},$$

with goodwill requests represented as *feature vectors* $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^m$ and observed human goodwill decisions as *labels* $y_i \in \mathcal{Y}$. This is exactly the type of data commonly assumed in the setting of supervised machine learning, where the goal is to learn an optimal predictor $h^* \in \mathcal{H}$ maximizing predictive accuracy, or, more generally, minimizing the expected loss (risk)

$$\mathcal{R}(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim P}\, l(y, h(\boldsymbol{x}))\,, \tag{1}$$

where $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function and the expectation is taken with respect to the data generating process P (a joint probability measure on $\mathcal{X} \times \mathcal{Y}$). Moreover, $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ is the set of predictors (mappings $\mathcal{X} \to \mathcal{Y}$) the learner can choose from; this set is also called the hypothesis space in machine learning.

As already said, the goodwill use case qualifies as what has recently been coined *prescriptive* machine learning [17]. In contrast to the common setting of *predictive* machine learning, the goal is not to predict some underlying ground-truth, but rather to learn models that stipulate appropriate decisions or actions to be taken in order to achieve a certain goal. In fact, in the case of goodwill, one may argue that there is nothing like a "right" or "true" monetary contribution, nor is a decision either right or wrong. Instead, a decision is more or less appropriate, fair for the customer and strategically opportune for the company. From this point of view, one may also question the idea of learning a model that seeks to mimic the human expert, taking her decisions as a target for prediction [34], all the more since these decisions appear to be biased. For example, we found that a decision of 50% contribution is somewhat overrepresented in the data, letting one suspect that this is often taken as a default choice for a partial cost coverage, even if it might not necessarily be the most appropriate percentage. In the following, we will nevertheless assume that mimicking the expert is a reasonable strategy, at least as a first step toward a data-driven goodwill assessment, leaving more elaborate approaches for future work.

Under this premise, the problem can essentially be tackled by methods for supervised learning, which, in one way or the

other, replace the true risk (1) as a target of optimization by the *empirical risk*

$$\mathcal{R}_{emp}(h) := \frac{1}{n}\sum_{i=1}^{n} l(y_i, h(\boldsymbol{x}_i)) .$$

As opposed to the true risk, which requires knowledge of P, the latter can be computed on the training data.

3 Uncertainty in automated decision making

Since $\mathcal{R}_{emp}(h)$ is only an estimation of the true risk $\mathcal{R}(h)$, the empirical risk minimizer

$$\hat{h} := \underset{h\in\mathcal{H}}{\arg\min}\, \mathcal{R}_{emp}(h)$$

(or the minimizer of any variant of the empirical risk) will at best approximate but not equal the true risk minimizing hypothesis

$$h^* := \underset{h\in\mathcal{H}}{\arg\min}\, \mathcal{R}(h) .$$

Consequently, there is uncertainty related to a presumably sub-optimal model $\hat{h}$, the prescriptions of which might not always be appropriate. Hence, in business processes like goodwill assessment, where wrong decisions might heavily impact customer satisfaction and also have a financial impact on the manufacturer, deploying prescriptive models without any safety mechanisms is hard to conceive.

From a risk minimizing perspective it is reasonable to equip the model with a *reject option* and to abstain from an automatic decision in case the uncertainty related to a query $\boldsymbol{x}$ is too high. Abstaining from decisions and trading off coverage for higher classification accuracy is also known as *selective classification* [11]. A standard *selective classifier* consists of a *classifier function* f and a binary *selection function* $g : \mathcal{X} \to \{0, 1\}$ which controls whether the classifier f abstains from a prediction or not:

$$(f, g)(\boldsymbol{x}) := \begin{cases} f(\boldsymbol{x}) & \text{if } g(x) = 1 \\ \varnothing & \text{if } g(x) = 0 \end{cases} .$$

In our specific assessment use case, since there is already a manual human assessment process in place, *rejection* means to forward the query to a human expert for a manual assessment. The whole assessment process could hereby be considered as a piecewise function $a(\boldsymbol{x})$, with the sub-functions $\hat{h}(\boldsymbol{x})$ and $m(\boldsymbol{x})$ for automatic prescriptive machine learning and manual human assessment, respectively:

$$a(\boldsymbol{x}) = (\hat{h}, m, g)(\boldsymbol{x}) := \begin{cases} \hat{h}(\boldsymbol{x}) & \text{if } g(\boldsymbol{x}) = 1 \\ m(\boldsymbol{x}) & \text{if } g(\boldsymbol{x}) = 0 \end{cases} .$$

Whether the input $\boldsymbol{x}$ is selected for prescription or not depends on a risk assessment with regard to $\boldsymbol{x}$ and $\hat{h}(\boldsymbol{x})$. In case the risk $\mathcal{R}_{\hat{h}}(\boldsymbol{x})$ associated with a query $\boldsymbol{x}$ exceeds a predefined risk threshold δ, the query is not supposed to be processed automatically and the selection function will make the system abstain:

$$g_\delta(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \mathcal{R}_{\hat{h}}(\boldsymbol{x}) \leq \delta \\ 0 & \text{otherwise} \end{cases} .$$

A tradeoff between reliability and degree of automation is inherent in an ML-enhanced assessment process $a(\boldsymbol{x})$. Since ML results produced by $\hat{h}(\boldsymbol{x})$ will most likely not be perfect all the time, there is a serious risk of wrong (maybe costly) ML decisions that might significantly impact the overall reliability of the process. This loss in reliability can be circumvented by shifting requests with high risk to human experts $m(\boldsymbol{x})$, which in turn will come at a loss of automation. In order to maximize the degree of automation while still maintaining sufficient reliability in the decision process, accurately quantifying the risk related to a request $\boldsymbol{x}$ is crucial. For business domains it is of great interest to find an optimal degree of automation vs. risk of inappropriate decisions depending on the criticality of the business process and its associated costs. This trade-off between risk and degree of automation is also known as the *risk-coverage (RC) trade-off* [11].

4 Reliable decision making using conformal prediction

In the following, we will outline our selective uncertainty-aware approach to automated decision making. We will start with enhancing our existing hierarchical model with conformal prediction, which allows us to quantify uncertainty associated to queries. In the next step, we will turn these uncertainties into risk values. Finally, we discuss how we can optimize the trade-off between risk and the degree of automation on the system level using multi-objective optimization. In the end, it is then up to a business decision maker (DM) to select a *Pareto-optimal* solution that best suits the use case at hand.

4.1 Conformal prediction for uncertainty quantification

One method that is widely used to quantify uncertainty is conformal prediction [3, 32, 36]. Unlike in a standard clas-

sification scenario, where a predictor outputs a single class (*point prediction*), conformal prediction outputs a *prediction set* $\Gamma^\epsilon(\boldsymbol{x})$ which is guaranteed to contain the correct label y with a probability of $1-\epsilon$, where $\epsilon > 0$ is a user-defined *significance level* or *error rate*. For instance, $\epsilon = 0.05$ means that the algorithm is allowed to make at most 5% invalid predictions on average. More formally, prediction sets $\Gamma^\epsilon(\boldsymbol{x})$ are guaranteed to fulfill the following property, which is also referred to as *marginal coverage*:

$$1-\epsilon \leq P(y \in \Gamma^\epsilon(\boldsymbol{x})) \leq 1-\epsilon+\frac{1}{n+1},$$

where n is the number of training examples seen by the learning algorithm so far.

The construction of prediction sets relies on so-called *nonconformity scores* $s(\boldsymbol{x}, y) \in \mathbb{R}$, which can be interpreted as a measure of plausibility of the input/output pair $(\boldsymbol{x}, y)$ in light of the data $\mathcal{D}$ seen so far: the higher the value $s(\boldsymbol{x}, y)$, the less the (hypothetical) data point $(\boldsymbol{x}, y)$ "fits" the (truly observed) training data. The standard inductive conformal prediction (ICP) algorithm consists of the following steps [1, 29, 30]:

1. Split the available data into a training, calibration, and test data set.
2. Induce a predictive model h on the training data.
3. Define a score function $\alpha = s(\boldsymbol{x}, y) \in \mathbb{R}$, where larger scores mean higher *non-conformity* of $(\boldsymbol{x}, y)$; for example, if h is a scoring classifier, $s(\boldsymbol{x}, y)$ could be given by the score assigned to y by $h(\boldsymbol{x})$.
4. Compute the critical value $\hat{q}$ as the $\frac{\lceil (n+1)(1-\epsilon) \rceil}{n}$ empirical quantile (which is essentially 1 - ϵ with a small correction) of the *true* calibration scores $\alpha_1 = s(\boldsymbol{x}_1, y_1), \ldots, \alpha_n = s(\boldsymbol{x}_n, y_n)$
5. Use the critical value $\hat{q}$ to calculate the prediction sets for new before unseen examples:

$$\Gamma^\epsilon(\boldsymbol{x}) = \{y : \alpha = s(\boldsymbol{x}, y) \leq \hat{q}\}$$

The value $\hat{q}$ plays the role of a p-value as known from statistical hypothesis testing. Such a p-value can also be associated with every candidate outcome:

$$p(\boldsymbol{x}, y) = \frac{\#\{i \in \{1, \ldots, n+1\} \mid \alpha_i \geq \alpha_{n+1} = s(\boldsymbol{x}, y)\}}{n+1}.$$

Thus, $p(\boldsymbol{x}, y)$ corresponds to the percentage of (real) data points that are at least as nonconforming as $(\boldsymbol{x}, y)$. Consequently, the smaller $p(\boldsymbol{x}, y)$, the less plausible y can be considered as an outcome for $\boldsymbol{x}$, and the p-values of all candidate outcomes $y \in \mathcal{Y}$ allows one to sort them from most plausible to least plausible.

The prediction set $\Gamma^\epsilon(\boldsymbol{x})$ is obtained by cutting p-values at the threshold $\hat{q}$, thereby dichotomising $\mathcal{Y}$ into plausible and implausible candidates. Ideally, $\Gamma^\epsilon(\boldsymbol{x})$ is a singleton set, suggesting that there is exactly one plausible outcome while all other can be excluded. This is a case in which the learner can decide in an unequivocal way. More generally, the larger $|\Gamma^\epsilon(\boldsymbol{x})|$, the more uncertain the learner is. Obviously, the size of $\Gamma^\epsilon(\boldsymbol{x})$ is also influenced by the error probability ϵ: The smaller ϵ, the larger $\Gamma^\epsilon(\boldsymbol{x})$ tends to be.

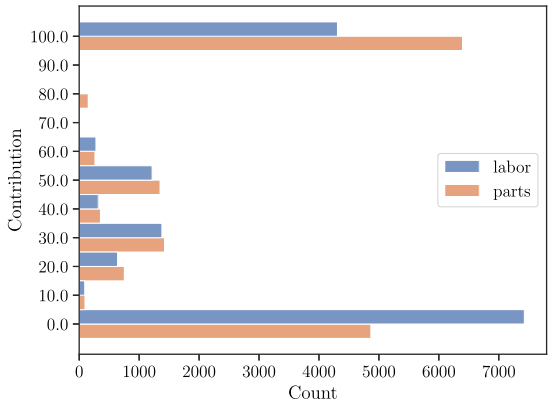
Interestingly, the prediction set can also be empty ($\Gamma^\epsilon(\boldsymbol{x}) = \emptyset$). This happens in cases where a query $\boldsymbol{x}$ cannot be combined into a sufficiently conforming tuple $(\boldsymbol{x}, y)$ with any of the candidates y, e.g., because $\boldsymbol{x}$ itself is an atypical case. Obviously, just like overly large prediction sets $\Gamma^\epsilon(\boldsymbol{x})$, empty predictions indicate a high level of uncertainty, suggesting to the learner that it might be better to abstain.

Let us finally make a remark on the error probability ϵ, which, as already mentioned, has a direct influence on the size of the prediction sets—and hence the probability that a learner may abstain from taking action. In conformal prediction, this value is normally quite small, with 0.1 and 0.05 being typical choices. Such values are also common in statistical hypothesis testing, so as to guarantee a low type-I error probability. While keeping the error probability low is reasonable in general, and indeed important in many applications, larger values of ϵ might be quite meaningful in applications such as goodwill assessment. Here, ϵ can also be seen as a parameter controlling the degree of automation and hence the workload of the human expert to whom ambiguous cases are transferred. In principle, ϵ can then also be tuned to the availability of human resources. Starting with a very small ϵ close to 0, all prediction sets will be full ($\Gamma^\epsilon(\boldsymbol{x}) = \mathcal{Y}$) and hence all cases rejected. By increasing ϵ step by step, the learner will become less cautious and exclude outcomes in a more aggressive way, thereby increasing the number of cases that can be decided automatically (and decreasing the workload of the human expert). If human resources are limited, this might be the only way to achieve the necessary level of automation.

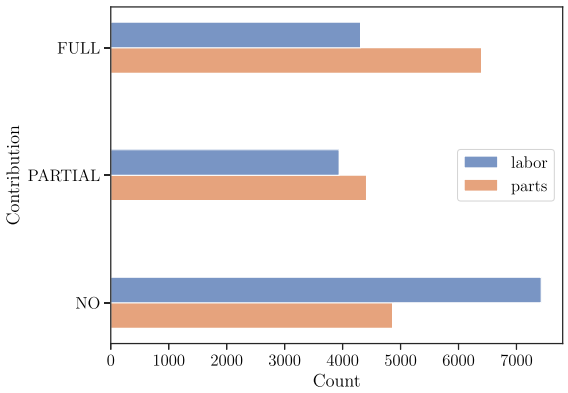
4.2 The hierarchical assessment model

For the model training step, we will re-use the hierarchical approach already outlined in [13]. It uses a *qualitative ranking layer* to predict the three main goodwill contribution ranks $\mathcal{Y}_{\text{rank}} = \{1, 2, 3\} = \{\text{NO, PARTIAL, FULL}\}$ and a subsequent *quantitative regression layer* for an exact prediction of the PARTIAL goodwill contributions ($\mathcal{Y}_{\text{partial}} = \{10, 20, \ldots, 90\}$).

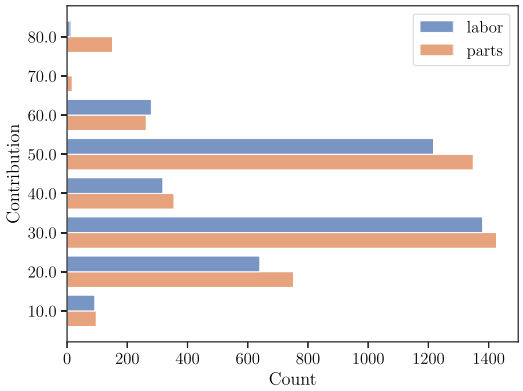
This hierarchical approach to goodwill assessment was chosen because the data is heavily imbalanced, with many 0 and 100% contributions on the one side and fewer, more widely distributed partial contributions on the other side [13]



(a) Raw distribution of contributions for one selected sales market.



(b) Collapsed main ranks distribution for ranking layer.



(c) Partial contributions distribution for regression layer.

Fig. 1 Distribution of the past contributions before and after the hierarchical restructuring

(cf. Fig. 1a). Combining the partial contribution data in the first layer counteracts this imbalance (cf. Fig. 1b).

Structuring the model hierarchically also makes sense from a risk assessment perspective, because errors in the qualitative ranking layer (e.g., NO vs. FULL contribution) potentially have a greater impact than errors in the quantitative regression layer (e.g., 50% vs. 80% contribution), both financially and on customer satisfaction.

In the hierarchical model, ranking is reduced to binary classifications using the framework presented in [25]:

$$r(\boldsymbol{x}) = 1 + \sum_{k=1}^{K-1} f(\boldsymbol{x}, k)\,. \tag{2}$$

Here, f is a binary predictor trained to answer the question whether the true rank of $\boldsymbol{x}$ exceeds k (in which case $f(\boldsymbol{x}, k) = 1$, otherwise 0). Data for training f is constructed from the original training data. To this end, $K-1$ new training examples are produced for each original training example $(\boldsymbol{x}, y)$, one for every k[1]:

$$\boldsymbol{x}^k = (\boldsymbol{x}, k), \quad y^k = [\![k < y]\!], \quad w_{y,k} = |C_{y,k} - C_{y,k+1}|\,.$$

Here, $w_{y,k}$ is the weight of the training example,[2] which is derived from the original cost-matrix: $C_{y,k}$ is the cost of predicting k when the ground-truth is y (see Implementation section for an example of a neutral cost matrix). Using this cost sensitive approach for training the models, different strategies can be implemented, e.g., customer friendly vs. cost oriented.

Figure 2 summarizes the architecture of our uncertainty-aware approach with each model layer being equipped with an additional risk assessment and reject option. The model can abstain from a decision when the risk assessment step

[1] $[\![\cdot]\!]$ denotes the indicator function returning 1 if the argument is true and 0 otherwise.

[2] The binary classifier used must hence be able to handle weighted examples.

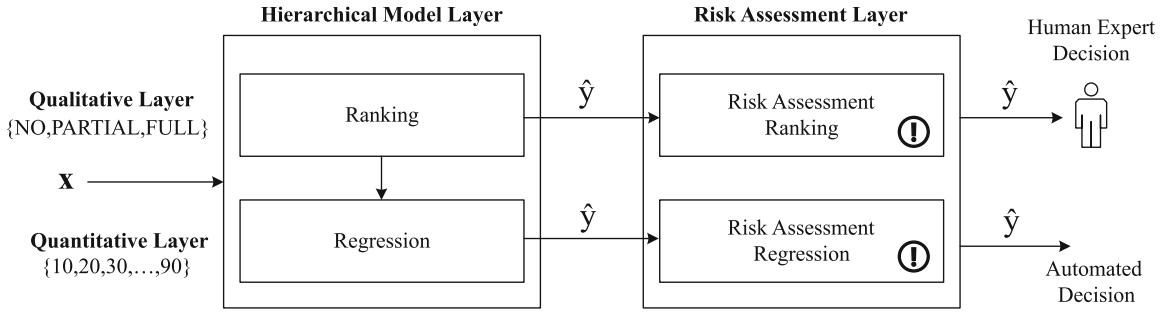


Fig. 2 Overview of uncertainty-aware goodwill assessments with reject option

indicates a too high risk for a wrong assessment. Rejecting a decision in our case means to forward the query to a human expert for manual assessment. Nonetheless, the model output can be used to assist the expert in the form of a decision support system (DSS). In this case, the human expert is in full control of the final decision but also gets the model's output presented to support her in the decision process.

4.3 Conformalizing the hierarchical model

A core engineering task, which has a major influence on the quality of conformal prediction, is to build a good nonconformity function that entails all known information about the data and the model. Based on the outputs of the nonconformity function, the critical value $\hat{q}$ that controls the outcomes to be put into the final prediction set is determined.

4.3.1 Conformalizing the ranking layer

Recall the binary predictor that we use to define the ranking function (2). We realize this predictor by training a probabilistic classifier, i.e., by setting $f(\boldsymbol{x}, k) = [\![p(y = 1 \mid \boldsymbol{x}, k) > 1/2]\!]$, where $p(y = 1 \mid \boldsymbol{x}, k)$ is the (predicted) probability that the rank of $\boldsymbol{x}$ exceeds k. To define a nonconformity score for the ranking layer, we refer to these probabilistic predictions:

$$s_{\text{rank}}(\boldsymbol{x}, y) := \left| \left(1 + \sum_{k=1}^{K-1} \hat{p}(y = 1 \mid \boldsymbol{x}, k) \right) - y \right| \in [0, K-1] .$$

The sum over probabilities yields a "soft" rank expressed in terms of a real (instead of an integer) number in $[1, K]$, and $s_{\text{rank}}(\boldsymbol{x}, y)$ is a measure of distance of that number to the rank y.

The prediction set for the ranking layer is given by

$$\Gamma^{\epsilon}_{\text{RA}}(\boldsymbol{x}) = \{ y \mid s_{\text{rank}}(\boldsymbol{x}, y) \leq \hat{q} \} \subseteq \{1, 2, 3\} ,$$

where $\hat{q}$ is the critical value obtained on the calibration data for the significance level ϵ.

4.3.2 Conformalizing the regression layer

Nonconformity scores for the regression layer can be obtained using quantile regression (QR), which is the standard approach to create a notion of uncertainty for real-valued problems [1, 31]. Depending on the significance level ϵ, a lower $(\epsilon/2)$ and an upper quantile $(1 - \epsilon/2)$ need to be determined. QR yields prediction intervals of the form $[\hat{t}_{\epsilon/2}(\boldsymbol{x}), \hat{t}_{1-\epsilon/2}(\boldsymbol{x})]$, and the width of these intervals serves as a heuristic notion of uncertainty. The score function can be defined as the projective distance of a candidate outcome y to the interval:

$$s_{\text{reg}}(\boldsymbol{x}, y) := \max \left\{ \hat{t}_{\epsilon/2}(\boldsymbol{x}) - y, y - \hat{t}_{1-\epsilon/2}(\boldsymbol{x}) \right\}$$

Note that $s_{\text{reg}}(\boldsymbol{x}, y)$ is negative for values y inside the interval and positive outside; the minimal value is obtained for the midpoint of the interval.

Using conformal prediction, the scores can then be calibrated as usual. The prediction interval for conformalized quantile regression is then given by

$$\Gamma^{\epsilon}_{\text{RE}}(\boldsymbol{x}) = \left[\hat{t}_{\epsilon/2}(\boldsymbol{x}) - \hat{q}, \hat{t}_{1-\epsilon/2}(\boldsymbol{x}) + \hat{q} \right] .$$

4.4 Risk quantification using conformal prediction

As already mentioned, in conformal prediction the uncertainty of the conformal predictor is quantified by the size of the prediction set. The higher the cardinality of the prediction set, or the width of the prediction interval in the case of regression, the higher the uncertainty. In the following, we make use of this notion of uncertainty to quantify the risk associated with a certain goodwill request being processed in an automated fashion by the prescriptive models.

4.4.1 Quantifying risk

To quantify the risk of wrong assessments in ranking (WARA), we make use of the conformal predictor's prediction set size $|\Gamma^{\epsilon}(\boldsymbol{x})|$, which is either 1, 2 or 3 (or 0 in the case of the empty set):

$$\mathcal{R}_{\text{WARA}}(\boldsymbol{x}) = \frac{|\Gamma^{\epsilon}_{\text{RA}}(\boldsymbol{x})|}{3}$$

Note that, if the conformal predictor for ranking outputs an empty set $\Gamma^{\epsilon}_{\text{RA}}(\boldsymbol{x}) = \emptyset$ we consider this as low risk query with $\mathcal{R}_{\text{WARA}}(\boldsymbol{x}) = 0$, since the model must anyway abstain from a decision.

The risk of wrong assessments in regression (WARE) is based on the conformal predictor's interval size normalized by the overall regression interval size (in our use case from 10 to 90 %):

$$\mathcal{R}_{\text{WARE}}(\boldsymbol{x}) = \min\left(\frac{\max\Gamma^{\epsilon}_{\text{RE}}(\boldsymbol{x}) - \min\Gamma^{\epsilon}_{\text{RE}}(\boldsymbol{x})}{80}, 1\right)$$

The interval cannot be empty in that sense but it can get arbitrarily small.

4.5 Selective uncertainty-aware automated decision making

To abstain from decisions in cases where the risk is too high, we need to define *selection functions* for the ranking and regression layer, respectively, as well as corresponding risk thresholds δ_{rank} and δ_{reg}. The empty prediction set is treated as an exception and also leads to abstention:

$$g_{\delta_{\text{rank}}}(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \Gamma^{\epsilon}_{\text{RA}}(\boldsymbol{x}) \neq \emptyset \\ & \wedge \mathcal{R}_{\text{WARA}}(\boldsymbol{x}) \leq \delta_{\text{rank}} \\ 0 & \text{otherwise} \end{cases}$$

$$g_{\delta_{\text{reg}}}(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \mathcal{R}_{\text{WARE}}(\boldsymbol{x}) \leq \delta_{\text{reg}} \\ 0 & \text{otherwise.} \end{cases}$$

We can now outline the complete uncertainty-aware assessment system $a(\boldsymbol{x})$ as follows. First, the query $\boldsymbol{x}$ is processed by the ranking layer $\hat{h}_{\text{rank}}$. If the selection function $g_{\delta_{\text{rank}}}(\boldsymbol{x})$ selects the input for decision, the result of $\hat{h}_{\text{rank}}(\boldsymbol{x})$ is considered valid. In the case of a PARTIAL contribution ($\hat{h}_{\text{rank}}(\boldsymbol{x}) = 2$), the query is passed on to the regression layer and further processed by the regression model $\hat{h}_{\text{reg}}$. In any case, if the ranking $g_{\delta_{\text{rank}}}$ or regression selection functions $g_{\delta_{\text{reg}}}$ abstain from a decision, the query is forwarded to a manual assessment $m(\boldsymbol{x})$ by a human expert:

$$a(\boldsymbol{x}) = (\hat{h}_{\text{rank}}, g_{\delta_{\text{rank}}}, \hat{h}_{\text{reg}}, g_{\delta_{\text{reg}}}, m)(\boldsymbol{x}) := \begin{cases} \hat{h}_{\text{rank}}(\boldsymbol{x}) & \text{if } g_{\delta_{\text{rank}}}(\boldsymbol{x}) = 1 \\ & \wedge(\hat{h}_{\text{rank}}(\boldsymbol{x}) = 1 \vee \hat{h}_{\text{rank}}(\boldsymbol{x}) = 3) \\ \hat{h}_{\text{reg}}(\boldsymbol{x}) & \text{if } g_{\delta_{\text{rank}}}(\boldsymbol{x}) = 1 \wedge \hat{h}_{\text{rank}}(\boldsymbol{x}) = 2 \\ & \wedge g_{\delta_{\text{reg}}}(\boldsymbol{x}) = 1 \\ m(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

4.6 The risk vs. degree of automation trade-off

Given a proper uncertainty quantification, there is an obvious trade-off between risk and degree of automation in decision support systems. The more risk of possibly suboptimal or inappropriate decisions one is willing to take, the higher the degree of automation of the system can be. This trade-off can be formalized in terms of a *multi-objective optimization* (MO) problem. Essentially, in our use case we seek to maximize the degree of automation while simultaneously minimizing the overall risk of wrong assessments.

In general, a MO problem can mathematically be formulated as follows [15]:

$$\begin{aligned} \min \quad & f(x) = \{f_1(x), \ldots, f_k(x)\} \\ \text{s.t.} \quad & x \in \Omega \end{aligned}$$

Usually, the goal is to find a *Pareto-optimal* solution. A solution $x^* \in \Omega$ is called *Pareto-optimal* if there is no other solution $x \in \Omega$, $x^* \neq x$, such that $f_i(x) \leq f_i(x^*)$ and $f_j(x) < f_j(x^*)$ for at least one j [15].

When a Pareto optimal solution is found, a *decision maker* (DM) can select the best solution from the *Pareto set* or *front*. The DM is supposed to be a domain expert and must be able to select the solution representing the best trade-off for the problem at hand.

Methods for solving MO problems are categorized according to when in the optimization process the DM contributes her expertise in finding the best trade-off. In *a priori* methods, the DM is asked for her preferences in advance. Her preferences are then taken into account during the optimization process to find a Pareto-optimal solution as close as possible to the specified preferences. In *a posteriori* methods, an approximation of the whole Pareto set is determined and presented to the DM. The DM can then select the best trade-off. In *interactive* methods, the DM's expertise and preferences are integrated into the optimization process and she can iteratively provide feedback.

When looking at our use case, we have four parameters that control the risk and the degree of automation of our assessment system: The threshold risk values (δ_{rank},δ_{reg}) and the conformal predictors' significance levels (ϵ_{rank}, ϵ_{reg}):

$$\boldsymbol{u} = \begin{pmatrix} \epsilon_{\text{rank}} \\ \delta_{\text{rank}} \\ \epsilon_{\text{reg}} \\ \delta_{\text{reg}} \end{pmatrix}$$

The three objectives we seek to optimize are the risk for ranking $\mathcal{R}_{\text{WARA}}$ and regression $\mathcal{R}_{\text{WARE}}$, as well as the overall degree of automation (DoA):

$$\boldsymbol{v} = \begin{pmatrix} \bar{\mathcal{R}}_{\text{WARA}}(\boldsymbol{u}) \\ \bar{\mathcal{R}}_{\text{WARE}}(\boldsymbol{u}) \\ \text{DoA}(\boldsymbol{u}) \end{pmatrix},$$

where

$$\bar{\mathcal{R}}_{\text{WARA}}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_{\text{WARA}}(\boldsymbol{x}_i \mid \boldsymbol{u}),$$
$$\bar{\mathcal{R}}_{\text{WARE}}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{R}_{\text{WARE}}(\boldsymbol{x}_i \mid \boldsymbol{u}).$$

Moreover, the DoA is defined as follows:

$$DoA(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} [\![g_{\delta_{\text{rank}}}(\boldsymbol{x}) = 1 \wedge \hat{h}_{\text{rank}}(\boldsymbol{x}) \in \{1, 3\}]\!] + [\![g_{\delta_{\text{rank}}}(\boldsymbol{x}) = 1 \wedge \hat{h}_{\text{rank}}(\boldsymbol{x}) = 2 \wedge g_{\delta_{\text{reg}}}(\boldsymbol{x}) = 1]\!]$$

Formally, our optimization problem can be formulated according to the equation below. The risk values ($\mathcal{R}_{\text{WARA}}$, $\mathcal{R}_{\text{WARE}}$) are supposed to be minimized, whereas the degree of automation (DoA) is supposed to be maximized. Moreover, all optimization parameters $\boldsymbol{u}$ are restricted to the interval $[0, 1]$.

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & \bar{\mathcal{R}}_{\text{WARA}}(\boldsymbol{u}) \\ \min_{\boldsymbol{u}} \quad & \bar{\mathcal{R}}_{\text{WARE}}(\boldsymbol{u}) \\ \max_{\boldsymbol{u}} \quad & \text{DoA}(\boldsymbol{u}) \\ \text{s.t.} \quad & 0 \leq \epsilon_{\text{rank}}, \delta_{\text{rank}}, \epsilon_{\text{reg}}, \delta_{\text{reg}} \leq 1 \end{aligned}$$

In the end, our overall goal is to offer the business DM a Pareto set of solutions from which she can choose the best trade-off in terms of risk and degree of automation. Explicating and clearly explaining this trade-off with a set of Pareto-optimal solutions makes the ML system more transparent to business DMs. This may also help to increase trust into the ML system, as the trade-off is known and can be controlled.

Table 1 Characteristics of the goodwill data set

Goodwill data set	
Overall data set size	15,397
Number of categorical features	14
Number of numeric features	8
Number of boolean features	2
Number of NO contributions (labor)	7,426
Number of PARTIAL contributions (labor)	3,940
Number of FULL contributions (labor)	4,309
Number of NO contributions (parts)	4,865
Number of PARTIAL contributions (parts)	4,412
Number of FULL contributions (parts)	6,398

5 Evaluation

In the following, we conduct several experiments using our approach as outlined in the previous section and the goodwill data set. We begin with a short description of the data set and some implementation details. Next, we evaluate the coverage and set sizes of our conformal predictors based on different significance levels. Subsequently, we identify Pareto-optimal solutions for our objective space (risk, degree of automation, accuracy) using random search. These Pareto-optimal solutions can then be used to identify a suitable trade-off by a decision maker.

5.1 The goodwill data set

The data set we will use to evaluate our approach is a goodwill data set of a BMW NSC. The features are the data contained in a goodwill request and the labels are the contributions assessed for labor and parts by the human experts. We will not treat the problem as a multi-label classification task, but instead build separate prescriptive conformal predictors for labor and part contributions, respectively. Table 1 summarizes the characteristics of the data set.

5.2 Implementation

To implement the ranking part of the hierarchical model according to [25], we make use of XGBoost [7] with the cost matrix shown in Table 2. Essentially, this is a neutral cost matrix that does not implement a certain strategy (e.g., customer friendly vs. cost oriented). In the case of *partial* ranks, the costs equal the absolute error of the regression layer and lie in the interval $[0, 80]$.

To implement the regression layer, as well as the quantile regression models for conformal prediction, we make use of a feed-forward neural network with two dense hidden layers and 512 neurons each. The model is trained for 200 epochs

Table 2 Cost matrix for the ranking layer

		Prescribed		
		NO	PARTIAL	FULL
Actual	NO	0	100	200
	PARTIAL	100	[0,80]	100
	FULL	200	100	0

with batch size 32. For quantile regression, we use the *pinball loss* function and for the regular regression layer the *mean absolute error* (mae) loss function.

Figure 3 depicts our conformal inference architecture in detail. It consists of three layers:

1. The *point prediction layer* contains the hierarchical goodwill assessment model already outlined in [13]. It outputs point predictions for goodwill requests without any uncertainty awareness.
2. The *conformal prediction layer* enhances the *point prediction layer* with inductive conformal predictors for the ranking and regression layers.
3. The *risk assessment layer* utilizes the prediction set and interval sizes output by the *conformal prediction layer* to quantify the risk associated with a request and either forwards the request to a human assessment or takes over the point prediction result as the result of the assessment.

5.3 Evaluation of conformal prediction

First, we evaluate our conformal prediction implementation on the goodwill data set of the NSC using ten-fold cross validation for several significance levels $\epsilon = \{0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.03, 0.02, 0.01\}$. During each iteration, we use approximately 690 examples (5%) of the training examples for calibration. The following plots then display the mean and the 95% confidence interval for the 10 folds.

Figure 4 shows the prediction set and interval sizes as well as the coverage of the ranking and regression layers for parts and labor contributions. As expected, smaller significance levels ϵ lead to higher coverage and also larger prediction sets and interval sizes. The coverage of a conformal predictor's prediction set (or interval size in the case of regression) can be calculated as follows:

$$C = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \in \Gamma^{\epsilon}(\boldsymbol{x}_i)\}$$

The mean value of the coverage $\bar{C}$ calculated during the ten folds should center around $1-\epsilon$, which is the case for ranking, e.g. $\epsilon = 0.2, \bar{C} = 0.78$ or $\epsilon = 0.7, \bar{C} = 0.275$. This is a good indicator for the correct implementation of conformal prediction. For regression, the coverage plot is not as accurate as for ranking but also displays a constant coverage increase for smaller significance levels. In addition, the prediction set and interval sizes stay small for a long time and only increase steeply for very small significance levels $\epsilon \leq 0.1$, which also underlines the accuracy of the conformal predictor and the quality of the score functions. The average prediction set size for ranking is hereby calculated as follows:

$$S = \frac{1}{n}\sum_{i=1}^{n} |\Gamma^{\epsilon}(\boldsymbol{x}_i)|.$$

In the case of regression, the spread of the interval is taken as the interval size:

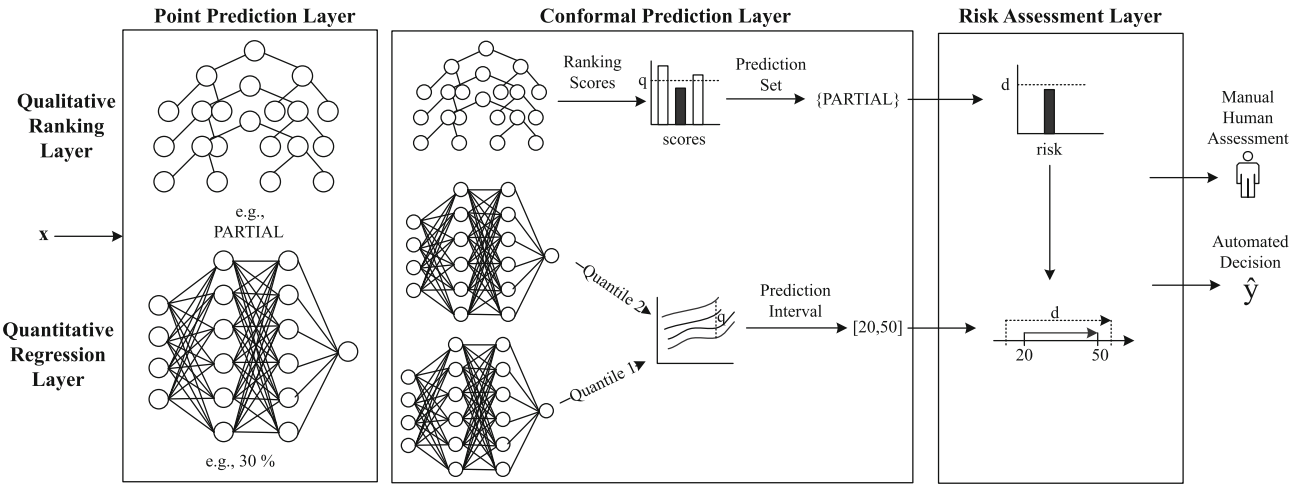
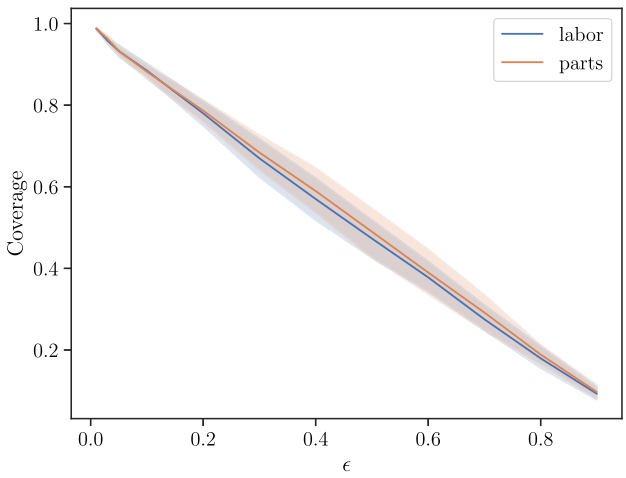
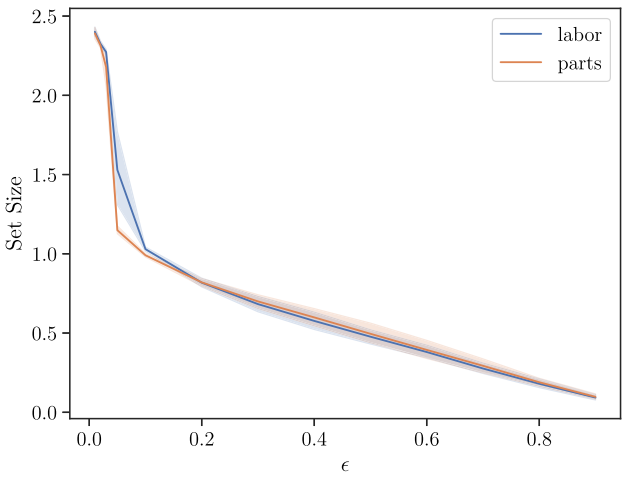


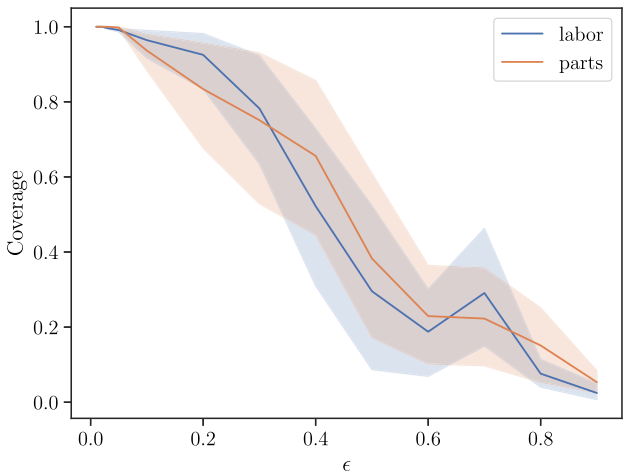
Fig. 3 Overview of the inference architecture



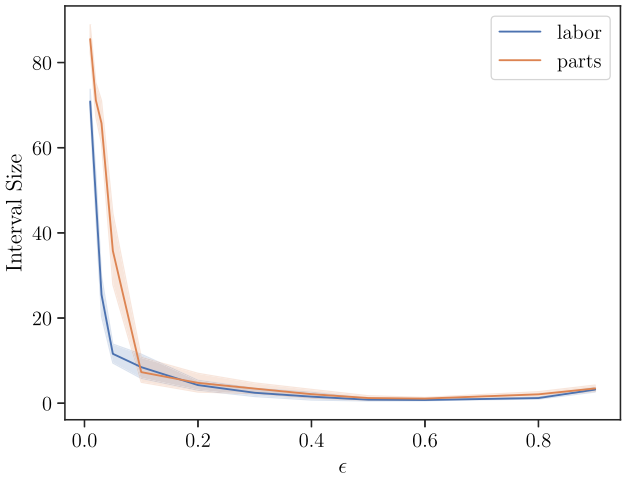
(a) Coverage of conformal predictor for ranking.



(b) Set sizes of conformal predictor for ranking.



(c) Coverage of conformal predictor for regression.



(d) Interval sizes of conformal predictor for regression.

Fig. 4 Coverage and set size plots for several significance levels ϵ

$$S = \frac{1}{n}\sum_{i=1}^{n} \max \Gamma^{\epsilon}(\boldsymbol{x}_i) - \min \Gamma^{\epsilon}(\boldsymbol{x}_i)\,.$$

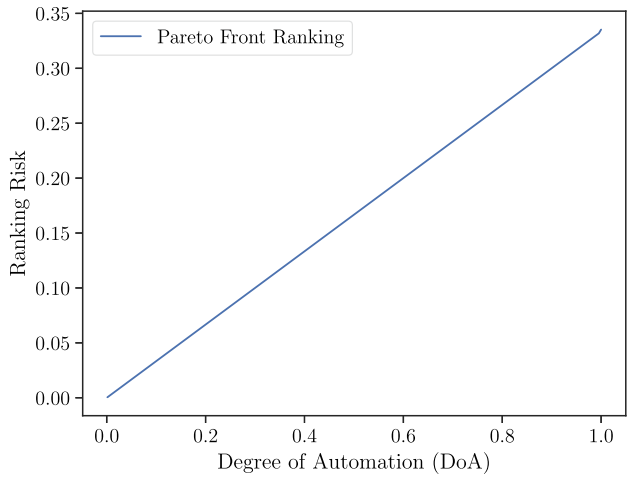
5.4 Evaluation of selective uncertainty-aware Pareto optimization

In order to identify good *a posteriori* trade-offs for our objectives, we perform a simple random search limited to 1000 iterations. Table 3 shows the *design space* used for randomly exploring the *objective space*. The values are hereby drawn from a uniform distribution.

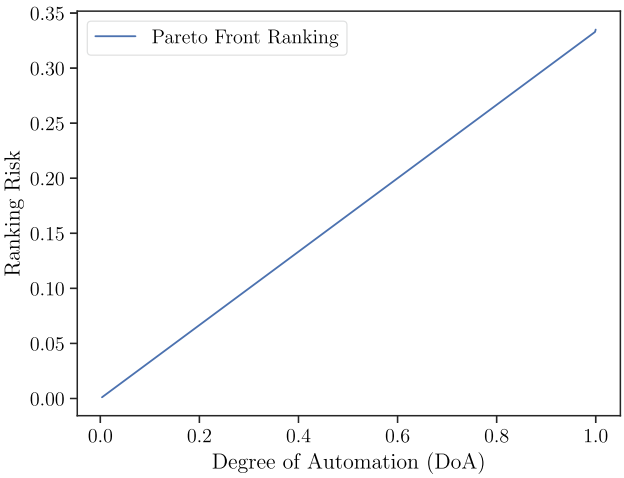
In each random search trial, we train the hierarchical model using the training data set (13,164 examples), then calibrate our conformal predictors with the calibration data set (693 examples) and evaluate our model's conformal and point predictions using the test set (1540 examples). Next, we determine the non-dominated points in our explored *objective space* forming the Pareto front of our multi objective

Table 3 Design space for randomly exploring the objective space (risk, accuracy, degree of automation)

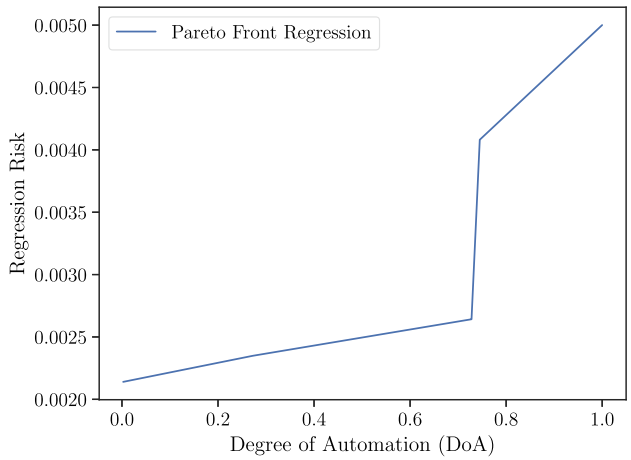
Design space - Random search	
ϵ_{rank}	$\{\epsilon_{\text{rank}} \in \mathbb{R} \,\vert\, 0 \leq \epsilon_{\text{rank}} \leq 1\}$
δ_{rank}	$\{\delta_{\text{rank}} \in \mathbb{R} \,\vert\, 0 \leq \delta_{\text{rank}} \leq 1\}$
ϵ_{reg}	$\{\epsilon_{\text{reg}} \in \mathbb{R} \,\vert\, 0 \leq \epsilon_{\text{reg}} \leq 1\}$
δ_{reg}	$\{\delta_{\text{reg}} \in \mathbb{R} \,\vert\, 0 \leq \delta_{\text{reg}} \leq 1\}$



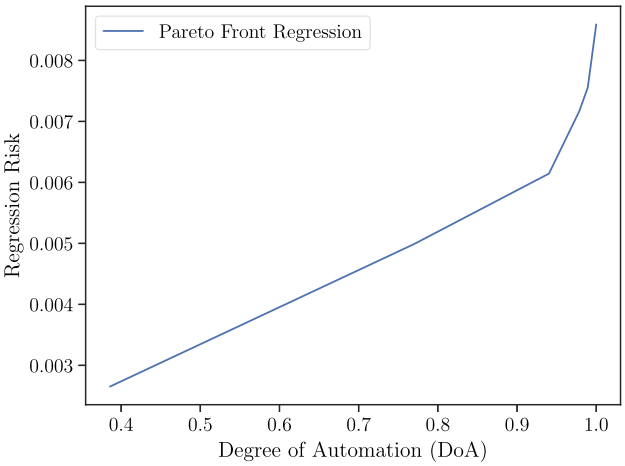
(a) Risk vs. DoA for the Labor Ranking Layer.



(b) Risk vs. DoA for the Parts Ranking Layer.



(c) Risk vs. DoA for the Labor Regression Layer.

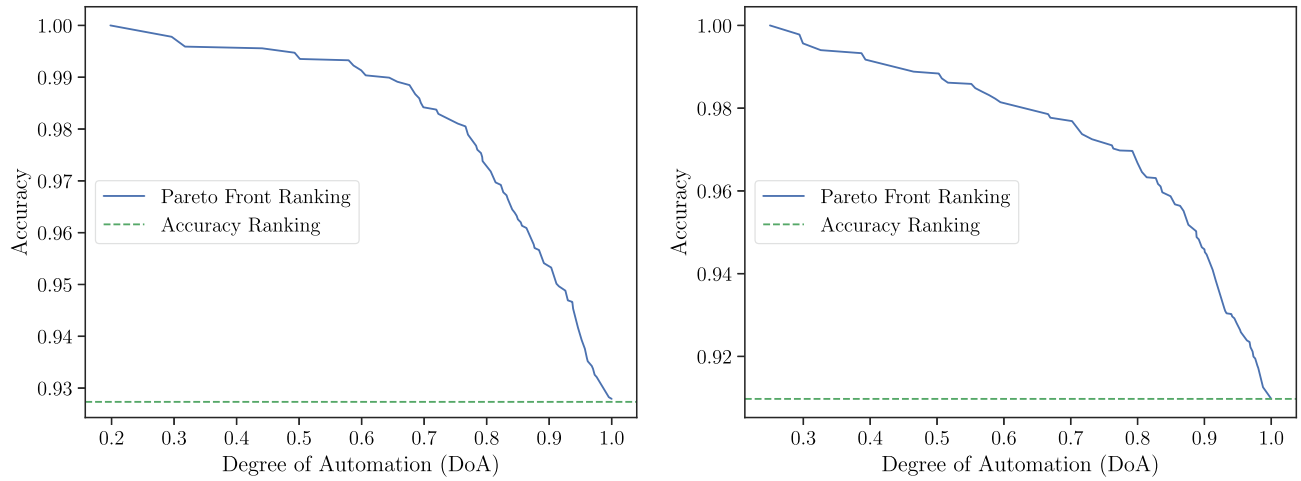


(d) Risk vs. DoA for the Parts Regression Layer.

Fig. 5 Trade-offs between risk and degree of automation (DoA)

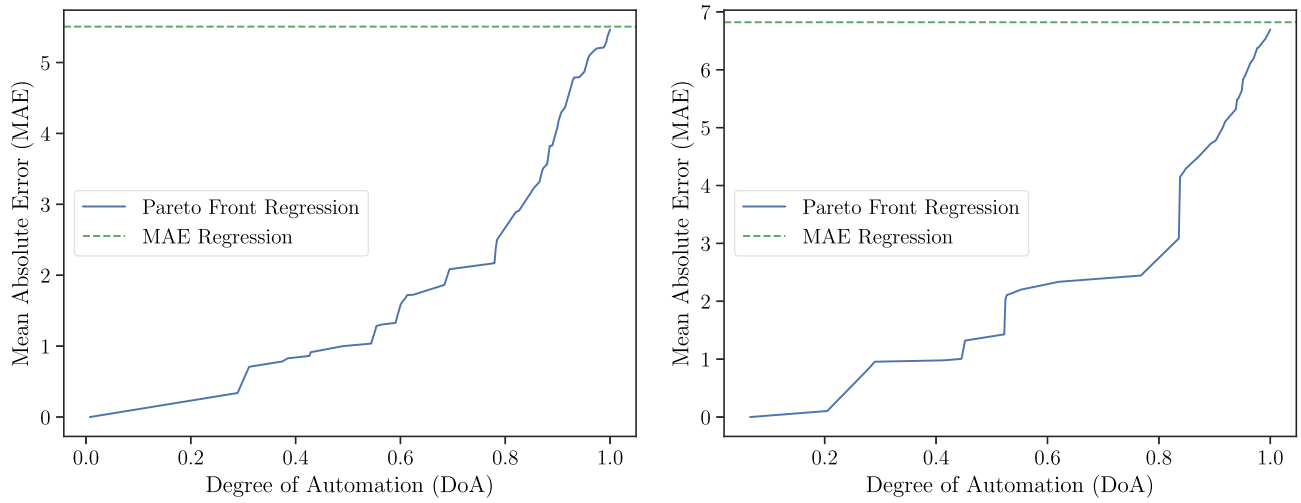
optimization problem. We hereby first look at the risk vs. degree of automation trade-off for ranking and regression in Fig. 5 also known as *risk coverage trade-off*. The degree of automation that is achievable in the ranking layer hereby linearly increases with increasing risk. Requests whose risk values exceed the given risk thresholds are hereby rejected and not considered for automatic processing. A similar behavior is visible for the regression layer, when looking at the Pareto set for the regression risk vs. degree of automation trade-off (cf. Fig. 5). However, the regression risk does not increase constantly. It first increases moderately and shoots up for higher degrees of automation. Nevertheless, higher risk goes hand in hand with higher degree of automation for both layers.

Since our calculated risk values based on conformal prediction outputs are rather abstract values, we also look at the accuracy vs. degree of automation trade-offs for the ranking layer in Fig. 6. As a baseline, we also show the overall accuracy of our ranking layer, which is 92.7% for labor and 90.97% for parts contributions respectively. The shown plots are very similar to *Accuracy-Rejection Curves* [27], but instead of plotting the amount of rejected queries in per cent we plot the amount of selected or processed queries accumulating in the degree of automation of the system. The accuracy of the ranking layer is monotone decreasing for increasing degrees of automation, which indicates that, by virtue of our conformal ranking predictor, the ranking layer is capable of quantifying its uncertainty well.



(a) Accuracy vs. DoA for Labor. **(b)** Accuracy vs. DoA for Parts.

Fig. 6 Trade-off between accuracy and degree of automation (DoA) for the ranking layer

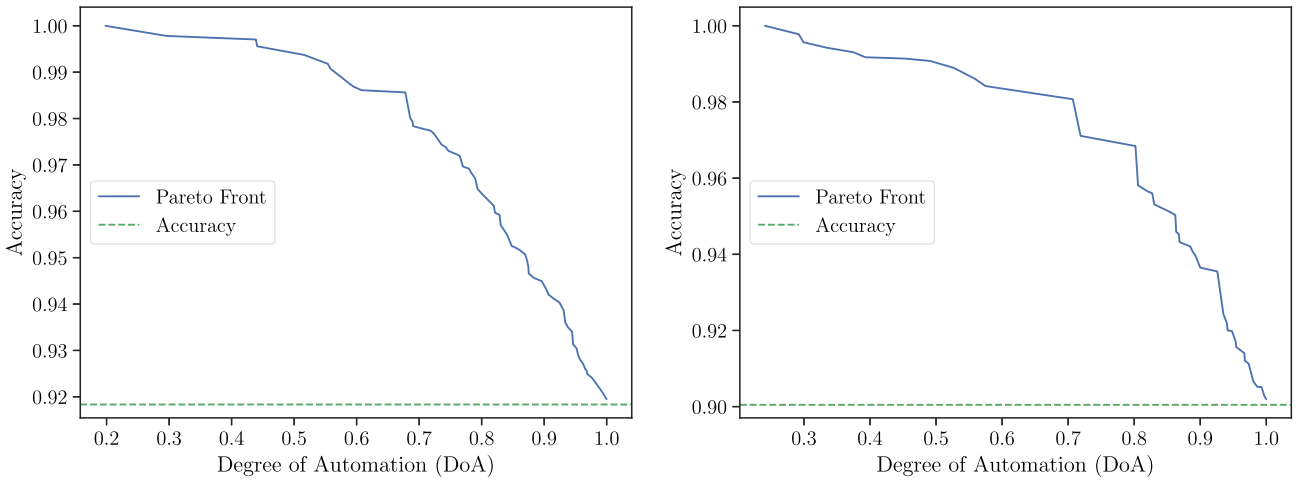


(a) MAE vs. DoA for Labor. **(b)** MAE vs. DoA for Parts.

Fig. 7 Trade-off between mean absolute error (MAE) and degree of automation (DoA) for the regression layer

When looking at the mean absolute error (mae) vs. degree of automation trade-off in the regression layer, we can also see a similar behavior (cf. Fig. 7). For increasing degrees of automation, the mean absolute error is monotone increasing, which also underpins the capability of the regression layer to quantify its uncertainty well. Abstaining randomly would in contrast lead to a flat curve. An overall MAE of 5.49 for labor and 6.67 for parts respectively in the regression layer can easily be undercut by reducing the degree of automation.

Figure 8 shows plots for the overall accuracy vs. degree of automation trade-offs of the hierarchical model as a whole, including the ranking layer as well as the regression layer. An accuracy of 100% is achievable with a degree of automation of 20%, which is however not a practically useful scenario. A degree of automation of 70% might be a good trade-off and leads to an accuracy of 98% for labor and parts, respectively, on the test data. In general, we can also see a clear monotonic decrease of the overall accuracy with increasing degree of automation which ensures the uncertainty quantification capability also of the overall hierarchical model. Looking at this trade-off, a business decision maker can select a practically reasonable solution. Whether degree of automation outweighs high accuracy requirements very much depends on the use case. As goodwill assessment is a process entailing financial risk, very high accuracy is definitely an important requirement. Since there is anyway a human assessment process in place, degree of automation is presumably a less important criterion than accuracy.



(a) Accuracy vs. DoA for Labor. **(b)** Accuracy vs. DoA for Parts.

Fig. 8 Overall trade-off between accuracy and degree of automation (DoA)

Table 4 Some selected accuracy vs. degree of automation trade-off values including the corresponding design space values for labor

ϵ_{rank}	δ_{rank}	ϵ_{reg}	δ_{reg}	Accuracy (ACC)	Degree of automation (DOA)
0.060531	0.476986	0.844075	0.168960	0.919481	1.000000
0.259388	0.317007	0.670188	0.086850	0.963563	0.801948
0.370768	0.123228	0.667563	0.489556	0.978363	0.690260
0.395235	0.000244	0.715322	0.843102	0.993711	0.516234
0.824623	0.433202	0.941667	0.162653	1.000000	0.198052

Table 5 Some selected accuracy vs. degree of automation trade-off values including the corresponding design space values for parts

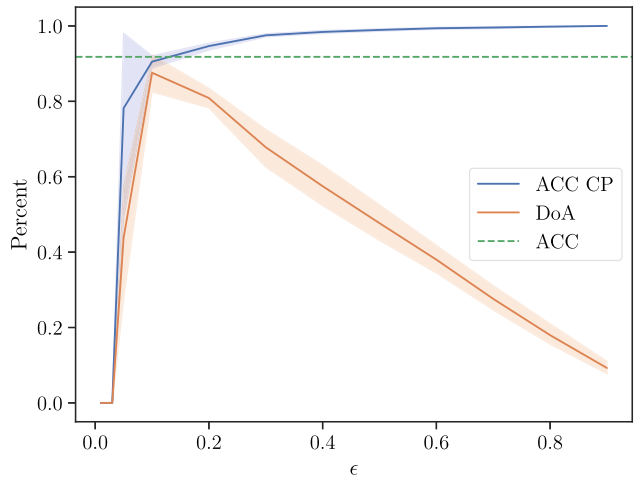
ϵ_{rank}	δ_{rank}	ϵ_{reg}	δ_{reg}	Accuracy (ACC)	Degree of automation (DOA)
0.085349	0.207636	0.728141	0.957674	0.901948	1.000000
0.214033	0.204153	0.338893	0.211275	0.958098	0.805844
0.196397	0.985753	0.783504	0.078822	0.980716	0.707143
0.313255	0.082391	0.619651	0.057391	0.990753	0.491558
0.764676	0.130926	0.948631	0.518402	1.000000	0.240909

Tables 4 and 5 show some selected accuracy vs. degree of automation trade-off values for parts and labor, respectively, including the corresponding design space values.

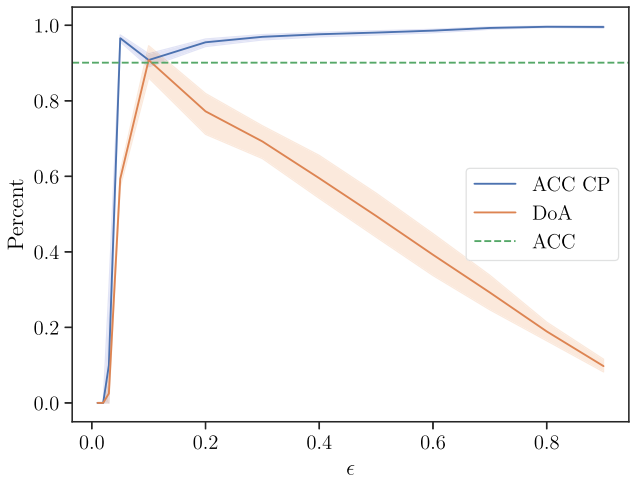
5.5 The effect of the significance level ϵ

In the following, we study the effect of the significance level ϵ on the achievable prescription accuracy and degree of automation. This can be done by fixing the risk thresholds for the ranking as well as the regression layer. A reasonable threshold for ranking might be $\delta_{\text{rank}} = \frac{1}{3}$, which essentially means that we only want to consider prediction sets for automated decision where the conformal ranking predictor is certain about the result. For regression, we might want to tolerate a risk of $\delta_{\text{reg}} = \frac{10}{80}$, which is an interval spread of 10%, otherwise we do not trust the result and want the case to be processed manually. Please note that these thresholds are exemplary thresholds and not universally applicable. They are specific to the problem of goodwill assessment and the proposed hierarchical model structure. In general, defining an optimal risk threshold is a task on its own which must also take the context of the application into account [38], as even an optimal risk-averse threshold does not reliably go in a particular direction [19]. In the case of goodwill assessment, the risk-averse decision maker [37] may also not want to miss out on reduced costs trough automation and take these into account when defining risk thresholds.

Figure 9 shows 10-fold cross validated mean plots for the conformal predictor's accuracy and the overall degree of automation depending on the significance level $\epsilon = \{0.9, 0.8, \ldots, 0.1, 0.05, 0.03, 0.02, 0.01\}$. As a baseline, we again show the overall accuracy (ACC) of the hierarchical model as a whole over all test data. One can see that

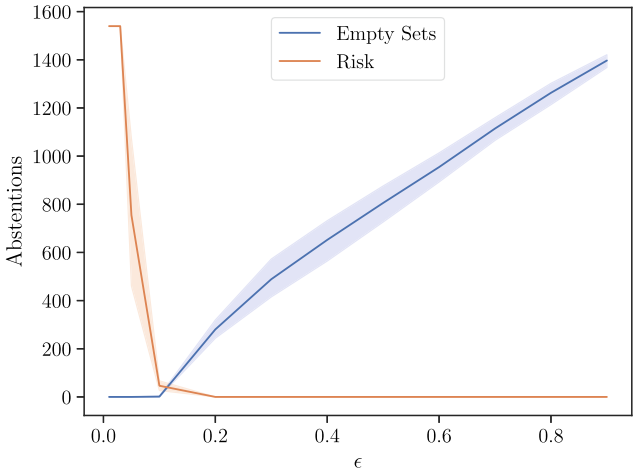


(a) Accuracy and Degree of Automation for labor.

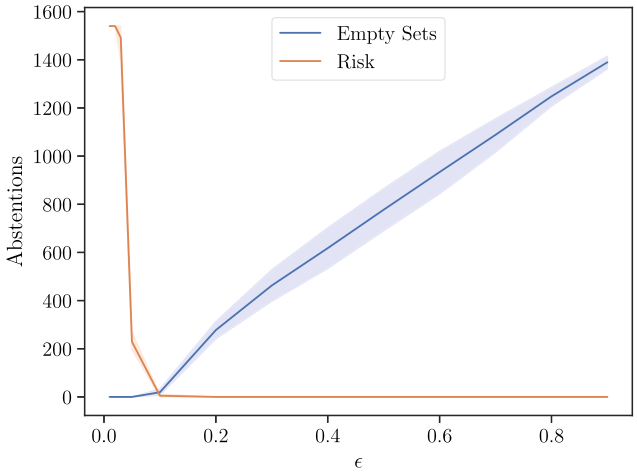


(b) Accuracy and Degree of Automation for Parts.

Fig. 9 Accuracy and degree of automation plots for $\delta_{\text{rank}} = \frac{1}{3}$ and $\delta_{\text{reg}} = \frac{10}{80}$ depending on ϵ



(a) Abstention types for labor.



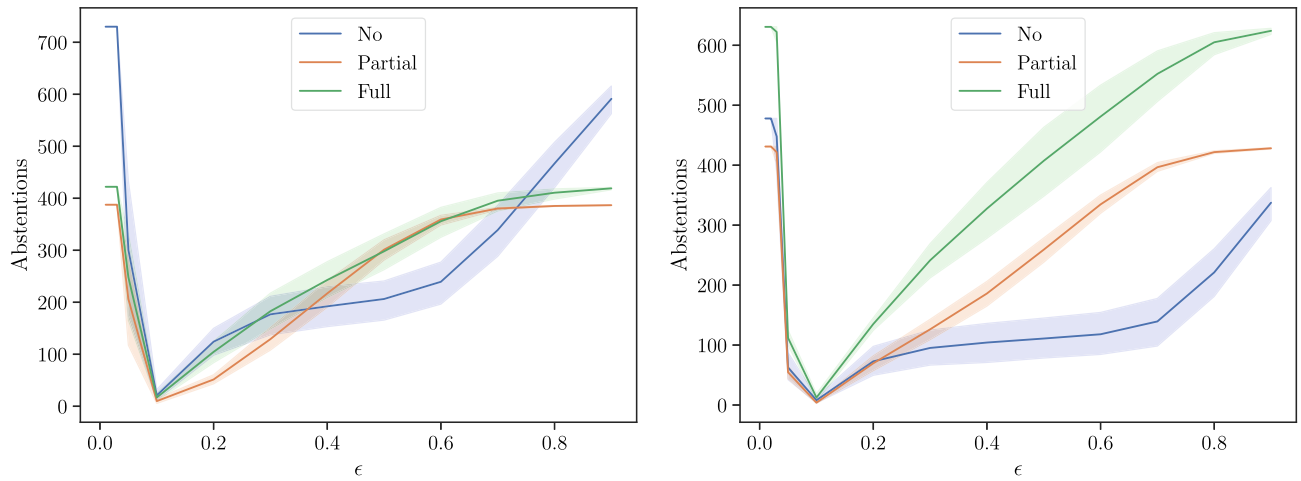
(b) Abstention types for parts.

Fig. 10 Abstention types of the conformal hierarchical predictor depending on ϵ

with decreasing ϵ the degree of automation (DoA) increases whereas the accuracy decreases (ACC CP). So if accuracy is important, ϵ values need to be rather large. If the degree of automation is important, ϵ values need to be rather low. At a certain ϵ value, accuracy and degree of automation drop of steeply, since the prediction sets and intervals become too large and exceed the predefined risk thresholds, which makes the model abstain completely from deciding requests.

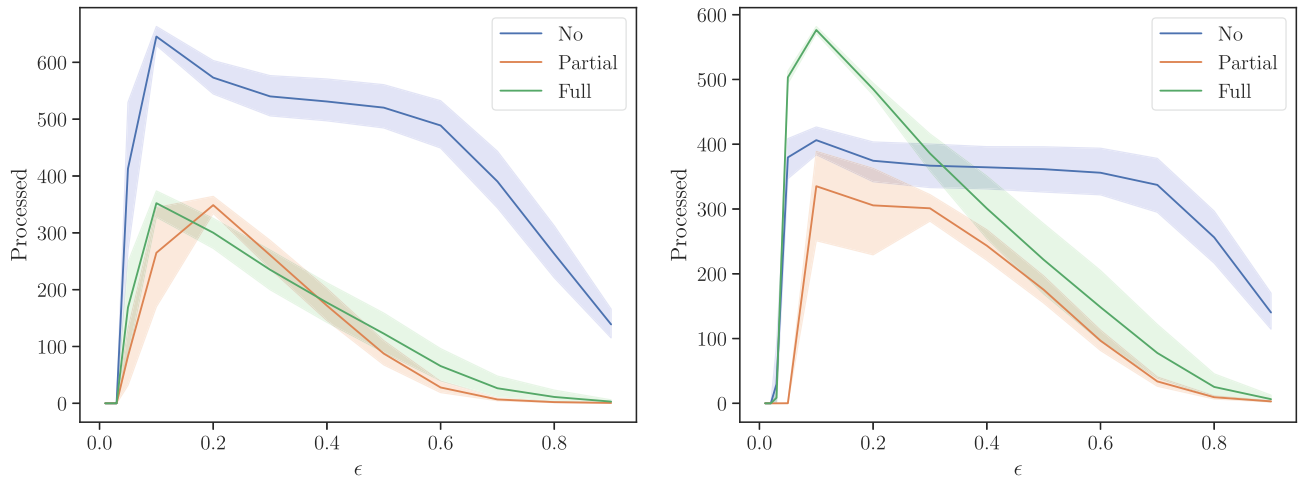
Figure 10 displays the corresponding reasons for abstentions depending on the significance level ϵ. For larger ϵ values, abstentions are exclusively caused by empty sets. In that case, few predicted cases fall below the required quantile threshold $\hat{q}$. For instance, if $\epsilon = 0.9$ only 10% ($1 - \epsilon = 1 - 0.9 = 0.1$) of the lowest scores are considered valid results and lie within the quantile $\hat{q} = 1 - \epsilon = 1 - 0.9 = 0.1$. With decreasing ϵ there are less and less empty prediction sets until the sets grow so large that abstentions are solely due to risk assessments. In the end, for $\epsilon \leq 0.03$, the conformal predictors only output non-unique prediction sets, which leads to complete abstention in our case due to our strict thresholds.

Figure 11 breaks down the abstentions by contribution type (no, partial, or full contribution). Abstentions for all types of contributions strictly decrease for decreasing ϵ values until the sets become too large, leading to complete abstention due to violation of the risk threshold. It is noticeable that for labor as well as part abstentions the *No* abstentions drop off steeper in the beginning. One may speculate that the *No* contributions have the smallest scores and are therefore overrepresented in the smaller score quantiles $\hat{q}$, e.g., with $\epsilon > 0.6$. Moreover, given this observation, at



(a) Abstention ranks for labor. **(b)** Abstention ranks for parts.

Fig. 11 Abstention ranks of the conformal hierarchical predictor depending on ϵ



(a) Processed ranks for labor. **(b)** Processed ranks for parts.

Fig. 12 Processed ranks of the conformal hierarchical predictor depending on ϵ

least a part of the *No* contribution assessments seems to be quite certain or obvious.

Figure 12 breaks down the processed contributions by their contribution type (no, partial, or full contribution). Processed hereby means that the predictor did not abstain from answering the particular request. Like for abstentions, it is visible that *No* contributions are processed preferentially. The *No* contribution scores seem to be overrepresented in the lower quantiles. Nevertheless, contributions strictly increase for all contribution types with increasing ϵ until complete risk abstention sets in.

Since the abstentions and processed contributions are not balanced, one could argue to use *class-balanced conformal prediction* [1] instead, where scores and quantiles are determined per class. However, given the use case at hand, there is not necessarily a need for class-balanced coverage. If manual work reduction is the main goal of introducing ML into the process, this coverage imbalance might have no negative impact at all, since there is no difference in effort known between the assessments of the different contributions. It could even be considered beneficial that *No* contributions are the most certain ones to be assessed automatically, as they also entail the least financial risk.

6 Conclusion and future work

We developed and evaluated an uncertainty-aware approach for automated decision making, in which conformal prediction is used to quantify the risk associated with ML prescriptions. As a use case, we looked at automated decision making for goodwill assessments in the automotive domain using a goodwill data set of a car manufacturer. Instead

of providing mathematical guarantees for limited risk, we emphasize the trade-off between risk and degree of automation, and how an *a posteriori* Pareto-optimal solution can be explored by a business decision maker to select the best trade-off for the particular business use case at hand.

To underpin the capability of conformal predictors to quantify uncertainty in a proper way, we present risk-coverage plots and accuracy-rejection curves. We also analyzed CP's significance level parameter ϵ and how it affects the number of empty prediction sets as well as the achievable accuracy and degree of automation of the system. Concretely, by abstaining to answer the 30% most risky or uncertain queries, our hierarchical predictor is capable of increasing its overall accuracy from 92 to 98% for labor and from 90 to 98% for parts contributions, respectively.

Achieving even higher accuracies is presumably not very reasonable, as this comes at a significant loss in degree of automation. Additionally, human decisions cannot be considered a consistent gold standard and might be biased in one or another direction. A certain amount of *aleatoric* uncertainty is supposedly irreducible in a human decision process and will remain. Nevertheless, the amount of wrongly prescribed contributions can be significantly reduced with our selective uncertainty-aware approach, which makes the introduction of ML in high-stake environments more feasible.

Proceeding from this well working uncertainty-aware approach to automated decision making, we plan to address three major challenges in the future:

1. *Explainability*: Making machine learning based goodwill prescriptions more accessible and transparent to IT and business decision makers is in our eyes of utmost importance to foster trust into the system, but also to fulfill internal revision audit requirements. We consider decision explanations equally important for both scenarios in which the machine learning models are supposed to be used (Automated Decision Making (ADM) or Decision Support System (DSS)). Therefore, we plan to investigate and satisfy the different explanation needs of our stakeholders using Explainable Artificial Intelligence (XAI) methods [5, 16, 26].
2. *Human-AI interaction*: How human experts are influenced by AI assisting their work or taking over some of their workload is another interesting and important aspect that needs to be followed up [4]. Overconfidence into the decision model by human experts and decision makers, also known as *automation bias* [24], as well as undue reluctance, also known as *algorithm aversion* [10], are issues to be evaluated and calibrated properly. Whether XAI can help in this trust calibration process, by making the reasoning process of machine learning models more transparent, is still an active area of research [21, 28, 35]. Moreover, there is also a recent line of research particularly focusing on the effect of providing set-valued predictions to human-AI teams instead of single predictions [2, 6].
3. *Weak supervision*: As already mentioned, human goodwill decisions cannot necessarily be taken as a gold standard. The data may contain concept drift and shift due to strategy changes in the assessment process over time or other human induced biases leading to *noisy labels*. Hence, past decisions should be considered and modeled as *weak* information about the target rather than an incontestable ground truth, suggesting the use of methods for weakly supervised learning [14, 39].

Author Contributions Conceptualization: Stefan Haas, Eyke Hüllermeier; Methodology: Stefan Haas, Eyke Hüllermeier; Software: Stefan Haas; Validation: Stefan Haas, Eyke Hüllermeier; Formal analysis: Stefan Haas, Eyke Hüllermeier; Investigation: Stefan Haas; Writing—Original Draft: Stefan Haas; Writing—Review and Editing: Stefan Haas, Eyke Hüllermeier; Visualization: Stefan Haas; Supervision: Eyke Hüllermeier; Project administration: Stefan Haas.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Due to the nature of the research, due to commercial supporting data is not available.

Declarations

Conflict of interest Stefan Haas reports an employment relationship with BMW AG. All other authors have no Conflict of interest.

References

1. Angelopoulos, A.N., Bates, S.: Conformal prediction: a gentle introduction. Found. Trends Mach. Learn. **16**(4), 494–591 (2023)
2. Babbar, V., Bhatt, U.: Weller A On the utility of prediction sets in human-ai teams. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, 23–29 July 2022, pp. 2457–2463. ijcai.org (2022)
3. Balasubramanian, V., Ho, S.S., Vovk, V.: Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications, 1st edn. Morgan Kaufmann Publishers Inc., San Francisco (2014)
4. Bondi, E., Koster, R., Sheahan, H., et al.: Role of human-ai interaction in selective prediction. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022 Virtual Event, February 22–March 1, 2022, pp. 5286–5294. AAAI Press (2022)

5. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. (JAIR) **70**, 245–317 (2021)
6. Campagner, A., Cabitza, F., Berjano, P., et al.: Three-way decision and conformal prediction: isomorphisms, differences and theoretical properties of cautious learning approaches. Inf. Sci. **579**, 347–367 (2021)
7. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August 13–17, 2016, pp. 785–794. ACM (2016)
8. Cortés-Ciriano, I., Bender, A.: Concepts and applications of conformal prediction in computational drug discovery (2019) CoRR abs/1908.03569. https://arxiv.org/abs/1908.03569
9. Dari, S., Hüllermeier, E.: Reliable driver gaze classification based on conformal prediction. In: Proceedings 30th Workshop Computational Intelligence (2020)
10. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. **144**(1), 114 (2015)
11. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. J. Mach. Learn. Res. (JMLR) **11**, 1605–1641 (2010)
12. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, June 19-24, 2016, JMLR Workshop and Conference Proceedings, vol. 48, pp. 1050–1059. JMLR.org (2016)
13. Haas, S., Hüllermeier, E.: A prescriptive machine learning approach for assessing goodwill in the automotive domain. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, September 19–23, 2022, Proceedings, Part VI. Lecture Notes in Computer Science, 13718, pp. 170–184. Springer (2022)
14. Haas, S., Hüllermeier, E.: Rectifying bias in ordinal observational data using unimodal label smoothing. In: Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track - European Conference, ECML PKDD 2023, Turin, September 18–22, 2023, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 14174, pp. 3–18. Springer (2023)
15. Hakanen, J., Allmendinger, R.: Multiobjective optimization and decision making in engineering sciences. Optim. Eng. **22**, 1031–1037 (2021)
16. Hong, S.R., Hullman, J., Bertini, E.: Human factors in model interpretability: Industry practices, challenges, and needs. Proceedings ACM Human Computer Interaction 4(CSCW):068:1–068:26 (2020)
17. Hüllermeier, E.: Prescriptive machine learning for automated decision making: Challenges and opportunities (2021) CoRR abs/2112.08268. https://arxiv.org/abs/2112.08268
18. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021)
19. Hupman, A.C.: Cutoff threshold decisions for classification algorithms with risk aversion. Decis. Anal. **19**(1), 63–78 (2022)
20. Javanmardi, A., Hüllermeier, E.: Conformal prediction intervals for remaining useful lifetime estimation(2022) CoRR abs/2212.14612. https://arxiv.org/abs/2212.14612
21. Kloker, A., Fleiß, J., Koeth, C., et al.: Caution or trust in ai? how to design xai in sensitive use cases? In: AMCIS 2022 Proceedings (2022)
22. Lahoti, P., Gummadi, P.K., Weikum, G.: Responsible model deployment via model-agnostic uncertainty learning. Mach. Learn. **112**(3), 939–970 (2023)
23. Lambrou, A., Papadopoulos, H., Kyriacou, E.C., et al.: Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction. In: Artificial Intelligence Applications and Innovations - 6th IFIP WG 12.5 International Conference, AIAI 2010, Larnaca, October 6–7, 2010. Proceedings, IFIP Advances in Information and Communication Technology, vol. 339, pp. 146–153. Springer (2010)
24. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors: J. Hum. Factors Ergon. Soc. **46**(1), 50–80 (2004)
25. Li, L., Lin, H.: Ordinal regression by extended binary classification. In: Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, December 4–7, 2006, pp. 865–872. MIT Press (2006)
26. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. **11**(3–4), 24:1-24:45 (2021)
27. Nadeem, M.S.A., Zucker, J., Hanczar, B.: Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In: Proceedings of the third International Workshop on Machine Learning in Systems Biology, MLSB 2009, Ljubljana, September 5-6, 2009, JMLR Proceedings, vol.8, pp. 65–81. JMLR.org (2010)
28. Panigutti, C., Beretta, A., Giannotti, F., et al.: Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In: CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, 29 April 2022–5 May 2022, pp. 568:1–568:9. ACM, (2022)
29. Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. In: Tools in Artificial Intelligence. IntechOpen, Rijeka, chap 18 (2008)
30. Papadopoulos, H., Vovk, V., Gammerman, A.: Conformal prediction with neural networks. In: 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), October 29-31, 2007, Patras, vol. 2, pp. 388–395. IEEE Computer Society (2007)
31. Romano, Y., Patterson, E., Candès, E.J.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, pp. 3538–3548 (2019)
32. Shafer, G., Vovk, V.: A tutorial on conformal prediction. J. Mach. Learn. Res. (JMLR) **9**, 371–421 (2008)
33. Shaker, M.H., Hüllermeier, E.: Aleatoric and epistemic uncertainty with random forests. In: Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, April 27-29, 2020, Proceedings, Lecture Notes in Computer Science, vol. 12080, pp. 444–456. Springer (2020)
34. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: Learning from logged bandit feedback. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, 6–11 July 2015, JMLR Workshop and Conference Proceedings, vol 37, pp. 814–823. JMLR.org (2015)
35. Vered, M., Livni, T., Howe, P.D.L., et al.: The effects of explanations on automation bias. Artif. Intell. **322**(103), 952 (2023)
36. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg (2005)
37. Werner, J.: Risk Aversion, pp. 1–6. Palgrave Macmillan UK, London (2016)
38. Wynants, L., Van Smeden, M., McLernon, D.J., et al.: Three myths about risk thresholds for prediction models. BMC Med. **17**(1), 1–7 (2019)
39. Zhou, Z.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**, 44–53 (2018)