International Journal of Approximate Reasoning 186 (2025) 109479

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Uncertainty quantification in ordinal classification: A comparison of measures

Stefan Haas [a,b,*], Eyke Hüllermeier [a,c]

[a] Institute of Informatics, LMU Munich, Germany
[b] BMW Group, Germany
[c] Munich Center for Machine Learning, Germany

ARTICLE INFO

Keywords:
Ordinal classification
Ordinal regression
Uncertainty quantification
Probabilistic classification
Consensus
Binary decomposition

ABSTRACT

Uncertainty quantification has received increasing attention in machine learning in the recent past, but the focus has mostly been on standard (nominal) classification and regression so far. In this paper, we address the question of how to quantify uncertainty in ordinal classification, where class labels have a natural (linear) order. We reckon that commonly used uncertainty measures such as Shannon entropy, confidence, or margin are not appropriate for the ordinal case. In our search for better measures, we draw inspiration from the social sciences literature, which offers various measures to assess so-called consensus or agreement in ordinal data. We argue that these measures, or, more specifically, the dual measures of dispersion or polarization, do have properties that qualify them as measures of uncertainty. Furthermore, inspired by binary decomposition techniques for multi-class classification in machine learning, we propose a new method that allows for turning any uncertainty measure into an ordinal uncertainty measure in a generic way. We evaluate all measures in an empirical study on twenty-three ordinal benchmark datasets, as well as in a real-world case study on automotive goodwill claim assessment. Our studies confirm that dispersion measures and our binary decomposition method surpass conventional (nominal) uncertainty measures.

1. Introduction

Supervised machine learning models are increasingly deployed for high-stakes automated decision making (ADM) in fields such as medicine or finance, which comes with the demand for reliable quantification of *predictive uncertainty* to prevent financial or reputational loss, or even loss of live. Information about the uncertainty related to the outcome $y \in \mathcal{Y}$ in a context specified by a query instance $\boldsymbol{x}_q$ could, for instance, be used to perform selective classification, also called classification with abstention or reject option [1,2], where highly uncertain queries are delegated to human experts. This in turn reduces the risk of wrong predictions and increases the overall accuracy of the predictor [3].

So far, the primary focus of predictive uncertainty quantification in machine learning has been on standard (probabilistic) classification, where a predictor outputs a probability distribution (vector) $\boldsymbol{p} = (p_1, \ldots, p_K)$ on the set of class labels $\mathcal{Y} = \{y_1, \ldots, y_K\}$, where $p_k = p(y_k)$ is the probability of y_k. The arguably most popular uncertainty measure in this case is Shannon entropy [4]:

* Corresponding author.
E-mail addresses: stefan.sh.haas@bmwgroup.com, stefan.haas@campus.lmu.de (S. Haas), eyke@lmu.de (E. Hüllermeier).

https://doi.org/10.1016/j.ijar.2025.109479
Received 6 August 2024; Received in revised form 29 April 2025; Accepted 22 May 2025

Available online 28 May 2025

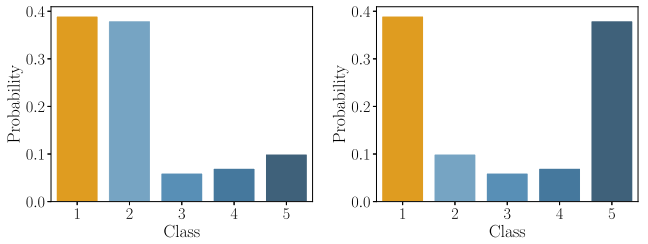


Fig. 1. Two very different distributions sharing the same Shannon entropy $H = 1.32$. In contrast, variance detects the higher dispersion on the right ($V = 3.25$) compared to the left ($V = 1.62$).

$$H(\boldsymbol{p}) := \mathbb{E}\big[-\log p(y)\big] = -\sum_{k=1}^{K} p(y_k) \log p(y_k)\,.$$

Typically, the class labels $y \in \mathcal{Y}$ are nominal categories, for example, different types of objects in image classification. However, there are real-world applications where $\mathcal{Y}$ corresponds to an *ordinal* scale, i.e., a natural (linear) order relation $y_1 \prec y_2 \prec \cdots \prec y_K$ can be defined on the class labels. Think of credit scoring with $\mathcal{Y} = \{\text{poor, fair, good, very good, excellent}\}$ or any other rating application, such as disease severity in medicine or employee performance evaluation in human resources. Since entropy is invariant against redistribution of probability mass, one may question the reasonableness of this measure in ordinal classification, where the dispersion of probability mass is an indicator for uncertainty. For an illustration, consider Fig. 1, where two very different predictive probability distributions are depicted that share the same entropy. Intuitively, the case on the right, with high probability for the two extreme outcomes, appears to be the more uncertain one. In credit scoring, for instance, it may suggest that the creditworthiness is either *poor* or *excellent*, but presumably nothing in-between. In this case, a wrong decision is likely to have more dramatic implications than mixing up, say, a *poor* and *fair* rating, like in the case on the left.

Since ordinal classification somewhat lies in-between classification and regression, one may also think of using uncertainty measures for regression, notably the variance, which is defined for continuous as well as discrete random variables [5,6]:

$$V(\boldsymbol{p}) := \sum_{k=1}^{K} p(y_k)(k-\mu)^2\,, \text{ with } \mu = \sum_{k=1}^{K} p(y_k) \cdot k\,. \tag{1}$$

Variance measures how far a set of numbers is spread out from their average value. Unlike entropy, it is not invariant against redistribution of probability mass (cf. Fig. 1). Note, however, that it assumes a *numerical* encoding of class labels. The common practice is to encode ordinal labels $y_1, \ldots, y_K$ as integers $1, \ldots, K$ [7], as we also did in (1), turning the ordinal scale $\mathcal{Y}$ into a cardinal (interval) scale with equal distances between the class labels. However, this is a critical assumption that is highly disputable and hard to justify theoretically. Practically, it may appear plausible in many cases, especially for Likert-type scales used in questionnaires and surveys.

For Likert scales, other measures have also been proposed in the social sciences literature: So-called *consensus* measures for ordinal data aim to determine the degree of consensus or agreement in survey data [8]. These measures are designed in a way to reach their respective maximum when all probability mass is concentrated on a single category, and their respective minimum for a distinct bimodal distribution, where the probability mass is equally allocated to the extreme ends of the ordinal scale. We believe that the corresponding complementary measures of *dispersion* or polarization are promising candidates for uncertainty quantification in ordinal classification. Similar to variance, they capture the degree of dispersion of a probability distribution or sample, while at the same time respecting the ordinal nature of the underlying scale. We will elucidate on this class of measures and their properties in Section 4.

In Section 5, we present a new class of measures, which are inspired by binary decomposition techniques for tackling polychotomous classification problems in machine learning [9]. Our approach allows for "lifting" any uncertainty measure applicable to a Bernoulli distribution (i.e., the case of binary classification) to a distribution on an ordinal scale. This includes established (nominal) uncertainty measures such as entropy and margin.

In general, our goal is to compare different measures for probabilistic ordinal classification according to their ability to capture uncertainty in a proper way (see Fig. 2 for a graphical overview of our approach). To this end, each candidate measure is used to quantify the uncertainty of predictions $p(y \mid \boldsymbol{x})$ over a set of (test) instances $\boldsymbol{x}$, and the suitability of the measure is then judged based on the performance achieved with the uncertainties in a downstream task, e.g. selective classification. For example, the uncertainties could be used to decide on a subset of the presumably most uncertain cases, on which the learner abstains, hoping to maximize the accuracy on the remaining (presumably less difficult) cases. The probabilities $p(y \mid \boldsymbol{x})$ themselves are obtained in a first step by training probabilistic predictive models, e.g., using proper scoring rules such as cross-entropy as loss functions.[1]

Our contributions can be summarized as follows:

[1] Proper scoring rules [10] are loss functions that are minimized (in expectation) by the true probabilities; broadly speaking, they incentivize the learner to predict probabilities in an unbiased way.

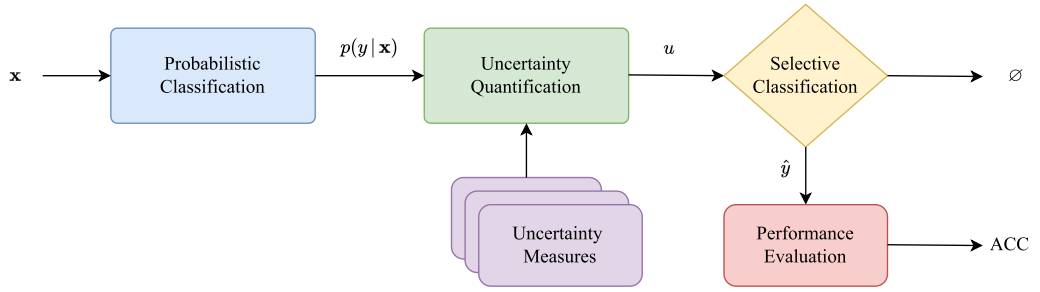


Fig. 2. Different uncertainty measures are evaluated for their ability to quantify uncertainty of predictions $p(y \mid \boldsymbol{x})$ over a set of (test) instances $\boldsymbol{x}$. The performance of these measures is assessed in a downstream selective classification task, where the learner abstains from uncertain cases ($\varnothing$) to maximize accuracy (ACC) on the remaining, less uncertain instances ($\hat{y}$).

- **Discussion of appropriate uncertainty measures for probabilistic ordinal classification:** After having introduced uncertainty representation through probability distributions over classes in Section 2, we revisit some uncertainty measures commonly used in machine learning in Section 3. In Section 4, we elaborate on properties that a good uncertainty measure for probabilistic ordinal classification should exhibit, and explain why common nominal measures such as confidence and entropy are not good candidates.
- **Proposal of using ordinal consensus measures for uncertainty quantification:** Also in Section 4, we introduce and advocate the usage of so-called ordinal consensus measures for quantifying uncertainty in ordinal classification by making use of their complementary dispersion measures. As previously stated, we consider these measures to be an ideal match for uncertainty quantification in ordinal classification.
- **Ordinal binary decomposition method for uncertainty quantification:** In Section 5, we show how any uncertainty measure, e.g., entropy or margin, can be turned into an ordinal uncertainty measure through decomposing the multi-class output into an ordered sequence of binary uncertainty quantification problems and aggregating the corresponding uncertainty degrees into an overall uncertainty score.
- **Empirical evaluation of uncertainty measures on ordinal benchmark datasets:** We validate our hypothesis that dispersion measures as well as our ordinal binary decomposition method are better candidates for quantifying uncertainty in ordinal classification than common nominal uncertainty measures through an extensive empirical evaluation on twenty-three ordinal benchmark datasets. Concretely, we calculate prediction rejection ratios (PRRs) and visualize rejection curves for the most common ordinal classification metrics accuracy (and its complementary misclassification rate), mean absolute error, and mean squared error.
- **Empirical evaluation of uncertainty measures on a real-world ADM use case:** Additionally, we conduct a case study on seven polarized automotive goodwill assessment datasets to further support our hypothesis through a real-world ADM use case.

2. Learning probabilistic predictors

We consider the setting of probabilistic supervised machine learning, in which a learner is given access to a set of training data

$$\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y},$$

with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^m$ a feature vector from an instance space $\mathcal{X}$, and $y_i \in \mathcal{Y}$ the corresponding class label or outcome from a set of outcomes $\mathcal{Y}$ that can be associated with an instance. In particular, we focus on the ordinal classification scenario, where $\mathcal{Y} = \{y_1, \ldots, y_K\}$ consist of a finite set of class labels equipped with a natural (linear) order relation:

$$y_1 \prec y_2 \prec \cdots \prec y_K.$$

Suppose a model or hypothesis space $\mathcal{H}$ to be given, where a hypothesis $h \in \mathcal{H}$ is a predictive model in the form of a mapping $\mathcal{X} \to \mathbb{P}(\mathcal{Y})$ from instances to probability distributions on outcomes. Assuming that training data as well as future (test) data is independently distributed according to an underlying (unknown) joint probability P on $\mathcal{X} \times \mathcal{Y}$, the goal in probabilistic supervised learning is to induce a hypothesis $h^* \in \mathcal{H}$ with low risk (expected loss)

$$R(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim P}\, l(h(\boldsymbol{x}), y) = \int_{\mathcal{X}\times\mathcal{Y}} l(h(\boldsymbol{x}), y)\, d\, P(\boldsymbol{x}, y),$$

where $l : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_+$ is a loss (error) function.

Training probabilistic predictors is typically accomplished by minimizing the (perhaps regularized) empirical risk

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

as an estimate of the true generalization performance, using loss functions such as proper scoring rules [10]. These have the nice theoretical property of incentivizing the learner to predict the correct conditional probabilities. Common examples of such loss functions include the log-loss and the Brier score. The empirical risk minimizer

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

serves as an approximation of the true risk minimizing hypothesis h^*. Given a query instance $\boldsymbol{x}_q \in \mathcal{X}$ as input, it produces a probabilistic prediction

$$\boldsymbol{p} = \hat{h}(\boldsymbol{x}_q) = (p(y_1), \ldots, p(y_K)) = (p_1, \ldots, p_K) \in \mathbb{P}(\mathcal{Y}) \tag{2}$$

as output, where p_k is the predicted probability for the k^{th} class y_k.

3. Uncertainty quantification for probabilistic predictors

Given a prediction (2), one might be interested in quantifying its uncertainty. In the literature, various measures have been proposed and are commonly used for that purpose. To simplify notation, we subsequently omit information about the query instance $\boldsymbol{x}_q$, which is supposed to be fixed. Following (2), we denote by $\boldsymbol{p}$ the probability distribution (vector) predicted for $\boldsymbol{x}_q$, and by $p(y_k)$ or simply p_k the probability assigned to class label y_k.

A very simple measure of predictive uncertainty, called confidence (CONF), is the gap between full certainty (a probability of 1) and the highest predicted probability [11]:

$$u_{\text{CONF}}(\boldsymbol{p}) = 1 - \max_{y_k \in \mathcal{Y}} p(y_k) = 1 - p_{(1)}\,,$$

where $(\cdot)$ is a permutation of $\{1, \ldots, K\}$ such that $p_{(1)} \geq p_{(2)} \geq \ldots \geq p_{(K)}$. Note that this measure implicitly assumes that, if the learner has to make a deterministic decision and commit to a single class label, it will indeed pick the one with highest probability. While this appears plausible, it might be rational to deviate from this decision in the case of cost-sensitive classification, where different mistakes may cause different costs.

Confidence only looks at the highest probability $p_{(1)}$ but largely ignores the remaining information provided by $\boldsymbol{p}$. Another simple approach, which at least incorporates the second largest probability, is to measure the margin (MARG) between the largest and second largest probability [11]:

$$u_{\text{MARG}}(\boldsymbol{p}) = 1 - \left(p_{(1)} - p_{(2)}\right)\,.$$

A larger difference between the two highest probabilities signifies lower uncertainty, whereas a smaller difference indicates higher uncertainty.

More information about the entire shape of $\boldsymbol{p}$ is captured by the (Shannon) entropy (ENT), a classical measure of uncertainty already discussed in the introduction. Broadly speaking, it quantifies the non-uniformity or "peakedness" [12] of a probability distribution:

$$u_{\text{ENT}}(\boldsymbol{p}) = -\sum_{k=1}^{K} p(y_k) \log\, p(y_k)\,,$$

with $0 \log 0 = 0$ by definition. Entropy is maximized by the uniform distribution $p_k \equiv 1/K$ and minimized by a Dirac delta-distribution that concentrates the entire probability mass on a single class — in this case, entropy is zero and indicates full certainty. Entropy is the de-facto standard for nominal classification in machine learning, where the uniform probability distribution is commonly associated with the least level of informedness or, equivalently, highest uncertainty.

As already outlined in the introduction, variance (VAR) is not maximized by a uniform distribution but measures the dispersion of a distribution in relation to its mean value μ:

$$u_{\text{VAR}}(\boldsymbol{p}) = \sum_{k=1}^{K} p(y_k) \cdot (y_k - \mu)^2 \quad \text{with} \quad \mu = \sum_{k=1}^{K} p(y_k) \cdot y_k \tag{3}$$

It is applicable to numeric data and a popular choice for quantifying uncertainty in regression [5,6]. Nevertheless, as already discussed, it is also applicable in ordinal classification, using an integer encoding of the labels from 1 to K.

4. Measuring consensus, polarization and agreement in ordinal data

The measures outlined in the previous section are well-established uncertainty measures in the field of machine learning. Other interesting measures have been proposed in the social sciences, albeit for a different purpose, namely, to assess agreement, consensus, concentration, dispersion, and polarization in ordinal data or ordered rating scales [8]. These measures are important tools for quantifying concentration or dispersion in Likert-scale surveys, ranging, for example, from "very strongly agree" to "very strongly disagree". First, we will examine some key properties of these ordinal measures, highlighting how they differ from the previously introduced

nominal measures, before presenting several examples of ordinal measures and how they can be used to measure uncertainty in ordinal classification.

4.1. Properties of ordinal measures

Despite their popularity in the social sciences, these ordinal measures have received limited attention in the machine learning community so far [13], although they posses several advantages over entropy and variance. For instance, in contrast to the latter, they vary between the meaningful bounds of 0 (maximum dispersion) and 1 (maximum concentration), which makes them easier to interpret [8]. Furthermore, they are designed to be less susceptible to outliers than standard deviation or variance, which are not only influenced by the dispersion of the distribution but also by its skewness [8,14]. This is particularly problematic when assessing dispersion for a distribution where the mean is located near one end of the scale. Because of their large difference from the mean, the few cases at the other end of the scale then strongly contribute to standard deviation or variance [15]. In general, these ordinal measures all fulfill the following properties as outlined by Aeppli and Ruedin [8]:

A1: Non-negativity: The measures are non-negative, meaning they assume values greater than or equal to 0. A value of 0 signifies the highest level of dispersion (or polarization), which occurs if and only if the probability mass is evenly split between the two extreme categories: $\boldsymbol{p} = (1/2, 0, \dots, 0, 1/2)$.

A2: Boundedness: The measures are upper-bounded by 1, meaning they assume values less than or equal to 1. A value of 1 represents the highest level of concentration (or consensus), occurring if and only if all probability mass is concentrated within a single category: $\boldsymbol{p} = (0, \dots, 0, 1, 0, \dots, 0)$.

A3: A uniform distribution $\boldsymbol{p} = (1/K, \dots, 1/K)$ yields a value that is strictly greater than 0 and strictly less than 1 (not necessarily 0.5).

We reckon that these properties of non-negativity, minimum and maximum dispersion (**A1**, **A2**) are also meaningful for uncertainty quantification in the context of ordinal classification. In particular, the highest degree of uncertainty should not be represented by a uniform distribution, as in standard nominal classification, but rather by a distribution that evenly splits the probability mass between the extreme categories.

Additional axioms can be required for uncertainty measures. The well-known Shannon entropy, for example, is characterized by continuity, symmetry, and additivity (in addition to non-negativity and maximum uncertainty). Except for additivity, these properties can also be considered for the ordinal case, albeit symmetry only makes sense in a very restricted form.

A4: Continuity: The uncertainty measure is a continuous function of the (predictive) probability distribution. Thus, small changes in the (predictive) probability distribution should only result in small changes in the uncertainty measure. This is crucial for the stability and robustness of the measure, ensuring that the uncertainty measure is not overly sensitive to minor perturbations in the (predictive) probability distribution caused by noise or slight variations in the input data.

A5: Invariance against reversal of the scale: This property ensures that the uncertainty measure, even if affected by the ordering of probabilities, is not affected by the direction of the ordinal scale. Formally, let $\boldsymbol{p} = (p_1, p_2, \dots, p_K)$ be a probability distribution on an ordinal scale $\mathcal{O} = \{1, 2, \dots, K\}$, and let $\sigma_{\leftrightarrow}$ denote the permutation defined by $\sigma_{\leftrightarrow}(k) = K - k + 1$. Then, we require that

$$u_{\mathrm{ORD}}(\boldsymbol{p}) = u_{\mathrm{ORD}}(\boldsymbol{p}_{\sigma_{\leftrightarrow}}),$$

where $\boldsymbol{p}_{\sigma_{\leftrightarrow}} = (p_{\sigma_{\leftrightarrow}(1)}, p_{\sigma_{\leftrightarrow}(2)}, \dots, p_{\sigma_{\leftrightarrow}(K)}) = (p_K, p_{K-1}, \dots, p_1)$. Note that this is a weaker form of invariance compared to common nominal measures like entropy, confidence, or margin, which are invariant to any permutation of the probabilities, i.e., $u(\boldsymbol{p}) = u(\boldsymbol{p}_\sigma)$ for any permutation σ. Since the focus of this axiom is on the exclusivity of invariance with respect to the reversal of the ordinal scale, any measure that is invariant to more than just the reversal of the ordinal scale violates this axiom.

4.2. Ordinal measures

Given that ordinal rating measures are specifically designed to capture the above characteristics, we believe that they are particularly well suited for quantifying uncertainty in ordinal classification. In the following, we introduce several such measures for ordinal data.

4.2.1. The measure by Leik

We begin with Leik's measure of ordinal consensus [16], which computes the dispersion D as a measure of ordinal consensus for a probability (relative frequency) distribution $\boldsymbol{p}$ with K categories using the cumulative distribution $F_k(\boldsymbol{p}) = \sum_{1 \le l \le k} p_l$:

$$D(\boldsymbol{p}) = \frac{2 \sum_{k=1}^{K} d_k}{K-1}, \text{ with } d_k = \begin{cases} F_k(\boldsymbol{p}) & \text{if } F_k(\boldsymbol{p}) \le 0.5 \\ 1 - F_k(\boldsymbol{p}) & \text{otherwise} \end{cases}.$$

In its original form, Leik's measure is a measure of dispersion. It ranges from 0 to 1, with 0 indicating no dispersion or maximal concentration, and 1 representing maximum dispersion or minimal concentration. When half of the probability mass is located at each extreme end of the ordinal scale, the measure reaches its maximum value of 1, indicating maximum dispersion or minimal

concentration or consensus. Conversely, when all the probability mass is concentrated on a single category, the measure takes the value 0, indicating minimal dispersion or maximal concentration or consensus. As outlined by Blair and Lacy [17], Leik's measure can also be transformed into a measure of concentration or consensus, in line with the above-listed properties:

$$C_1(\boldsymbol{p}) = 1 - D(\boldsymbol{p}) = \frac{\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5|}{(K-1)/2}. \tag{4}$$

Formally, the following proposition can be shown.

Proposition 4.1. *The measure C_1 satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

All proofs of the results presented in this paper can be found in Appendix A.

4.2.2. The measure by Blair and Lacy

Furthermore, Blair and Lacy also introduce a squared version of the measure [17]:

$$C_2(\boldsymbol{p}) = \frac{\sum_{k=1}^{K-1} (F_k(\boldsymbol{p}) - 0.5)^2}{(K-1)/4}, \tag{5}$$

which uses Euclidean distance instead of L_1-distance to measure the distance between the cumulative probability F_k and 0.5. Hence, the following proposition also holds.

Proposition 4.2. *The measure C_2 satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

Both Blair and Lacy's and Leik's measure can be considered as members of a family of measures that follow a similar construction principle and operate on cumulative probabilities F_k:

$$\text{Concentration} = \frac{D}{D_{\max}},$$

where D represents the measure of dispersion or concentration and $D_{\max}$ serves as a normalization factor. The purpose of $D_{\max}$ is to scale the measure to a range between 0 and 1, allowing for easier interpretation and comparison. The complementary measure of dispersion is then given by

$$\text{Measure of dispersion} = 1 - \frac{D}{D_{\max}}.$$

4.2.3. The measure by Tastle and Wierman

A different approach is taken by Tastle and Wierman, who expand on the Shannon entropy to define a measure of consensus as follows [18]:

$$\text{Cns}(\boldsymbol{p}) = 1 + \sum_{k=1}^{K} p_k \log_2\left(1 - \frac{|k-\mu|}{K-1}\right), \tag{6}$$

where $\mu = \sum_k p_k \cdot k$ is the expected value and (like in the case of Shannon entropy) $0 \cdot \log_2(0) = 0$ by definition. Unlike the previous measures it does not operate on cumulative probabilities but relies, like standard deviation or variance, on the distance to the mean μ to measure the dispersion of the distribution. Tastle and Wierman also consider the measure $\text{Dnt}(\boldsymbol{p}) = 1 - \text{Cns}(\boldsymbol{p})$, which they call dissention. Nonetheless, the following proposition is also valid.

Proposition 4.3. *The measure Cns satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

4.2.4. The measure by Van der Eijk

Another popular measure of agreement (or consensus) in ordered rating scales is the measure by Van der Eijk, which is introduced and thoroughly explained in a procedural form in [14]. In terms of a single formula, it can be written as follows:

$$A(\boldsymbol{p}) = \sum_{k=1}^{K} \underbrace{|S_k| \cdot \left(p_{(k)} - p_{(k-1)}\right)}_{w} \cdot \underbrace{\left(1 - \frac{|S_k| - 1}{K-1}\right)}_{V} \cdot \underbrace{\left(\frac{(K-2) \cdot |TU(S_k)| - (K-1) \cdot |TDU(S_k)|}{(K-2) \cdot (|TU(S_k)| + |TDU(S_k)|)}\right)}_{U}, \tag{7}$$

Table 1
This table illustrates the calculation of $A(\boldsymbol{p})$ for the exemplary bimodal five-class probability distribution $\boldsymbol{p} = (0.45, 0.15, 0.0, 0.0, 0.4)$, with $A(\boldsymbol{p}) = \sum_{k=1}^{K} w_k \cdot V_k \cdot U_k = \sum_{k=1}^{K} w_k \cdot A_k = -0.575$ (cf. Fig. 3).

k	$\lvert S_k \rvert$	$p_{(k)} - p_{(k-1)}$	$\lvert TDU(S) \rvert$	$\lvert TU(S) \rvert$	w	V	U	A
3	3	0.15	4	2	0.45	0.5	$-0.5\overline{5}$	$-0.2\overline{7}$
4	2	0.25	3	0	0.5	0.75	$-1.\overline{3}$	-1.0
5	1	0.05	0	0	0.05	1.0	1.0	1.0

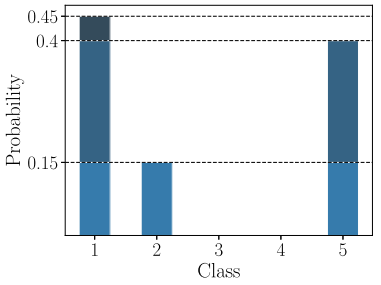


Fig. 3. Illustration of how Van der Eijk's measure of agreement reduces a probability distribution $\boldsymbol{p} = (0.45, 0.15, 0.0, 0.0, 0.4)$ horizontally into different layers based on the difference between the k-th and $(k-1)$-th smallest probabilities ($p_{(k)} - p_{(k-1)}$). The overall level of agreement is then an aggregation of the layer-wise levels of agreement weighted by the amount of probability mass of the particular layer.

where $(\cdot)$ is a permutation[2] such that $p_{(1)} \leq \ldots \leq p_{(K)}$. Moreover, $S_j = \{k \mid p_k \geq p_{(j)}\}$ is the set of ranks k whose probability p_k exceeds the j^{th}-largest probability $p_{(j)}$,

$$TDU(S) = \left\{(i,j,k) \mid 1 \leq i < j < k \leq K,\, i,k \in S,\, j \notin S\right\}$$

counts the number of rank triples in S that violate unimodality (the "in-between" probability p_j is lower than both p_i and p_k), and

$$TU(S) = \left\{(i,j,k) \mid 1 \leq i < j < k \leq K,\, (i,j \in S,\, k \notin S) \vee (j,k \in S,\, i \notin S)\right\}$$

counts the number of rank triples in S that are unimodal (where either p_i is lower than p_j and p_k or p_k is lower than p_i and p_j). Note that, $U = 1$ by definition if $|TDU(S)| = 0$ and $|TU(S)| = 0$, which is the case for uniform or Dirac distributions.

Fig. 3 illustrates how Van der Eijk's approach reduces the assessment of a distribution to the assessment of subsets of ordinal ranks, namely by decomposing the distribution "horizontally" into several layers. For each layer, a measure of agreement is obtained by counting the number of rank triplets that agree and disagree with unimodality, respectively. The layer-wise agreement values are then aggregated into an overall agreement score, weighted by the overall probability mass of each layer. Table 1 displays the corresponding layer-wise calculations for the probability distribution $\boldsymbol{p} = (0.45, 0.15, 0.0, 0.0, 0.4)$.

Van der Eijk's agreement measure ranges between -1 (maximal dispersion) to $+1$ (maximal concentration) and also assigns a meaningful value of 0 to the uniform distribution. To make the measure of agreement A fulfill the above properties (cf. Section 4.1), it can be scaled to the interval $[0, 1]$ as follows:

$$C_A(\boldsymbol{p}) = 1 + \frac{A(\boldsymbol{p})}{2}, \tag{8}$$

with a uniform distribution then resulting in a value of 0.5.

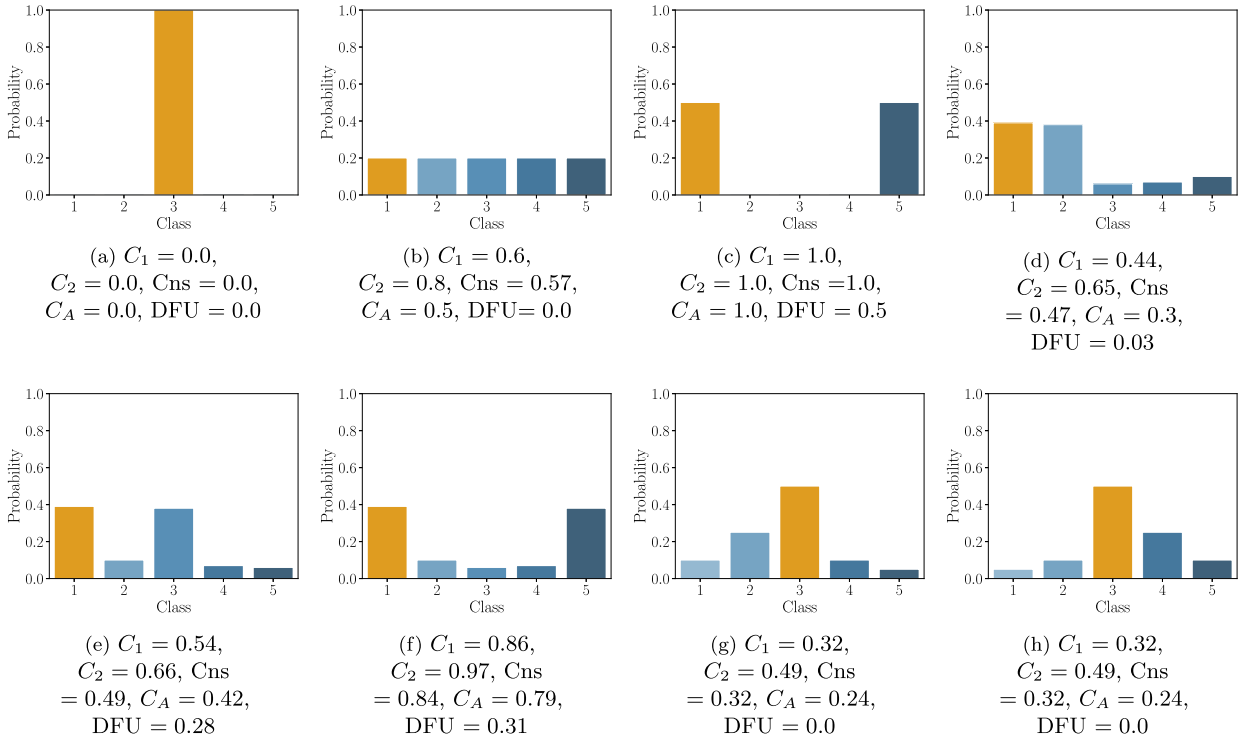
Formally, we can also show that the measure satisfies the axioms presented in Section 4.1.

Proposition 4.4. *The measure C_A satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

4.3. The measure by Pavlopoulos and Likas

In contrast to the previous measures, Koudenburg et al. [15] propose a data-driven approach to measuring opinion polarization (as the opposite of consensus). They introduce an opinion polarization index derived from survey data, which offers valuable insights into the characteristics of polarized opinion distributions. They develop their index in an empirical way, namely by training a regression model on exemplary distributions that were previously rated by 58 international experts in terms of the degree of polarization. By

[2] We set $p_{(0)} = 0$ by definition.



(a) $C_1 = 0.0$, $C_2 = 0.0$, Cns $= 0.0$, $C_A = 0.0$, DFU $= 0.0$

(b) $C_1 = 0.6$, $C_2 = 0.8$, Cns $= 0.57$, $C_A = 0.5$, DFU$= 0.0$

(c) $C_1 = 1.0$, $C_2 = 1.0$, Cns $=1.0$, $C_A = 1.0$, DFU $= 0.5$

(d) $C_1 = 0.44$, $C_2 = 0.65$, Cns $= 0.47$, $C_A = 0.3$, DFU $= 0.03$

(e) $C_1 = 0.54$, $C_2 = 0.66$, Cns $= 0.49$, $C_A = 0.42$, DFU $= 0.28$

(f) $C_1 = 0.86$, $C_2 = 0.97$, Cns $= 0.84$, $C_A = 0.79$, DFU $= 0.31$

(g) $C_1 = 0.32$, $C_2 = 0.49$, Cns $= 0.32$, $C_A = 0.24$, DFU $= 0.0$

(h) $C_1 = 0.32$, $C_2 = 0.49$, Cns $= 0.32$, $C_A = 0.24$, DFU $= 0.0$

Fig. 4. Results of the different ordinal consensus based uncertainty measures $u_{\text{Consensus}}$ and DFU on different simulated five-class probability distributions.

leveraging this expertise, Koudenburg et al. are able to create a quantitative measure that captures the level of polarization within a given dataset. It is important to note that the opinion polarization index derived by Koudenburg et al. has a limitation in that it is designed specifically for datasets with five categories. Consequently, its applicability is limited to situations where the response options are constrained to this particular number of categories.

Building upon the collected survey data and findings by Koudenburg et al. [15], Pavlopoulos and Likas [19] propose another measure to assess opinion polarization, called the distance from unimodality (DFU) measure. This measure has demonstrated a strong correlation with expert ratings in terms of polarized distributions. The DFU measure focuses on capturing the presence of opinion clusters, which Koudenburg et al. identified as one of the primary sources of polarization alongside extremity and distance [15]. In contrast to the regression model developed by Koudenburg et al. [15], DFU is generally applicable and not limited to five categories:

$$\text{DFU}(\boldsymbol{p}) = \max\{d_1, \ldots, d_K\} \quad \text{with} \tag{9}$$

$$d_k = \begin{cases} p_k - p_{k+1} & \text{if} \quad 1 \leq k < m \\ 0 & \text{if} \quad k = m \\ p_k - p_{k-1} & \text{if} \quad m < k \leq K \end{cases},$$

where m is the mode[3] of the distribution $\boldsymbol{p} = (p_1, \ldots, p_K)$. In case of a unimodal distribution, DFU will be 0 and indicate no polarization at all (cf. Fig. 4). In contrast, if DFU is greater than 0, it indicates a multimodal distribution containing opinion clusters and hence some sort of polarization. The DFU measure is also particularly interesting for the case of ordinal classification, as unimodality of the predicted output probabilities is often mentioned as a requirement for proper probabilistic ordinal classification [20,21]. Hence, violation of this property may be an indicator of increased uncertainty. However, DFU does not satisfy all axioms defined in Section 4.1 and is not able to quantify the "peakedness" of unimodal distributions, which questions its usefulness for uncertainty quantification in ordinal classification.

Proposition 4.5. *Under the assumption of a single mode m, the measure DFU satisfies axioms **A4** and **A5**, but violates axioms **A1**, **A2**, and **A3**.*

[3] In the case where $\boldsymbol{p}$ has several modes, m is taken as the smallest (left-most) one.

4.4. Ordinal uncertainty quantification using consensus measures

The measures (8), (4) and (5) introduced, respectively, by Van der Eijk [14], Leik [16], and Blair and Lacy [17] do not assume equal distances between categories. This is in contrast to the consensus measure (6) introduced by Tastle and Wierman [18], which treats ordinal scales as if they were interval scales [15]. Treating ordinal scales as interval scales is a common practice when analyzing Likert scale survey data, which is the primary application of the presented measures. In this context, the assumption of equal distances between categories allows for a simplified quantitative interpretation and analysis of the data including calculation of standard deviation or variance. The assumption of equal distances is also quite common in ordinal classification, which makes all quantitative measures also applicable to the ordinal classification setting [7].

In summary, the consensus measures $C \in \{C_1, C_2, \mathrm{Cns}, C_A\}$ proposed by Leik [16], Blair and Lacy [17], Tastle and Wierman [18], and Van der Eijk [14] give rise to a generic consensus-based uncertainty quantification framework for probabilistic ordinal classification, suggesting a consensus-based uncertainty measure u_{CONS} that is obtained by turning consensus into a complementary measure of dispersion:

$$u_{\mathrm{CONS}}(\boldsymbol{x}_q) = 1 - C(p(y \mid \boldsymbol{x}_q)) \ .$$

The DFU measure (9), which represents a distinct approach, can be directly applied to quantify uncertainty in probabilistic ordinal classification.

Fig. 4 compares the different consensus measures, plugged into the generic uncertainty measure u_{CONS}, over eight simulated probability distributions, including the two distributions leading to the upper and lower bound values of 0 and 1 as well as the uniform distribution. DFU is also shown though it conceptionally differs significantly from the other measures.

4.5. Variance

Unlike the other uncertainty measures presented in Section 3, variance (3) satisfies the axioms defined in Section 4.1.

Proposition 4.6. *The measure VAR satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

Unlike variance, entropy, confidence, and margin violate axioms **A1** and **A3**, as they are maximized or minimized by a uniform distribution and are not constrained by the extreme bimodal distribution. Furthermore, they are not exclusively invariant under the reversal of the ordinal scale but are invariant to any rank permutations, which violates axiom **A5**. Overall, these violations make them theoretically less suitable for uncertainty quantification in ordinal classification, similar to DFU (9).

5. Binary decomposition for uncertainty quantification in ordinal classification

In machine learning, binary reduction techniques are used to tackle multinomial classification tasks with binary classifiers. Such techniques reduce a single multinomial problem to a set of binary classification problems. At prediction time, a query instance is submitted to each of the binary models, and the predictions produced by the models are combined into a prediction for the original multinomial problem. The most straightforward and arguably simplest reduction scheme is the one-vs-rest decomposition, where one binary classifier is trained per class, with the task to separate that class from all other classes [22].

In the case of ordinal classification, the most natural reduction to the binary case is achieved through binary splits of the ordinal scale, separating a lower part $\{y_1, \ldots, y_m\}$ of the scale from an upper part $\{y_{m+1}, \ldots, y_K\}$ [23,24]. Indeed, if the ordinal structure on the class labels is reflected in the corresponding class-conditional distributions, these binary problems are presumably easier to solve than those produced by other splits [25].

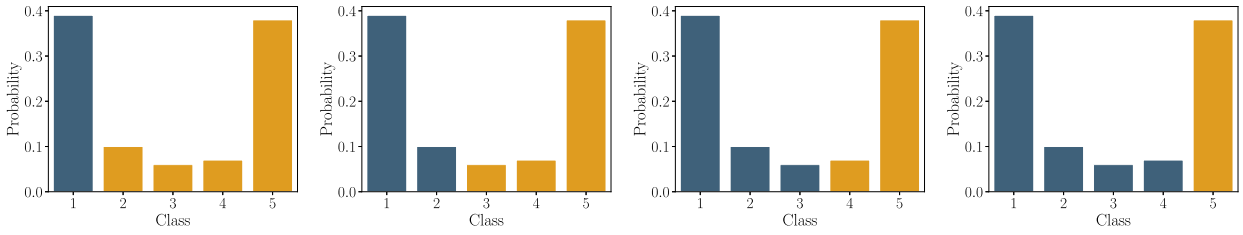
The principle of binary reduction can also be applied to uncertainty quantification [26]. In the ordinal case, it suggests a measure of the form

$$u_{\mathrm{ORD}}(\boldsymbol{p}) = \sum_{k=1}^{K-1} u_{\mathrm{BIN}}\left(\sum_{i=1}^{k} p_i, \sum_{j=k+1}^{K} p_j \right), \tag{10}$$

where u_{BIN} is any uncertainty measure applicable to the binary case, i.e., an appropriate measure of uncertainty for Bernoulli distributions (see Fig. 5 for an illustration). We call u_{BIN} the generator of u_{ORD}. Examples of generators include established measures such as entropy and margin, which are invariant to probability mass re-distribution in their original (multinomial) form.

The measure (10) is plausible in the following sense: The more bi- or multimodal the distribution $\boldsymbol{p}$, and the greater the distance between the modes, the more "uncertain split" can be produced, and the higher the sum on the right-hand side becomes. In this regard, the measure is very much in line with the dispersion measures discussed in the previous section, in particular with the principle proposed by Van der Eijk (8) [14]. Formally, the following lemma can be shown very easily.

Lemma 5.1. *Let u_{BIN} be any generator that is maximized by a uniform probability distribution $\boldsymbol{p}_{BIN} = (1/2, 1/2)$. Then, the measure (10) is maximized by the bimodal distribution $\boldsymbol{p} = (1/2, 0, \ldots, 0, 1/2)$. Likewise, let u_{BIN} be any generator that is minimized by $\boldsymbol{p}_{BIN} = (0, 1)$ and $\boldsymbol{p}_{BIN} = (1, 0)$. Then, the measure (10) is also minimal on the Dirac distributions.*



(a) Class 1 vs. Classes 2 to 5.

(b) Classes 1 and 2 vs. Classes 3 to 5.

(c) Classes 1 to 3 vs. Classes 4 and 5.

(d) Classes 1 to 4 vs. Class 5.

Fig. 5. Five class example of an ordinal binary decomposition.

Furthermore, the measure (10) is also invariant toward reversal of the ordinal scale, provided u_{BIN} is symmetric (which is a property that most uncertainty measures satisfy when being applied to a Bernoulli distribution, including entropy, variance, margin, and confidence).

Lemma 5.2. *Under the assumption of symmetry for the generator u_{BIN}, consider a probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ and its reversal $\boldsymbol{p}_{\sigma_\leftrightarrow} = (p_K, p_{K-1}, \ldots, p_1)$ on an ordinal scale $\mathcal{O} = \{1, 2, \ldots, K\}$. Then, $\boldsymbol{p}$ and $\boldsymbol{p}_{\sigma_\leftrightarrow}$ result in the same uncertainty: $u_{ORD}(\boldsymbol{p}) = u_{ORD}(\boldsymbol{p}_{\sigma_\leftrightarrow})$.*

Overall, the following proposition can be deduced from the above lemmas.

Proposition 5.1. *Under the assumptions of symmetry and continuity for the generator u_{BIN}, the measure u_{ORD} satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5**.*

Interestingly, several existing measures are recovered as a special case of the binary decomposition method, with a suitable choice of the generator.

Proposition 5.2. *A normalized version of the binary decomposition method with margin as generator reduces to the complementary dispersion measure D_1 for the measure C_1 in (4).*

Proposition 5.3. *A normalized version of the binary decomposition method with variance as generator reduces to the complementary dispersion measure D_2 for the measure C_2 in (5).*

Although aggregating the binary uncertainty estimates using the sum (10) appears natural, other aggregation functions $F : \mathbb{R}^K \to \mathbb{R}$ are also conceivable and may even enable further connections to existing measures, as well as more nuanced uncertainty quantification in the ordinal case. In principle, all functions lower-bounded by the minimum and upper-bounded by the maximum, the so-called averaging operators [27], could be considered as candidates. The simplest extension of (10) is a weighted sum

$$u_{\text{WORD}}(\boldsymbol{p}) = \sum_{k=1}^{K-1} w_k \cdot u_{\text{BIN}}\left(\sum_{i=1}^{k} p_i, \sum_{j=k+1}^{K} p_j\right), \quad \text{where} \sum_{k=1}^{K-1} w_k = 1,\ w_k \geq 0, \tag{11}$$

with non-negative weights $w_1, \ldots, w_{K-1}$. For instance, there is often an interest in ordinal classification to improve the reliability in deciding the extreme cases, the first and last class on the ordinal scale, as deciding those wrongly may have the most severe consequences [28]. This can be accomplished by making w_1 and w_{K-1} higher than the other weights.

Another interesting class of (parametrized) aggregation functions is the ordered weighted average (OWA), which interpolates between the minimum and maximum [29]:

$$F(a_1, \ldots, a_K) = \sum_{k=1}^{K} w_k\, b_k, \tag{12}$$

where b_k is the k-th largest of the input values in $\boldsymbol{a}$, and $\boldsymbol{w}$ a vector of non-negative weights summing to one. Note that the minimum is obtained for $w_K = 1$, the maximum for $w_1 = 1$, and the standard arithmetic mean for $w_1 = \ldots = w_K = 1/K$.

Although many different aggregations of the binary uncertainty estimates are conceivable and worth investigating in future work, we will stick to the sum as the most generic one for the rest of this paper.

Table 2
Twenty-three common ordinal benchmark datasets used for evaluating the different uncertainty measures.

Dataset	# instances	# features	# classes
Grub Damage	155	8	4
Obesity	2,111	16	7
CMC	1,473	9	3
New Thyroid	215	5	3
Balance Scale	625	4	3
Automobile	205	25	7
Eucalyptus	736	19	5
TAE	151	5	3
Heart (CLE)	303	13	5
SWD	1,000	10	4
ERA	1,000	4	9
ESL	488	4	9
LEV	1,000	4	5
Red Wine	1,599	11	6
White Wine	4,898	11	7
Triazines	186	60	5
Machine CPU	209	6	10
Auto MPG	392	7	10
Boston Housing	506	13	5
Pyrimidines	74	27	10
Abalone	4,177	8	10
Wisconsin Breast Cancer	194	32	5
Stocks Domain	950	9	5

6. Experiments with ordinal benchmark datasets

In this section, we evaluate the previously introduced uncertainty measures on common tabular ordinal benchmark datasets.[4] The focus is on how well these measures are capable of quantifying uncertainty in the ordinal case and improving the reliability of decision making.

6.1. Choice of base learner and datasets

For our evaluation, we rely on gradient boosted tree (GBT) models as base learners instead of neural networks, as tree-based models represent the state of the art for tabular data, and this type of data is common in high-risk ADM environments like finance or medicine [30,31] (refer to Appendix B for additional experiments using a multi-layer perceptron (MLP)). Concretely, we utilize the LightGBM instantiation of GBTs [32] with the cross-entropy (CE) loss for multi-class classification:

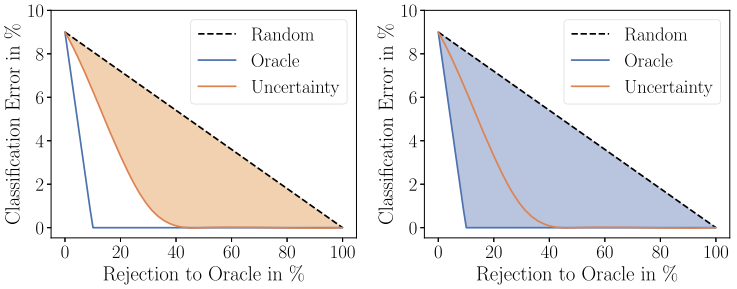
$$l_{CE}(\boldsymbol{y}, \boldsymbol{p}) = -\sum_{k=1}^{K} y_k \log(p_k), \tag{13}$$

where $\boldsymbol{y}$ is a one-hot (0/1) encoded vector with y_k being 1 for the true class y and 0 for the rest of the classes, and $\boldsymbol{p}$ the predictive probability distribution. This approach enables us to obtain conditional probability distributions $p(y \mid \boldsymbol{x})$, which serve as the foundation for evaluating various uncertainty measures. Moreover, CE is also a proper scoring rule, which encourages the model to output probability distributions that reflect the true underlying probabilities of the data [10].

As will be detailed further below, common ordinal classification metrics or losses, such as accuracy, mean absolute error, or quadratic weighted kappa (QWK) [33] will be used for evaluating predictive performance in the end. One may wonder, therefore, why cross-entropy (13) should be used for training, instead of targeting any of these losses directly or using other popular ordinal losses like squared earth mover's distance (EMD^2), which take the ordinal structure into account during training [34]. The reason is that such losses, while tailored to producing good ordinal predictions, do not incentivize an unbiased prediction of true probabilities (they are not proper scoring rules). Instead, as discussed by de la Torre et al. for the QWK loss [33] and Liu et al. [35], they tend to bias the predictive probabilities toward unimodality. Furthermore, in ordinal classification, a common theme is to explicitly constrain predictive output probabilities to unimodality [20,21]. However, the enforcement of unimodal output probabilities can be too restrictive, a notion recently recognized with the introduction of quasi-unimodal distributions. These distributions only enforce unimodality in the vicinity of the true class, offering a more nuanced approach [36]. By sidestepping these constraints, our aim is to uncover the natural structure of ordinal predictive probability distributions through the use of an unbiased proper scoring rule, without the imposition of strong, potentially unrealistic assumptions (refer to Appendix C for additional experiments illustrating the superiority of the CE loss as a proper scoring rule over ordinal predictors when it comes to uncertainty quantification).

Table 2 presents the attributes of the twenty-three ordinal benchmark datasets utilized for our evaluation, which are widely recognized within the realm of ordinal classification research [37,38]. These datasets are characterized by variability in size, number

[4] The source code is available at https://github.com/stefanahaas41/uncertainty-quantification-probabilistic-ordinal-classification.



(a) Shaded area is $AR_{uncertainty}$. (b) Shaded area is AR_{oracle}.

Fig. 6. Example Prediction Rejection Curves [51].

of features, and class distributions, offering a robust foundation for a thorough assessment of various uncertainty quantification measures.

In terms of preprocessing the datasets for the experimental evaluation, all categorical features were one-hot (0/1) encoded and the ordinal labels $y_1, \ldots, y_k$ were integer encoded from $1, \ldots, K$.

6.2. Experimental setup

To compare the different uncertainty measures on the different datasets we compute prediction rejection ratios (PRRs) [39] for different classifier performance evaluation metrics using 10-fold cross validation. The PRR is calculated on the basis of rejection curves [40,41], where first the predictive uncertainties of the test dataset are determined based on an uncertainty measure and then queries are successively rejected with descending predictive uncertainty. If the uncertainty quantification works properly this should result in a monotone increasing or, depending on the selected performance metric, decreasing rejection curve. When calculating PRRs, the assumption is that rejected queries are delegated to an oracle that will answer queries correctly. Concretely, the PRR of an uncertainty measure is calculated by measuring the area between the uncertainty measure's rejection curve and a random rejection curve which in expectation is a straight line—$AR_{uncertainty}$ (cf. Fig. 6a). This value is then normalized by the area between the perfect oracle (ORC) rejection curve and the random rejection line—AR_{oracle} (cf. Fig. 6b):

$$PRR = \frac{AR_{uncertainty}}{AR_{oracle}} = \frac{AU_{uncertainty} - AU_{random}}{AU_{oracle} - AU_{random}}$$

Consequently, a PRR of 1 indicates perfect rejection whereas a value of 0 indicates random rejection. The area between the rejection curves AR can be calculated by making use of the area under the curve (AUC) metric with $AU = 1 - AUC$, which essentially calculates the area above the rejection curve [6,42]. The PRR can also become negative, which indicates worse than random uncertainty quantification.

To calculate a PRR, one also needs to select a performance evaluation metric for the classifier. In the realm of ordinal classification, accuracy (ACC), mean absolute error (MAE), and QWK appear to be the most popular performance metrics [7,43–46]. While the QWK requires a complete confusion matrix, which can be problematic for small datasets and at the tail of the rejection curve, the mean squared error (MSE) serves as a suitable alternative. MSE not only emphasizes larger errors but is also a well-established metric for evaluating performance in ordinal classification contexts [43,47–50]. To make all rejection curves go in the same direction, we measure the misclassification rate (MCR) (also known as mean zero-one error (MZE)) instead of ACC, as is commonly done [6,39,42]:

$$\text{MCR} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)$$

Similar to the approach outlined by Kotsiantis and Pintelas [52], we determine the final prediction $\hat{y}$ of the probabilistic predictor according to Bayesian decision theory, i.e., we take a decision that minimizes the expected loss (Bayes risk). The optimal policy that minimizes the risk is also called the Bayes estimator. Given our performance measures MCR, MAE and MSE we have three corresponding losses (l_{01}, l_1, l_2) that need to be minimized given the posterior predictive probabilities over the ordinal classes in order to take the decision with the least associated risk:

$$\hat{y} = \arg\min_{\hat{y} \in \mathcal{Y}} R(\hat{y} \mid \boldsymbol{x}) = \arg\min_{\hat{y} \in \mathcal{Y}} \mathbb{E}_{p(y \mid \boldsymbol{x})}[l(\hat{y}, y)] = \arg\min_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} l(\hat{y}, y) \cdot p(y \mid \boldsymbol{x}).$$

Furthermore, we also include the Bayesian risk associated with a certain prediction $\hat{y}$ based on l_1 and l_2 losses as baseline uncertainty measures in our set of evaluated uncertainty measures [52]:

$$R_{l_1}(\hat{y} \mid \boldsymbol{x}) = \mathbb{E}_{p(y \mid \boldsymbol{x})}[l_1(\hat{y}, y)] = \sum_{y \in \mathcal{Y}} |\hat{y} - y| \cdot p(y \mid \boldsymbol{x}),$$

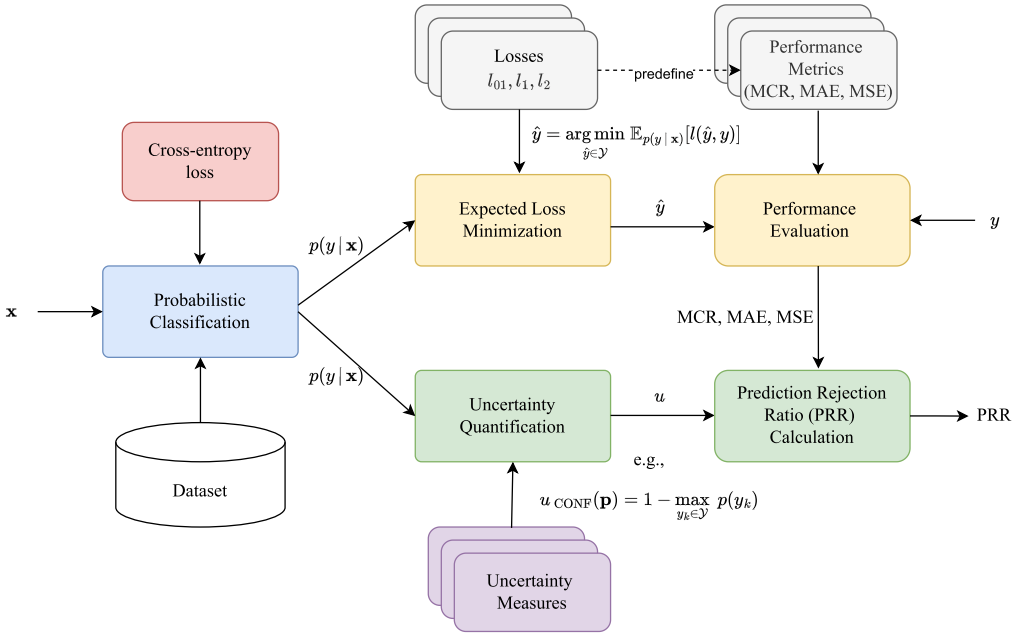


Fig. 7. Overview of the experimental approach: Final predictions $\hat{y}$ are derived using various Bayes estimators, while the predictive uncertainty u is quantified using different uncertainty measures. All these measures utilize the unbiased and realistic predictive probability distribution $\boldsymbol{p} = p(y \mid \boldsymbol{x})$ obtained using cross-entropy loss as a proper scoring rule. Eventually, the PRR values are calculated based on the quantified uncertainty and the obtained performance metrics predefined by the respective losses.

$$R_{l_2}(\hat{y} \mid \boldsymbol{x}) = \mathbb{E}_{p(y \mid \boldsymbol{x})}[l_2(\hat{y}, y)] = \sum_{y \in \mathcal{Y}} (\hat{y} - y)^2 \cdot p(y \mid \boldsymbol{x}) .$$

The risk associated with the l_{01} loss is already covered by the u_{CONF} uncertainty measure which calculates the probability of making an incorrect decision:

$$R_{l_{01}}(\hat{y} \mid \boldsymbol{x}) = \mathbb{E}_{p(y \mid \boldsymbol{x})}[l_{01}(\hat{y}, y)] = 1 - \arg\max_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) .$$

Fig. 7 graphically illustrates the experimental approach employed to calculate the PRR values for various Bayes estimators, performance metrics, and uncertainty measures. These calculations are based on unbiased and realistic predictive probability distributions obtained through cross-entropy loss as a proper scoring rule.

6.3. Results and analysis

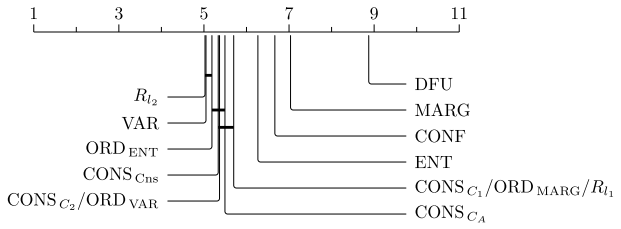
Table 3 displays the overall PRR results of a 10-fold cross validation on the selected ordinal benchmark datasets. In total, we evaluate fourteen uncertainty measures: CONF, MARG, ENT, VAR, CONS_{Cns} [18], CONS_{C_1} [16], CONS_{C_2} [17], ORD_{ENT}, ORD_{MARG}, ORD_{VAR}, R_{l_1}, R_{l_2}, CONS_{C_A} [14] and DFU [19]. The first three measures do not take into account the dispersion of the output probability distribution and are common nominal classification uncertainty measures, whereas the rest of the measures can be considered dispersion measures, with DFU as a special case focusing on the detection of non-unimodal distributions, respectively opinion clusters. As one can see, there is no overall clear winner at first sight, and the performance of a measure appears to depend on the data.

However, overall when considering MCR, MAE and MSE, dispersion measures have an edge over CONF, MARG and ENT, when looking at the critical difference (CD) diagram in Fig. 8a. The groups of best performing uncertainty measures solely consists of measures that take the dispersion of the probability distribution into account, and there is a statistically significant difference between dispersion measures compared to nominal classification measures. Interestingly, when looking only at MCR or the exact hit rate, there is no statistically significant difference between all measures (excluding DFU) (cf. Fig. 8b). One may have expected that nominal classification measures have an advantage here.

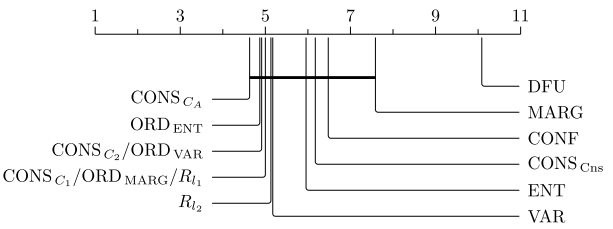
When considering the distance of the errors by looking at MAE and MSE, the best performing group consists of VAR and the Bayes risk for the l_2 loss (R_{l_2}), followed by the rest of the dispersion measures (cf. Fig. 8f). As expected, nominal classification measures fail in taking the error distance into account and are not competitive when it comes to distance-based errors. Though VAR and R_{l_2} perform best when it comes to taking the error distance into account, they do not perform so well when it comes to the exact hit-rate based on MCR. This behavior is also visible for CONS_{Cns}, which, just like VAR, also measures the dispersion of the distribution with regard to the mean. Other measures like CONS_{C_2} or ORD_{ENT} seem to strike a better balance between categorical classification accuracy (hit rate) and minimum distance-based error. As already proven in Section 5, CONS_{C_2} and ORD_{VAR} as well as CONS_{C_1}

Table 3
PRRs for the different uncertainty measures and ordinal benchmark datasets using 10-fold cross-validation with LightGBM as base learner.

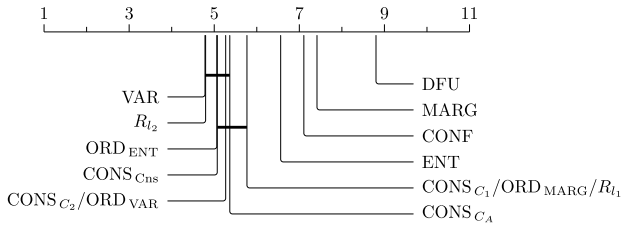
Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Cns}$	$CONS_{C_1}$	$CONS_{C_3}$	$CONS_{C_4}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{l_1}	R_{l_2}
Triazines	MCR	0.1368 ± 0.3073	0.1235 ± 0.2903	0.1788 ± 0.2762	0.1889 ± 0.2491	0.178 ± 0.26	0.1863 ± 0.2821	0.1666 ± 0.2632	0.1933 ± 0.2847	0.0669 ± 0.1725	0.1824 ± 0.271	0.1863 ± 0.2821	0.1666 ± 0.2632	0.1863 ± 0.2821	**0.2052** ± 0.2561
	MAE	0.1582 ± 0.3453	0.1411 ± 0.3361	0.238 ± 0.3072	0.3176 ± 0.2388	0.2896 ± 0.2725	0.2559 ± 0.3217	0.2531 ± 0.2891	0.2582 ± 0.3164	0.0427 ± 0.3001	0.276 ± 0.2974	0.2559 ± 0.3217	0.2531 ± 0.2891	0.2559 ± 0.3217	**0.3307** ± 0.2532
	MSE	0.1146 ± 0.3485	0.1287 ± 0.3558	0.206 ± 0.3239	0.3454 ± 0.2633	0.3133 ± 0.2826	0.2325 ± 0.3468	0.2483 ± 0.3128	0.229 ± 0.3475	-0.0017 ± 0.4853	0.2845 ± 0.327	0.2325 ± 0.3468	0.2483 ± 0.3128	0.2325 ± 0.3468	**0.3632** ± 0.2467
Machine CPU	MCR	0.7118 ± 0.1656	0.6856 ± 0.1807	0.7361 ± 0.1422	0.7976 ± 0.1446	0.7692 ± 0.1482	0.7626 ± 0.1573	0.775 ± 0.156	0.7814 ± 0.1505	0.5421 ± 0.3997	0.7846 ± 0.1356	0.7626 ± 0.1573	0.775 ± 0.156	0.7626 ± 0.1573	**0.7998** ± 0.1371
	MAE	0.6349 ± 0.1503	0.5975 ± 0.1694	0.6685 ± 0.1402	**0.7762** ± 0.1369	0.7184 ± 0.1313	0.7105 ± 0.1255	0.7298 ± 0.1258	0.7429 ± 0.1263	0.5784 ± 0.4175	0.7516 ± 0.1143	0.7105 ± 0.1255	0.7298 ± 0.1258	0.7105 ± 0.1255	0.7746 ± 0.1249
	MSE	0.5662 ± 0.1707	0.518 ± 0.1867	0.6018 ± 0.1658	**0.7541** ± 0.1427	0.6661 ± 0.1432	0.6561 ± 0.1402	0.6817 ± 0.1382	0.7021 ± 0.1357	0.5902 ± 0.4384	0.7184 ± 0.1289	0.6561 ± 0.1402	0.6817 ± 0.1382	0.6561 ± 0.1402	0.7478 ± 0.1287
Auto MPG	MCR	0.345 ± 0.1364	0.3317 ± 0.14	0.3658 ± 0.123	**0.4126** ± 0.0988	0.386 ± 0.1138	0.3829 ± 0.1137	0.3931 ± 0.1053	0.3909 ± 0.1048	0.1828 ± 0.1159	0.4037 ± 0.0973	0.3829 ± 0.1137	0.3931 ± 0.1053	0.3829 ± 0.1137	0.4029 ± 0.1083
	MAE	0.3485 ± 0.116	0.3307 ± 0.1206	0.3617 ± 0.1116	**0.4582** ± 0.1076	0.4402 ± 0.1063	0.4264 ± 0.0963	0.4389 ± 0.107	0.4353 ± 0.0951	0.2575 ± 0.2245	0.4469 ± 0.0982	0.4264 ± 0.0963	0.4389 ± 0.107	0.4264 ± 0.0963	0.4544 ± 0.1128
	MSE	0.3474 ± 0.2539	0.3301 ± 0.2549	0.3438 ± 0.2316	**0.4973** ± 0.1856	0.4795 ± 0.2055	0.4486 ± 0.2024	0.4683 ± 0.2057	0.4637 ± 0.1961	0.3548 ± 0.3224	0.4799 ± 0.1895	0.4486 ± 0.2024	0.4683 ± 0.2057	0.4486 ± 0.2024	0.4889 ± 0.1944
Pyrimidines	MCR	0.1434 ± 0.5549	0.1395 ± 0.596	**0.1839** ± 0.5618	-0.1734 ± 0.3926	-0.0529 ± 0.455	0.0512 ± 0.4001	-0.0086 ± 0.3905	-0.0123 ± 0.4874	0.0558 ± 0.3954	-0.0278 ± 0.3872	0.0512 ± 0.4001	-0.0086 ± 0.3905	0.0512 ± 0.4001	-0.1734 ± 0.3926
	MAE	0.106 ± 0.3315	0.0434 ± 0.3076	0.3371 ± 0.2252	0.2449 ± 0.3379	0.2608 ± 0.3647	0.3479 ± 0.2953	0.3486 ± 0.2926	**0.3663** ± 0.2219	0.2396 ± 0.3515	0.2957 ± 0.3093	0.3479 ± 0.2953	0.3486 ± 0.2926	0.3479 ± 0.2953	0.2358 ± 0.3439
	MSE	0.1688 ± 0.5395	-0.0235 ± 0.4597	0.2901 ± 0.6089	0.5872 ± 0.3016	0.5745 ± 0.2872	0.5694 ± 0.2538	**0.5965** ± 0.2379	0.5212 ± 0.3031	0.3067 ± 0.3976	0.5785 ± 0.2933	0.5694 ± 0.2538	0.5965 ± 0.2379	0.5694 ± 0.2538	0.5945 ± 0.2933
Abalone	MCR	0.2629 ± 0.0303	0.2422 ± 0.0302	0.2783 ± 0.0392	0.2874 ± 0.035	0.2889 ± 0.0259	0.2872 ± 0.027	0.2854 ± 0.0334	0.282 ± 0.0298	0.0461 ± 0.0823	0.2857 ± 0.038	0.2872 ± 0.027	0.2854 ± 0.0334	0.2872 ± 0.027	**0.2922** ± 0.0284
	MAE	0.2447 ± 0.0466	0.2118 ± 0.0479	0.2925 ± 0.0474	**0.3215** ± 0.0458	0.3039 ± 0.0454	0.295 ± 0.0474	0.3065 ± 0.0455	0.2991 ± 0.0426	0.1025 ± 0.0952	0.3159 ± 0.0456	0.295 ± 0.0474	0.3065 ± 0.0455	0.295 ± 0.0474	0.318 ± 0.0473
	MSE	0.2132 ± 0.0949	0.1706 ± 0.0931	0.2833 ± 0.0882	**0.3221** ± 0.0835	0.2779 ± 0.0978	0.2642 ± 0.0972	0.2932 ± 0.0867	0.286 ± 0.0852	0.1217 ± 0.1079	0.3149 ± 0.0826	0.2642 ± 0.0972	0.2932 ± 0.0867	0.2642 ± 0.0972	0.3039 ± 0.0945
Boston Housing	MCR	0.3612 ± 0.2102	0.3652 ± 0.2086	0.364 ± 0.2154	**0.3705** ± 0.2144	0.3623 ± 0.2119	0.363 ± 0.2126	0.3653 ± 0.2124	0.367 ± 0.2143	-0.0357 ± 0.2781	0.3688 ± 0.2096	0.363 ± 0.2126	0.3653 ± 0.2124	0.363 ± 0.2126	0.3667 ± 0.212
	MAE	0.355 ± 0.2066	0.3586 ± 0.2058	0.359 ± 0.2116	**0.3769** ± 0.2074	0.3616 ± 0.207	0.3616 ± 0.2062	0.3641 ± 0.2058	0.3694 ± 0.2064	-0.0224 ± 0.2827	0.3713 ± 0.2028	0.3616 ± 0.2062	0.3641 ± 0.2058	0.3616 ± 0.2062	0.3718 ± 0.2041
	MSE	0.3393 ± 0.1991	0.3425 ± 0.1993	0.3446 ± 0.2026	**0.378** ± 0.1882	0.3529 ± 0.1897	0.3515 ± 0.1892	0.3542 ± 0.1881	0.3647 ± 0.1865	0.0081 ± 0.3165	0.3668 ± 0.186	0.3515 ± 0.1892	0.3542 ± 0.1881	0.3515 ± 0.1892	0.3714 ± 0.1842
Stocks Domain	MCR	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
	MAE	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
	MSE	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
Wisconsin Breast Cancer	MCR	0.2093 ± 0.3534	0.139 ± 0.3264	**0.2763** ± 0.3239	0.1976 ± 0.3016	0.1699 ± 0.3346	0.1838 ± 0.3426	0.2121 ± 0.3204	0.1557 ± 0.3397	-0.0961 ± 0.3344	0.2493 ± 0.29	0.1838 ± 0.3426	0.2121 ± 0.3204	0.1838 ± 0.3426	0.2008 ± 0.324
	MAE	0.1418 ± 0.2491	0.0913 ± 0.2475	0.2018 ± 0.2611	0.2263 ± 0.2923	0.2296 ± 0.2845	0.2228 ± 0.2627	0.2328 ± 0.2609	0.1713 ± 0.2517	0.1391 ± 0.2426	0.2394 ± 0.2555	0.2228 ± 0.2627	0.2328 ± 0.2609	0.2228 ± 0.2627	**0.251** ± 0.2897
	MSE	0.1149 ± 0.2686	0.0691 ± 0.2501	0.1634 ± 0.2834	0.1357 ± 0.3245	0.1429 ± 0.3008	0.1429 ± 0.2893	0.1528 ± 0.2964	0.1027 ± 0.2623	0.1731 ± 0.2083	0.1465 ± 0.3104	0.1429 ± 0.2893	0.1528 ± 0.2964	0.1429 ± 0.2893	**0.1806** ± 0.3045
Obesity	MCR	0.8883 ± 0.0845	0.8866 ± 0.0855	0.8874 ± 0.0843	**0.894** ± 0.0775	0.888 ± 0.0867	0.8893 ± 0.0829	0.8896 ± 0.0823	0.8872 ± 0.0834	0.5076 ± 0.3054	0.8892 ± 0.0819	0.8893 ± 0.0829	0.8896 ± 0.0823	0.8893 ± 0.0829	0.8933 ± 0.0775
	MAE	0.8856 ± 0.0819	0.8841 ± 0.083	0.8848 ± 0.0819	**0.8913** ± 0.0752	0.8855 ± 0.0844	0.8867 ± 0.0805	0.8871 ± 0.08	0.8846 ± 0.081	0.5047 ± 0.3041	0.8867 ± 0.0796	0.8867 ± 0.0805	0.8871 ± 0.08	0.8867 ± 0.0805	0.8906 ± 0.0751
	MSE	0.8839 ± 0.0808	0.8826 ± 0.0819	0.8831 ± 0.0809	**0.8896** ± 0.0744	0.8838 ± 0.0834	0.8851 ± 0.0795	0.8855 ± 0.0791	0.8828 ± 0.08	0.5054 ± 0.3047	0.8852 ± 0.0787	0.8851 ± 0.0795	0.8855 ± 0.0791	0.8851 ± 0.0795	0.8889 ± 0.0743
CMC	MCR	0.3143 ± 0.0738	**0.3146** ± 0.0775	0.306 ± 0.0665	0.2399 ± 0.0435	0.2282 ± 0.0461	0.2851 ± 0.0585	0.2772 ± 0.0532	0.2678 ± 0.0486	0.009 ± 0.069	0.2745 ± 0.0529	0.2851 ± 0.0585	0.2772 ± 0.0532	0.2851 ± 0.0585	0.2226 ± 0.0441
	MAE	0.2113 ± 0.0546	0.2218 ± 0.063	0.1973 ± 0.0436	0.2889 ± 0.0717	0.2926 ± 0.0705	0.2754 ± 0.0636	0.278 ± 0.0658	0.2778 ± 0.0695	0.0222 ± 0.146	0.2774 ± 0.0659	0.2754 ± 0.0636	0.278 ± 0.0658	0.2754 ± 0.0636	**0.2964** ± 0.0666
	MSE	0.0308 ± 0.0515	0.0411 ± 0.0578	0.0303 ± 0.0416	0.1405 ± 0.0795	0.1576 ± 0.0736	0.0807 ± 0.0541	0.089 ± 0.0602	0.1018 ± 0.0801	-0.0426 ± 0.1895	0.0907 ± 0.0601	0.0807 ± 0.0541	0.089 ± 0.0602	0.0807 ± 0.0541	**0.1739** ± 0.0704
Grub Damage	MCR	0.2406 ± 0.239	0.2157 ± 0.2222	0.2767 ± 0.2586	0.286 ± 0.2756	0.2384 ± 0.317	0.2553 ± 0.2915	0.3038 ± 0.3327	0.2871 ± 0.3163	0.072 ± 0.1933	**0.3176** ± 0.3154	0.2553 ± 0.2915	0.3038 ± 0.3327	0.2553 ± 0.2915	0.2359 ± 0.274
	MAE	0.0922 ± 0.2525	0.0739 ± 0.2451	0.1287 ± 0.2708	**0.2431** ± 0.2965	0.2121 ± 0.3041	0.1577 ± 0.2975	0.2169 ± 0.3508	0.2045 ± 0.3586	0.1267 ± 0.2378	0.2254 ± 0.3252	0.1577 ± 0.2975	0.2169 ± 0.3508	0.1577 ± 0.2975	0.2417 ± 0.2513
	MSE	0.159 ± 0.3607	0.1237 ± 0.3671	0.1764 ± 0.3528	0.2793 ± 0.2708	0.2606 ± 0.2436	0.1871 ± 0.329	0.2361 ± 0.3313	0.2159 ± 0.3824	0.1544 ± 0.2695	0.2375 ± 0.3425	0.1871 ± 0.329	0.2361 ± 0.3313	0.1871 ± 0.329	**0.2986** ± 0.2125
New Thyroid	MCR	0.9822 ± 0.0288	0.9822 ± 0.0288	0.9822 ± 0.0288	**1.0** ± 0.0	1.0 ± 0.0	0.9875 ± 0.02	0.9875 ± 0.02	1.0 ± 0.0	0.5203 ± 0.5054	0.9875 ± 0.02	0.9875 ± 0.02	0.9875 ± 0.02	0.9875 ± 0.02	1.0 ± 0.0
	MAE	0.9742 ± 0.0462	0.9742 ± 0.0462	0.9742 ± 0.0462	**0.9969** ± 0.0076	0.9969 ± 0.0076	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9969 ± 0.0076	0.5421 ± 0.4852	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9938 ± 0.0153
	MSE	0.969 ± 0.0582	0.969 ± 0.0582	0.969 ± 0.0582	**0.9949** ± 0.0125	0.9949 ± 0.0125	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9949 ± 0.0125	0.5561 ± 0.475	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9898 ± 0.0251
Balance Scale	MCR	0.8648 ± 0.0996	0.8627 ± 0.0937	0.8551 ± 0.1164	0.8642 ± 0.0703	0.8531 ± 0.0686	0.8602 ± 0.0818	0.8679 ± 0.077	**0.8817** ± 0.0669	0.0997 ± 0.2413	0.8729 ± 0.0743	0.8602 ± 0.0818	0.8679 ± 0.077	0.8602 ± 0.0818	0.8509 ± 0.068
	MAE	0.8072 ± 0.0953	0.8051 ± 0.0924	0.7949 ± 0.1074	0.8327 ± 0.0746	0.8217 ± 0.0767	0.8141 ± 0.076	0.8247 ± 0.0745	**0.8483** ± 0.0756	0.132 ± 0.2571	0.8309 ± 0.0713	0.8141 ± 0.076	0.8247 ± 0.0745	0.8141 ± 0.076	0.8206 ± 0.0706
	MSE	0.8032 ± 0.1032	0.8016 ± 0.1039	0.801 ± 0.1033	0.8303 ± 0.0607	0.8264 ± 0.0678	0.8142 ± 0.0754	0.8221 ± 0.0667	**0.8392** ± 0.055	0.0736 ± 0.2428	0.8278 ± 0.0593	0.8142 ± 0.0754	0.8221 ± 0.0667	0.8142 ± 0.0754	0.8336 ± 0.0733
Automobile	MCR	0.6564 ± 0.3921	0.6546 ± 0.3793	0.671 ± 0.3805	0.6911 ± 0.3496	0.6885 ± 0.3621	0.6832 ± 0.3676	0.6885 ± 0.3674	0.6965 ± 0.3608	0.4635 ± 0.17	0.6904 ± 0.3522	0.6832 ± 0.3676	0.6885 ± 0.3674	0.6832 ± 0.3676	**0.7036** ± 0.3567
	MAE	0.6215 ± 0.3772	0.6182 ± 0.3645	0.6284 ± 0.3672	0.6653 ± 0.3294	0.6581 ± 0.3442	0.6468 ± 0.3538	0.6498 ± 0.3541	0.6634 ± 0.3455	0.4099 ± 0.302	0.654 ± 0.3408	0.6468 ± 0.3538	0.6498 ± 0.3541	0.6468 ± 0.3538	**0.6803** ± 0.3373
	MSE	0.6003 ± 0.3861	0.5927 ± 0.3738	0.5974 ± 0.3762	0.6436 ± 0.3331	0.6299 ± 0.3495	0.6185 ± 0.3629	0.6187 ± 0.3637	0.6366 ± 0.3519	0.3452 ± 0.4262	0.6231 ± 0.3512	0.6185 ± 0.3629	0.6187 ± 0.3637	0.6185 ± 0.3629	**0.6582** ± 0.3392
Eucalyptus	MCR	0.447 ± 0.0761	0.4476 ± 0.077	0.4428 ± 0.0876	0.4439 ± 0.0876	**0.4571** ± 0.0833	0.4556 ± 0.0773	0.4503 ± 0.0846	0.4466 ± 0.08	0.0345 ± 0.136	0.442 ± 0.0872	0.4556 ± 0.0773	0.4503 ± 0.0846	0.4556 ± 0.0773	0.4523 ± 0.0818
	MAE	0.428 ± 0.0517	0.4267 ± 0.0577	0.4298 ± 0.0573	0.4369 ± 0.0488	**0.4446** ± 0.0545	0.4412 ± 0.0495	0.4393 ± 0.053	0.4365 ± 0.0507	0.0231 ± 0.1391	0.4341 ± 0.0497	0.4412 ± 0.0495	0.4393 ± 0.053	0.4412 ± 0.0495	0.4429 ± 0.0489
	MSE	0.3957 ± 0.0815	0.3962 ± 0.0863	0.4022 ± 0.0912	0.4182 ± 0.0819	0.4203 ± 0.0844	0.4158 ± 0.0788	0.417 ± 0.087	0.4164 ± 0.0842	0.0129 ± 0.1227	0.4139 ± 0.0829	0.4158 ± 0.0788	0.417 ± 0.087	0.4158 ± 0.0788	**0.4212** ± 0.0816
TAE	MCR	0.1154 ± 0.2666	0.0897 ± 0.2809	0.1422 ± 0.2643	0.1044 ± 0.3305	0.0589 ± 0.3033	0.154 ± 0.3105	0.1526 ± 0.3076	0.1108 ± 0.3173	**0.1902** ± 0.1997	0.1566 ± 0.333	0.154 ± 0.3105	0.1526 ± 0.3076	0.154 ± 0.3105	0.0423 ± 0.2976
	MAE	0.1 ± 0.2511	0.0766 ± 0.2699	0.133 ± 0.187	0.1766 ± 0.2973	0.1485 ± 0.2889	**0.2121** ± 0.2494	0.1958 ± 0.2452	0.1821 ± 0.3006	0.1478 ± 0.2066	0.1973 ± 0.2525	0.2121 ± 0.2494	0.1958 ± 0.2452	0.2121 ± 0.2494	0.1328 ± 0.3007
	MSE	0.0208 ± 0.3281	-0.0067 ± 0.3443	0.008 ± 0.2383	0.2304 ± 0.3097	0.2082 ± 0.2995	0.1955 ± 0.3094	0.2084 ± 0.2665	0.1936 ± 0.3486	**0.3243** ± 0.1727	0.2187 ± 0.2662	0.1955 ± 0.3094	0.2084 ± 0.2665	0.1955 ± 0.3094	0.1849 ± 0.2956
Heart (CLE)	MCR	0.5456 ± 0.1668	0.5276 ± 0.1618	**0.5763** ± 0.1566	0.5709 ± 0.1325	0.5534 ± 0.1482	0.5541 ± 0.1492	0.5583 ± 0.1375	0.5747 ± 0.1458	0.1375 ± 0.2534	0.5744 ± 0.1347	0.5541 ± 0.1492	0.5583 ± 0.1375	0.5541 ± 0.1492	0.5703 ± 0.142
	MAE	0.5029 ± 0.1576	0.4786 ± 0.1532	0.5285 ± 0.1573	**0.5575** ± 0.105	0.5344 ± 0.1225	0.5187 ± 0.1361	0.5298 ± 0.1238	0.5391 ± 0.1208	0.0889 ± 0.2527	0.5458 ± 0.1153	0.5187 ± 0.1361	0.5298 ± 0.1238	0.5187 ± 0.1361	0.5536 ± 0.1185
	MSE	0.3833 ± 0.1841	0.3494 ± 0.1832	0.4062 ± 0.1854	**0.4899** ± 0.1379	0.4479 ± 0.1438	0.4021 ± 0.1471	0.429 ± 0.1592	0.4304 ± 0.1412	0.0288 ± 0.2581	0.453 ± 0.1497	0.4021 ± 0.1471	0.429 ± 0.1592	0.4021 ± 0.1471	0.483 ± 0.1496
SWD	MCR	0.1983 ± 0.0991	0.1859 ± 0.1042	0.1848 ± 0.086	0.1949 ± 0.0967	0.2006 ± 0.102	**0.2019** ± 0.1052	0.1948 ± 0.0963	0.2018 ± 0.0977	0.043 ± 0.1066	0.1938 ± 0.0905	0.2019 ± 0.1052	0.1948 ± 0.0963	0.2019 ± 0.1052	0.1988 ± 0.1088
	MAE	0.1143 ± 0.0971	0.1152 ± 0.0931	0.1172 ± 0.0867	0.1372 ± 0.1034	0.1405 ± 0.103	0.1252 ± 0.1113	0.1285 ± 0.1002	0.1275 ± 0.1068	0.052 ± 0.1113	0.1312 ± 0.0953	0.1252 ± 0.1113	0.1285 ± 0.1002	0.1252 ± 0.1113	**0.1423** ± 0.1089
	MSE	0.1017 ± 0.0935	0.101 ± 0.0816	0.1154 ± 0.0946	0.1331 ± 0.1141	0.1401 ± 0.0978	0.1134 ± 0.0979	0.1166 ± 0.1073	0.1071 ± 0.1135	0.0498 ± 0.1022	0.1239 ± 0.1023	0.1134 ± 0.0979	0.1166 ± 0.1073	0.1134 ± 0.0979	**0.1403** ± 0.1032
ERA	MCR	0.1708 ± 0.0604	**0.2059** ± 0.0941	0.1377 ± 0.0886	0.0468 ± 0.0662	0.0801 ± 0.0635	0.131 ± 0.0633	0.101 ± 0.066	0.1777 ± 0.0575	0.0901 ± 0.1104	0.085 ± 0.0646	0.131 ± 0.0633	0.101 ± 0.066	0.131 ± 0.0633	0.0598 ± 0.0802
	MAE	0.0353 ± 0.0965	0.0409 ± 0.1273	0.0069 ± 0.076	0.0138 ± 0.0923	0.0477 ± 0.0839	0.0597 ± 0.0668	0.0358 ± 0.0806	0.0439 ± 0.0688	**0.0756** ± 0.1181	0.0131 ± 0.0887	0.0597 ± 0.0668	0.0358 ± 0.0806	0.0597 ± 0.0668	0.0221 ± 0.0972
	MSE	0.0586 ± 0.0932	**0.0725** ± 0.1066	0.0088 ± 0.1241	-0.0023 ± 0.1514	0.0222 ± 0.1352	0.0283 ± 0.1189	0.0174 ± 0.141	0.0241 ± 0.1206	0.0421 ± 0.1256	0.0031 ± 0.1457	0.0283 ± 0.1189	0.0174 ± 0.141	0.0283 ± 0.1189	0.0016 ± 0.152
ESL	MCR	0.2079 ± 0.1653	0.2119 ± 0.1854	0.2164 ± 0.1122	0.2738 ± 0.1225	**0.2953** ± 0.1443	0.2487 ± 0.1418	0.2484 ± 0.1228	0.2275 ± 0.1257	0.0123 ± 0.1516	0.2467 ± 0.1141	0.2487 ± 0.1418	0.2484 ± 0.1228	0.2487 ± 0.1418	0.2817 ± 0.1347
	MAE	0.1738 ± 0.1687	0.1778 ± 0.1887	0.182 ± 0.1401	**0.2887** ± 0.1285	0.2864 ± 0.1415	0.2305 ± 0.1371	0.2343 ± 0.1158	0.2197 ± 0.1146	0.0469 ± 0.1569	0.2524 ± 0.1218	0.2305 ± 0.1371	0.2343 ± 0.1158	0.2305 ± 0.1371	0.287 ± 0.1346
	MSE	0.1048 ± 0.2408	0.1098 ± 0.2698	0.1089 ± 0.2014	**0.275** ± 0.1536	0.2374 ± 0.1863	0.1689 ± 0.1811	0.18 ± 0.1441	0.1718 ± 0.1428	0.1026 ± 0.1814	0.227 ± 0.1351	0.1689 ± 0.1811	0.18 ± 0.1441	0.1689 ± 0.1811	0.2592 ± 0.1613
LEV	MCR	0.1559 ± 0.1345	0.1398 ± 0.1402	0.1571 ± 0.1182	0.1539 ± 0.11	0.1582 ± 0.1268	0.1605 ± 0.13	**0.1652** ± 0.1202	0.1619 ± 0.1271	-0.0153 ± 0.1251	0.1493 ± 0.1169	0.1605 ± 0.13	0.1652 ± 0.1202	0.1605 ± 0.13	0.1542 ± 0.1235
	MAE	0.1302 ± 0.1181	0.1079 ± 0.121	0.1426 ± 0.1144	0.1397 ± 0.1001	0.1311 ± 0.1096	0.1342 ± 0.1134	**0.1483** ± 0.1077	0.1457 ± 0.1143	0.0062 ± 0.1436	0.137 ± 0.1091	0.1342 ± 0.1134	0.1483 ± 0.1077	0.1342 ± 0.1134	0.1315 ± 0.1047
	MSE	0.1375 ± 0.1169	0.1048 ± 0.1145	0.1607 ± 0.1268	0.1546 ± 0.1063	0.1297 ± 0.1088	0.1394 ± 0.1076	0.161 ± 0.1116	**0.1666** ± 0.1204	0.0264 ± 0.182	0.1572 ± 0.1206	0.1394 ± 0.1076	0.161 ± 0.1116	0.1394 ± 0.1076	0.1335 ± 0.101
Red Wine	MCR	0.4037 ± 0.0738	0.3982 ± 0.0732	0.4169 ± 0.0708	0.4228 ± 0.0698	0.4074 ± 0.0677	0.4109 ± 0.0691	0.4146 ± 0.0687	0.4155 ± 0.0694	-0.052 ± 0.0356	**0.4237** ± 0.0688	0.4109 ± 0.0691	0.4146 ± 0.0687	0.4109 ± 0.0691	0.4171 ± 0.0673
	MAE	0.4021 ± 0.0719	0.396 ± 0.0727	0.4217 ± 0.0682	0.4321 ± 0.0659	0.4133 ± 0.0645	0.4144 ± 0.0667	0.4205 ± 0.0651	0.419 ± 0.0672	-0.062 ± 0.0413	**0.4322** ± 0.0661	0.4144 ± 0.0667	0.4205 ± 0.0651	0.4144 ± 0.0667	0.4251 ± 0.0639
	MSE	0.39 ± 0.0898	0.3827 ± 0.0925	0.4162 ± 0.0818	**0.4296** ± 0.0694	0.4049 ± 0.0783	0.4028 ± 0.082	0.4126 ± 0.076	0.4078 ± 0.0812	-0.0647 ± 0.0781	0.4283 ± 0.0737	0.4028 ± 0.082	0.4126 ± 0.076	0.4028 ± 0.082	0.4203 ± 0.0729
White Wine	MCR	0.3642 ± 0.0537	0.3583 ± 0.0509	0.3545 ± 0.053	0.3388 ± 0.0519	0.354 ± 0.0485	**0.3673** ± 0.0508	0.3583 ± 0.0495	0.3672 ± 0.0527	0.0477 ± 0.0641	0.3472 ± 0.053	0.3673 ± 0.0508	0.3583 ± 0.0495	0.3673 ± 0.0508	0.3584 ± 0.0487
	MAE	0.3479 ± 0.058	0.338 ± 0.0542	0.3495 ± 0.0604	0.338 ± 0.0566	0.3465 ± 0.0535	0.3564 ± 0.0555	0.3532 ± 0.0572	**0.3587** ± 0.0575	0.058 ± 0.0602	0.3454 ± 0.0592	0.3564 ± 0.0555	0.3532 ± 0.0572	0.3564 ± 0.0555	0.3523 ± 0.0507
	MSE	0.339 ± 0.0698	0.3249 ± 0.0674	0.3493 ± 0.0716	0.3409 ± 0.0662	0.3423 ± 0.0628	0.3498 ± 0.0641	0.3516 ± 0.0694	**0.3554** ± 0.0687	0.0574 ± 0.0624	0.3475 ± 0.0691	0.3498 ± 0.0641	0.3516 ± 0.0694	0.3498 ± 0.0641	0.3506 ± 0.0589



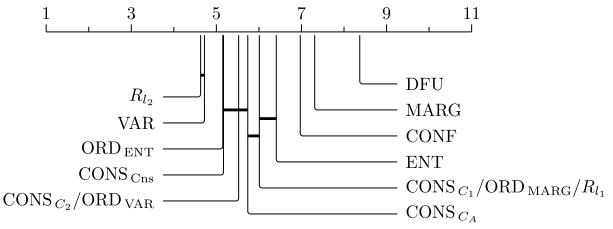
(a) Overall result for MCR, MAE and MSE.



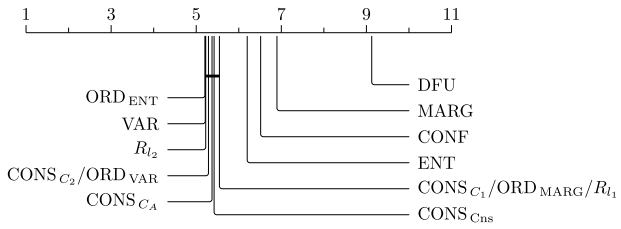
(b) Result only for MCR.



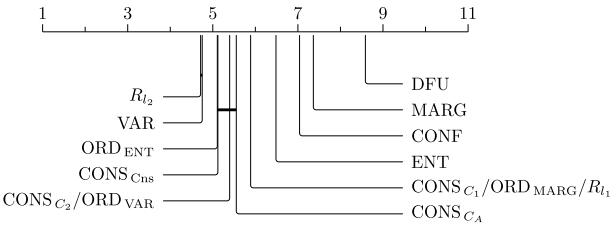
(c) Result for MAE.



(d) Result for MSE.



(e) Result for MCR and MAE.



(f) Result for MAE and MSE.

Fig. 8. Critical difference (CD) diagrams (https://github.com/mirkobunse/critdd) for the evaluated uncertainty measures over all performance metrics and datasets based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon test with LightGBM as base learner. Groups of uncertainty measures that are not significantly different (at p = 0.05) are connected [53,54].

and ORD_{MARG} are equivalent and thus lead to the same results in terms of their PRRs. Interestingly, from an empirical perspective, CONS_{C_1} and ORD_{MARG} appear to be equivalent to the Bayes risk with l_1 loss, R_{l_1}, also yielding the same results.

The DFU measure performs worst on all performance metrics and is often close or even worse than random rejection, which indicates that the probabilistic output of the predictor is mostly unimodal. Given unimodal probability distributions, DFU is not able to quantify any uncertainty at all, which might explain its poor performance on the considered datasets. If the predictor outputs mostly unimodal distributions, as indicated by DFU, one could also expect that taking the distance into account when quantifying uncertainty does not play such a role. However, the results of our experiment suggest the opposite. Even when the output is mostly unimodal, taking the distance into account does matter.

Furthermore, this experiment shows that our hypothesis indeed seems warranted and is further underpinned with additional experiments using a multi-layer perceptron (MLP) as the base learner in Appendix B. In ordinal probabilistic classification, uncertainty seems to be indeed maximal if all probability mass is equally allocated to the extreme ends of the ordinal scale. This is in contrast to the standard assumption of a uniform distribution representing maximal uncertainty.

By looking at exemplary rejection curves, we can further illustrate the superiority or at least competitiveness of dispersion measures compared to common uncertainty measures like entropy, margin, and confidence (cf. Fig. 9).

7. Case study: automotive goodwill claim assessment

In the following, we evaluate the different uncertainty measures on seven real-world goodwill claim assessment datasets of a car manufacturer (cf. Table 4) with the goal to predict appropriate monetary contributions for parts and labor repair costs on an interval scale from 0 to 100% binned to 10% steps ($\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$). Since goodwill claim assessment can be considered a high-stakes process, needing to balance customer-satisfaction and financial interests, well functioning predictive uncertainty quantification is of utmost importance. Furthermore, as goodwill requests are to a large extend assessed manually by human experts at the moment [55], it is also a perfect use case for selective classification [1] in which uncertain requests are still delegated to human experts, while trivial or clear cases are supposed to be processed automatically through automated decision making [3].

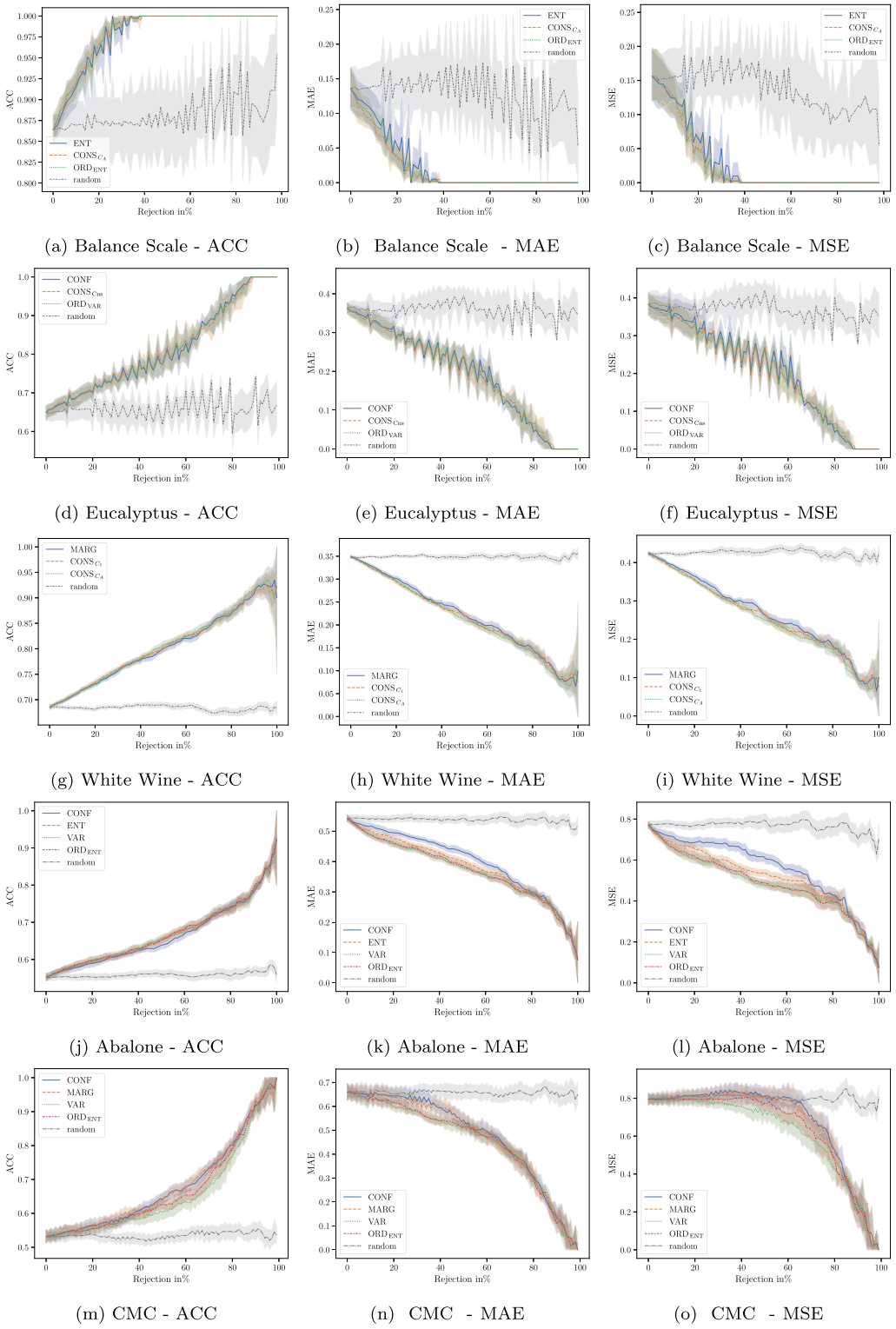
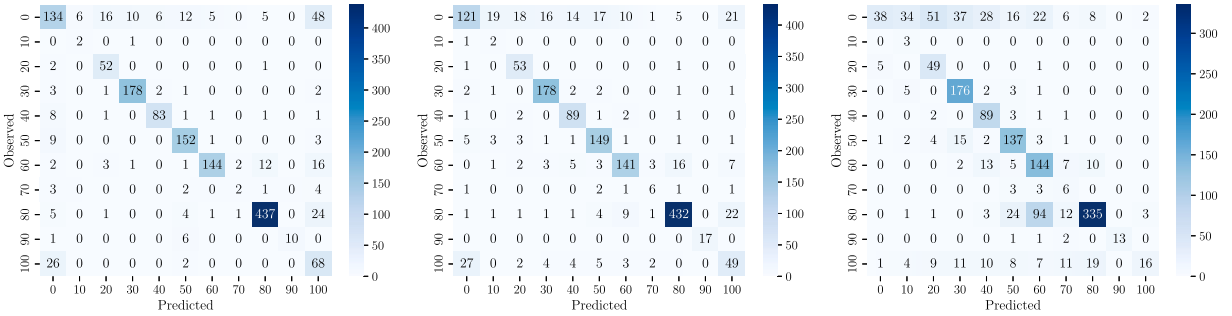


Fig. 9. Exemplary rejection curves for five of the ordinal benchmark datasets (Balance Scale, Eucalyptus, White Wine, Abalone and CMC).

Table 4
Goodwill claim assessment dataset sizes. All datasets have 26 features (18 categorical and 8 numeric) and a single label with 11 classes ($\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$).

Market	A	B	C	D	E	F	G
# Instances	9,127	7,636	21,209	19,066	174,008	9,127	9,945



(a) Cross-entropy (CE) loss (ACC: 0.83, MAE: 8.874, MSE: 642.633).

(b) Ordinal Regression approach [57] (ACC: 0.811, MAE: 8.251, MSE: 546.169).

(c) l_2 (mean squared error) loss (ACC: 0.671, MAE: 10.144, MSE: 442.502).

Fig. 10. Confusion matrices for goodwill claim assessment using different losses.

7.1. Datasets

The different goodwill claim assessment datasets are taken from different national sales markets and reflect the different goodwill assessment strategies of the national sales companies (NSC) of the car manufacturer. The attributes of the data instances entail information about the vehicle and the case, for instance, vehicle age, mileage, requested costs, defect code, whether the vehicle was regularly serviced, etc. [55]. Table 4 summarizes some characteristics of the datasets used for our evaluation. The sizes of the datasets vary heavily depending on the size of the sales market. In general, the datasets are in most cases heavily imbalanced [55], with the majority of instances falling into the extremes of no (0%) and full contributions (100%). This characteristic also polarizes the datasets in terms of decision outcomes. Given the variability in human goodwill judgment, it is crucial to recognize that observed decisions may not always be consistent. It is essential to account for this variability through unbiased predictive probability distributions and appropriate uncertainty quantification methodologies. For model training, the data is split into training and test data with a ratio of 80/20, where the test data contains the most recent 20% of the data.

7.2. Experimental setup

The problem of goodwill claim assessment can either be treated as an (ordinal) classification problem with 11 classes or a regression problem where predictions are rounded to the closest 10% step. Treating it as a classification problem using cross entropy loss results in a higher accuracy compared to treating it as a regression problem with L_2 loss (cf. Fig. 10). This increased accuracy however comes at the price of more substantial errors (e.g., 0 vs. 100%) manifested in a higher MSE. As already mentioned, this trade-off between categorical classification accuracy (hit rate) and minimum distance-based errors is inherent in ordinal classification and makes it a distinct problem [56]. There are many dedicated ordinal classification methods that try to represent this trade-off between accuracy and error spread on the loss level during training time and hence lie somewhere in the middle between classification and regression [33,34,45,47]. However, usually these methods have some drawbacks. For instance, the methods presented in [23] and [57] only provide deterministic predictions without uncertainty representation. This limitation can be critical in applications where understanding the uncertainty of predictions is essential, like in goodwill claim assessment. Additionally, as discussed in the previous Section 6, constraining predictive probability distributions to unimodality—explicitly [20,21], or implicitly [33,34,47]—negatively impacts uncertainty quantification as the probabilities are biased (cf. Appendix C). This is because unimodal constraints oversimplify the underlying predictive distributions by smoothing out the probabilities of distant classes, thereby leading to an underestimation of the true uncertainty present in the data. In the context of non-continuous ordinal rating data, such as that examined in our case study, truthful probability reporting is essential for an accurate representation of uncertainty. Constraining predictive probabilities to unimodality can obscure the true nature of the data, particularly when the underlying distribution is inherently polarized or multimodal. By allowing for polarized predictive probability distributions, we can better capture the full spectrum of uncertainty inherent in ordinal assessments.

Considering this, we again intentionally disregard the ordinal structure during the training phase by employing cross-entropy loss, which as a strictly proper scoring rule provides unbiased probabilistic predictions [10] and enables quantifiable uncertainty. Given that the historic goodwill claim assessment data used for our study is observational data with human decision makers acting as teachers, we deliberately want to account for potential biases by not constraining the model in any way that would veil those.

Similar to our previous study on common ordinal benchmark datasets, our goal is then to find an uncertainty measure that post-hoc takes this ordinal structure into account, with a specific focus on reducing substantial errors.

Since the data is of mid-size tabular nature, we again rely on GBTs for our evaluation implementation. Concretely, we make use of eXtreme Gradient Boosting (XGBoost) in that case [58].

7.3. Results and analysis

Table 5 shows the PRRs of the different uncertainty measures for MCR, MAE and MSE on seven goodwill assessment datasets split by the task of predicting labor or parts contributions.

Overall, when considering all performance metrics (MCR, MAE and MSE), we have a similar picture as in the previous benchmark study with measures taking distance into account outperforming standard nominal classification measures (cf. Table 6). However, in contrast to the previous study, standard nominal classification uncertainty measures outperform the other measures when focusing on the exact hit-rate through MCR. Nonetheless, when the focus is on reducing the error spread, indicated by MAE and MSE, VAR as well as CONS measures clearly outperform ENT, MARG and CONF.

Also, the binary decomposition method performs very competitive and even outperforms variance on MAE and MSE with entropy as binary base measure. Again, VAR, R_{l_2} and CONS_{Cns} shine on MAE and MSE, but perform poorly on MCR. Similar to the previous findings, other consensus and ordinal binary decomposition-based measures like CONS_{C_1}, CONS_{C_2} or ORD_{VAR} appear to strike a better balance between categorical classification accuracy (hit rate) and minimum distance-based error.

Interestingly, DFU does not come in last when looking at particular measures (e.g., only MCR or MAE and MSE), which is an indicator for non-unimodal predictive probability distributions output by the predictor. Compared to the previous study, there seems to be a more pronounced difference between classification accuracy and distance-based error, supposedly triggered by the non-unimodal predictive output probabilities of the predictor. On the goodwill claim assessment datasets it becomes even clearer that the binary decomposition method as well as the consensus measures (maybe apart from CONS_{Cns}) strike a better balance between accuracy and minimal distance-based error (cf. Tables 7 and 8).

Fig. 11 shows some exemplary rejection curves for which the above findings are clearly visible. Variance as well as consensus and ordinal binary decomposition-based measures have a clear advantage over ENT or CONF when looking at MSE or MAE. However, when solely looking at ACC, ENT or CONF are competitive or even better.

Tables 9 and 10 show corresponding performance metrics for rejections from 0% up to 50% in 10% steps for the overall best performing uncertainty measure (CONS_{C_2}, ORD_{VAR}). As can be seen, performance metrics ACC, MAE, MSE and QWK improve when rejecting uncertain queries including the domain-specific relevant cost metrics – underpayment, overpayment and total costs. Underpayment indicates how much the model would contribute less than the human experts and overpayment, the other way around. The total costs deviation (TOTAL) is then just the sum of the two.

Tables 9 and 10 also display the respective thresholds for the particular rejection percentages which are bound between 0 and 1. These thresholds could be used in a downstream selective classification [2] approach where a classifier $\hat{h}(\boldsymbol{x})$ rejects queries depending on a binary selection function $g(\boldsymbol{x})$, which will either indicate selection $g(\boldsymbol{x}) = 1$ or abstention $g(\boldsymbol{x}) = 0$:

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

Whether the function suggests to select the query for automated processing or abstention depends on the risk $\mathcal{R}_{\hat{h}}(\boldsymbol{x})$ associated with the query. If the calculated risk is below a given threshold δ, like the ones shown in Tables 9 and 10, the function will suggest selection:

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

As already stated, selective classification in combination with a consensus or ordinal binary decomposition-based uncertainty measure is an effective strategy to increase reliability in automated goodwill claim decisions. Concretely, using a consensus or binary decomposition-based measure will lead to an increase in hit-rate as well as a reduction in error distances, since it considers both aspects in a balanced way. Hence, employing a consensus or binary decomposition-based measure accounts for potentially polarized predictive probabilities that the learner may have picked up from the likely biased historic expert decisions.

8. Conclusion and future work

In this work, we have introduced and evaluated several uncertainty quantification measures with regards to their capability of quantifying uncertainty in probabilistic ordinal classification. We argued that the highest uncertainty in probabilistic ordinal classification should be represented by a distinct bimodal distribution, in which all probability mass is equally concentrated at the extreme ends of the ordinal scale, and the lowest uncertainty when all probability mass is allocated to a single class label. This is in contrast to nominal classification, where a uniform distribution typically indicates the highest degree of uncertainty. We also argued that complementary dispersion measures of so called consensus measures, originating from the social sciences, as well as our newly proposed ordinal binary decomposition method, in which uncertainty quantification is reduced to an ordered sequence of binary uncertainty quantification problems, best capture these distributions.

Table 5
PRRs for different measures over goodwill claim assessment data.

Dataset	Metric	CONF	MARG	ENT	VAR	CONS_{Cns}	CONS_{C_1}	CONS_{C_2}	CONS_{C_4}	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{l_1}	R_{l_2}
Market A (Parts)	ACC	0.7364	0.7245	**0.747**	0.6468	0.6541	0.6972	0.6943	0.6956	0.67	0.6916	0.6972	0.6943	0.6972	0.6484
	MAE	0.6416	0.6351	0.6433	0.685	0.6827	0.6731	0.6767	0.6707	0.6188	0.6821	0.6731	0.6767	0.6731	**0.6852**
	MSE	0.5102	0.504	0.5138	**0.6331**	0.6237	0.5799	0.5882	0.5778	0.5007	0.6	0.5799	0.5882	0.5799	0.6324
Market A (Labor)	ACC	0.8192	0.8097	**0.8284**	0.7238	0.7247	0.7807	0.7783	0.7785	0.7738	0.7743	0.7807	0.7783	0.7807	0.7253
	MAE	0.6886	0.6826	0.6924	0.7136	0.7071	0.7127	0.7169	0.709	0.6737	**0.72**	0.7127	0.7169	0.7127	0.7141
	MSE	0.5694	0.5643	0.5713	**0.6553**	0.6432	0.6212	0.6294	0.6169	0.5632	0.6372	0.6212	0.6294	0.6212	0.6552
Market B (Parts)	ACC	0.6434	0.6432	0.6398	0.6391	0.6563	**0.6567**	0.6562	0.6501	0.6475	0.6531	0.6567	0.6562	0.6567	0.64
	MAE	0.6163	0.6222	0.5991	0.713	**0.7193**	0.6899	0.6944	0.6903	0.6526	0.6988	0.6899	0.6944	0.6899	0.713
	MSE	0.5136	0.5246	0.4914	0.6858	**0.6877**	0.6268	0.636	0.6349	0.5718	0.6468	0.6268	0.636	0.6268	0.6852
Market B (Labor)	ACC	**0.7791**	0.7768	0.7775	0.7333	0.744	0.765	0.7617	0.759	0.7319	0.7566	0.765	0.7617	0.765	0.7348
	MAE	0.7762	0.7775	0.7684	0.833	**0.8336**	0.8248	0.8284	0.8256	0.7854	0.8311	0.8248	0.8284	0.8248	0.8332
	MSE	0.7326	0.7356	0.7241	**0.845**	0.8427	0.8125	0.8206	0.8161	0.7559	0.8281	0.8125	0.8206	0.8125	0.8449
Market C (Parts)	ACC	**0.6029**	0.6007	0.5893	0.4702	0.5127	0.5494	0.5347	0.5489	0.4028	0.5158	0.5494	0.5347	0.5494	0.4761
	MAE	0.5725	0.5562	0.5847	0.6038	0.6086	0.6074	0.6085	0.6083	0.5329	**0.6098**	0.6074	0.6085	0.6074	0.604
	MSE	0.4268	0.4059	0.453	**0.62**	0.5883	0.5418	0.5579	0.5399	0.5032	0.5796	0.5418	0.5579	0.5418	0.618
Market C (Labor)	ACC	0.7816	0.7796	**0.7818**	0.6888	0.7002	0.7348	0.7304	0.7284	0.716	0.7235	0.7348	0.7304	0.7348	0.6895
	MAE	0.8013	0.802	0.7933	0.7979	0.7989	0.8076	0.8071	0.8063	0.798	0.8059	**0.8076**	0.8071	0.8076	0.7978
	MSE	0.8225	0.8243	0.8138	**0.8589**	0.8568	0.8531	0.855	0.8531	0.8339	0.8568	0.8531	0.855	0.8531	0.8587
Market D (Parts)	ACC	**0.6803**	0.6734	0.6749	0.5005	0.5175	0.6057	0.5924	0.602	0.5424	0.5805	0.6057	0.5924	0.6057	0.5028
	MAE	0.5147	0.5206	0.5021	0.5206	0.5193	0.5589	0.5575	**0.5662**	0.5348	0.5555	0.5589	0.5575	0.5589	0.5212
	MSE	0.412	0.4202	0.4025	0.494	0.4814	0.4936	0.4994	**0.5079**	0.4907	0.5044	0.4936	0.4994	0.4936	0.494
Market D (Labor)	ACC	**0.754**	0.753	0.7511	0.6227	0.6409	0.6995	0.6919	0.6911	0.6588	0.6771	0.6995	0.6919	0.6995	0.6265
	MAE	0.7623	0.7587	0.763	0.7557	0.7553	**0.7752**	0.7749	0.7731	0.7586	0.7725	0.7752	0.7749	0.7752	0.7561
	MSE	0.7285	0.7229	0.7322	**0.7759**	0.7721	0.7689	0.772	0.766	0.7408	0.7754	0.7689	0.772	0.7689	0.7755
Market E (Parts)	ACC	0.6081	**0.6099**	0.5927	0.5794	0.5791	0.6015	0.5989	0.6	0.6055	0.5956	0.6015	0.5989	0.6015	0.5794
	MAE	0.6042	0.6067	0.5881	0.6056	0.6045	0.612	**0.6124**	0.6105	0.6041	0.612	0.612	0.6124	0.612	0.6056
	MSE	0.5163	0.5181	0.508	0.5398	0.5388	0.532	0.5349	0.5304	0.5169	0.537	0.532	0.5349	0.532	**0.5399**
Market E (Labor)	ACC	0.6188	**0.6223**	0.6014	0.5908	0.5929	0.6141	0.6102	0.6135	0.6217	0.6056	0.6141	0.6102	0.6141	0.5908
	MAE	0.6183	0.6224	0.6006	0.6206	0.6211	0.6268	0.6265	**0.6275**	0.6231	0.6255	0.6268	0.6265	0.6268	0.6207
	MSE	0.5731	0.5776	0.5618	0.5991	0.5986	0.5909	0.5936	0.5928	0.5798	0.5952	0.5909	0.5936	0.5909	**0.5992**
Market F (Parts)	ACC	0.7364	0.7245	**0.747**	0.6468	0.6541	0.6972	0.6943	0.6956	0.67	0.6916	0.6972	0.6943	0.6972	0.6484
	MAE	0.6416	0.6351	0.6433	0.685	0.6827	0.6731	0.6767	0.6707	0.6188	0.6821	0.6731	0.6767	0.6731	**0.6852**
	MSE	0.5102	0.504	0.5138	**0.6331**	0.6237	0.5799	0.5882	0.5778	0.5007	0.6	0.5799	0.5882	0.5799	0.6324
Market F (Labor)	ACC	0.8192	0.8097	**0.8284**	0.7238	0.7247	0.7807	0.7783	0.7785	0.7738	0.7743	0.7807	0.7783	0.7807	0.7253
	MAE	0.6886	0.6826	0.6924	0.7136	0.7071	0.7127	0.7169	0.709	0.6737	**0.72**	0.7127	0.7169	0.7127	0.7141
	MSE	0.5694	0.5643	0.5713	**0.6553**	0.6432	0.6212	0.6294	0.6169	0.5632	0.6372	0.6212	0.6294	0.6212	0.6552
Market G (Parts)	ACC	**0.7319**	0.7194	0.7286	0.6533	0.6633	0.7113	0.7031	0.7013	0.6483	0.697	0.7113	0.7031	0.7113	0.6536
	MAE	0.5769	0.5704	0.5783	0.6637	**0.6688**	0.6628	0.6661	0.6463	0.5446	0.6655	0.6628	0.6661	0.6628	0.6637
	MSE	0.4605	0.4546	0.4644	0.6388	0.6344	0.5869	0.6032	0.5758	0.451	0.6115	0.5869	0.6032	0.5869	**0.6389**
Market G (Labor)	ACC	0.6665	0.6642	**0.6691**	0.6379	0.637	0.6568	0.6535	0.6518	0.6595	0.6496	0.6568	0.6535	0.6568	0.6376
	MAE	0.6771	0.6766	0.6756	0.6745	0.6739	0.6789	**0.6789**	0.6772	0.6752	0.6781	0.6789	0.6789	0.6789	0.6741
	MSE	0.6092	0.6094	0.606	0.6235	0.6229	0.6175	0.6202	0.6178	0.6095	0.6216	0.6175	0.6202	0.6175	**0.6236**

Table 6
Ranks of measures for MCR, MAE and MSE on goodwill assessment.

Rank	Measure	Avg. Rank	Avg. PRR
1	CONS_{C_2}	5.82 ± 1.92	0.6678 ± 0.0871
1	ORD_{VAR}	5.82 ± 1.92	0.6678 ± 0.0871
2	ORD_{ENT}	6.0 ± 3.1	0.6685 ± 0.0866
3	CONS_{C_1}	6.31 ± 2.19	0.6665 ± 0.0879
3	ORD_{MARG}	6.31 ± 2.19	0.6665 ± 0.0879
3	R_{l_1}	6.31 ± 2.19	0.6665 ± 0.0879
4	R_{l_2}	7.17 ± 5.1	0.66 ± 0.0885
5	CONS_{Cns}	7.49 ± 4.79	0.6605 ± 0.0872
6	VAR	7.54 ± 5.49	0.6595 ± 0.0888
7	CONS_{C_A}	7.68 ± 2.37	0.6645 ± 0.087
8	CONF	8.83 ± 4.98	0.6455 ± 0.1104
9	MARG	8.96 ± 4.58	0.6426 ± 0.1098
10	ENT	9.48 ± 4.95	0.6431 ± 0.1116
11	DFU	11.29 ± 2.9	0.6285 ± 0.1018

Table 7
Ranks of measures for MCR on goodwill assessment.

Rank	Measure	Avg. Rank	Avg. PRR
1	CONF	2.29 ± 2.3	0.7127 ± 0.0759
2	MARG	3.14 ± 2.38	0.7079 ± 0.0729
3	ENT	3.79 ± 4.35	0.7112 ± 0.0834
4	CONS_{C_1}	4.86 ± 0.86	0.6822 ± 0.0716
4	ORD_{MARG}	4.86 ± 0.86	0.6822 ± 0.0716
4	R_{l_1}	4.86 ± 0.86	0.6822 ± 0.0716
5	CONS_{C_A}	7.86 ± 1.1	0.6782 ± 0.0706
6	CONS_{C_2}	8.0 ± 0.85	0.677 ± 0.0739
6	ORD_{VAR}	8.0 ± 0.85	0.677 ± 0.0739
7	DFU	9.79 ± 3.96	0.6516 ± 0.0953
8	ORD_{ENT}	9.86 ± 0.86	0.6704 ± 0.0765
9	CONS_{Cns}	11.64 ± 2.41	0.643 ± 0.0718
10	R_{l_2}	12.57 ± 0.55	0.6342 ± 0.0774
11	VAR	13.5 ± 0.68	0.6327 ± 0.0781

Table 8
Ranks of measures for MAE and MSE on goodwill assessment.

Rank	Measure	Avg. Rank	Avg. PRR
1	ORD_{ENT}	4.07 ± 1.64	0.6675 ± 0.0925
2	R_{l_2}	4.46 ± 4.07	0.6729 ± 0.0921
3	VAR	4.55 ± 4.23	0.673 ± 0.0921
4	CONS_{C_2}	4.73 ± 1.26	0.6632 ± 0.0939
4	ORD_{VAR}	4.73 ± 1.26	0.6632 ± 0.0939
5	CONS_{Cns}	5.41 ± 4.31	0.6693 ± 0.094
6	CONS_{C_1}	7.04 ± 2.3	0.6586 ± 0.0953
6	ORD_{MARG}	7.04 ± 2.3	0.6586 ± 0.0953
6	R_{l_1}	7.04 ± 2.3	0.6586 ± 0.0953
7	CONS_{C_A}	7.59 ± 2.81	0.6577 ± 0.0945
8	MARG	11.88 ± 1.68	0.6099 ± 0.1115
9	DFU	12.04 ± 1.86	0.617 ± 0.1047
10	CONF	12.11 ± 1.31	0.6119 ± 0.1105
11	ENT	12.32 ± 1.72	0.609 ± 0.1093

Table 9
Exemplary rejection thresholds for market B using CONS_{C_2} or ORD_{VAR} (parts).

Rejection	ACC	MAE	MSE	QWK	UNDERPAYMENT	OVERPAYMENT	TOTAL	THRESHOLD
0%	0.821	9.352	686.444	0.645	-163,946.17	53,778.43	-110,167.74	1.0
10%	0.86	6.249	412.382	0.756	-38,432.59	113,479.52	75,046.93	0.594
20%	0.902	4.166	267.86	0.826	-16,730.41	80,686.62	63,956.21	0.293
30%	0.931	2.797	178.16	0.869	-7,942.01	50,347.16	42,405.15	0.142
40%	0.945	2.286	150.054	0.881	-5,925.01	38,091.96	32,166.95	0.071
50%	0.958	1.582	95.72	0.922	-2,450.01	12,347.42	9,897.41	0.033

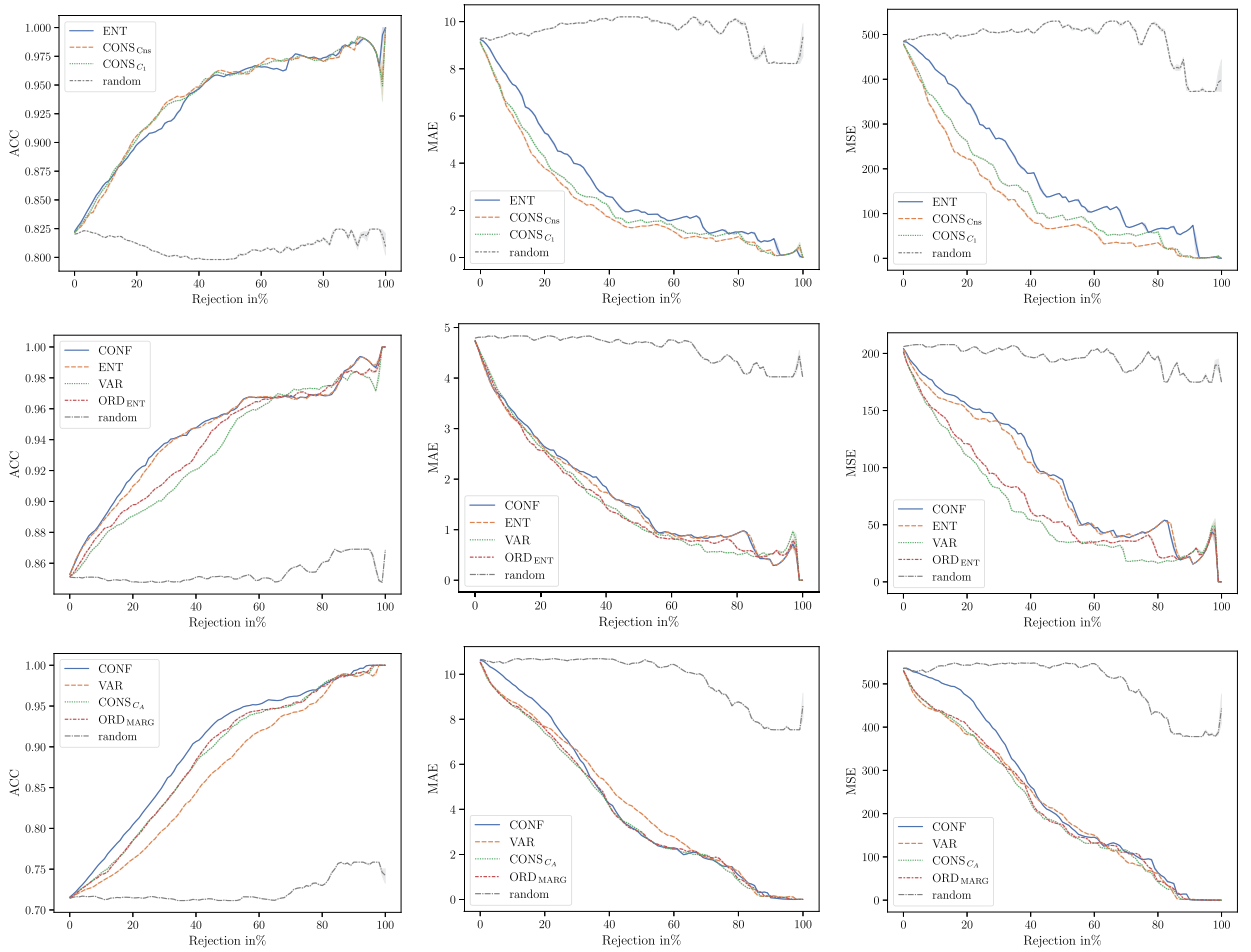


Fig. 11. Exemplary rejection curves for three of the goodwill claim assessment datasets displaying conventional uncertainty measures like entropy, margin and variance in comparison to consensus and ordinal binary decomposition-based measures.

Table 10
Exemplary rejection thresholds for market B using CONS_{C_2} or ORD_{VAR} (labor).

Rejection	ACC	MAE	MSE	QWK	UNDERPAYMENT	OVERPAYMENT	TOTAL	THRESHOLD
0%	0.886	7.092	588.147	0.585	-32,296.75	17,068.6	-15,228.15	1.0
10%	0.925	3.44	236.713	0.77	-8,989.2	15,985.3	6,996.1	0.538
20%	0.959	1.554	101.953	0.881	-3,871.6	6,668.28	2,796.68	0.203
30%	0.978	0.753	47.119	0.939	-1,741.0	3,080.66	1,339.66	0.055
40%	0.99	0.303	16.901	0.975	-1,371.0	743.0	-628.0	0.017
50%	0.992	0.272	17.51	0.969	-1,061.0	307.0	-754.0	0.006

With regard to the investigated uncertainty measures, we can draw the following conclusions from our evaluations on twenty-three ordinal benchmark datasets and a case study on seven automotive goodwill claim assessment datasets:

- Overall, when simultaneously looking at hit-rate and error distances (indicated by MCR, MAE and MSE), variance, the proposed ordinal binary decomposition method, and complementary dispersion measures of consensus measures outperform standard nominal classification uncertainty measures like entropy, margin and confidence when it comes to uncertainty quantification for probabilistic ordinal classification. This also supports our hypothesis that maximal uncertainty is expressed by a distinct bimodal distribution in ordinal classification.
- This is also the case when the predictive output probabilities are of unimodal nature, as indicated by low DFU measurements in our benchmark study. One might expect that distance may not be overly relevant in this case, and nominal classification measures should perform at least competitive to measures taking distance and the ordinal structure into account.

- When only looking at the distance of errors (indicated by MAE and MSE), the observation that dispersion measures, including variance and the binary decomposition method, outperform nominal measures is further enforced.
- Nominal classification uncertainty measures like entropy, margin, and confidence are competitive when it comes to misclassification rate and may outperform distance-based measures for multimodal outputs, as shown in our case study on automotive goodwill claim assessment.
- When it comes to preventing distance-based errors, measured by MAE and MSE, VAR and R_{l_2} perform very well. However, when it comes to reducing the misclassification rate, they are less effective.
- Complementary dispersion measures of consensus measures as well as the proposed ordinal binary decomposition method seem to strike a better balance between categorical classification accuracy (hit rate) and distance-based errors compared to standard nominal uncertainty measures and variance. Hence, they appear to best reflect this inherent trade-off of between accuracy and error distance in ordinal classification.
- In any case, an uncertainty measure in ordinal classification should consider error distance. If larger errors are supposed to be minimized, as indicated by MSE, VAR and R_{l_2} are most effective. If the exact hit-rate is equally important to error distance minimization, as indicated by MCR and MAE, the ordinal binary decomposition method, as well as complementary dispersion measures of consensus measures, strike a good balance. The usage of nominal uncertainty measures is only warranted in cases where the focus is solely on the exact hit-rate, as indicated by MCR, which is usually not the case in ordinal classification. According to our experiments, this guideline applies to datasets exhibiting unimodal as well as polarized prior class distributions, though the difference between nominal and dispersion measures is more pronounced for multimodal predictive distributions. Moreover, we recommend the usage of cross-entropy loss as a proper scoring rule over dedicated ordinal losses in ordinal classification to ensure unbiased uncertainty quantification.

An interesting direction for future work on the quantification of uncertainty in probabilistic ordinal classification is to separate total uncertainty into its aleatoric and epistemic parts [59], and to investigate whether this can be accomplished with the consensus measures presented in this paper or the ordinal binary decomposition method. This distinction is not possible on the basis of standard first-order probabilities as used in this work, however, and calls for more expressive representations (such as second-order distributions). Moreover, it might be interesting to evaluate further probabilistic base classifiers and datasets (e.g. image datasets) and study the effect of probability calibration [60] on the investigated uncertainty measures.

CRediT authorship contribution statement

Stefan Haas: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Eyke Hüllermeier:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stefan Haas reports a relationship with Bayerische Motoren Werke AG that includes: employment.

Appendix A. Proofs

Proof of Proposition 4.1. We prove that the measure C_1 satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5** of Section 4.1.

A1: Given the bimodal distribution $\boldsymbol{p} = (1/2, 0, \dots, 0, 1/2)$ on $\mathcal{O} = \{1, 2, \dots, K\}$, the cumulative probabilities will be $\boldsymbol{F} = (1/2, \dots, 1/2, 1)$. This minimizes the numerator of C_1 with $\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5| = \sum_{k=1}^{K-1} 0 = 0$. Thus, $C_1(\boldsymbol{p}) = \frac{0}{(K-1)/2} = 0$, which is the lower bound of the C_1 measure.

A2: Given a Dirac distribution of the form $\boldsymbol{p} = (0, \dots, 0, 1, 0, \dots, 0)$ on $\mathcal{O} = \{1, 2, \dots, K\}$, the cumulative probabilities will be $\boldsymbol{F} = (0, \dots, 0, 1, \dots, 1)$. This maximizes the numerator of C_1 with $\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5| = \sum_{k=1}^{K-1} 0.5 = \frac{1}{2}(K-1)$, because $|p - 0.5|$ is upper-bounded by $\frac{1}{2}$ for $0 \leq p \leq 1$. Thus, $C_1(\boldsymbol{p}) = \frac{(K-1)/2}{(K-1)/2} = 1$, which is the upper bound of the C_1 measure.

A3: This directly follows from **A1** and **A2**.

A4: This is satisfied as the individual components that make up C_1 are all continuous functions of $\boldsymbol{p}$.

A5: Given a probability distribution $\boldsymbol{p} = (p_1, p_2, \dots, p_K)$ and its reversal $\boldsymbol{p}_{\sigma_\leftrightarrow} = (p_K, p_{K-1}, \dots, p_1)$ on an ordinal scale $\mathcal{O} = \{1, 2, \dots, K\}$. To show that $C_1(\boldsymbol{p}) = C_1(\boldsymbol{p}_{\sigma_\leftrightarrow})$ one needs to show that

$$\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5| = \sum_{k=1}^{K-1} |F_k(\boldsymbol{p}_{\sigma_\leftrightarrow}) - 0.5|.$$

Given the following relationship $F_k(\boldsymbol{p}_{\sigma_\leftrightarrow}) = \sum_{j=1}^{k} p_{\sigma_\leftrightarrow(j)} = \sum_{j=1}^{k} p_{K-j+1} = 1 - \sum_{j=1}^{K-k} p_j = 1 - F_{K-k}(\boldsymbol{p})$, we have:

$$|F_k(\boldsymbol{p}_{\sigma_\leftrightarrow}) - 0.5| = |(1 - F_{K-k}(\boldsymbol{p})) - 0.5| = |F_{K-k}(\boldsymbol{p}) - 0.5|.$$

Next, given the commutative property of summation $\sum_{k=1}^{K-1} F_k(\boldsymbol{p}) = \sum_{k=1}^{K-1} F_{K-k}(\boldsymbol{p})$, with $F_{K-k}(\boldsymbol{p})$ being the cumulative probabilities of $\boldsymbol{p}$ in reversed order, we then have

$$\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}_{\sigma_\leftrightarrow}) - 0.5| = \sum_{k=1}^{K-1} |F_{K-k}(\boldsymbol{p}) - 0.5| = \sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5|.$$

From this we can conclude that $C_1(\boldsymbol{p}) = C_1(\boldsymbol{p}_{\sigma_\leftrightarrow})$. □

Proof of Proposition 4.2. We prove that the measure C_2 satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5** of Section 4.1. The proof is analogous to the proof of Proposition 4.1. □

Proof of Proposition 4.3. We prove that the measure Cns satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5** of Section 4.1.

Tastle and Wierman demonstrate that their Cns measure produces a single value ranging from 0 for complete disagreement to 1 for complete agreement. This essentially validates axioms **A1**-**A3** [18]. Therefore, we will focus on axioms **A4** and **A5** in this discussion.

A4: For the logarithm to be defined, its argument must be strictly positive, i.e.

$$0 < 1 - \frac{|k-\mu|}{K-1}.$$

Since k ranges between 1 and K, and μ lies in the interval $[1, K]$, $|k-\mu|$ will always be $\leq K-1$. The only case where the argument could be 0 is $k = K$ and $\mu = 1$. However, if $\mu = 1$, then $p_1 = 1$ and $p_2 = \ldots = p_K = 0$, so that the sum in (6) reduces to the first summand, which evaluates to 0 (by definition), so that $\text{Cns}(\boldsymbol{p}) = 1$.

Since $\log_2(x)$ is continuous for $x > 0$ and $\lim_{x \downarrow 0} x \cdot \log_2(x) = 0$, and the rest of the terms in (6) are all continuous functions of $\boldsymbol{p}$, we can conclude that Cns is a continuous function of $\boldsymbol{p}$.

A5: Given a probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ and its reversal $\boldsymbol{p}_{\sigma_\leftrightarrow} = (p_K, p_{K-1}, \ldots, p_1)$ on an ordinal scale $\mathcal{O} = \{1, 2, \ldots, K\}$. One needs to show that $\text{Cns}(\boldsymbol{p}) = \text{Cns}(\boldsymbol{p}_{\sigma_\leftrightarrow})$. Given the relationship

$$\begin{aligned}
\mu_{\sigma_\leftrightarrow} &= \sum_{k=1}^{K} k \cdot p_{K-k+1} = \sum_{k=1}^{K} (K-k+1) \cdot p_k \\
&= (K+1) \sum_{k=1}^{K} p_k - \sum_{k=1}^{K} p_k \cdot k \\
&= (K+1) - \mu
\end{aligned}$$

between the expected values $\mu_{\sigma_\leftrightarrow}$ of $\boldsymbol{p}_{\sigma_\leftrightarrow}$ and μ of $\boldsymbol{p}$ respectively, as well as the commutative property of summation $\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} p_{K-k+1}$, we have

$$\begin{aligned}
\text{Cns}(\boldsymbol{p}_{\sigma_\leftrightarrow}) &= 1 + \sum_{k=1}^{K} p_{\sigma_\leftrightarrow(k)} \log_2\left(1 - \frac{|k - \mu_{\sigma_\leftrightarrow}|}{K-1}\right) \\
&= 1 + \sum_{k=1}^{K} p_{K-k+1} \log_2\left(1 - \frac{|k - ((K+1)-\mu)|}{K-1}\right) \\
&= 1 + \sum_{k=1}^{K} p_{K-k+1} \log_2\left(1 - \frac{|(K-k+1)-\mu|}{K-1}\right) \\
&= 1 + \sum_{k=1}^{K} p_k \log_2\left(1 - \frac{|k-\mu|}{K-1}\right) \\
&= \text{Cns}(\boldsymbol{p}).
\end{aligned}$$

Hence, the Cns measure is invariant against reversal of the ordinal scale. □

Proof of Proposition 4.4. We prove that the measure C_A satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5** of Section 4.1.

A1: The extreme bimodal distribution will minimize the term U with the maximum possible number of unimodality violations for triples $|TDU(S)| = K-2$. Given this and $|S_1| = 2$, we have

$$A(\boldsymbol{p}) = \left(1 - \frac{1}{K-1}\right) \cdot \left(\frac{-(K-1)\cdot(K-2)}{(K-2)^2}\right) \tag{A.1}$$

$$= \left(\frac{(K-2)}{(K-1)}\right) \cdot \left(\frac{-(K-1)\cdot(K-2)}{(K-2)^2}\right)$$
$$= -\frac{(K-2)^2}{(K-2)^2} = -1.$$

We omit the term w here, which is $w = |S_1| \cdot 0.5 = 2 \cdot 0.5 = 1$. Following this, we can conclude that A is minimized by the extreme bimodal distribution with the lower bound -1. In turn, C_A will normalize A to have the lower bound 0.

A2: Since a Dirac distribution will maximize each term of A (7), with $w = 1$, $V = \left(1 - \frac{|S_1|-1}{K-1}\right) = \left(1 - \frac{1-1}{K-1}\right) = 1$, and $U = 1$ by definition, we can conclude that A (7) as well as C_A (8) are maximized by a Dirac distribution with the upper bound 1.

A3: A uniform distribution will lead to $w = |S_1| \cdot 1/K = K \cdot 1/K = 1$, $V = \left(1 - \frac{|S_1|-1}{K-1}\right) = \left(1 - \frac{K-1}{K-1}\right) = 0$, and $U = 1$ by definition. Hence, A will be 0 for the uniform distribution and 0.5 for the normalized version C_A.

A4: The measure A is a finite sum of products of continuous functions. Since the sum and product of continuous functions are also continuous, A and C_A are continuous.

A5: Given the commutative property of addition and multiplication, A is invariant against reversal of the ordinal scale when this property holds for all its terms (w, V, and U).

- The weight term w is invariant against reversal of the ordinal scale, as it is calculated based on the difference between adjacent sorted probabilities $\left(p_{(k)} - p_{(k-1)}\right)$ and the number of categories being equal to or greater than the probability p_k ($|S_k|$). Hence, this term is even invariant to any permutation of the probabilities.
- This also applies to the term $V = \left(1 - \frac{|S_k|-1}{K-1}\right)$ as it will not be affected by any permutation.
- The term U depends on the counting of rank triples $|TDU(S)|$ and $|TU(S)|$. Since triples are invariant against reversal of the ordinal scale, U is also invariant against reversal of the ordinal scale.

Since each term (w, V, and U) is invariant against reversal of the ordinal scale, we can conclude that A and C_A are also invariant against reversal of the ordinal scale. □

Proof of Proposition 4.5. Under the assumption of a single mode m, we prove that the measure DFU satisfies Axioms **A4** and **A5**, but violates Axioms **A1**, **A2**, and **A3** of Section 4.1. Notably, the measure DFU would need to be scaled to lie within the range $[0, 1]$, and Axioms **A1** and **A2** are violated in their inverted form.

A1: This axiom is violated as the extreme bimodal distribution is not the only distribution leading to the upper bound of 0.5 for DFU. For example,

$$\mathrm{DFU}\left(\left(\frac{1}{2}, 0, \ldots, 0, \frac{1}{2}\right)\right) = \mathrm{DFU}\left(\left(\frac{1}{2}, 0, \frac{1}{2}, 0, \ldots, 0\right)\right) = 0.5.$$

A2: This axiom is violated as DFU does not distinguish between unimodal distributions and their degree of "peakedness." For example,

$$\mathrm{DFU}((0, \ldots, 0.2, 0.6, 0.2, \ldots, 0)) = \mathrm{DFU}((0, \ldots, 0, 1, 0, \ldots, 0)) = 0.$$

Hence, Dirac distributions are not the only distributions that lead to the lower bound of 0 for DFU.

A3: This is violated, since the uniform distribution, as a unimodal distribution, leads to the same lower bound of 0 for DFU as the Dirac distribution:

$$\mathrm{DFU}\left(\left(\frac{1}{K}, \ldots, \frac{1}{K}\right)\right) = \mathrm{DFU}((0, \ldots, 0, 1, 0, \ldots, 0)) = 0.$$

A4: Since each d_k is continuous and the maximum of a finite set of continuous functions is also continuous, we can conclude that DFU is continuous.

A5: Given a probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ and its reversal $\boldsymbol{p}_{\sigma_\leftrightarrow} = (p_K, p_{K-1}, \ldots, p_1)$ on an ordinal scale $\mathcal{O} = \{1, 2, \ldots, K\}$. One needs to show that $\mathrm{DFU}(\boldsymbol{p}) = \mathrm{DFU}(\boldsymbol{p}_{\sigma_\leftrightarrow})$ by demonstrating that the calculated distances d_k and $d_{\sigma_\leftrightarrow}(k)$ are the same, with

$$d_{\sigma_\leftrightarrow}(k) = \begin{cases} p_{\sigma_\leftrightarrow}(k) - p_{\sigma_\leftrightarrow}(k+1) = p_{K-k+1} - p_{K-k} & \text{if} \quad 1 \le k < m \\ 0 & \text{if} \quad k = m \\ p_{\sigma_\leftrightarrow}(k) - p_{\sigma_\leftrightarrow}(k-1) = p_{K-k+1} - p_{K-k+2} & \text{if} \quad m < k \le K \end{cases}. \tag{A.2}$$

Since $d_{\sigma_\leftrightarrow}(k) = p_{K-k+1} - p_{K-k}$ and $d_k = p_k - p_{k-1}$ are the same pairwise distances in reversed order, just like $d_{\sigma_\leftrightarrow}(k) = p_{K-k+1} - p_{K-k+2}$ and $d_k = p_k - p_{k+1}$, we can conclude that the measured pairwise distances of d_k and $d_{\sigma_\leftrightarrow}(k)$ are the same (in reversed order). Due to the fact that the max operator on a set of distances is invariant to any permutations, we can further conclude that DFU is invariant against reversal of the ordinal scale with $\mathrm{DFU}(\boldsymbol{p}) = \mathrm{DFU}(\boldsymbol{p}_{\sigma_\leftrightarrow})$. Please note that this only holds for the existence of a single mode m. In the case of multiple modes, where the leftmost mode is taken as the mode m, this axiom may be violated. □

Proof of Proposition 4.6. We prove that the measure u_{VAR} satisfies axioms **A1**, **A2**, **A3**, **A4**, and **A5** of Section 4.1. Note that Axioms **A1** and **A2** are satisfied in their inverted form, and $u_{,\mathrm{VAR}}$ would need to be scaled to lie within the range $[0, 1]$.

A1: Popoviciu's inequality on variances provides an upper bound for the variance of any bounded probability distribution. Specifically, if an ordinal variable takes values in the interval $[1, K]$, then the variance satisfies:

$$u_{\text{VAR}} \leq \frac{1}{4}(K-1)^2.$$

Equality holds if and only if the distribution is bimodal with half of the probability mass at each of the extreme values 1 and K. Hence, the extreme bimodal distribution $\boldsymbol{p} = (1/2, 0, \ldots, 0, 1/2)$ exclusively maximizes u_{VAR}.

A2: For a Dirac distribution $\boldsymbol{p} = (0, \ldots, 0, 1, 0, \ldots, 0)$ with $p_j = 1$ for some $j \in \{1, \ldots, K\}$ and $p_k = 0$ for all $k \neq j$. The expected value of the distribution is $\mu = \sum_{k=1}^{K} p_k \cdot k = (0 \cdot k) + \ldots + (0 \cdot k) + (1 \cdot j) + (0 \cdot k) + \ldots + (0 \cdot k) = j$. Substituting $\mu = j$ into the variance formula, we get: $u_{\text{VAR}}(\boldsymbol{p}) = \sum_{k=1}^{K} p_k \cdot (k-\mu)^2 = 0 \cdot (k-j)^2 + \ldots + 0 \cdot (k-j)^2 + 1 \cdot (j-j)^2 + 0 \cdot (k-j)^2 + \ldots + 0 \cdot (k-j)^2 = 0$. Since the variance u_{VAR} is zero for a Dirac distribution, and variance is non-negative, this is the minimum possible value. Therefore, u_{VAR} is exclusively minimized by a Dirac distribution.

A3: This directly follows from **A1** and **A2**.

A4: This trivially holds true.

A5: Given the relationship $\mu_{\sigma_\leftrightarrow} = \sum_{k=1}^{K}(K-k+1) \cdot p_k = \sum_{k=1}^{K} K \cdot p_k + p_k - \sum_{k=1}^{K} p_k \cdot k = (K+1)\sum_{k=1}^{K} p_k - \sum_{k=1}^{K} p_k \cdot k = (K+1) - \mu$ between the expected values $\mu_{\sigma_\leftrightarrow}$ of $\boldsymbol{p}_{\sigma_\leftrightarrow}$ and μ of $\boldsymbol{p}$ respectively, as well as the commutative property of summation, we have:

$$\begin{aligned} u_{\text{VAR}}(\boldsymbol{p}_{\sigma_\leftrightarrow}) &= \sum_{k=1}^{K} p_{\sigma_\leftrightarrow}(k) \cdot (k - \mu_{\sigma_\leftrightarrow})^2 \\ &= \sum_{k=1}^{K} p_{(K-k+1)} \cdot (k - ((K+1) - \mu))^2 \\ &= \sum_{k=1}^{K} p_{(K-k+1)} \cdot ((K-k+1) - \mu)^2 \\ &= \sum_{k=1}^{K} p_k \cdot (k-\mu)^2 \\ &= u_{\text{VAR}}(\boldsymbol{p}). \end{aligned} \tag{A.3}$$

Hence, u_{VAR} is invariant against reversal of the ordinal scale. □

Proof of Lemma 5.1. Given the bimodal distribution $\boldsymbol{p} = (1/2, 0, \ldots, 0, 1/2)$ on $\mathcal{Y} = \{y_1, \ldots, y_k\}$, each binary reduction in (10) is of the form $\boldsymbol{p}_{\text{BIN}} = (1/2, 1/2)$. Likewise, given a Dirac distribution $\boldsymbol{p} = (0, \ldots, 0, 1, 0, \ldots, 0)$, each binary reduction is of the form $\boldsymbol{p}_{\text{BIN}} = (0, 1)$ or $\boldsymbol{p}_{\text{BIN}} = (1, 0)$. □

Proof of Lemma 5.2. Assuming symmetry for the generator u_{BIN}, with $u_{\text{BIN}}(p_1, p_2) = u_{\text{BIN}}(p_2, p_1)$ for $\boldsymbol{p} = (p_1, p_2)$ and given the commutative property of addition, the following holds:

$$\begin{aligned} u_{\text{ORD}}(\boldsymbol{p}_{\sigma_\leftrightarrow}) &= \sum_{k=1}^{K-1} u_{\text{BIN}}\left(\sum_{i=1}^{k} p_{\sigma_\leftrightarrow}(i), \sum_{j=k+1}^{K} p_{\sigma_\leftrightarrow}(j)\right) \\ &= \sum_{k=1}^{K-1} u_{\text{BIN}}\left(\sum_{i=1}^{k} p_{K-i+1}, \sum_{j=k+1}^{K} p_{K-j+1}\right) \\ &= \sum_{k=1}^{K-1} u_{\text{BIN}}\left(\sum_{i=K-k+1}^{K} p_i, \sum_{j=1}^{K-k} p_j\right) \\ &= \sum_{k=1}^{K-1} u_{\text{BIN}}\left(\sum_{i=1}^{k} p_i, \sum_{j=k+1}^{K} p_j\right) \\ &= u_{\text{ORD}}(\boldsymbol{p}) \quad \square \end{aligned} \tag{A.4}$$

Proof of Proposition 5.1. The fact that u_{ORD} satisfies axioms **A1**, **A2**, and **A3** directly follows from Lemma 5.1 (though **A1** and **A2** are satisfied in inverted non-normalized form, which in turn makes u_{ORD} directly applicable to uncertainty quantification). Additionally, axiom **A5** follows from Lemma 5.2. Given that the generator u_{BIN} is continuous, we can also conclude that u_{ORD} is continuous, since a finite sum of continuous functions is also continuous, which satisfies axiom **A4**. □

Proof of Proposition 5.2. The proof starts by defining the normalized version of the binary decomposition method with margin as the generator and shows the equivalence to the complementary dispersion measure D_1 by simplifying the expression step-by-step.

The key step is to recognize that the margin generator leads to the absolute difference between cumulative probabilities and their complement, which directly relates to the C_1 measure:

$$\begin{aligned}
D_1(\boldsymbol{p}) &= \frac{1}{(K-1)} \sum_{k=1}^{K-1} u_{\text{MARG}}\left(\sum_{i=1}^{k} p_i, \sum_{j=k+1}^{K} p_j\right) \\
&= \frac{1}{(K-1)} \sum_{k=1}^{K-1} 1 - \left|\sum_{i=1}^{k} p_i - \sum_{j=k+1}^{K} p_j\right| \\
&= 1 - \frac{\sum_{k=1}^{K-1} |\sum_{i=1}^{k} p_i - \sum_{j=k+1}^{K} p_j|}{(K-1)} \\
&= 1 - \frac{\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - (1 - F_k(\boldsymbol{p}))|}{(K-1)} \\
&= 1 - \frac{\sum_{k=1}^{K-1} |2F_k(\boldsymbol{p}) - 1|/2}{(K-1)/2} \\
&= 1 - \frac{\sum_{k=1}^{K-1} |F_k(\boldsymbol{p}) - 0.5|}{(K-1)/2} \\
&= 1 - C_1(\boldsymbol{p}) \quad \square
\end{aligned} \tag{A.5}$$

Proof of Proposition 5.3. The proof begins by defining the normalized version of the binary decomposition method with variance as the generator and then demonstrates the equivalence to the complementary dispersion measure D_2 by simplifying the expression step-by-step. The key step is to recognize that the variance generator leads to the product of cumulative probabilities and their complements, which directly relates to the C_2 measure:

$$\begin{aligned}
D_2(\boldsymbol{p}) &= \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} u_{\text{VAR}}\left(\sum_{i=1}^{k} p_i, \sum_{j=k+1}^{K} p_j\right) \\
&= \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \left(\sum_{i=1}^{k} p_i \cdot \sum_{j=k+1}^{K} p_j\right) \\
&= \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} F_k(\boldsymbol{p})(1 - F_k(\boldsymbol{p})) \\
&= 1 - \left(1 - \frac{\sum_{k=1}^{K-1} F_k(\boldsymbol{p})(1 - F_k(\boldsymbol{p}))}{(K-1)/4}\right) \\
&= 1 - \left(1 + \frac{\sum_{k=1}^{K-1} F_k(\boldsymbol{p})(F_k(\boldsymbol{p}) - 1)}{(K-1)/4}\right) \\
&= 1 - \frac{\sum_{k=1}^{K-1} F_k(\boldsymbol{p})(F_k(\boldsymbol{p}) - 1) + 0.25}{(K-1)/4} \\
&= 1 - \frac{\sum_{k=1}^{K-1} F_k(\boldsymbol{p})^2 - F_k(\boldsymbol{p}) + 0.25}{(K-1)/4} \\
&= 1 - \frac{\sum_{k=1}^{K-1} (F_k(\boldsymbol{p}) - 0.5)^2}{(K-1)/4} \\
&= 1 - C_2(\boldsymbol{p}) \quad \square
\end{aligned} \tag{A.6}$$

Appendix B. Prediction rejection ratios (PRRs) with multi-layer perceptron (MLP) as base learner

In this section, we present additional experimental results using a multi-layer perceptron (MLP) [61] with CE loss as the base learner instead of GBTs (cf. Section 6). Refer to Table B.11 for the parameters of the feed-forward network. Additionally, in addition to one-hot (0/1) encoding categorical features and integer encoding the labels, all features were also standardized.

The obtained ranks for the different uncertainty measures based on the measured PRR values resemble those of GBTs, with measures taking distance into account significantly surpassing common nominal measures on these tabular ordinal benchmark datasets, as visible in the CD diagrams in Fig. B.12 and the detailed results in Table B.12.

Table B.11
MLP parameters [61].

Parameter	Value
Hidden Layer Sizes	[128, 64]
Activation Function	ReLU
Solver	Adam
Maximum Epochs	200
Batch Size	200
L2 Regularization (alpha)	1e-04
Learning Rate	1e-03

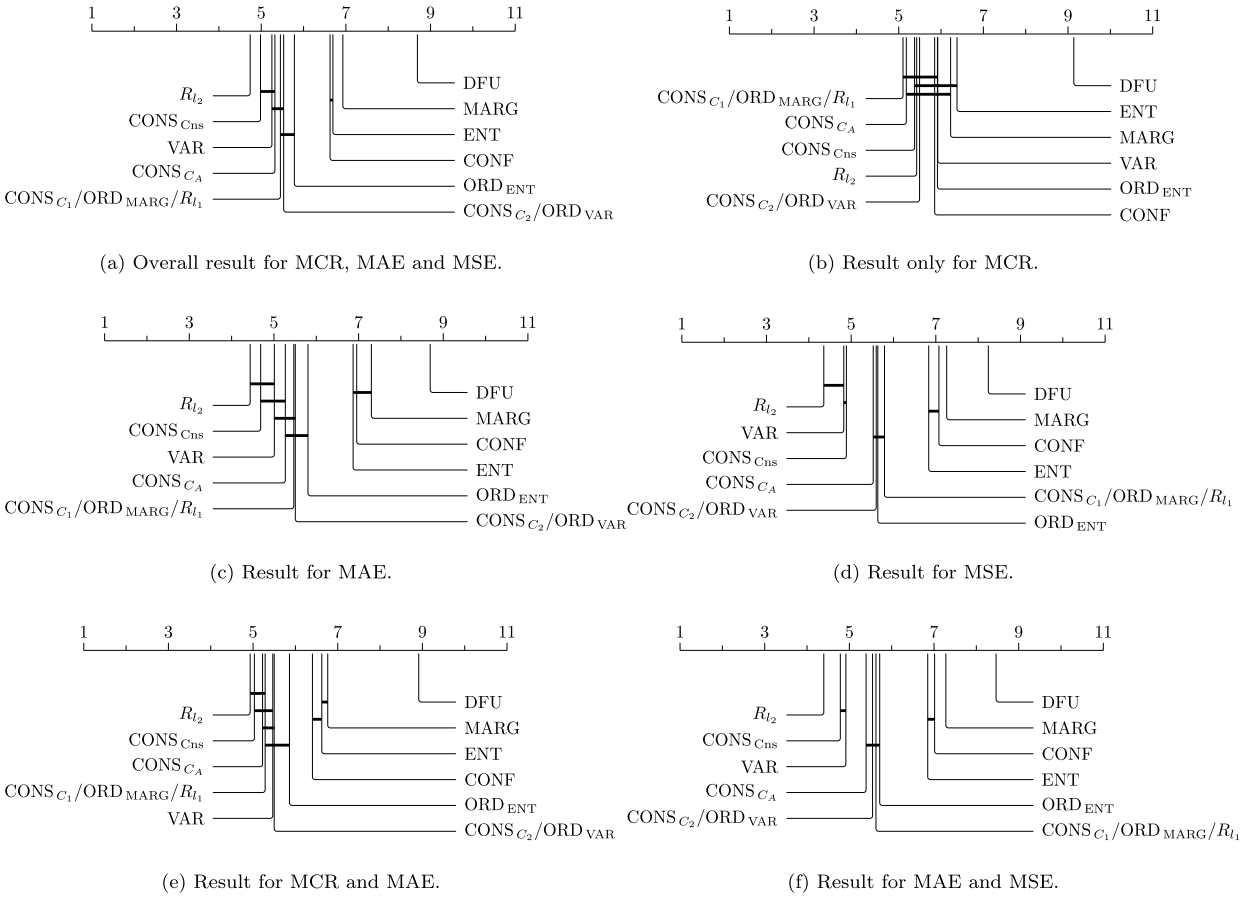


Fig. B.12. Critical difference (CD) diagrams for the evaluated uncertainty measures over all performance metrics and datasets based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon test with an MLP as the base learner. Groups of uncertainty measures that are not significantly different (at $p = 0.05$) are connected [53,54].

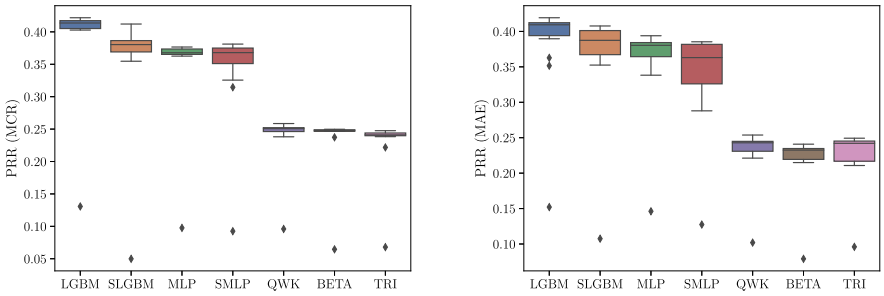
Appendix C. Comparison of prediction rejection ratios (PRRs) for different predictors

In this section, we want to evaluate the influence of the base learner on uncertainty quantification in ordinal classification. To do this, we compare the PRR values obtained for various predictors on the tabular ordinal benchmark datasets over all uncertainty measures. Keep in mind that the PRR is independent of the predictive performance of the predictor and solely assesses the quality of the uncertainty quantification [51]. We compare the following diverse set of predictors: LightGBM with CE loss (LGBM) [32], *A Simple Approach to Ordinal Classification* [24] with LGBM and CE loss as binary base learner (SLGBM), MLP with CE loss (MLP) [61], *A Simple Approach to Ordinal Classification* [24] with MLP and CE loss as binary base learner (SMLP), MLP with QWK loss (QWK) [33,62,63], MLP with ordinal soft labeling based on triangular distributions (TRI) [62–64], and MLP with ordinal soft labeling based on the beta distribution (BETA) [62,63,65]. The listed predictors cover a broad range of ordinal methods we want to compare to the standard CE loss as a proper scoring rule.

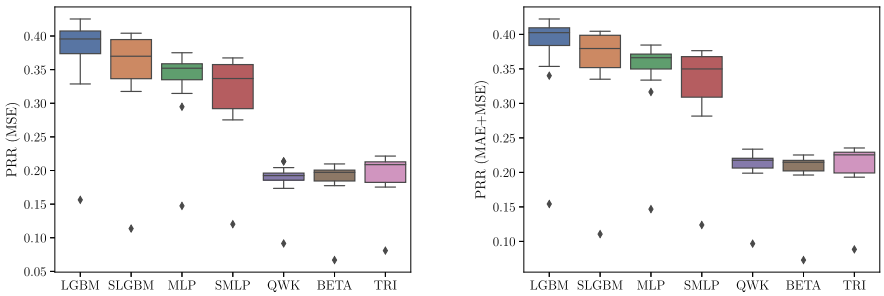
To allow for a fair comparison of the different neural network-based predictors, we chose the same configurations as in Appendix B for the MLP, SMLP, QWK, BETA, and TRI predictors (cf. Table B.11). Since our primary interest is in uncertainty quantification, and not predictive performance, we deliberately do not perform any further hyperparameter tuning.

Table B.12
PRRs for the different uncertainty measures and ordinal benchmark datasets using 10-fold cross-validation with an MLP as the base learner.

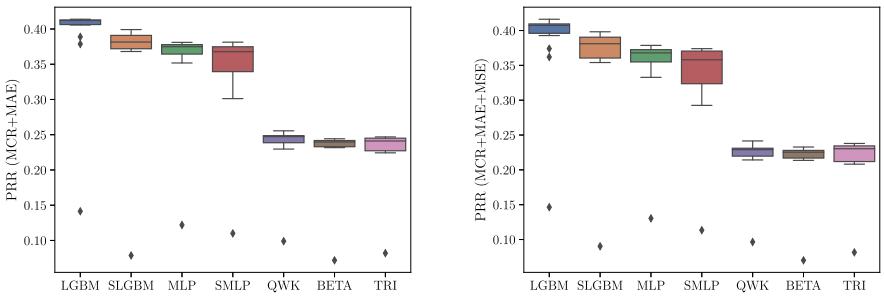
Dataset	Metric	CONF	MARG	ENT	VAR	CONS_{Cns}	CONS_{C_1}	CONS_{C_2}	CONS_{C_4}	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{I_1}	R_{I_2}
Triazines	ACC	0.0721 ± 0.2426	0.041 ± 0.2415	0.0953 ± 0.2401	0.0868 ± 0.2936	0.0877 ± 0.2796	0.0854 ± 0.2672	**0.0959** ± 0.2816	0.0952 ± 0.2779	0.0736 ± 0.278	0.0938 ± 0.2822	0.0854 ± 0.2672	0.0959 ± 0.2816	0.0854 ± 0.2672	0.0884 ± 0.2803
	MAE	0.1833 ± 0.1919	0.1504 ± 0.1925	0.2145 ± 0.1663	0.2388 ± 0.1931	0.238 ± 0.1938	0.2184 ± 0.1993	0.2317 ± 0.2079	0.234 ± 0.2007	**0.2681** ± 0.2501	0.2317 ± 0.2066	0.2184 ± 0.1993	0.2317 ± 0.2079	0.2184 ± 0.1993	0.2429 ± 0.1916
	MSE	0.2114 ± 0.1935	0.1877 ± 0.1898	0.2484 ± 0.1808	0.2775 ± 0.2289	0.2757 ± 0.2146	0.2448 ± 0.205	0.2633 ± 0.2147	0.2625 ± 0.2114	**0.3055** ± 0.2256	0.2668 ± 0.218	0.2448 ± 0.205	0.2633 ± 0.2147	0.2448 ± 0.205	0.2761 ± 0.2301
Machine CPU	ACC	0.6171 ± 0.19	0.5714 ± 0.1991	0.6573 ± 0.185	**0.7228** ± 0.1246	0.6786 ± 0.1568	0.6829 ± 0.1624	0.6926 ± 0.1732	0.6958 ± 0.1677	0.6628 ± 0.1603	0.6905 ± 0.1614	0.6829 ± 0.1624	0.6926 ± 0.1732	0.6829 ± 0.1624	0.7079 ± 0.1307
	MAE	0.5758 ± 0.2401	0.5025 ± 0.2575	0.6554 ± 0.1853	**0.727** ± 0.1673	0.6945 ± 0.1841	0.6816 ± 0.1894	0.7023 ± 0.1854	0.6926 ± 0.1854	0.7058 ± 0.1062	0.7028 ± 0.1829	0.6816 ± 0.1894	0.7023 ± 0.1854	0.6816 ± 0.1894	0.722 ± 0.1745
	MSE	0.5313 ± 0.3124	0.4457 ± 0.3216	0.601 ± 0.203	0.7024 ± 0.1965	0.6861 ± 0.1958	0.6707 ± 0.2066	0.6818 ± 0.2018	0.6782 ± 0.1973	0.7041 ± 0.1034	0.6804 ± 0.2025	0.6707 ± 0.2066	0.6818 ± 0.2018	0.6707 ± 0.2066	**0.7051** ± 0.2007
Auto MPG	ACC	0.3533 ± 0.1097	0.3429 ± 0.0983	0.3501 ± 0.1542	0.379 ± 0.1612	0.3779 ± 0.1143	0.3847 ± 0.1068	0.3695 ± 0.1411	0.3759 ± 0.1327	0.1373 ± 0.2025	0.3665 ± 0.156	0.3847 ± 0.1068	0.3695 ± 0.1411	0.3847 ± 0.1068	**0.3874** ± 0.1283
	MAE	0.3419 ± 0.1405	0.3004 ± 0.129	0.3697 ± 0.1409	0.4164 ± 0.1584	0.413 ± 0.158	0.4086 ± 0.137	0.4081 ± 0.1456	0.4179 ± 0.1334	0.2236 ± 0.172	0.4028 ± 0.1464	0.4086 ± 0.137	0.4081 ± 0.1456	0.4086 ± 0.137	**0.4205** ± 0.1612
	MSE	0.3459 ± 0.1433	0.2847 ± 0.1365	0.3703 ± 0.1653	0.4122 ± 0.2452	0.4119 ± 0.2525	0.405 ± 0.2056	0.4079 ± 0.2318	**0.4221** ± 0.2119	0.2215 ± 0.2276	0.3971 ± 0.2248	0.405 ± 0.2056	0.4079 ± 0.2318	0.405 ± 0.2056	0.4123 ± 0.2541
Pyrimidines	ACC	0.0514 ± 0.3327	**0.0549** ± 0.2403	-0.0738 ± 0.341	-0.0941 ± 0.4573	-0.0652 ± 0.4214	-0.0254 ± 0.3109	-0.0883 ± 0.3331	-0.0018 ± 0.3369	-0.0649 ± 0.2524	-0.0741 ± 0.3557	-0.0254 ± 0.3109	-0.0883 ± 0.3331	-0.0254 ± 0.3109	-0.1133 ± 0.437
	MAE	0.2048 ± 0.3698	0.1852 ± 0.3425	0.1806 ± 0.4419	0.2266 ± 0.5205	0.2178 ± 0.4683	0.201 ± 0.4205	0.1724 ± 0.4429	**0.2422** ± 0.4212	0.2082 ± 0.2977	0.2157 ± 0.4665	0.201 ± 0.4205	0.1724 ± 0.4429	0.201 ± 0.4205	0.2063 ± 0.5018
	MSE	0.2272 ± 0.4368	0.2051 ± 0.4218	0.2463 ± 0.4471	0.3193 ± 0.4848	0.3118 ± 0.4018	0.2788 ± 0.358	0.2691 ± 0.3797	**0.32** ± 0.3752	0.2847 ± 0.3739	0.3093 ± 0.4146	0.2788 ± 0.358	0.2691 ± 0.3797	0.2788 ± 0.358	0.2961 ± 0.4655
Abalone	ACC	0.3465 ± 0.0652	0.3126 ± 0.0704	0.3285 ± 0.0615	0.3189 ± 0.054	0.3437 ± 0.0625	**0.3528** ± 0.0639	0.335 ± 0.0618	0.345 ± 0.0612	0.0225 ± 0.0667	0.3213 ± 0.0564	0.3528 ± 0.0639	0.335 ± 0.0618	0.3528 ± 0.0639	0.3387 ± 0.0547
	MAE	0.3674 ± 0.0522	0.2913 ± 0.0572	0.3936 ± 0.0461	0.3967 ± 0.0377	0.3998 ± 0.0441	0.398 ± 0.0448	0.4008 ± 0.0431	0.4034 ± 0.0425	0.0819 ± 0.0834	0.3963 ± 0.0401	0.398 ± 0.0448	0.4008 ± 0.0431	0.398 ± 0.0448	**0.406** ± 0.037
	MSE	0.4004 ± 0.072	0.2756 ± 0.06	0.4719 ± 0.0692	0.4913 ± 0.0619	0.4715 ± 0.0652	0.4561 ± 0.0669	0.4815 ± 0.0661	0.4772 ± 0.0688	0.1573 ± 0.0858	0.4871 ± 0.065	0.4561 ± 0.0669	0.4815 ± 0.0661	0.4561 ± 0.0669	**0.4916** ± 0.0581
Boston Housing	ACC	0.4113 ± 0.142	0.4189 ± 0.1396	0.4113 ± 0.1409	0.4367 ± 0.13	**0.4483** ± 0.1321	0.4315 ± 0.1295	0.431 ± 0.1275	0.4315 ± 0.1252	-0.1139 ± 0.2308	0.427 ± 0.1293	0.4315 ± 0.1295	0.431 ± 0.1275	0.4315 ± 0.1295	0.4461 ± 0.1289
	MAE	0.3897 ± 0.173	0.3949 ± 0.1696	0.3986 ± 0.1749	0.4435 ± 0.1656	0.4467 ± 0.1616	0.428 ± 0.1621	0.4296 ± 0.1612	0.4343 ± 0.1584	-0.0644 ± 0.2905	0.4282 ± 0.1668	0.428 ± 0.1621	0.4296 ± 0.1612	0.428 ± 0.1621	**0.4489** ± 0.1616
	MSE	0.3188 ± 0.2236	0.3241 ± 0.2253	0.3343 ± 0.2331	**0.408** ± 0.2107	0.3989 ± 0.2064	0.376 ± 0.2099	0.3823 ± 0.214	0.3917 ± 0.203	-0.0197 ± 0.3391	0.3847 ± 0.2187	0.376 ± 0.2099	0.3823 ± 0.214	0.376 ± 0.2099	0.4078 ± 0.2025
Stocks Domain	ACC	0.7104 ± 0.053	**0.712** ± 0.0521	0.7064 ± 0.0556	0.7045 ± 0.0532	0.7098 ± 0.0527	0.7101 ± 0.0529	0.7084 ± 0.0537	0.7088 ± 0.0532	0.0667 ± 0.1669	0.7052 ± 0.0544	0.7101 ± 0.0529	0.7084 ± 0.0537	0.7101 ± 0.0529	0.7068 ± 0.0526
	MAE	0.7053 ± 0.0621	**0.707** ± 0.0615	0.7013 ± 0.064	0.6994 ± 0.062	0.7048 ± 0.0618	0.705 ± 0.062	0.7034 ± 0.0627	0.7038 ± 0.0622	0.0772 ± 0.1834	0.7001 ± 0.0629	0.705 ± 0.062	0.7034 ± 0.0627	0.705 ± 0.062	0.7017 ± 0.0615
	MSE	0.7091 ± 0.0685	**0.7107** ± 0.0679	0.7051 ± 0.0704	0.7034 ± 0.0679	0.7086 ± 0.0682	0.7088 ± 0.0684	0.7072 ± 0.0691	0.7076 ± 0.0686	0.0855 ± 0.1749	0.704 ± 0.0692	0.7088 ± 0.0684	0.7072 ± 0.0691	0.7088 ± 0.0684	0.7056 ± 0.0679
Wisconsin Breast Cancer	ACC	0.2284 ± 0.2257	**0.2549** ± 0.2251	0.2476 ± 0.2515	0.254 ± 0.2291	0.2267 ± 0.2278	0.2302 ± 0.2322	0.2487 ± 0.2331	0.2511 ± 0.229	0.1021 ± 0.3648	0.2544 ± 0.2374	0.2302 ± 0.2322	0.2487 ± 0.2331	0.2302 ± 0.2322	0.2538 ± 0.2289
	MAE	0.1399 ± 0.2296	0.1505 ± 0.246	0.172 ± 0.2467	0.1946 ± 0.2395	0.1767 ± 0.2442	0.1674 ± 0.2337	0.1724 ± 0.2446	0.1721 ± 0.2422	0.0559 ± 0.179	0.1751 ± 0.2368	0.1674 ± 0.2337	0.1724 ± 0.2446	0.1674 ± 0.2337	**0.2016** ± 0.2371
	MSE	0.0228 ± 0.205	0.036 ± 0.2232	0.035 ± 0.2139	0.06 ± 0.1771	0.0353 ± 0.1877	0.0209 ± 0.1798	0.0328 ± 0.1921	0.0363 ± 0.1797	-0.0146 ± 0.1863	0.0339 ± 0.1872	0.0209 ± 0.1798	0.0328 ± 0.1921	0.0209 ± 0.1798	**0.0705** ± 0.1775
Obesity	ACC	0.6996 ± 0.1197	0.698 ± 0.1203	0.7004 ± 0.1201	0.7153 ± 0.1169	0.7098 ± 0.119	0.7074 ± 0.1187	0.7077 ± 0.1192	0.7086 ± 0.1189	0.3128 ± 0.0759	0.7091 ± 0.1189	0.7074 ± 0.1187	0.7077 ± 0.1192	0.7074 ± 0.1187	**0.7159** ± 0.1167
	MAE	0.6877 ± 0.1297	0.6861 ± 0.1301	0.6889 ± 0.1301	**0.709** ± 0.1246	0.7009 ± 0.1278	0.698 ± 0.1273	0.6986 ± 0.1281	0.7005 ± 0.1265	0.3406 ± 0.087	0.7003 ± 0.1279	0.698 ± 0.1273	0.6986 ± 0.1281	0.698 ± 0.1273	0.7089 ± 0.1237
	MSE	0.6509 ± 0.1803	0.6491 ± 0.1798	0.6524 ± 0.1807	**0.681** ± 0.1699	0.6683 ± 0.1758	0.6644 ± 0.1753	0.6655 ± 0.1763	0.6689 ± 0.1719	0.3695 ± 0.101	0.668 ± 0.1762	0.6644 ± 0.1753	0.6655 ± 0.1763	0.6644 ± 0.1753	0.6799 ± 0.1681
CMC	ACC	**0.3125** ± 0.0578	0.3075 ± 0.0594	0.3115 ± 0.0543	0.262 ± 0.0796	0.2523 ± 0.0808	0.2958 ± 0.0755	0.2962 ± 0.0787	0.2786 ± 0.0802	0.0065 ± 0.0625	0.2935 ± 0.0778	0.2958 ± 0.0755	0.2962 ± 0.0787	0.2958 ± 0.0755	0.247 ± 0.076
	MAE	0.1692 ± 0.0588	0.1823 ± 0.0587	0.1524 ± 0.0582	0.2856 ± 0.0732	**0.2938** ± 0.0728	0.2546 ± 0.0736	0.2608 ± 0.075	0.2647 ± 0.0743	0.0438 ± 0.1734	0.2614 ± 0.0727	0.2546 ± 0.0736	0.2607 ± 0.0749	0.2546 ± 0.0736	0.2921 ± 0.0774
	MSE	-0.0501 ± 0.0514	-0.0409 ± 0.0548	-0.0534 ± 0.0453	0.1019 ± 0.0932	**0.1219** ± 0.0901	0.0199 ± 0.0719	0.0328 ± 0.0772	0.0634 ± 0.0976	-0.0279 ± 0.1918	0.0378 ± 0.0784	0.0199 ± 0.0719	0.0328 ± 0.0772	0.0199 ± 0.0719	0.1215 ± 0.095
Grub Damage	ACC	0.1264 ± 0.2738	0.1306 ± 0.2719	0.1273 ± 0.2484	0.1327 ± 0.2264	0.1339 ± 0.2385	0.1372 ± 0.2516	0.133 ± 0.2534	0.1337 ± 0.2487	0.1091 ± 0.22	**0.1482** ± 0.2439	0.1372 ± 0.2516	0.133 ± 0.2534	0.1372 ± 0.2516	0.1351 ± 0.2253
	MAE	0.1719 ± 0.2461	0.1692 ± 0.2485	0.1861 ± 0.2196	0.2433 ± 0.1971	0.2411 ± 0.209	0.2157 ± 0.224	0.2259 ± 0.2219	0.2269 ± 0.2252	0.2375 ± 0.2462	**0.2466** ± 0.2182	0.2157 ± 0.224	0.2259 ± 0.2219	0.2157 ± 0.224	0.2389 ± 0.1984
	MSE	0.1743 ± 0.2098	0.1619 ± 0.2202	0.2294 ± 0.2574	0.2845 ± 0.2276	0.2798 ± 0.2236	0.2446 ± 0.2161	0.263 ± 0.2269	0.26 ± 0.2334	0.219 ± 0.2945	**0.2866** ± 0.2238	0.2446 ± 0.2161	0.263 ± 0.2269	0.2446 ± 0.2161	0.2736 ± 0.2305
New Thyroid	ACC	**0.9789** ± 0.0422	0.9789 ± 0.0422	0.9789 ± 0.0422	0.9342 ± 0.0607	0.9448 ± 0.0467	0.9789 ± 0.0422	0.9543 ± 0.0625	0.9543 ± 0.0625	0.099 ± 0.6567	0.9543 ± 0.0625	0.9789 ± 0.0422	0.9543 ± 0.0625	0.9789 ± 0.0422	0.9648 ± 0.0445
	MAE	0.9621 ± 0.0385	0.9621 ± 0.0385	0.9621 ± 0.0385	0.942 ± 0.0548	0.9558 ± 0.0393	**0.9793** ± 0.0415	0.9621 ± 0.054	0.9621 ± 0.054	0.2441 ± 0.6591	0.9621 ± 0.054	0.9793 ± 0.0415	0.9621 ± 0.054	0.9793 ± 0.0415	0.9759 ± 0.0301
	MSE	0.9877 ± 0.0245	0.9877 ± 0.0245	0.9877 ± 0.0245	0.9585 ± 0.0394	0.9585 ± 0.0394	**1.0** ± 0.0	0.9785 ± 0.0283	0.9785 ± 0.0283	0.2286 ± 0.6949	0.9785 ± 0.0283	1.0 ± 0.0	0.9785 ± 0.0283	1.0 ± 0.0	0.9785 ± 0.0283
Balance Scale	ACC	0.9794 ± 0.0227	0.9732 ± 0.0324	**0.9917** ± 0.0116	0.9917 ± 0.0116	0.9753 ± 0.0325	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9884 ± 0.0164	0.2731 ± 0.2291	0.9917 ± 0.0116	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9794 ± 0.0227	0.9794 ± 0.0227
	MAE	0.9794 ± 0.0227	0.9691 ± 0.0308	**0.9917** ± 0.0116	0.9917 ± 0.0116	0.9753 ± 0.0325	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9884 ± 0.0164	0.2783 ± 0.2339	0.9917 ± 0.0116	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9794 ± 0.0227	0.9794 ± 0.0227
	MSE	0.97 ± 0.0301	0.9584 ± 0.0341	**0.9866** ± 0.0173	0.9849 ± 0.0204	0.9654 ± 0.0378	0.97 ± 0.0301	0.9786 ± 0.0233	0.98 ± 0.0311	0.2127 ± 0.2086	0.9866 ± 0.0173	0.97 ± 0.0301	0.9786 ± 0.0233	0.97 ± 0.0301	0.97 ± 0.0301
Automobile	ACC	0.6087 ± 0.169	0.6077 ± 0.1778	0.6159 ± 0.1666	0.6461 ± 0.1815	**0.6588** ± 0.1802	0.6278 ± 0.1757	0.6325 ± 0.1764	0.6399 ± 0.189	0.1865 ± 0.2471	0.6442 ± 0.1863	0.6278 ± 0.1757	0.6325 ± 0.1764	0.6278 ± 0.1757	0.6411 ± 0.1811
	MAE	0.5782 ± 0.1452	0.5744 ± 0.1714	0.593 ± 0.1223	**0.6597** ± 0.1187	0.6561 ± 0.1286	0.617 ± 0.1351	0.6179 ± 0.1331	0.6414 ± 0.1419	0.2662 ± 0.2664	0.6351 ± 0.1404	0.617 ± 0.1351	0.6179 ± 0.1331	0.617 ± 0.1351	0.6551 ± 0.1172
	MSE	0.4706 ± 0.2127	0.466 ± 0.2488	0.4894 ± 0.1886	**0.5927** ± 0.1659	0.5713 ± 0.1828	0.5192 ± 0.1923	0.518 ± 0.1916	0.5562 ± 0.199	0.3207 ± 0.3228	0.5419 ± 0.1982	0.5192 ± 0.1923	0.518 ± 0.1916	0.5192 ± 0.1923	0.5923 ± 0.164
Eucalyptus	ACC	0.3781 ± 0.1218	0.3764 ± 0.1245	0.3764 ± 0.1173	0.3867 ± 0.1089	**0.3923** ± 0.1165	0.3863 ± 0.1159	0.3868 ± 0.1139	0.3844 ± 0.1112	0.0567 ± 0.0865	0.3859 ± 0.1117	0.3863 ± 0.1159	0.3868 ± 0.1139	0.3863 ± 0.1159	0.3894 ± 0.1104
	MAE	0.3684 ± 0.0981	0.3666 ± 0.1014	0.3686 ± 0.0905	0.387 ± 0.0828	**0.391** ± 0.091	0.3835 ± 0.0887	0.3842 ± 0.0886	0.382 ± 0.0865	0.0736 ± 0.0971	0.3843 ± 0.0853	0.3835 ± 0.0887	0.3842 ± 0.0886	0.3835 ± 0.0887	0.39 ± 0.0846
	MSE	0.2739 ± 0.1054	0.2747 ± 0.1089	0.2716 ± 0.0992	0.2976 ± 0.1026	0.2986 ± 0.1075	0.2887 ± 0.1006	0.29 ± 0.1019	0.2887 ± 0.104	0.0735 ± 0.1641	0.2917 ± 0.1012	0.2887 ± 0.1006	0.29 ± 0.1019	0.2887 ± 0.1006	**0.3006** ± 0.1046
TAE	ACC	0.3197 ± 0.2681	0.3155 ± 0.27	**0.3699** ± 0.3052	0.2785 ± 0.2782	0.2337 ± 0.2417	0.3422 ± 0.2758	0.3452 ± 0.2962	0.3234 ± 0.3136	0.3133 ± 0.2544	0.3564 ± 0.3155	0.3422 ± 0.2758	0.3452 ± 0.2962	0.3422 ± 0.2758	0.2573 ± 0.2612
	MAE	0.2438 ± 0.2996	0.2337 ± 0.2839	0.2808 ± 0.354	0.2726 ± 0.3876	0.249 ± 0.3392	0.2676 ± 0.369	0.292 ± 0.3907	0.2932 ± 0.4059	**0.3828** ± 0.1444	0.3023 ± 0.4138	0.2676 ± 0.369	0.292 ± 0.3907	0.2676 ± 0.369	0.2512 ± 0.3604
	MSE	0.1901 ± 0.334	0.1741 ± 0.3109	0.2691 ± 0.3972	0.2444 ± 0.3838	0.2108 ± 0.3279	0.2213 ± 0.3685	0.2483 ± 0.3975	0.2431 ± 0.4024	**0.391** ± 0.2021	0.262 ± 0.4331	0.2213 ± 0.3685	0.2483 ± 0.3975	0.2213 ± 0.3685	0.2111 ± 0.3477
Heart (CLE)	ACC	0.5042 ± 0.1372	0.4765 ± 0.1319	0.5423 ± 0.1514	**0.5516** ± 0.1179	0.5317 ± 0.1152	0.5205 ± 0.1377	0.5324 ± 0.1272	0.5333 ± 0.1196	0.2361 ± 0.2612	0.5478 ± 0.1196	0.5205 ± 0.1377	0.5324 ± 0.1272	0.5205 ± 0.1377	0.546 ± 0.1148
	MAE	0.4347 ± 0.1905	0.4092 ± 0.1871	0.4843 ± 0.2058	**0.5233** ± 0.1548	0.5016 ± 0.1648	0.4802 ± 0.1861	0.4981 ± 0.174	0.4963 ± 0.164	0.2681 ± 0.2767	0.5125 ± 0.1676	0.4802 ± 0.1861	0.4981 ± 0.174	0.4802 ± 0.1861	0.5108 ± 0.1607
	MSE	0.3049 ± 0.2016	0.2876 ± 0.2024	0.3412 ± 0.2095	**0.4115** ± 0.1716	0.38 ± 0.1712	0.3476 ± 0.1871	0.3711 ± 0.1771	0.373 ± 0.1709	0.2217 ± 0.2833	0.3891 ± 0.1731	0.3476 ± 0.1871	0.3711 ± 0.1771	0.3476 ± 0.1871	0.3917 ± 0.1703
SWD	ACC	0.2174 ± 0.1341	0.2035 ± 0.1389	0.1873 ± 0.1239	0.2008 ± 0.1225	0.2119 ± 0.1324	0.2214 ± 0.1411	0.2015 ± 0.1292	0.2091 ± 0.1325	0.0415 ± 0.1136	0.195 ± 0.1249	0.2212 ± 0.141	0.2016 ± 0.1292	**0.2217** ± 0.1412	0.2164 ± 0.1402
	MAE	0.1613 ± 0.0897	0.1616 ± 0.1101	0.1315 ± 0.0816	0.1562 ± 0.0793	0.1705 ± 0.0774	0.1764 ± 0.0918	0.1526 ± 0.0846	0.1544 ± 0.0916	0.0227 ± 0.0968	0.1458 ± 0.082	0.1763 ± 0.0915	0.1525 ± 0.0848	0.1764 ± 0.0918	**0.1841** ± 0.0968
	MSE	0.138 ± 0.1001	0.1384 ± 0.1007	0.1135 ± 0.1014	0.1427 ± 0.0936	0.1584 ± 0.0762	0.153 ± 0.089	0.1307 ± 0.0999	0.1253 ± 0.1041	0.0126 ± 0.1016	0.1272 ± 0.0969	0.1528 ± 0.0887	0.1306 ± 0.1	0.1529 ± 0.0891	**0.1749** ± 0.0957
ERA	ACC	0.1484 ± 0.0781	**0.1821** ± 0.1054	0.1121 ± 0.0708	0.0047 ± 0.1062	0.0403 ± 0.1209	0.0952 ± 0.118	0.071 ± 0.1183	0.1488 ± 0.0731	0.0504 ± 0.0738	0.0481 ± 0.1096	0.0958 ± 0.1172	0.0735 ± 0.1179	0.098 ± 0.1156	0.012 ± 0.113
	MAE	0.0219 ± 0.067	-0.0038 ± 0.0893	-0.0113 ± 0.0854	-0.009 ± 0.1556	0.0116 ± 0.1469	0.0083 ± 0.1243	-0.0087 ± 0.1333	0.0104 ± 0.0947	**0.0581** ± 0.1277	-0.0255 ± 0.1349	0.0086 ± 0.1243	-0.008 ± 0.1332	0.0107 ± 0.1228	-0.0021 ± 0.1593
	MSE	0.0228 ± 0.105	-0.0043 ± 0.0952	-0.0061 ± 0.1183	0.0341 ± 0.1995	0.033 ± 0.1767	0.0109 ± 0.1596	0.0102 ± 0.1652	0.0079 ± 0.1307	**0.0503** ± 0.147	0.0059 ± 0.1697	0.0108 ± 0.1598	0.0112 ± 0.1659	0.0135 ± 0.1591	0.0382 ± 0.197
ESL	ACC	0.2265 ± 0.2031	**0.2292** ± 0.2049	0.1954 ± 0.201	0.1928 ± 0.1872	0.2216 ± 0.1887	0.225 ± 0.2076	0.2092 ± 0.2057	0.2139 ± 0.2056	-0.0048 ± 0.2012	0.1885 ± 0.1915	0.225 ± 0.2076	0.2092 ± 0.2057	0.225 ± 0.2076	0.2206 ± 0.1947
	MAE	0.2114 ± 0.1969	**0.2177** ± 0.2004	0.1775 ± 0.2006	0.1707 ± 0.1858	0.2013 ± 0.1796	0.2084 ± 0.2015	0.1904 ± 0.2018	0.1956 ± 0.2042	-0.0446 ± 0.2235	0.1648 ± 0.1918	0.2084 ± 0.2015	0.1904 ± 0.2018	0.2084 ± 0.2015	0.1981 ± 0.1837
	MSE	0.185 ± 0.2661	**0.1878** ± 0.2696	0.1476 ± 0.2701	0.1347 ± 0.2503	0.166 ± 0.2407	0.1767 ± 0.2708	0.1584 ± 0.2668	0.1643 ± 0.2749	-0.081 ± 0.2366	0.1261 ± 0.2597	0.1767 ± 0.2708	0.1584 ± 0.2668	0.1767 ± 0.2708	0.1582 ± 0.2513
LEV	ACC	0.1576 ± 0.0858	0.1371 ± 0.099	0.1584 ± 0.0962	0.1709 ± 0.1113	0.1746 ± 0.0791	0.1735 ± 0.0821	0.1724 ± 0.0888	0.1692 ± 0.0769	-0.0128 ± 0.118	0.1661 ± 0.1051	0.1735 ± 0.0821	0.1724 ± 0.0888	0.1734 ± 0.0822	**0.1812** ± 0.0897
	MAE	0.1523 ± 0.0756	0.1271 ± 0.0828	0.1669 ± 0.1055	**0.1795** ± 0.1115	0.1684 ± 0.0867	0.1716 ± 0.081	0.1785 ± 0.0988	0.1753 ± 0.0792	0.0202 ± 0.1325	0.1757 ± 0.1132	0.1716 ± 0.0808	0.1785 ± 0.0989	0.1716 ± 0.0807	0.1788 ± 0.0948
	MSE	0.1662 ± 0.0855	0.1304 ± 0.0807	0.1946 ± 0.132	**0.2056** ± 0.1313	0.1778 ± 0.1046	0.1914 ± 0.0939	0.2033 ± 0.1257	0.2047 ± 0.1066	0.0296 ± 0.1797	0.2051 ± 0.136	0.191 ± 0.0936	0.2033 ± 0.1259	0.191 ± 0.0936	0.1943 ± 0.105
Red Wine	ACC	0.2949 ± 0.0705	0.2941 ± 0.0744	0.2807 ± 0.0583	0.2831 ± 0.053	**0.3107** ± 0.0682	0.3036 ± 0.0657	0.296 ± 0.0606	0.2913 ± 0.0652	-0.025 ± 0.0925	0.2788 ± 0.0575	0.3036 ± 0.0657	0.296 ± 0.0606	0.3036 ± 0.0657	0.3038 ± 0.0582
	MAE	0.2456 ± 0.0787	0.2418 ± 0.0808	0.2458 ± 0.0718	0.2611 ± 0.0745	**0.2849** ± 0.0845	0.2673 ± 0.0815	0.2679 ± 0.0803	0.2535 ± 0.0818	-0.043 ± 0.0956	0.2527 ± 0.0751	0.2673 ± 0.0815	0.2679 ± 0.0803	0.2673 ± 0.0815	0.2797 ± 0.0782
	MSE	0.2058 ± 0.106	0.1998 ± 0.0991	0.2168 ± 0.1211	0.2462 ± 0.1319	**0.268** ± 0.1345	0.2398 ± 0.12	0.2464 ± 0.1341	0.222 ± 0.1211	-0.065 ± 0.0991	0.2315 ± 0.1322	0.2398 ± 0.12	0.2464 ± 0.1341	0.2398 ± 0.12	0.2666 ± 0.1306
White Wine	ACC	0.2471 ± 0.029	0.2377 ± 0.0318	0.2371 ± 0.0392	0.2402 ± 0.0404	0.2479 ± 0.0315	**0.2547** ± 0.0304	0.2457 ± 0.0374	0.2502 ± 0.0351	0.054 ± 0.0578	0.2396 ± 0.0409	0.2547 ± 0.0304	0.2457 ± 0.0374	0.2547 ± 0.0304	0.2489 ± 0.0354
	MAE	0.2233 ± 0.0486	0.2076 ± 0.0495	0.2295 ± 0.0552	0.2443 ± 0.0444	0.248 ± 0.0394	0.2478 ± 0.0427	0.2436 ± 0.0455	0.2451 ± 0.0449	0.0553 ± 0.0534	0.2406 ± 0.0472	0.2478 ± 0.0427	0.2436 ± 0.0455	0.2478 ± 0.0427	**0.2539** ± 0.0421
	MSE	0.1913 ± 0.0902	0.1695 ± 0.0873	0.2175 ± 0.0973	0.2519 ± 0.0725	0.2478 ± 0.0707	0.2355 ± 0.0775	0.2414 ± 0.0788	0.2389 ± 0.0811	0.0567 ± 0.0689	0.2443 ± 0.0782	0.2355 ± 0.0775	0.2414 ± 0.0788	0.2355 ± 0.0775	**0.2586** ± 0.0717



(a) PRRs for MCR. (b) PRRs for MAE.



(c) PRRs for MSE. (d) PRRs for MAE and MSE combined.



(e) PRRs for MCR and MAE combined. (f) PRRs for MCR, MAE and MSE combined.

Fig. C.13. PRR values obtained over all tabular ordinal benchmark datasets and uncertainty measures grouped by underlying base learner.

Fig. C.13 shows the PRR values obtained over all datasets and uncertainty measures for the different predictors, depicted by different performance measures (MCR, MAE, and MSE). In general, LGBM is able to obtain the highest PRR values, which is no surprise as GBTs are known to outperform neural networks on tabular datasets and are able to better deal with this modality. Furthermore, one can clearly see that the usage of CE loss is beneficial when it comes to uncertainty quantification over the simple ordinal approach in terms of uncertainty quantification, as manifested in higher PRR values (LGBM vs. SLGBM and MLP vs. SMLP), though the simple ordinal approach improves predictive performance (cf. Table C.13). Moreover, specific ordinal losses like QWK and the unimodal soft labeling approaches (BETA and TRI) lead to substantially smaller PRR values overall, and in particular for MSE, as they tend to bias predictive probabilities towards unimodality [33]. This loss of information appears to negatively affect uncertainty quantification and justifies our usage of the cross-entropy loss as a proper scoring rule over dedicated ordinal losses for the purpose of uncertainty quantification in ordinal classification.

Table C.13 displays the average results of the different predictors over all datasets in terms of predictive performance (ACC, 1-OFF, MAE, MSE, and QWK) as well as calibration (negative log-likelihood (NLL), Brier Score (BS), and expected calibration error (ECE)). In summary, LGBM and SLGBM generally perform well across most metrics. They exhibit the best accuracy, calibration, and reasonable error rates. SLGBM improves on distance-based errors (MAE, MSE, and QWK) compared to LGBM but worsens calibration in terms of NLL. MLP and SMLP show competitive accuracy and QWK, though having higher NLL and slightly worse calibration compared to LGBM and SLGBM. SMLP improves on distance-based errors (MAE, MSE, and QWK) compared to MLP at the cost of calibration (NLL, BS, and ECE). QWK has good QWK but lower accuracy and higher error rates compared to other models. BETA and TRI generally perform worse across most metrics, but still show some competitive aspects in specific areas. In general, ordinal methods exhibit

Table C.13
Average performance and calibration of the different predictors over the tabular ordinal benchmark datasets.

Predictor	ACC (↑)	1-OFF (↑)	MAE (↓)	MSE (↓)	QWK (↑)	NLL (↓)	BS (↓)	ECE (↓)
LGBM	**0.627** ±0.196	0.898 ±0.114	0.526 ±0.378	0.961 ±1.031	0.673 ±0.246	**1.145** ±0.591	0.520 ±0.244	0.071 ±0.047
SLGBM	0.625 ±0.198	**0.906** ±0.108	**0.506** ±0.352	**0.851** ±0.848	**0.689** ±0.233	1.693 ±1.145	**0.517** ±0.238	**0.069** ±0.044
MLP	0.620 ±0.197	0.895 ±0.116	0.529 ±0.363	0.948 ±0.942	0.664 ±0.262	1.419 ±1.039	0.552 ±0.295	0.081 ±0.072
SMLP	0.621 ±0.201	0.901 ±0.115	0.513 ±0.354	0.877 ±0.857	0.681 ±0.247	2.281 ±2.009	0.564 ±0.308	0.085 ±0.072
QWK	0.578 ±0.189	0.891 ±0.115	0.584 ±0.360	1.062 ±0.989	0.682 ±0.222	1.745 ±0.849	0.647 ±0.266	0.103 ±0.053
BETA	0.611 ±0.192	0.892 ±0.115	0.549 ±0.365	1.028 ±1.022	0.636 ±0.252	1.892 ±1.168	0.598 ±0.279	0.094 ±0.047
TRI	0.596 ±0.192	0.886 ±0.113	0.573 ±0.370	1.078 ±1.041	0.613 ±0.256	2.229 ±1.305	0.646 ±0.285	0.107 ±0.053

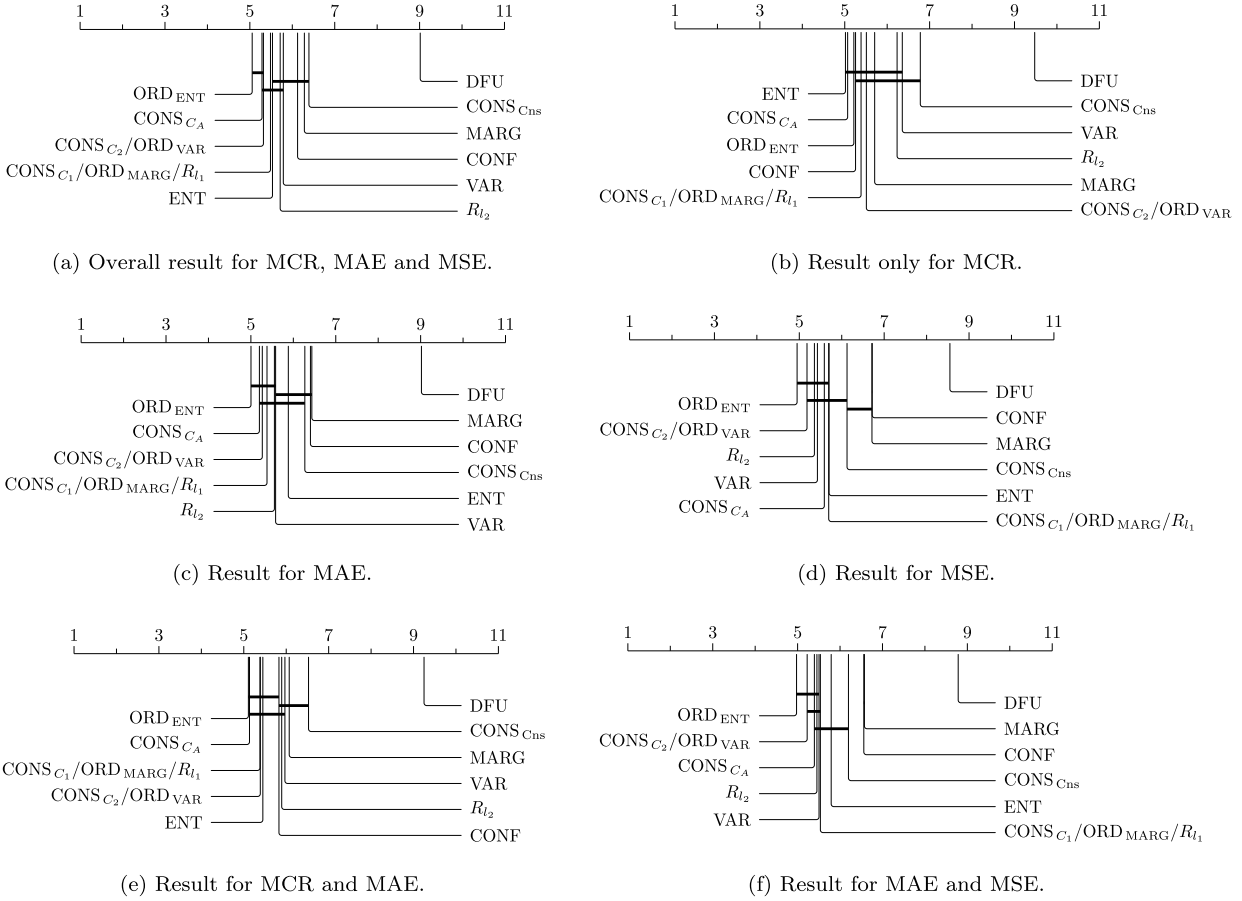


Fig. D.14. Critical difference (CD) diagrams for the evaluated uncertainty measures over all performance metrics and datasets based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon test with the base learner *A Simple Approach to Ordinal Classification* and LightGBM as the binary base learner [24]. Groups of uncertainty measures that are not significantly different (at $p = 0.05$) are connected [53,54].

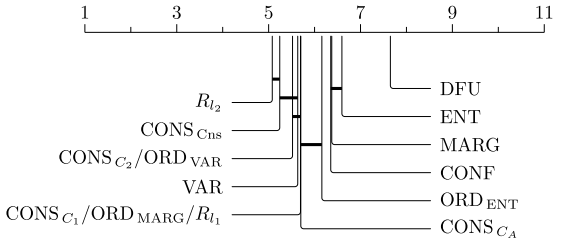
larger calibration issues in relation to cross-entropy loss, as indicated by higher NLL, BS, and ECE values. This appears to negatively impact uncertainty quantification in ordinal classification and in turn leads to smaller PRR values.

Appendix D. Prediction rejection ratios (PRRs) with a simple approach to ordinal classification as base learner

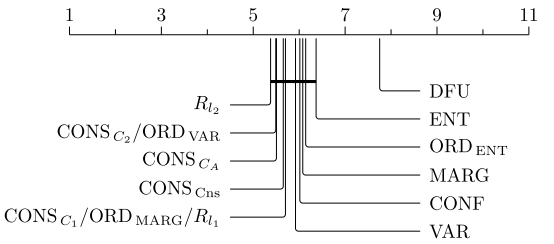
In this section, we present additional experimental results using *A Simple Approach to Ordinal Classification* with LightGBM as a binary base learner (SLGBM) [24] instead of LightGBM with CE loss (cf. Section 6). As shown in Appendix C, the simple approach to ordinal classification leads to increased predictive performance at the cost of worsened uncertainty quantification, indicated by smaller PRR values compared to LightGBM with CE loss. This is also visible when looking at the CD diagrams in Fig. D.14. The results are not as significant as for GBTs and MLPs with CE loss (cf. Section 6 and Appendix B), as the ordinal approach leads to biased predictive probabilities in which predictive probability distributions are squashed (cf. Appendix C). Nonetheless, the superiority of certain measures depending on the performance metric is still visible, though there is more overlap than when using CE loss and the measures become more interchangeable. When the goal is to decrease distance-based errors, the ordinal binary decomposition method, VAR, R_{l_2}, and complementary dispersion measures of consensus measures still outperform nominal measures in most cases

Table D.14
PRRs for the different uncertainty measures and ordinal benchmark datasets using 10-fold cross-validation with the base learner *A Simple Approach to Ordinal Classification* and LightGBM as the binary base learner [24].

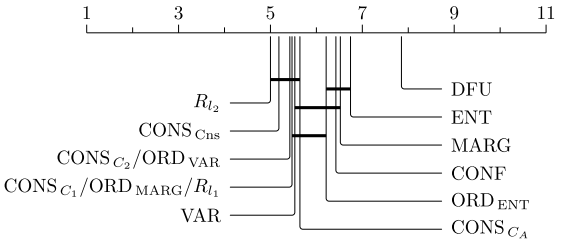
Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Cos}$	$CONS_{C_1}$	$CONS_{C_2}$	$CONS_{C_3}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Triazines	ACC	0.3444 ± 0.331	0.3432 ± 0.2928	**0.4018** ± 0.2813	0.3489 ± 0.2471	0.3972 ± 0.2174	0.3413 ± 0.3261	0.3515 ± 0.2948	0.3794 ± 0.226	0.0071 ± 0.1513	0.3722 ± 0.222	0.3827 ± 0.2743	0.3787 ± 0.2265	0.3927 ± 0.2242	0.3666 ± 0.2238
	MAE	0.2563 ± 0.2416	0.2331 ± 0.2046	0.338 ± 0.2525	0.3636 ± 0.2006	**0.385** ± 0.1921	0.3291 ± 0.2917	0.3061 ± 0.2722	0.3606 ± 0.1772	0.1165 ± 0.1875	0.3836 ± 0.1768	0.3439 ± 0.2303	0.375 ± 0.1944	0.3718 ± 0.1805	0.3725 ± 0.1905
	MSE	0.1884 ± 0.3759	0.2177 ± 0.3509	0.2509 ± 0.3581	0.3817 ± 0.2666	**0.3943** ± 0.269	0.2173 ± 0.4127	0.2091 ± 0.4334	0.3482 ± 0.2922	0.25 ± 0.2676	0.3921 ± 0.2593	0.2622 ± 0.3958	0.3815 ± 0.273	0.3566 ± 0.277	0.3893 ± 0.2654
Machine CPU	ACC	0.639 ± 0.1914	0.6399 ± 0.1738	0.7071 ± 0.1781	**0.7668** ± 0.1362	0.7043 ± 0.2021	0.4518 ± 0.3364	0.405 ± 0.3554	0.7139 ± 0.1786	0.5077 ± 0.1273	0.7572 ± 0.1356	0.5269 ± 0.284	0.7363 ± 0.1564	0.7215 ± 0.1755	0.7492 ± 0.1563
	MAE	0.5524 ± 0.1557	0.5478 ± 0.1504	0.638 ± 0.1581	**0.7296** ± 0.1249	0.6555 ± 0.1846	0.3762 ± 0.2585	0.3241 ± 0.2626	0.6438 ± 0.1872	0.5864 ± 0.0946	0.712 ± 0.1301	0.4369 ± 0.2371	0.6877 ± 0.1429	0.6557 ± 0.1574	0.7108 ± 0.1365
	MSE	0.4621 ± 0.188	0.4875 ± 0.1616	0.588 ± 0.189	**0.7014** ± 0.1268	0.6304 ± 0.1652	0.1857 ± 0.3343	0.1563 ± 0.3141	0.578 ± 0.1907	0.6551 ± 0.0884	0.6731 ± 0.1501	0.2328 ± 0.3625	0.6519 ± 0.1475	0.6052 ± 0.1618	0.6743 ± 0.1362
Auto MPG	ACC	0.2725 ± 0.1863	0.2954 ± 0.1675	0.3395 ± 0.1484	**0.3793** ± 0.1211	0.3474 ± 0.1428	0.2528 ± 0.1956	0.2793 ± 0.1521	0.2882 ± 0.1779	0.1172 ± 0.118	0.3696 ± 0.1206	0.2513 ± 0.1999	0.363 ± 0.1335	0.3571 ± 0.1452	0.3779 ± 0.1396
	MAE	0.2812 ± 0.2049	0.3234 ± 0.1718	0.3775 ± 0.1528	**0.4342** ± 0.1442	0.3979 ± 0.171	0.2372 ± 0.2128	0.2323 ± 0.1697	0.3178 ± 0.2029	0.2215 ± 0.1111	0.4262 ± 0.1329	0.2432 ± 0.2321	0.4168 ± 0.1491	0.4054 ± 0.1641	0.4313 ± 0.1597
	MSE	0.2725 ± 0.229	0.3232 ± 0.1706	0.4008 ± 0.1546	**0.46** ± 0.1529	0.416 ± 0.1991	0.2423 ± 0.2716	0.3194 ± 0.2078	0.4141 ± 0.1884	0.2652 ± 0.2259	0.4523 ± 0.1428	0.2615 ± 0.2658	0.4343 ± 0.1704	0.4191 ± 0.1803	0.4566 ± 0.1639
Pyrimidines	ACC	0.0804 ± 0.6058	0.0748 ± 0.6133	0.1257 ± 0.5547	-0.2077 ± 0.5077	-0.2619 ± 0.5521	0.1448 ± 0.4364	**0.176** ± 0.3584	-0.1413 ± 0.4168	-0.258 ± 0.533	-0.2053 ± 0.5285	-0.0906 ± 0.4701	-0.2053 ± 0.5285	-0.2619 ± 0.5521	-0.2404 ± 0.5333
	MAE	-0.0195 ± 0.3923	0.0843 ± 0.4628	0.113 ± 0.4605	0.2075 ± 0.3835	0.1812 ± 0.4435	0.3186 ± 0.2779	**0.3416** ± 0.2707	0.1706 ± 0.3623	0.2656 ± 0.388	0.1742 ± 0.42	0.1407 ± 0.3829	0.1799 ± 0.4235	0.1768 ± 0.4376	0.1953 ± 0.3967
	MSE	-0.1639 ± 0.461	-0.0718 ± 0.5038	-0.1121 ± 0.5742	0.3513 ± 0.3546	0.3279 ± 0.3972	0.2575 ± 0.2952	0.3234 ± 0.3886	0.2676 ± 0.3339	**0.4493** ± 0.303	0.3009 ± 0.366	0.2457 ± 0.3709	0.3086 ± 0.3714	0.3375 ± 0.3868	0.3427 ± 0.3709
Abalone	ACC	0.2914 ± 0.0384	0.2771 ± 0.0403	0.3219 ± 0.0394	0.3282 ± 0.04	0.3323 ± 0.0426	0.2845 ± 0.0427	0.2778 ± 0.0386	0.3158 ± 0.035	0.0792 ± 0.0651	0.3315 ± 0.0389	0.2929 ± 0.0392	0.3346 ± 0.0368	0.3349 ± 0.0399	**0.3371** ± 0.0436
	MAE	0.2803 ± 0.0539	0.2583 ± 0.0608	0.3568 ± 0.0487	0.3784 ± 0.0488	0.3692 ± 0.0588	0.2894 ± 0.0557	0.2953 ± 0.0504	0.3391 ± 0.0475	0.1111 ± 0.0617	0.3812 ± 0.0478	0.2976 ± 0.0521	0.3769 ± 0.0483	0.3626 ± 0.0551	**0.382** ± 0.0564
	MSE	0.289 ± 0.09	0.2579 ± 0.1074	0.3902 ± 0.0628	0.4107 ± 0.0615	0.3793 ± 0.0822	0.2953 ± 0.0818	0.3282 ± 0.0668	0.3714 ± 0.0749	0.1062 ± 0.0876	**0.4148** ± 0.0579	0.3115 ± 0.08	0.3991 ± 0.0641	0.3635 ± 0.0787	0.4006 ± 0.0721
Boston Housing	ACC	0.4176 ± 0.0668	0.4223 ± 0.0794	0.4316 ± 0.0912	0.4253 ± 0.1106	0.4211 ± 0.1179	0.4173 ± 0.0849	0.4286 ± 0.1013	0.4254 ± 0.098	-0.0082 ± 0.1467	**0.432** ± 0.1082	0.411 ± 0.0706	0.4257 ± 0.1027	0.4252 ± 0.1027	0.4242 ± 0.1104
	MAE	0.4034 ± 0.0855	0.4301 ± 0.0788	0.4452 ± 0.089	0.4486 ± 0.1249	0.4439 ± 0.131	0.4129 ± 0.0798	0.4382 ± 0.0966	0.443 ± 0.1008	0.0422 ± 0.1563	**0.4521** ± 0.1139	0.4022 ± 0.0789	0.445 ± 0.1079	0.4444 ± 0.108	0.4437 ± 0.1188
	MSE	0.3636 ± 0.1957	0.4266 ± 0.0727	0.4482 ± 0.0926	**0.4612** ± 0.1576	0.4571 ± 0.1625	0.3929 ± 0.077	0.4141 ± 0.0794	0.4489 ± 0.1154	0.0791 ± 0.2128	0.461 ± 0.1354	0.3959 ± 0.0974	0.4528 ± 0.128	0.4523 ± 0.1281	0.4523 ± 0.1392
Stocks Domain	ACC	**0.7065** ± 0.0664	0.7064 ± 0.0672	0.7059 ± 0.0665	0.6947 ± 0.07	0.6747 ± 0.0847	0.7051 ± 0.0667	0.7044 ± 0.0679	0.7015 ± 0.0702	-0.0067 ± 0.1894	0.7029 ± 0.0691	0.7053 ± 0.0661	0.7029 ± 0.0688	0.7029 ± 0.0688	0.6945 ± 0.0701
	MAE	**0.7065** ± 0.0664	0.7064 ± 0.0672	0.7059 ± 0.0665	0.6947 ± 0.07	0.6747 ± 0.0847	0.7051 ± 0.0667	0.7044 ± 0.0679	0.7015 ± 0.0702	-0.0067 ± 0.1894	0.7029 ± 0.0691	0.7053 ± 0.0661	0.7029 ± 0.0688	0.7029 ± 0.0688	0.6945 ± 0.0701
	MSE	**0.7065** ± 0.0664	0.7064 ± 0.0672	0.7059 ± 0.0665	0.6947 ± 0.07	0.6747 ± 0.0847	0.7051 ± 0.0667	0.7044 ± 0.0679	0.7015 ± 0.0702	-0.0067 ± 0.1894	0.7029 ± 0.0691	0.7053 ± 0.0661	0.7029 ± 0.0688	0.7029 ± 0.0688	0.6945 ± 0.0701
Wisconsin Breast Cancer	ACC	**0.1233** ± 0.2423	-0.0096 ± 0.247	0.1123 ± 0.2854	-0.0967 ± 0.1389	-0.1131 ± 0.1432	0.0667 ± 0.2359	0.0545 ± 0.1849	-0.0055 ± 0.1634	-0.1043 ± 0.307	-0.0063 ± 0.112	0.0443 ± 0.2065	-0.092 ± 0.1617	-0.0863 ± 0.1794	-0.0887 ± 0.1608
	MAE	0.1199 ± 0.1988	0.0826 ± 0.2138	0.0836 ± 0.2307	0.0398 ± 0.1371	0.0374 ± 0.1227	0.1043 ± 0.2183	0.1118 ± 0.2039	0.0884 ± 0.1946	0.0659 ± 0.1899	0.0937 ± 0.1531	**0.1254** ± 0.2048	0.0232 ± 0.1601	0.0314 ± 0.1713	0.0572 ± 0.1152
	MSE	**0.0992** ± 0.2526	0.0384 ± 0.2588	0.0378 ± 0.3074	-0.0685 ± 0.1844	-0.0662 ± 0.1553	0.0452 ± 0.193	0.0546 ± 0.1889	-0.053 ± 0.2368	-0.0602 ± 0.1715	-0.0362 ± 0.1929	-0.0123 ± 0.229	-0.086 ± 0.204	-0.0988 ± 0.2159	-0.0458 ± 0.1529
Obesity	ACC	0.7982 ± 0.1539	0.8303 ± 0.0756	0.8343 ± 0.0721	0.82 ± 0.1074	0.7907 ± 0.0945	0.8015 ± 0.1463	0.8324 ± 0.081	0.8331 ± 0.0786	0.1358 ± 0.4281	**0.8393** ± 0.0777	0.8021 ± 0.146	0.838 ± 0.0781	0.8382 ± 0.0779	0.82 ± 0.1074
	MAE	0.7982 ± 0.1539	0.8303 ± 0.0756	0.8343 ± 0.0721	0.82 ± 0.1074	0.7907 ± 0.0945	0.8015 ± 0.1463	0.8324 ± 0.081	0.8331 ± 0.0786	0.1358 ± 0.4281	**0.8393** ± 0.0777	0.8021 ± 0.146	0.838 ± 0.0781	0.8382 ± 0.0779	0.82 ± 0.1074
	MSE	0.7605 ± 0.1608	0.8296 ± 0.0791	0.8337 ± 0.0756	0.8198 ± 0.109	0.7903 ± 0.0957	0.7502 ± 0.1756	0.7882 ± 0.1211	0.8308 ± 0.0796	0.1835 ± 0.3726	**0.8388** ± 0.0806	0.7507 ± 0.1755	0.8375 ± 0.0811	0.8376 ± 0.0809	0.8198 ± 0.109
CMC	ACC	**0.3399** ± 0.0651	0.3357 ± 0.0678	0.3382 ± 0.0669	0.2419 ± 0.042	0.2201 ± 0.0423	0.3138 ± 0.0644	0.3069 ± 0.0628	0.2786 ± 0.0606	-0.0362 ± 0.1	0.2988 ± 0.0556	0.3074 ± 0.059	0.289 ± 0.0555	0.2986 ± 0.0584	0.2099 ± 0.0415
	MAE	0.2239 ± 0.0704	0.2391 ± 0.0771	0.2119 ± 0.0637	0.2891 ± 0.0656	0.2865 ± 0.0626	0.2948 ± 0.0667	0.2989 ± 0.0617	0.2808 ± 0.0724	-0.0293 ± 0.1628	0.2961 ± 0.068	**0.3003** ± 0.0702	0.2942 ± 0.0691	0.2976 ± 0.0725	0.2893 ± 0.0585
	MSE	0.0634 ± 0.0722	0.0697 ± 0.0795	0.0719 ± 0.0673	0.17 ± 0.0948	0.1869 ± 0.0907	0.1107 ± 0.094	0.1245 ± 0.0998	0.1192 ± 0.0972	-0.0952 ± 0.1877	0.1265 ± 0.0931	0.1168 ± 0.0896	0.1316 ± 0.0879	0.1194 ± 0.0885	**0.2043** ± 0.093
Grub Damage	ACC	0.2809 ± 0.2247	**0.2976** ± 0.227	0.2448 ± 0.2082	0.1141 ± 0.3624	0.0947 ± 0.3833	0.2639 ± 0.2414	0.2638 ± 0.2727	0.2111 ± 0.3474	-0.0055 ± 0.2739	0.2033 ± 0.3505	0.2488 ± 0.279	0.1963 ± 0.3222	0.235 ± 0.2864	0.0953 ± 0.3424
	MAE	0.1638 ± 0.264	0.2086 ± 0.2671	0.1447 ± 0.3095	0.1212 ± 0.3109	0.1079 ± 0.3171	0.1827 ± 0.2542	**0.2281** ± 0.2998	0.1781 ± 0.3121	0.0497 ± 0.2587	0.1748 ± 0.3169	0.1748 ± 0.2955	0.1826 ± 0.3023	0.1911 ± 0.2567	0.108 ± 0.2746
	MSE	0.1094 ± 0.2826	0.2109 ± 0.3376	0.0582 ± 0.2972	0.227 ± 0.2776	0.2187 ± 0.2946	0.2002 ± 0.2799	0.2615 ± 0.2899	0.2705 ± 0.2812	0.0437 ± 0.2798	0.2588 ± 0.2653	0.2308 ± 0.319	0.2719 ± 0.2805	**0.2801** ± 0.302	0.2315 ± 0.2703
New Thyroid	ACC	0.896 ± 0.157	0.9476 ± 0.0892	0.9408 ± 0.0908	0.9408 ± 0.0711	0.9182 ± 0.1224	0.789 ± 0.339	**0.9487** ± 0.0472	0.9408 ± 0.0711	0.2949 ± 0.5865	0.9484 ± 0.0741	0.9333 ± 0.0877	0.9408 ± 0.0711	0.9408 ± 0.0711	0.9408 ± 0.0711
	MAE	0.87 ± 0.2191	0.9359 ± 0.1171	0.9291 ± 0.1177	0.9364 ± 0.0807	0.9216 ± 0.1145	0.7794 ± 0.3368	**0.9441** ± 0.0536	0.9364 ± 0.0807	0.3318 ± 0.578	0.944 ± 0.0838	0.9265 ± 0.1031	0.9364 ± 0.0807	0.9364 ± 0.0807	0.9364 ± 0.0807
	MSE	0.8619 ± 0.2193	0.935 ± 0.0946	0.9282 ± 0.0952	0.9385 ± 0.0596	0.9229 ± 0.0867	0.8262 ± 0.177	0.9425 ± 0.057	0.9385 ± 0.0596	0.2709 ± 0.6553	**0.9461** ± 0.0634	0.9296 ± 0.0738	0.9385 ± 0.0596	0.9385 ± 0.0596	0.9385 ± 0.0596
Balance Scale	ACC	0.9095 ± 0.0488	**0.9175** ± 0.0463	0.9173 ± 0.0452	0.8951 ± 0.0526	0.8727 ± 0.0496	0.89 ± 0.07	0.9091 ± 0.0496	0.9113 ± 0.0455	0.0937 ± 0.2037	0.9169 ± 0.0448	0.9099 ± 0.0471	0.9061 ± 0.0508	0.9042 ± 0.05	0.8852 ± 0.054
	MAE	0.8796 ± 0.0455	0.8979 ± 0.043	0.8862 ± 0.0369	0.8856 ± 0.0514	0.8644 ± 0.0567	0.8636 ± 0.0573	0.8861 ± 0.0437	**0.902** ± 0.0467	0.1053 ± 0.2077	0.9005 ± 0.04	0.8957 ± 0.042	0.895 ± 0.0494	0.896 ± 0.0475	0.8788 ± 0.0554
	MSE	0.8745 ± 0.043	0.8604 ± 0.0491	**0.8888** ± 0.0435	0.8191 ± 0.0448	0.7966 ± 0.0617	0.8155 ± 0.0406	0.8248 ± 0.0418	0.8333 ± 0.0407	0.0083 ± 0.2812	0.826 ± 0.0375	0.825 ± 0.0395	0.8324 ± 0.0464	0.8342 ± 0.0445	0.8106 ± 0.0539
Automobile	ACC	0.5938 ± 0.3061	0.626 ± 0.3008	0.645 ± 0.3108	0.6149 ± 0.2664	0.5155 ± 0.2572	0.5211 ± 0.3616	0.5578 ± 0.3288	**0.6476** ± 0.3092	0.1302 ± 0.3885	0.6231 ± 0.2946	0.5489 ± 0.3375	0.6069 ± 0.2876	0.6167 ± 0.2973	0.5812 ± 0.2733
	MAE	0.5551 ± 0.3184	0.5937 ± 0.3108	0.6239 ± 0.305	0.6357 ± 0.2241	0.5477 ± 0.2091	0.5189 ± 0.3384	0.5576 ± 0.3016	**0.6395** ± 0.2912	0.061 ± 0.3653	0.6216 ± 0.2614	0.5193 ± 0.3382	0.6069 ± 0.2636	0.6167 ± 0.274	0.5999 ± 0.231
	MSE	0.5523 ± 0.3519	0.58 ± 0.3334	0.6289 ± 0.2979	**0.6665** ± 0.1373	0.5832 ± 0.1447	0.5471 ± 0.2608	0.5536 ± 0.2412	0.6401 ± 0.2701	0.0442 ± 0.3769	0.6275 ± 0.206	0.5305 ± 0.3263	0.6209 ± 0.226	0.6271 ± 0.238	0.6323 ± 0.1515
Eucalyptus	ACC	0.4037 ± 0.0762	0.3994 ± 0.0785	0.4017 ± 0.0856	0.4028 ± 0.1062	0.4008 ± 0.0941	0.4067 ± 0.0758	**0.4178** ± 0.0845	0.4088 ± 0.0988	-0.0087 ± 0.2287	0.4013 ± 0.097	0.4037 ± 0.0865	0.4039 ± 0.0996	0.4041 ± 0.0936	0.4054 ± 0.0996
	MAE	0.4095 ± 0.0928	0.4057 ± 0.0945	0.411 ± 0.0967	0.4161 ± 0.1051	0.4106 ± 0.1001	0.4159 ± 0.084	**0.4265** ± 0.0947	0.4186 ± 0.1016	-0.0228 ± 0.213	0.4127 ± 0.0987	0.4125 ± 0.0934	0.4129 ± 0.1046	0.4126 ± 0.099	0.4174 ± 0.0992
	MSE	0.3851 ± 0.125	0.3954 ± 0.1179	0.4022 ± 0.1178	**0.4195** ± 0.0962	0.4094 ± 0.1041	0.3783 ± 0.126	0.3894 ± 0.1217	0.4159 ± 0.1021	-0.0641 ± 0.2024	0.4142 ± 0.1006	0.4042 ± 0.1122	0.4104 ± 0.108	0.4092 ± 0.1034	0.4192 ± 0.0951
TAE	ACC	0.1154 ± 0.2124	0.1027 ± 0.193	**0.1715** ± 0.2559	-0.0107 ± 0.3177	-0.0461 ± 0.332	0.0628 ± 0.2931	0.0391 ± 0.3437	0.0272 ± 0.29	0.0511 ± 0.2645	0.0425 ± 0.3377	0.0493 ± 0.256	0.0223 ± 0.3198	0.0342 ± 0.2571	-0.0234 ± 0.3243
	MAE	0.0652 ± 0.1604	0.0206 ± 0.2036	0.1432 ± 0.1816	0.1597 ± 0.3708	0.1396 ± 0.3833	0.173 ± 0.2686	0.1686 ± 0.3629	0.1347 ± 0.3226	**0.2179** ± 0.2802	0.1696 ± 0.3575	0.1675 ± 0.2477	0.1479 ± 0.3521	0.1528 ± 0.2616	0.1271 ± 0.3892
	MSE	-0.0277 ± 0.3016	0.0228 ± 0.3622	-0.0445 ± 0.2275	0.2649 ± 0.3391	0.2688 ± 0.3864	0.1822 ± 0.3476	0.1871 ± 0.3506	0.1977 ± 0.3925	**0.4212** ± 0.2296	0.1915 ± 0.344	0.216 ± 0.3755	0.2015 ± 0.3613	0.2197 ± 0.3815	0.2362 ± 0.3878
Heart (CLE)	ACC	0.4938 ± 0.107	0.4901 ± 0.0857	0.5439 ± 0.097	0.5259 ± 0.1571	0.5029 ± 0.1545	0.4476 ± 0.1332	0.4964 ± 0.1376	0.5374 ± 0.11	0.0689 ± 0.2364	**0.5471** ± 0.1273	0.516 ± 0.1126	0.5306 ± 0.1216	0.5281 ± 0.1059	0.5081 ± 0.158
	MAE	0.4686 ± 0.1179	0.4428 ± 0.115	0.5333 ± 0.0712	0.5278 ± 0.1402	0.5061 ± 0.133	0.4266 ± 0.1631	0.4774 ± 0.1243	0.5287 ± 0.1041	0.0969 ± 0.2217	**0.5521** ± 0.1081	0.5252 ± 0.1182	0.5274 ± 0.1138	0.5284 ± 0.1051	0.5185 ± 0.1512
	MSE	0.407 ± 0.1142	0.3509 ± 0.1251	0.4795 ± 0.0647	0.4637 ± 0.1573	0.4323 ± 0.1513	0.3643 ± 0.2137	0.4081 ± 0.1541	0.4506 ± 0.123	0.0306 ± 0.2634	**0.4814** ± 0.1383	0.4413 ± 0.1053	0.4563 ± 0.1362	0.4533 ± 0.1338	0.4487 ± 0.1661
SWD	ACC	0.2027 ± 0.1023	0.1854 ± 0.1123	0.1957 ± 0.0955	0.2092 ± 0.1024	0.2072 ± 0.1082	0.2072 ± 0.1068	0.2049 ± 0.1001	**0.213** ± 0.096	0.0443 ± 0.1095	0.206 ± 0.0985	0.2072 ± 0.1068	0.2049 ± 0.1001	0.2072 ± 0.1068	0.2007 ± 0.1128
	MAE	0.1218 ± 0.067	0.1243 ± 0.0689	0.1211 ± 0.0847	**0.1451** ± 0.094	0.1408 ± 0.0912	0.1302 ± 0.0767	0.1323 ± 0.0911	0.1358 ± 0.0855	0.0571 ± 0.109	0.1364 ± 0.0895	0.1302 ± 0.0767	0.1323 ± 0.0911	0.1302 ± 0.0767	0.1417 ± 0.0876
	MSE	0.1003 ± 0.0865	0.0942 ± 0.0803	0.1263 ± 0.0937	**0.1604** ± 0.1054	0.1547 ± 0.0914	0.1248 ± 0.0919	0.1322 ± 0.1031	0.1155 ± 0.1138	0.0475 ± 0.1079	0.1411 ± 0.1035	0.1248 ± 0.0919	0.1322 ± 0.1031	0.1248 ± 0.0919	0.1571 ± 0.1045
ERA	ACC	0.145 ± 0.0764	**0.1631** ± 0.1352	0.139 ± 0.088	0.0273 ± 0.0782	0.0608 ± 0.0721	0.1085 ± 0.084	0.0753 ± 0.0743	0.1566 ± 0.0729	0.0469 ± 0.1274	0.0722 ± 0.0718	0.1074 ± 0.0838	0.0746 ± 0.0731	0.1072 ± 0.0851	0.0449 ± 0.0899
	MAE	0.0313 ± 0.1002	0.0428 ± 0.1289	0.0051 ± 0.0707	0.0084 ± 0.0952	0.0426 ± 0.0855	**0.059** ± 0.0672	0.0338 ± 0.0818	0.0404 ± 0.0722	0.0533 ± 0.0973	0.01 ± 0.0894	0.0586 ± 0.0671	0.0322 ± 0.0831	0.0564 ± 0.0671	0.0157 ± 0.1
	MSE	0.0642 ± 0.1	**0.0759** ± 0.1138	0.0143 ± 0.1226	-0.0027 ± 0.148	0.0207 ± 0.1353	0.0295 ± 0.1167	0.0129 ± 0.1351	0.0251 ± 0.1188	0.0288 ± 0.1204	0.0049 ± 0.1455	0.0286 ± 0.1166	0.0132 ± 0.1338	0.0262 ± 0.1168	0.0036 ± 0.1469
ESL	ACC	0.1369 ± 0.1794	0.1782 ± 0.1608	0.1845 ± 0.1283	0.239 ± 0.1247	**0.2586** ± 0.1314	0.1533 ± 0.1808	0.1507 ± 0.178	0.1627 ± 0.1369	0.0555 ± 0.2048	0.2179 ± 0.1327	0.1492 ± 0.1847	0.2301 ± 0.1308	0.2228 ± 0.1369	0.25 ± 0.137
	MAE	0.138 ± 0.2302	0.2111 ± 0.1742	0.2394 ± 0.1422	0.2935 ± 0.1249	**0.3062** ± 0.1319	0.1449 ± 0.2442	0.1446 ± 0.2371	0.128 ± 0.2315	0.1212 ± 0.2284	0.2754 ± 0.1381	0.1418 ± 0.2484	0.2837 ± 0.1319	0.2762 ± 0.1416	0.3014 ± 0.1357
	MSE	0.0993 ± 0.2802	0.1637 ± 0.2088	0.1902 ± 0.1843	0.2312 ± 0.1625	**0.2358** ± 0.1721	0.0944 ± 0.2985	0.1118 ± 0.2891	0.0929 ± 0.2663	0.1276 ± 0.1967	0.2148 ± 0.1801	0.0938 ± 0.3007	0.218 ± 0.1735	0.2071 ± 0.1863	0.2325 ± 0.1796
LEV	ACC	0.1518 ± 0.1201	0.1344 ± 0.1197	0.155 ± 0.1079	0.1535 ± 0.1016	0.1613 ± 0.1041	0.1589 ± 0.1157	0.1693 ± 0.111	0.1628 ± 0.114	0.0157 ± 0.1443	0.148 ± 0.1048	0.159 ± 0.1158	**0.1704** ± 0.1093	0.1615 ± 0.1113	0.159 ± 0.1026
	MAE	0.1279 ± 0.1166	0.1056 ± 0.1166	0.1493 ± 0.1074	0.1466 ± 0.0982	0.1412 ± 0.0981	0.1348 ± 0.1092	0.1608 ± 0.106	0.1541 ± 0.1072	0.0437 ± 0.1439	0.1434 ± 0.1012	0.1348 ± 0.1092	**0.1622** ± 0.1046	0.1374 ± 0.105	0.1383 ± 0.095
	MSE	0.1396 ± 0.1115	0.1078 ± 0.1165	0.1808 ± 0.1198	0.1736 ± 0.1026	0.1499 ± 0.1047	0.1478 ± 0.1099	0.1861 ± 0.1094	0.1832 ± 0.1144	0.0734 ± 0.1849	0.1763 ± 0.1118	0.1479 ± 0.1099	**0.1879** ± 0.1089	0.1501 ± 0.1057	0.1469 ± 0.1041
Red Wine	ACC	0.3548 ± 0.0749	0.3469 ± 0.0739	0.3606 ± 0.0729	0.3459 ± 0.0829	0.3402 ± 0.0825	0.351 ± 0.0847	0.3492 ± 0.0839	0.3612 ± 0.0797	-0.0378 ± 0.0327	0.3601 ± 0.0776	**0.3612** ± 0.0778	0.3524 ± 0.0764	0.3539 ± 0.0772	0.3496 ± 0.0828
	MAE	0.3434 ± 0.0819	0.3359 ± 0.0819	0.3545 ± 0.0766	0.3456 ± 0.0765	0.3331 ± 0.0881	0.341 ± 0.0886	0.3426 ± 0.085	0.3562 ± 0.0835	-0.048 ± 0.0394	**0.3582** ± 0.0765	0.3525 ± 0.0838	0.3481 ± 0.0795	0.347 ± 0.0832	0.3462 ± 0.0836
	MSE	0.3262 ± 0.113	0.3199 ± 0.114	0.3441 ± 0.1096	0.3443 ± 0.1043	0.3197 ± 0.1252	0.3271 ± 0.1069	0.3331 ± 0.1053	0.3463 ± 0.1105	-0.0448 ± 0.0715	**0.3529** ± 0.1065	0.3353 ± 0.1117	0.3388 ± 0.1124	0.3342 ± 0.1141	0.3399 ± 0.1113
White Wine	ACC	0.3459 ± 0.073	0.3339 ± 0.0746	0.3201 ± 0.061	0.2759 ± 0.052	0.3 ± 0.0557	0.3435 ± 0.0632	0.3289 ± 0.0612	0.3461 ± 0.0596	0.0563 ± 0.0606	0.2948 ± 0.0514	**0.3483** ± 0.0644	0.3084 ± 0.0536	0.3271 ± 0.0598	0.302 ± 0.0487
	MAE	0.3316 ± 0.0714	0.3192 ± 0.0671	0.3155 ± 0.0682	0.2742 ± 0.0581	0.2941 ± 0.0527	0.3324 ± 0.0639	0.3222 ± 0.0685	**0.3382** ± 0.0681	0.0534 ± 0.0555	0.2937 ± 0.0584	0.3365 ± 0.0655	0.3048 ± 0.0594	0.3192 ± 0.0586	0.2962 ± 0.0494
	MSE	0.3034 ± 0.0728	0.2981 ± 0.0723	0.3221 ± 0.0885	0.2956 ± 0.0839	0.3074 ± 0.0715	0.3152 ± 0.0659	0.3105 ± 0.0695	0.3276 ± 0.0794	0.05 ± 0.0725	0.3158 ± 0.0862	0.3089 ± 0.0662	0.3229 ± 0.0854	**0.3283** ± 0.0787	0.3099 ± 0.0695



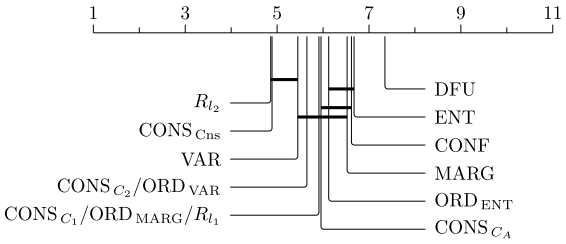
(a) Overall result for MCR, MAE and MSE.



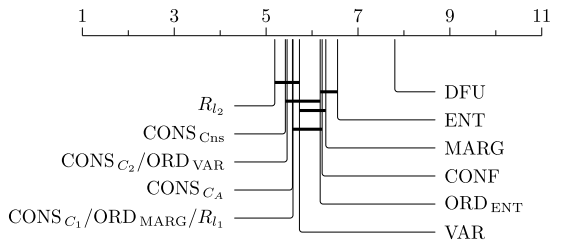
(b) Result only for MCR.



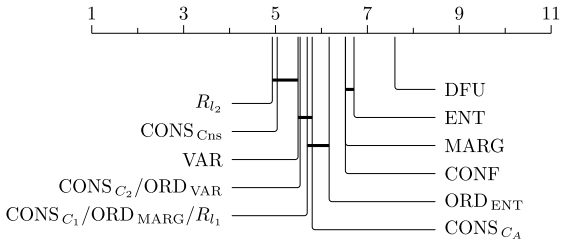
(c) Result for MAE.



(d) Result for MSE.



(e) Result for MCR and MAE.



(f) Result for MAE and MSE.

Fig. E.15. Critical difference (CD) diagrams for the evaluated uncertainty measures over all performance metrics and datasets based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon test with an MLP and QWK loss. Groups of uncertainty measures that are not significantly different (at $p = 0.05$) are connected [53,54].

(cf. Fig. D.14c, Fig. D.14d, and Fig. D.14f). Moreover, the ordinal binary decomposition method again seems to strike a better balance than VAR and R_{l_2} when it comes to the trade-off between exact hit-rate and minimization of distance-based errors, even more so since the error distances are less due to the squashed predictive probability distributions (cf. Fig. D.14a and Fig. D.14e). Table D.14 displays the detailed PRR results for all uncertainty measures and datasets using SLGBM.

Appendix E. Prediction rejection ratios (PRRs) with quadratic weighted kappa (QWK) as the loss function

In this section, we present additional experimental results using an MLP with QWK [33] as the loss function instead of CE (cf. Sections 6 and Appendix B). Again, refer to Table B.11 for the parameters of the MLP. As shown in Appendix C, QWK leads to increased predictive performance in terms of QWK over CE loss at the cost of worsened uncertainty quantification, indicated by smaller PRR values. Overall, nominal measures are still significantly outperformed by measures taking distance into account (cf. Fig. E.15a), though results are, similar to the simple ordinal approach (cf. Appendix D), not as significant as with CE loss anymore. The superiority of measures taking distance into account is still particularly visible for MAE and MSE (cf. Fig. E.15f) and also overall (cf. Fig. E.15a). However, in general, just like in Appendix D, the different uncertainty measures have become more interchangeable due to the biased squashed predictive probability distributions. This again demonstrates the advantage of CE loss for uncertainty quantification in ordinal classification. Table E.15 displays the detailed PRR results for all uncertainty measures and datasets using QWK loss.

Data availability

Some datasets used are publicly available. Some datasets are confidential.

Table E.15
PRRs for the different uncertainty measures and ordinal benchmark datasets using 10-fold cross-validation with an MLP and QWK loss [33].

Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Cris}$	$CONS_{C_1}$	$CONS_{C_2}$	$CONS_{C_4}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{I_1}	R_{I_2}
Triazines	ACC	0.2064 ± 0.3448	0.2067 ± 0.3594	0.2138 ± 0.2875	0.07 ± 0.4394	0.095 ± 0.4334	0.0979 ± 0.4061	0.1188 ± 0.4056	**0.214** ± 0.3077	-0.0423 ± 0.3118	0.1157 ± 0.3964	0.0979 ± 0.4061	0.1188 ± 0.4056	0.0979 ± 0.4061	0.0989 ± 0.4238
	MAE	0.2958 ± 0.3199	0.247 ± 0.3199	**0.3137** ± 0.2742	0.27 ± 0.2947	0.258 ± 0.3152	0.3132 ± 0.335	0.3068 ± 0.3174	0.2787 ± 0.2986	0.16 ± 0.1645	0.3061 ± 0.3072	0.3132 ± 0.335	0.3068 ± 0.3174	0.3132 ± 0.335	0.2178 ± 0.3358
	MSE	0.2251 ± 0.2632	0.1913 ± 0.2675	0.2279 ± 0.2376	0.2554 ± 0.3023	**0.2571** ± 0.2861	0.2416 ± 0.2851	0.2408 ± 0.2853	0.1877 ± 0.2737	0.0944 ± 0.3228	0.2417 ± 0.2842	0.2416 ± 0.2851	0.2408 ± 0.2853	0.2416 ± 0.2851	0.212 ± 0.2814
Machine CPU	ACC	0.3461 ± 0.3784	0.3499 ± 0.3806	0.3397 ± 0.369	0.3932 ± 0.3681	0.3779 ± 0.3895	0.3692 ± 0.3824	0.3731 ± 0.3783	0.3866 ± 0.3848	**0.4284** ± 0.4006	0.3874 ± 0.378	0.3692 ± 0.3824	0.3731 ± 0.3783	0.3692 ± 0.3824	0.3952 ± 0.3673
	MAE	0.2744 ± 0.4141	0.283 ± 0.4246	0.2693 ± 0.4063	0.4364 ± 0.399	0.3984 ± 0.4515	0.3683 ± 0.4256	0.3755 ± 0.4233	0.3793 ± 0.4227	**0.4712** ± 0.3978	0.3858 ± 0.4175	0.3683 ± 0.4256	0.3755 ± 0.4233	0.3683 ± 0.4256	0.4293 ± 0.4071
	MSE	0.1971 ± 0.5017	0.2076 ± 0.5189	0.2077 ± 0.4949	0.393 ± 0.5023	0.3325 ± 0.5526	0.2908 ± 0.5059	0.302 ± 0.5101	0.2965 ± 0.5098	**0.4114** ± 0.4518	0.3139 ± 0.495	0.2908 ± 0.5059	0.302 ± 0.5101	0.2908 ± 0.5059	0.3811 ± 0.5103
Auto MPG	ACC	0.2142 ± 0.0869	0.2102 ± 0.0985	0.2133 ± 0.0786	0.2709 ± 0.0984	0.2591 ± 0.0925	0.2445 ± 0.0863	0.2497 ± 0.0963	0.2552 ± 0.0945	**0.304** ± 0.1506	0.255 ± 0.0769	0.2445 ± 0.0863	0.2497 ± 0.0963	0.2445 ± 0.0863	0.2719 ± 0.1006
	MAE	0.2152 ± 0.0912	0.2136 ± 0.0889	0.2053 ± 0.1024	0.2672 ± 0.1198	0.267 ± 0.1135	0.2546 ± 0.1035	0.2533 ± 0.1133	0.2521 ± 0.1117	**0.2985** ± 0.1773	0.2516 ± 0.1034	0.2546 ± 0.1035	0.2533 ± 0.1133	0.2546 ± 0.1035	0.2667 ± 0.1248
	MSE	0.071 ± 0.1635	0.0788 ± 0.1682	0.0735 ± 0.1698	0.1084 ± 0.1659	0.1128 ± 0.1773	0.0978 ± 0.1618	0.0953 ± 0.1635	0.0865 ± 0.16	**0.1788** ± 0.2043	0.089 ± 0.1604	0.0978 ± 0.1618	0.0953 ± 0.1635	0.0978 ± 0.1618	0.1085 ± 0.1719
Pyrimidines	ACC	-0.0073 ± 0.4975	-0.1132 ± 0.6203	0.1058 ± 0.4352	0.0813 ± 0.5463	0.0681 ± 0.4439	0.0399 ± 0.48	0.0602 ± 0.5042	0.0288 ± 0.4958	**0.1838** ± 0.4793	0.096 ± 0.4534	0.0399 ± 0.48	0.0602 ± 0.5042	0.0399 ± 0.48	0.12 ± 0.5264
	MAE	0.0766 ± 0.3678	0.0113 ± 0.3587	0.1021 ± 0.3763	0.1421 ± 0.4204	0.0351 ± 0.4102	0.0005 ± 0.4567	0.0587 ± 0.4062	0.0816 ± 0.3881	0.0123 ± 0.3409	0.0446 ± 0.3888	0.0005 ± 0.4567	0.0587 ± 0.4062	0.0005 ± 0.4567	**0.1502** ± 0.3569
	MSE	-0.0537 ± 0.4374	0.0454 ± 0.4705	-0.0608 ± 0.4675	0.0444 ± 0.5467	-0.0369 ± 0.549	-0.0982 ± 0.5931	-0.0565 ± 0.5747	-0.079 ± 0.4904	-0.0349 ± 0.4706	-0.0667 ± 0.4583	-0.0982 ± 0.5931	-0.0565 ± 0.5747	-0.0982 ± 0.5931	**0.0655** ± 0.5233
Abalone	ACC	0.167 ± 0.0855	0.1668 ± 0.0843	0.1648 ± 0.0847	0.1774 ± 0.0953	0.1739 ± 0.0902	0.1736 ± 0.0905	0.1727 ± 0.0906	0.176 ± 0.094	0.0672 ± 0.1249	0.1733 ± 0.0929	0.1736 ± 0.0905	0.1727 ± 0.0906	0.1736 ± 0.0904	**0.1783** ± 0.0964
	MAE	0.1533 ± 0.1251	0.1531 ± 0.125	0.1515 ± 0.1227	0.1624 ± 0.1299	0.1611 ± 0.1291	0.1597 ± 0.1289	0.159 ± 0.1281	0.1605 ± 0.1296	0.0483 ± 0.1313	0.1591 ± 0.1286	0.1597 ± 0.1289	0.159 ± 0.128	0.1597 ± 0.1289	**0.1637** ± 0.1324
	MSE	0.1213 ± 0.1743	0.1202 ± 0.1755	0.123 ± 0.1706	0.1272 ± 0.1667	0.127 ± 0.1727	0.1248 ± 0.1721	0.1244 ± 0.1695	0.1241 ± 0.1686	0.0066 ± 0.1428	0.125 ± 0.1673	0.1248 ± 0.1721	0.1244 ± 0.1694	0.1248 ± 0.1721	**0.1285** ± 0.1716
Boston Housing	ACC	0.3587 ± 0.2232	0.3592 ± 0.2235	0.3508 ± 0.2276	0.3876 ± 0.2397	**0.4095** ± 0.2474	0.3961 ± 0.2371	0.3905 ± 0.2399	0.3925 ± 0.2366	-0.1767 ± 0.211	0.3817 ± 0.2364	0.3961 ± 0.2371	0.3905 ± 0.2399	0.3961 ± 0.2371	0.405 ± 0.2434
	MAE	0.3153 ± 0.1943	0.3218 ± 0.1958	0.3149 ± 0.2118	0.3544 ± 0.2373	**0.3782** ± 0.2384	0.3592 ± 0.2233	0.3583 ± 0.2345	0.3579 ± 0.2288	-0.1223 ± 0.1858	0.35 ± 0.2354	0.3592 ± 0.2233	0.3583 ± 0.2345	0.3592 ± 0.2233	0.378 ± 0.2371
	MSE	0.2328 ± 0.2327	0.2406 ± 0.2368	0.2342 ± 0.2424	0.2766 ± 0.2461	**0.2986** ± 0.2511	0.2668 ± 0.2415	0.274 ± 0.2557	0.2732 ± 0.2497	-0.0354 ± 0.2498	0.2703 ± 0.2497	0.2668 ± 0.2415	0.274 ± 0.2557	0.2668 ± 0.2415	0.295 ± 0.2492
Stocks Domain	ACC	-0.0551 ± 0.2565	-0.0574 ± 0.2644	-0.0631 ± 0.2404	-0.053 ± 0.2428	-0.047 ± 0.2598	-0.0496 ± 0.2543	-0.0558 ± 0.245	-0.0496 ± 0.2459	**0.1443** ± 0.4103	-0.0566 ± 0.2387	-0.0496 ± 0.2543	-0.0558 ± 0.245	-0.0496 ± 0.2543	-0.0433 ± 0.2529
	MAE	-0.0607 ± 0.2625	-0.063 ± 0.2705	-0.0687 ± 0.2464	-0.0586 ± 0.2489	-0.0526 ± 0.266	-0.0552 ± 0.2602	-0.0614 ± 0.251	-0.0552 ± 0.2517	**0.1491** ± 0.4081	-0.0622 ± 0.2449	-0.0552 ± 0.2602	-0.0614 ± 0.251	-0.0552 ± 0.2602	-0.0489 ± 0.2587
	MSE	-0.0684 ± 0.2664	-0.0711 ± 0.2735	-0.0727 ± 0.2564	-0.0626 ± 0.2584	-0.0587 ± 0.2723	-0.0625 ± 0.2644	-0.0668 ± 0.2585	-0.062 ± 0.2563	**0.1453** ± 0.3992	-0.0659 ± 0.2547	-0.0625 ± 0.2644	-0.0668 ± 0.2585	-0.0625 ± 0.2644	-0.0559 ± 0.2632
Wisconsin Breast Cancer	ACC	0.3744 ± 0.2153	0.3546 ± 0.1934	**0.4269** ± 0.2601	0.3228 ± 0.274	0.3032 ± 0.2687	0.376 ± 0.2436	0.3822 ± 0.2527	0.4052 ± 0.261	-0.1898 ± 0.4986	0.3649 ± 0.2516	0.376 ± 0.2436	0.3822 ± 0.2527	0.376 ± 0.2436	0.3069 ± 0.276
	MAE	0.191 ± 0.1801	**0.1966** ± 0.1695	0.1942 ± 0.1926	0.0741 ± 0.1887	0.075 ± 0.1805	0.1458 ± 0.1712	0.1345 ± 0.1853	0.1689 ± 0.1899	-0.1297 ± 0.1813	0.1142 ± 0.1763	0.1458 ± 0.1712	0.1345 ± 0.1853	0.1458 ± 0.1712	0.0764 ± 0.1738
	MSE	0.1998 ± 0.2633	**0.2296** ± 0.2588	0.1721 ± 0.2311	0.0225 ± 0.1439	0.0185 ± 0.1416	0.0824 ± 0.1765	0.0811 ± 0.1635	0.1384 ± 0.1951	-0.1149 ± 0.2102	0.0659 ± 0.1546	0.0824 ± 0.1765	0.0811 ± 0.1635	0.0824 ± 0.1765	0.044 ± 0.1345
Obesity	ACC	0.1729 ± 0.5138	0.1941 ± 0.5197	0.1362 ± 0.5001	0.2807 ± 0.565	0.2802 ± 0.5589	0.2408 ± 0.5429	0.2411 ± 0.5438	0.258 ± 0.5524	**0.39** ± 0.4871	0.2298 ± 0.5408	0.2408 ± 0.5429	0.2411 ± 0.5438	0.2408 ± 0.5429	0.2855 ± 0.5659
	MAE	0.1709 ± 0.5167	0.1921 ± 0.5222	0.1353 ± 0.5031	0.2771 ± 0.5663	0.2771 ± 0.5606	0.2385 ± 0.5453	0.2389 ± 0.5461	0.2556 ± 0.5541	**0.3865** ± 0.4887	0.2281 ± 0.5429	0.2385 ± 0.5453	0.2389 ± 0.5461	0.2385 ± 0.5453	0.282 ± 0.5673
	MSE	0.1159 ± 0.4862	0.1388 ± 0.4905	0.0826 ± 0.4744	0.2341 ± 0.5299	0.2315 ± 0.5256	0.1898 ± 0.5108	0.1908 ± 0.5113	0.2104 ± 0.518	**0.3379** ± 0.4419	0.1812 ± 0.5072	0.1898 ± 0.5108	0.1908 ± 0.5113	0.1898 ± 0.5108	0.2383 ± 0.5316
CMC	ACC	0.2667 ± 0.0778	0.263 ± 0.0759	**0.2707** ± 0.0855	0.2492 ± 0.0853	0.2431 ± 0.0843	0.2596 ± 0.0859	0.26 ± 0.0836	0.2579 ± 0.0829	0.1854 ± 0.0799	0.2597 ± 0.0848	0.2596 ± 0.0859	0.2601 ± 0.0836	0.2597 ± 0.0859	0.2455 ± 0.0873
	MAE	0.1852 ± 0.1029	0.1816 ± 0.1021	0.1908 ± 0.1074	0.2039 ± 0.1056	0.2028 ± 0.1047	0.1973 ± 0.1068	0.1995 ± 0.1044	0.197 ± 0.1035	0.0643 ± 0.0686	0.2003 ± 0.1049	0.1973 ± 0.1068	0.1996 ± 0.1044	0.1974 ± 0.1068	**0.2048** ± 0.1095
	MSE	0.0434 ± 0.0994	0.0431 ± 0.0993	0.0489 ± 0.0961	0.0898 ± 0.1005	**0.0915** ± 0.1005	0.0636 ± 0.099	0.0709 ± 0.0977	0.0711 ± 0.0983	-0.0183 ± 0.0753	0.0736 ± 0.0983	0.0636 ± 0.099	0.0709 ± 0.0977	0.0637 ± 0.099	0.0904 ± 0.1039
Grub Damage	ACC	0.2412 ± 0.2663	**0.2611** ± 0.2905	0.2268 ± 0.2241	0.1932 ± 0.2722	0.2001 ± 0.2989	0.2141 ± 0.2712	0.1897 ± 0.2733	0.1887 ± 0.276	0.22 ± 0.1992	0.1908 ± 0.2722	0.2141 ± 0.2712	0.1897 ± 0.2733	0.2141 ± 0.2712	0.2213 ± 0.2655
	MAE	0.2292 ± 0.2824	0.214 ± 0.2849	0.2246 ± 0.2639	**0.3042** ± 0.2489	0.2949 ± 0.2512	0.2411 ± 0.2645	0.2644 ± 0.2558	0.2401 ± 0.2607	0.1989 ± 0.2189	0.2656 ± 0.2472	0.2411 ± 0.2645	0.2644 ± 0.2558	0.2411 ± 0.2645	0.2855 ± 0.2482
	MSE	0.1439 ± 0.269	0.1316 ± 0.2577	0.1671 ± 0.279	**0.2846** ± 0.315	0.2774 ± 0.2835	0.1826 ± 0.2849	0.2294 ± 0.2958	0.1653 ± 0.2883	0.1511 ± 0.3198	0.2282 ± 0.3022	0.1826 ± 0.2849	0.2294 ± 0.2958	0.1826 ± 0.2849	0.2192 ± 0.2714
New Thyroid	ACC	0.9571 ± 0.0413	0.9646 ± 0.0301	0.9471 ± 0.0475	0.9451 ± 0.0782	0.925 ± 0.0743	0.9646 ± 0.0301	0.9545 ± 0.04	**0.9651** ± 0.0435	0.2908 ± 0.2841	0.9651 ± 0.0435	0.9646 ± 0.0301	0.9545 ± 0.04	0.9646 ± 0.0301	0.9625 ± 0.0582
	MAE	0.9535 ± 0.0336	0.949 ± 0.0347	0.9501 ± 0.0457	0.9646 ± 0.0512	0.9445 ± 0.053	0.9636 ± 0.0381	0.957 ± 0.0398	**0.9708** ± 0.0258	0.4562 ± 0.2503	0.9708 ± 0.0258	0.9636 ± 0.0381	0.957 ± 0.0398	0.9636 ± 0.0381	0.9622 ± 0.0487
	MSE	0.9375 ± 0.0642	0.9216 ± 0.0802	0.9405 ± 0.0867	0.9727 ± 0.041	0.956 ± 0.0465	0.9664 ± 0.0535	0.9605 ± 0.0513	**0.9742** ± 0.0271	0.5876 ± 0.3271	0.9742 ± 0.0271	0.9664 ± 0.0535	0.9605 ± 0.0513	0.9664 ± 0.0535	0.9742 ± 0.0271
Balance Scale	ACC	0.9392 ± 0.0679	0.9299 ± 0.0656	0.9387 ± 0.0487	0.9363 ± 0.0516	0.7979 ± 0.1368	0.9392 ± 0.0679	0.9411 ± 0.047	**0.9437** ± 0.0449	0.1831 ± 0.2245	0.9387 ± 0.0487	0.9392 ± 0.0679	0.9411 ± 0.047	0.9392 ± 0.0679	0.9369 ± 0.0666
	MAE	**0.9468** ± 0.0666	0.9386 ± 0.0656	0.934 ± 0.0472	0.932 ± 0.0496	0.815 ± 0.1357	0.9468 ± 0.0666	0.9403 ± 0.044	0.9425 ± 0.0421	0.0392 ± 0.4347	0.934 ± 0.0472	0.9468 ± 0.0666	0.9403 ± 0.044	0.9468 ± 0.0666	0.9448 ± 0.0657
	MSE	**0.9427** ± 0.0642	0.9345 ± 0.0626	0.934 ± 0.0472	0.932 ± 0.0496	0.8115 ± 0.1389	0.9427 ± 0.0642	0.9403 ± 0.044	0.9425 ± 0.0421	0.0444 ± 0.44	0.934 ± 0.0472	0.9427 ± 0.0642	0.9403 ± 0.044	0.9427 ± 0.0642	0.9407 ± 0.0631
Automobile	ACC	0.4039 ± 0.2083	0.3575 ± 0.208	0.3944 ± 0.2328	0.3871 ± 0.2428	0.4057 ± 0.2179	**0.4281** ± 0.2125	0.4231 ± 0.2194	0.4127 ± 0.2217	0.0335 ± 0.3054	0.3808 ± 0.2369	0.4281 ± 0.2125	0.4231 ± 0.2194	0.4281 ± 0.2125	0.4003 ± 0.2364
	MAE	0.3791 ± 0.2187	0.3395 ± 0.225	0.372 ± 0.2478	0.3745 ± 0.2551	0.3852 ± 0.2347	**0.4015** ± 0.2241	0.3989 ± 0.2347	0.394 ± 0.2299	0.097 ± 0.3128	0.3641 ± 0.2539	0.4015 ± 0.2241	0.3989 ± 0.2347	0.4015 ± 0.2241	0.387 ± 0.2443
	MSE	0.3813 ± 0.2548	0.3604 ± 0.2789	0.3551 ± 0.2827	0.3715 ± 0.298	0.3683 ± 0.2773	**0.3826** ± 0.2643	0.3766 ± 0.2904	0.3725 ± 0.2869	0.1954 ± 0.2853	0.354 ± 0.307	0.3826 ± 0.2643	0.3766 ± 0.2904	0.3826 ± 0.2643	0.3773 ± 0.2784
Eucalyptus	ACC	0.3706 ± 0.1	0.3641 ± 0.0943	0.3607 ± 0.0943	0.3735 ± 0.1111	0.3752 ± 0.1054	**0.3767** ± 0.1003	0.3694 ± 0.1033	0.3723 ± 0.1057	0.0896 ± 0.1217	0.3661 ± 0.1113	0.3767 ± 0.1003	0.3694 ± 0.1033	0.3767 ± 0.1003	0.3766 ± 0.1052
	MAE	0.3778 ± 0.0866	0.3629 ± 0.0738	0.3819 ± 0.0776	**0.4029** ± 0.0927	0.3974 ± 0.0946	0.3945 ± 0.0873	0.3924 ± 0.0872	0.3917 ± 0.095	0.1264 ± 0.1097	0.3924 ± 0.0949	0.3945 ± 0.0873	0.3924 ± 0.0872	0.3945 ± 0.0873	0.4002 ± 0.0925
	MSE	0.3194 ± 0.0993	0.3053 ± 0.0959	0.3345 ± 0.0896	**0.3644** ± 0.0946	0.3479 ± 0.1054	0.3384 ± 0.1037	0.3444 ± 0.0972	0.3395 ± 0.1054	0.1595 ± 0.0999	0.3507 ± 0.0928	0.3384 ± 0.1037	0.3444 ± 0.0972	0.3384 ± 0.1037	0.3534 ± 0.1009
TAE	ACC	**0.1403** ± 0.3363	0.1152 ± 0.3569	0.1364 ± 0.3115	0.0605 ± 0.3617	0.0442 ± 0.401	0.0644 ± 0.42	0.097 ± 0.3412	0.1167 ± 0.3487	-0.0477 ± 0.3118	0.097 ± 0.3412	0.0644 ± 0.42	0.097 ± 0.3412	0.0644 ± 0.42	0.0897 ± 0.3479
	MAE	0.0243 ± 0.2865	0.0335 ± 0.3057	0.0089 ± 0.2685	0.014 ± 0.2789	0.0152 ± 0.3019	0.0397 ± 0.3007	0.0267 ± 0.2648	0.0185 ± 0.3052	-0.0786 ± 0.3735	0.0303 ± 0.2685	0.0397 ± 0.3007	0.0267 ± 0.2648	0.0397 ± 0.3007	**0.0877** ± 0.2685
	MSE	-0.0714 ± 0.3229	-0.0577 ± 0.3411	-0.0884 ± 0.3012	-0.0719 ± 0.307	-0.043 ± 0.31	-0.0322 ± 0.3198	-0.0697 ± 0.2965	-0.068 ± 0.3337	-0.1015 ± 0.4367	-0.0661 ± 0.3009	-0.0322 ± 0.3198	-0.0697 ± 0.2965	-0.0322 ± 0.3198	**0.0429** ± 0.3208
Heart (CLE)	ACC	0.3971 ± 0.2242	0.3758 ± 0.2277	0.4059 ± 0.2198	0.3473 ± 0.2027	0.3149 ± 0.2126	0.3762 ± 0.2097	0.3763 ± 0.2036	**0.4244** ± 0.1774	0.2186 ± 0.1934	0.3847 ± 0.2019	0.3762 ± 0.2097	0.3763 ± 0.2036	0.3762 ± 0.2097	0.3389 ± 0.2074
	MAE	0.3405 ± 0.238	0.317 ± 0.2501	0.3509 ± 0.239	0.3213 ± 0.2522	0.2859 ± 0.2681	0.3332 ± 0.2451	0.3315 ± 0.2426	**0.3703** ± 0.2013	0.1925 ± 0.2498	0.3411 ± 0.2326	0.3332 ± 0.2451	0.3315 ± 0.2426	0.3332 ± 0.2451	0.3088 ± 0.2491
	MSE	0.3101 ± 0.2305	0.2789 ± 0.2503	0.326 ± 0.2351	0.3161 ± 0.2696	0.272 ± 0.2845	0.3063 ± 0.2504	0.3089 ± 0.2511	**0.3437** ± 0.2111	0.2026 ± 0.335	0.3205 ± 0.2384	0.3063 ± 0.2504	0.3089 ± 0.2511	0.3063 ± 0.2504	0.2966 ± 0.2636
SWD	ACC	0.1853 ± 0.0955	0.1928 ± 0.0975	0.1722 ± 0.0739	0.1864 ± 0.0816	**0.1954** ± 0.0952	0.1877 ± 0.0964	0.1825 ± 0.0829	0.1772 ± 0.0831	0.0163 ± 0.0923	0.174 ± 0.0749	0.1877 ± 0.0964	0.1825 ± 0.0829	0.1877 ± 0.0964	0.1934 ± 0.0952
	MAE	0.1879 ± 0.0969	0.1963 ± 0.0975	0.1741 ± 0.0787	0.1905 ± 0.0856	**0.2024** ± 0.1009	0.192 ± 0.0984	0.1858 ± 0.0854	0.1792 ± 0.0865	0.0065 ± 0.0943	0.1762 ± 0.0803	0.192 ± 0.0984	0.1858 ± 0.0854	0.192 ± 0.0984	0.1979 ± 0.0985
	MSE	0.1876 ± 0.1183	0.1969 ± 0.118	0.1732 ± 0.106	0.1928 ± 0.1096	**0.2075** ± 0.1266	0.194 ± 0.1201	0.1863 ± 0.1086	0.1785 ± 0.1101	-0.0019 ± 0.1167	0.1759 ± 0.1067	0.194 ± 0.1201	0.1863 ± 0.1086	0.194 ± 0.1201	0.2005 ± 0.1199
ERA	ACC	0.1439 ± 0.1203	0.1255 ± 0.1142	0.1652 ± 0.1106	**0.181** ± 0.111	0.1387 ± 0.1281	0.1545 ± 0.1149	0.1672 ± 0.1127	0.1735 ± 0.0989	0.1034 ± 0.1044	0.1795 ± 0.1103	0.1545 ± 0.1149	0.1672 ± 0.1127	0.1545 ± 0.1149	0.1654 ± 0.1269
	MAE	0.0426 ± 0.148	0.0423 ± 0.1443	0.0398 ± 0.1399	0.1478 ± 0.1265	0.1388 ± 0.1559	0.1036 ± 0.1505	0.1063 ± 0.1452	0.1105 ± 0.1401	**0.167** ± 0.0988	0.1046 ± 0.1339	0.1036 ± 0.1505	0.1063 ± 0.1452	0.1036 ± 0.1505	0.1438 ± 0.1379
	MSE	-0.039 ± 0.1963	-0.0261 ± 0.2045	-0.0584 ± 0.1814	0.0828 ± 0.14	0.0793 ± 0.1813	0.0264 ± 0.1908	0.0184 ± 0.1749	0.0263 ± 0.181	**0.1836** ± 0.1313	0.013 ± 0.1649	0.0264 ± 0.1908	0.0184 ± 0.1749	0.0264 ± 0.1908	0.0844 ± 0.1492
ESL	ACC	0.0922 ± 0.1431	0.0879 ± 0.1542	0.0755 ± 0.1455	0.097 ± 0.1717	0.0977 ± 0.1914	0.1017 ± 0.1664	0.1024 ± 0.1743	0.1084 ± 0.1705	-0.0491 ± 0.1696	0.0842 ± 0.1722	0.1017 ± 0.1664	0.102 ± 0.1746	0.1013 ± 0.1668	**0.1137** ± 0.1794
	MAE	0.0744 ± 0.178	0.0707 ± 0.187	0.0574 ± 0.183	0.0717 ± 0.2019	0.0802 ± 0.2176	0.0818 ± 0.1963	0.0806 ± 0.2044	0.0837 ± 0.1997	-0.0963 ± 0.1582	0.0615 ± 0.2035	0.0818 ± 0.1963	0.0803 ± 0.2047	0.0814 ± 0.1967	**0.0906** ± 0.2084
	MSE	0.0146 ± 0.2086	0.0123 ± 0.2136	-0.0044 ± 0.2241	0.0094 ± 0.232	0.0276 ± 0.2318	0.023 ± 0.2204	0.0221 ± 0.23	0.02 ± 0.2223	-0.1868 ± 0.1086	0.0027 ± 0.2366	0.023 ± 0.2204	0.0218 ± 0.2303	0.0227 ± 0.2208	**0.029** ± 0.2298
LEV	ACC	0.0772 ± 0.1122	0.0904 ± 0.1233	0.0619 ± 0.1076	0.0669 ± 0.0991	**0.0993** ± 0.1215	0.0809 ± 0.1056	0.077 ± 0.1038	0.06 ± 0.1008	-0.036 ± 0.0888	0.0591 ± 0.101	0.0809 ± 0.1056	0.077 ± 0.1038	0.0809 ± 0.1056	0.0771 ± 0.1016
	MAE	0.0489 ± 0.1018	0.062 ± 0.1128	0.0362 ± 0.0973	0.0457 ± 0.0837	**0.076** ± 0.1055	0.0557 ± 0.0914	0.052 ± 0.0907	0.0379 ± 0.0838	-0.0023 ± 0.0945	0.0358 ± 0.0883	0.0557 ± 0.0914	0.052 ± 0.0907	0.0557 ± 0.0914	0.0542 ± 0.0862
	MSE	0.0118 ± 0.1021	0.0242 ± 0.1114	0.0029 ± 0.0967	0.0199 ± 0.0879	**0.0469** ± 0.0996	0.0233 ± 0.0901	0.0203 ± 0.091	0.0096 ± 0.0818	0.0443 ± 0.1486	0.0067 ± 0.0911	0.0233 ± 0.0901	0.0203 ± 0.091	0.0233 ± 0.0901	0.0257 ± 0.0871
Red Wine	ACC	0.1695 ± 0.0622	**0.1698** ± 0.0614	0.1679 ± 0.0622	0.1643 ± 0.0597	0.1679 ± 0.0623	0.1673 ± 0.0614	0.1681 ± 0.0615	0.1661 ± 0.0622	-0.0272 ± 0.1224	0.1668 ± 0.0614	0.1679 ± 0.0615	0.1676 ± 0.0613	0.1675 ± 0.0613	0.1646 ± 0.06
	MAE	**0.1529** ± 0.0615	0.1527 ± 0.0607	0.1501 ± 0.0633	0.1459 ± 0.0593	0.1505 ± 0.0622	0.1507 ± 0.0605	0.1507 ± 0.0622	0.1495 ± 0.0609	-0.0433 ± 0.1172	0.1488 ± 0.062	0.1513 ± 0.0608	0.1502 ± 0.0617	0.1508 ± 0.0604	0.1477 ± 0.0582
	MSE	**0.1281** ± 0.0717	0.1271 ± 0.0701	0.124 ± 0.076	0.1192 ± 0.0708	0.1248 ± 0.0716	0.1259 ± 0.07	0.1251 ± 0.0738	0.1249 ± 0.072	-0.0696 ± 0.1164	0.1225 ± 0.0744	0.1265 ± 0.0706	0.1246 ± 0.073	0.126 ± 0.0698	0.1224 ± 0.0676
White Wine	ACC	0.0931 ± 0.0476	0.0922 ± 0.0474	0.0932 ± 0.0465	0.0927 ± 0.0455	0.0928 ± 0.047	0.0933 ± 0.0472	0.0936 ± 0.047	**0.0942** ± 0.0473	0.0219 ± 0.0575	0.0929 ± 0.0459	0.0934 ± 0.0473	0.0934 ± 0.0469	0.0933 ± 0.0472	0.0926 ± 0.0465
	MAE	0.0977 ± 0.0545	0.0965 ± 0.0543	0.0981 ± 0.0543	0.0993 ± 0.0545	0.0985 ± 0.0548	0.0988 ± 0.0547	0.0992 ± 0.0548	**0.1** ± 0.0552	0.027 ± 0.0587	0.0989 ± 0.0545	0.0988 ± 0.0548	0.099 ± 0.0549	0.0988 ± 0.0548	0.099 ± 0.0548
	MSE	0.1065 ± 0.0756	0.1049 ± 0.0753	0.1076 ± 0.0765	**0.1118** ± 0.0798	0.1092 ± 0.078	0.1092 ± 0.0775	0.1099 ± 0.078	0.111 ± 0.0788	0.0405 ± 0.0631	0.1104 ± 0.079	0.1092 ± 0.0774	0.1097 ± 0.0781	0.1092 ± 0.0774	0.1109 ± 0.0789

References

[1] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 4878–4887.

[2] K. Hendrickx, L. Perini, D.V. der Plas, W. Meert, J. Davis, Machine learning with a reject option: a survey, Mach. Learn. 113 (5) (2024) 3073–3110, https://doi.org/10.1007/S10994-024-06534-X.

[3] S. Haas, E. Hüllermeier, Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests, Int. J. Data Sci. Anal. (2024) 1–17.

[4] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.

[5] S. Depeweg, J.M. Hernández-Lobato, F. Doshi-Velez, S. Udluft, Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning, in: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 1192–1201.

[6] A. Malinin, L. Prokhorenkova, A. Ustimenko, Uncertainty in gradient boosting via ensembles, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021, OpenReview.net, 2021.

[7] P.A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, C. Hervás-Martínez, Ordinal regression methods: survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2016) 127–146, https://doi.org/10.1109/TKDE.2015.2457911.

[8] C. Aeppli, D. Ruedin, How to Measure Agreement, Consensus, and Polarization in Ordinal Data, SocArXiv syzbr, Center for Open Science, 2022.

[9] E. Allwein, R. Schapire, Y. Singer, Reducing multiclass to binary: a unifying approach for margin classifiers, J. Mach. Learn. Res. 1 (2001) 113–141.

[10] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. 102 (477) (2007) 359–378.

[11] V.-L. Nguyen, M.H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Mach. Learn. 111 (1) (2022) 89–122.

[12] D. Dubois, E. Hüllermeier, Comparing probability measures using possibility theory: a notion of relative peakedness, Int. J. Approx. Reason. 45 (2) (2007) 364–385.

[13] J. Schulz, R. Poyiadzi, R. Santos-Rodriguez, Uncertainty quantification of surrogate explanations: an ordinal consensus approach, arXiv preprint arXiv:2111.09121, 2021.

[14] C. Van der Eijk, Measuring agreement in ordered rating scales, Qual. Quant. 35 (2001) 325–341.

[15] N. Koudenburg, H.A. Kiers, Y. Kashima, A new opinion polarization index developed by integrating expert judgments, Front. Psychol. 12 (2021) 738258.

[16] R.K. Leik, A measure of ordinal consensus, Pac. Sociol. Rev. 9 (2) (1966) 85–90.

[17] J. Blair, M.G. Lacy, Statistics of ordinal variation, Sociol. Methods Res. 28 (3) (2000) 251–280.

[18] W.J. Tastle, M.J. Wierman, Consensus and dissention: a measure of ordinal dispersion, Int. J. Approx. Reason. 45 (3) (2007) 531–545.

[19] J. Pavlopoulos, A. Likas, Distance from unimodality for the assessment of opinion polarization, Cogn. Comput. 15 (2) (2023) 731–738.

[20] J.F.P. da Costa, H. Alonso, J.S. Cardoso, The unimodal model for the classification of ordinal data, Neural Netw. 21 (1) (2008) 78–91.

[21] C. Beckham, C. Pal, Unimodal probability distributions for deep ordinal classification, in: International Conference on Machine Learning, PMLR, 2017, pp. 411–419.

[22] R. Rifkin, A. Klautau, In defense of one-vs-all classification, J. Mach. Learn. Res. 5 (2004) 101–141.

[23] L. Li, H. Lin, Ordinal regression by extended binary classification, in: Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4–7, 2006, in: Advances in Neural Information Processing Systems, vol. 19, MIT Press, 2006, pp. 865–872.

[24] E. Frank, M.A. Hall, A simple approach to ordinal classification, in: Machine Learning: EMCL 2001, 12th European Conference on Machine Learning, Proceedings, Freiburg, Germany, September 5–7, 2001, in: Lecture Notes in Computer Science, vol. 2167, Springer, 2001, pp. 145–156.

[25] J.C. Hühn, E. Hüllermeier, Is an ordinal class structure useful in classifier learning?, Int. J. Data Min. Model. Manag. 1 (1) (2008) 45–67, https://doi.org/10.1504/IJDMMM.2008.022537.

[26] Y. Sale, P. Hofman, T. Löhr, L. Wimmer, T. Nagler, E. Hüllermeier, Label-wise aleatoric and epistemic uncertainty quantification, in: Proc. UAI, Conference on Uncertainty in Artificial Intelligence, 2024.

[27] R. Mesiar, A. Kolesárová, T. Calvo, M. Komorníková, A review of aggregation functions, in: H.B. Sola, F. Herrera, J. Montero (Eds.), Fuzzy Sets and Their Extensions: Representation, Aggregation and Models - Intelligent Systems from Decision Making to Data Mining, Web Intelligence and Computer Vision, in: Studies in Fuzziness and Soft Computing, vol. 220, Springer, 2008, pp. 121–144.

[28] V.M. Vargas, P.A. Gutiérrez, J. Barbero-Gómez, C. Hervás-Martínez, Improving the classification of extreme classes by means of loss regularisation and generalised beta distributions, CoRR, arXiv:2407.12417, 2024, https://doi.org/10.48550/ARXIV.2407.12417.

[29] R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Trans. Syst. Man Cybern. 18 (1) (1988) 183–190, https://doi.org/10.1109/21.87068.

[30] R. Shwartz-Ziv, A. Armon, Tabular data: deep learning is not all you need, Inf. Fusion 81 (2022) 84–90.

[31] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022, in: Advances in Neural Information Processing Systems, vol. 35, 2022, http://papers.nips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 3146–3154.

[33] J. de La Torre, D. Puig, A. Valls, Weighted kappa loss function for multi-class classification of ordinal data in deep learning, Pattern Recognit. Lett. 105 (2018) 144–154, https://doi.org/10.1016/J.PATREC.2017.05.018.

[34] L. Hou, C. Yu, D. Samaras, Squared earth mover's distance-based loss for training deep neural networks, CoRR, arXiv:1611.05916, 2016, arXiv:1611.05916.

[35] X. Liu, F. Fan, L. Kong, Z. Diao, W. Xie, J. Lu, J. You, Unimodal regularized neuron stick-breaking for ordinal classification, Neurocomputing 388 (2020) 34–44, https://doi.org/10.1016/J.NEUCOM.2020.01.025.

[36] T. Albuquerque, R. Cruz, J.S. Cardoso, Quasi-unimodal distributions for ordinal classification, Mathematics 10 (6) (2022) 980.

[37] J. Vanschoren, J.N. van Rijn, B. Bischl, L. Torgo, OpenML: networked science in machine learning, SIGKDD Explor. 15 (2) (2013) 49–60, https://doi.org/10.1145/2641190.2641198.

[38] D. Dua, C. Graff, UCI machine learning repository, http://archive.ics.uci.edu/ml, 2017.

[39] A. Malinin, B. Mlodozeniec, M. Gales, Ensemble distribution distillation, arXiv preprint arXiv:1905.00076, 2019.

[40] M.S.A. Nadeem, J. Zucker, B. Hanczar, Accuracy-rejection curves (arcs) for comparing classification methods with a reject option, in: S. Dzeroski, P. Geurts, J. Rousu (Eds.), Proceedings of the Third International Workshop on Machine Learning in Systems Biology, MLSB 2009, Ljubljana, Slovenia, September 5–6, 2009, in: JMLR Proceedings, vol. 8, JMLR.org, 2010, pp. 65–81, http://proceedings.mlr.press/v8/nadeem10a.html.

[41] J.C. Huhn, E. Hüllermeier, FR3: a fuzzy rule learner for inducing reliable classifiers, IEEE Trans. Fuzzy Syst. 17 (1) (2009) 138–149, https://doi.org/10.1109/TFUZZ.2008.2005490.

[42] P. Lahoti, K. Gummadi, G. Weikum, Responsible model deployment via model-agnostic uncertainty learning, Mach. Learn. 112 (3) (2023) 939–970.

[43] L. Gaudette, N. Japkowicz, Evaluation methods for ordinal classification, in: Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009, Proceedings 22, Kelowna, Canada, May 25–27, 2009, Springer, May 2009, pp. 207–210.

[44] A.E. Yilmaz, H. Demirhan, Weighted kappa measures for ordinal multi-class classification performance, Appl. Soft Comput. 134 (2023) 110020, https://doi.org/10.1016/J.ASOC.2023.110020.
[45] V.M. Vargas, P.A. Gutiérrez, C. Hervas-Martinez, Cumulative link models for deep ordinal classification, Neurocomputing 401 (2020) 48–58.
[46] R. Rosati, L. Romeo, V.M. Vargas, P.A. Gutiérrez, C. Hervás-Martínez, E. Frontoni, A novel deep ordinal classification approach for aesthetic quality control classification, Neural Comput. Appl. 34 (14) (2022) 11625–11639.
[47] F. Castagnos, M. Mihelich, C. Dognin, A simple log-based loss function for ordinal text classification, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12–17, 2022, in: International Committee on Computational Linguistics, 2022, pp. 4604–4609.
[48] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, P.A. Gutiérrez, Metrics to guide a multi-objective evolutionary algorithm for ordinal classification, Neurocomputing 135 (2014) 21–31, https://doi.org/10.1016/J.NEUCOM.2013.05.058.
[49] S. Haas, E. Hüllermeier, Rectifying bias in ordinal observational data using unimodal label smoothing, in: G.D.F. Morales, C. Perlich, N. Ruchansky, N. Kourtellis, E. Baralis, F. Bonchi (Eds.), Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track - European Conference, ECML PKDD 2023, Proceedings, Part VI, Turin, Italy, September 18–22, 2023, in: Lecture Notes in Computer Science, vol. 14174, Springer, 2023, pp. 3–18.
[50] S. Baccianella, A. Esuli, F. Sebastiani, Evaluation measures for ordinal regression, in: Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy, November 30–December 2, 2009, IEEE Computer Society, 2009, pp. 283–287.
[51] A. Malinin, Uncertainty estimation in deep learning with application to spoken language assessment, Ph.D. thesis, 2019.
[52] S.B. Kotsiantis, P.E. Pintelas, A cost sensitive technique for ordinal classification problems, in: Methods and Applications of Artificial Intelligence, Third Helenic Conference on AI, SETN 200, Proceedings, Samos, Greece, May 5–8, 2004, in: Lecture Notes in Computer Science, vol. 3025, Springer, 2004, pp. 220–229.
[53] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[54] A. Benavoli, G. Corani, F. Mangili, Should we really use post-hoc tests based on mean-ranks?, J. Mach. Learn. Res. 17 (1) (2016) 152–161.
[55] S. Haas, E. Hüllermeier, A prescriptive machine learning approach for assessing goodwill in the automotive domain, in: M. Amini, S. Canu, A. Fischer, T. Guns, P.K. Novak, G. Tsoumakas (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Proceedings, Part VI, France, September 19–23, 2022, in: Lecture Notes in Computer Science, vol. 13718, Springer, Grenoble, 2022, pp. 170–184.
[56] S. Kramer, G. Widmer, B. Pfahringer, M. de Groeve, Prediction of ordinal classes using regression trees, Fundam. Inform. 47 (1–2) (2001) 1–13.
[57] J. Cheng, Z. Wang, G. Pollastri, A neural network approach to ordinal regression, in: Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, Part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1–6, 2008, IEEE, 2008, pp. 1279–1284.
[58] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016, ACM, 2016, pp. 785–794.
[59] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506, https://doi.org/10.1007/S10994-021-05946-3.
[60] T. de Menezes e Silva Filho, H. Song, M. Perelló-Nieto, R. Santos-Rodríguez, M. Kull, P.A. Flach, Classifier calibration: a survey on how to assess and improve predicted class probabilities, Mach. Learn. 112 (9) (2023) 3211–3260, https://doi.org/10.1007/S10994-023-06336-7.
[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[62] F. Bérchez-Moreno, V.M. Vargas, R. Ayllón-Gavilán, D. Guijo-Rubio, C. Hervás-Martínez, J.C. Fernández, P.A. Gutiérrez, dlordinal: a python package for deep ordinal classification, CoRR, arXiv:2407.17163, 2024, https://doi.org/10.48550/ARXIV.2407.17163.
[63] M. Tietz, T.J. Fan, D. Nouri, B. Bossan, Skorch developers, skorch: a scikit-learn compatible neural network library that wraps PyTorch, https://skorch.readthedocs.io/en/stable/, Jul. 2017.
[64] V.M. Vargas, P.A. Gutiérrez, J. Barbero-Gómez, C. Hervás-Martínez, Soft labelling based on triangular distributions for ordinal classification, Inf. Fusion 93 (2023) 258–267, https://doi.org/10.1016/J.INFFUS.2023.01.003.
[65] V.M. Vargas, P.A. Gutiérrez, C. Hervás-Martínez, Unimodal regularisation based on beta distribution for deep ordinal regression, Pattern Recognit. 122 (2022) 108310, https://doi.org/10.1016/J.PATCOG.2021.108310.