Reliable Conformal Prediction for Ordinal Classification Using the Ranked Probability Score

Stefan Haas[1,3] **Luca Killmaier**[1] **Alireza Javanmardi**[1,2] **Eyke Hüllermeier**[1,2,4]

[1]Institute of Informatics, LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany,
[3]BMW Group, Munich, Germany
[4]German Centre for Artificial Intelligence (DFKI, DSA), Kaiserslautern, Germany

Abstract

Ordinal classification (OC) arises in high-stakes domains such as medicine and finance, where uncertainty quantification must account for the severity of ordinal errors. Conformal prediction (CP) provides distribution-free prediction sets with marginal coverage guarantees; however, its practical effectiveness depends critically on the choice of nonconformity function. We introduce a CP method for ordinal classification based on the ranked probability score (RPS), a proper scoring rule defined over cumulative predictive distributions. Although it reflects ordinal risk quite naturally, it has largely been neglected in conformal ordinal prediction (COP). When used as a measure of nonconformity, RPS yields median-centered contiguous prediction sets by construction. The method is model-agnostic, supports both assessed and grouped ordered categorical outcomes, and permits efficient implementation compared to greedy interval selection procedures. Across multiple ordinal image and tabular datasets, RPS-based CP produces contiguous prediction sets and strikes a favorable balance between prediction set width and the magnitude of ordinal miscoverage relative to existing CP methods.

1 INTRODUCTION

Ordinal classification (OC), also known as ordinal regression in statistics [McCullagh, 1980], refers to classification problems in which class labels exhibit a natural linear order. Representative applications include medical diagnosis [Albuquerque et al., 2021], age estimation [Cao et al., 2020], and credit risk assessment [Hirk et al., 2019]. Despite its prevalence in high-stakes domains, most work in OC has primarily focused on improving point prediction performance [Shi et al., 2023, Nachmani et al., 2025, Polat et al., 2025], whereas uncertainty quantification (UQ) has attracted attention only recently [Haas and Hüllermeier, 2025, 2026]. Here, uncertainty for a query $\boldsymbol{x}_q$ is typically represented by the conditional predictive distribution over ordered labels, $p(y \mid \boldsymbol{x}_q)$, produced by a probabilistic predictor.

Alternatively, uncertainty can be represented by set-valued predictions. Conformal prediction (CP) offers a principled, model-agnostic framework for post-hoc construction of prediction sets [Shafer and Vovk, 2008, Vovk et al., 2005, 1999, Angelopoulos and Bates, 2021]. It calibrates an underlying (heuristic) uncertainty estimate to achieve finite-sample, distribution-free marginal coverage at a user-specified miscoverage rate α. Instead of committing to a single label $y \in \mathcal{Y}$, CP outputs a set $\mathcal{C}_\alpha(\boldsymbol{x}_q) \subseteq \mathcal{Y}$ of plausible labels, whose size reflects predictive uncertainty at the query $\boldsymbol{x}_q$. However, producing sets in conformal ordinal prediction (COP) that are both informative and consistent with the ordinal structure of the label space remains challenging.

A key requirement for COP is that prediction sets be contiguous [Lu et al., 2022, Xu et al., 2023, Dey et al., 2023]. For example, consider age estimation [Shi et al., 2023, Yun et al., 2024] from images, where a latent continuous variable (e.g., age) is discretized into ordered categories, with $\mathcal{Y} = \{\texttt{baby}, \texttt{child}, \texttt{teenager}, \texttt{adult}, \texttt{senior}\}$. A valid prediction set should only contain adjacent age categories, e.g., $\mathcal{C}_\alpha(\boldsymbol{x}_q) = \{\texttt{child}, \texttt{teenager}, \texttt{adult}\}$, whereas non-contiguous sets such as $\mathcal{C}_\alpha(\boldsymbol{x}_q) = \{\texttt{child}, \texttt{senior}\}$ may appear unreasonable.

Another important category, alongside the previously described *grouped* ordered categorical variables, is the class of *assessed* ordered categorical variables [Anderson, 1984], in which human experts assign labels, as in financial risk assessment or medical survival prognosis. In these contexts, errors tend to be inherently larger due to inter-expert disagreement, which makes maintaining contiguity even more crucial for accurately capturing uncertainty. For example, consider physician opinions regarding survival

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1882–1912.

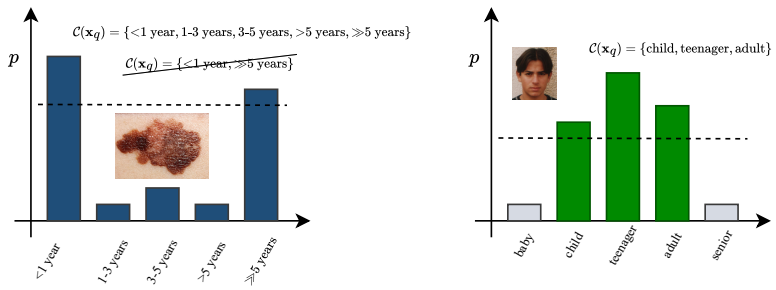


Figure 1: (Left) Illustration of an assessed ordered categorical variable (survival prognosis for melanoma [National Cancer Institute, 1985]) with extreme disagreement between physicians. To faithfully quantify uncertainty, the prediction set must be contiguous and include all intermediate classes between the conflicting assessments. (Right) Illustration of a grouped ordered categorical variable (age estimation), where unimodal predictive modeling is well justified and naturally leads to contiguous prediction sets [Lanitis et al., 2002, Panis et al., 2016].

prognosis for stage IV melanoma, which may be polarized, resulting in large clusters at $\texttt{<1 year}$ (very pessimistic group) and $\gg\texttt{5 years}$ (optimistic group influenced by immunotherapy outcomes). To properly quantify uncertainty in such cases, the prediction set should not be limited to the two most frequent categories, i.e., $\mathcal{C}_\alpha(\boldsymbol{x}_q) = \{\texttt{<1 year}, \gg\texttt{5 years}\}$; instead, it should encompass the entire range of plausible categories, $\mathcal{C}_\alpha(\boldsymbol{x}_q) = \{\texttt{<1 year}, \texttt{1–3 years}, \texttt{3–5 years}, \texttt{>5 years}, \gg\texttt{5 years}\}$. Otherwise, uncertainty measured by the size of the prediction set will be severely underestimated (see Figure 1).

Another important aspect of COP that has received limited attention is the *severity of miscoverage* when the true label lies outside the prediction set. For instance, in financial risk assessment, if the true label is $y = \texttt{very high}$ while $\mathcal{C}_\alpha(x) = \{\texttt{low}, \texttt{moderate}\}$, the resulting miscoverage is substantial and could lead to catastrophic risk misestimation. Ideally, when coverage fails, the true label should lie as close as possible to the boundary of the prediction set, minimizing the impact of miscoverage. Under this criterion, a larger set such as $\mathcal{C}_\alpha(x) = \{\texttt{low}, \texttt{moderate}, \texttt{high}\}$ may be preferable to a smaller set $\mathcal{C}_\alpha(x) = \{\texttt{low}, \texttt{moderate}\}$, even though both fail to cover the true label and the larger set is less efficient in the classical CP sense. This observation motivates uncertainty quantification methods for OC that account not only for coverage and efficiency, but also for the ordinal distance incurred under miscoverage.

These challenges motivate a model-agnostic conformal approach to OC that can be combined with arbitrary loss functions, accommodates both grouped and assessed ordered targets, and produces meaningful contiguous prediction sets that accurately quantify uncertainty. Moreover, such a method should leverage the entire predictive probability distribution in an unbiased and computationally efficient manner when constructing prediction sets. These requirements are not met by existing approaches (cf. Section 2).

In this paper, we advocate the ranked probability score (RPS) [Epstein, 1969] as a nonconformity measure for conformal prediction in OC. RPS is a proper scoring rule [Gneiting and Raftery, 2007] for ordinal outcomes that incentivizes truthful probability estimation and explicitly accounts for the linear structure of the label space. Despite being well-established in the forecasting literature [Murphy, 1970, 1971], it has only recently been recognized as a theoretically grounded metric for evaluating probabilistic ordinal classifiers in machine learning [Galdran, 2023]. To the best of our knowledge, RPS has not yet been proposed as a nonconformity measure for CP in OC. The main contributions of this paper are as follows:

- We propose RPS as a proper, model-agnostic nonconformity measure for CP in OC.
- We provide theoretical guarantees for desirable properties in OC: RPS-based conformal prediction sets (i) **satisfy marginal coverage**, (ii) **are nested** with respect to the miscoverage level (α), and (iii) **are contiguous**.
- Furthermore, we show that RPS-based prediction sets directly **optimize ordinal risk under oracle conditional coverage**, measured as set-based l_1 error, in contrast to mode-centered approaches which primarily target set efficiency.
- We show that RPS-based conformal prediction is computationally efficient, scaling linearly with both the number of labels and the number of calibration points.
- Finally, we empirically validate our approach on ordinal image and tabular datasets, showing that median-centered RPS-based prediction sets strike a favorable balance between interval width and ordinal miscoverage magnitude.

2 RELATED WORK

Ordinal Classification addresses the problem of predicting discrete ordered labels as commonly encountered in

many high-stakes domains, including medicine [Dorado-Moreno et al., 2017, Prodeau et al., 2019, Tariq et al., 2025] and finance [Hirk et al., 2019]. Unlike multinomial classification, where class labels are unordered, OC must account for the inherent order among classes, implying that misclassification costs typically increase with an increasing gap between predicted label $\hat{y}$ and true label y. At the same time, OC differs from regression in that the labels are discrete rather than continuous, and the underlying measurement scale is ordinal rather than cardinal. Thus, strictly speaking, there is no natural notion of distance. In spite of this, encoding the class labels by integers $1, \ldots, K$ and using distance-based losses such as $|\hat{y} - y|$ is common practice.

Recent work in OC has largely focused on improving predictive performance, often by minimizing distance-based losses such as mean absolute error [Gaudette and Japkowicz, 2009] or quadratic weighted kappa [Cohen, 1968]. Existing approaches can be broadly categorized into (i) *unimodal soft-labeling methods* [Diaz and Marathe, 2019, Liu et al., 2020, Haas and Hüllermeier, 2023, Vargas et al., 2022, 2023a,b], (ii) *ordinal loss functions* [Hou et al., 2016, de La Torre et al., 2018, Castagnos et al., 2022, Albuquerque et al., 2022, Nachmani et al., 2025, Polat et al., 2025], and (iii) *explicit unimodality constraints* [da Costa et al., 2008, Beckham and Pal, 2017, Dey et al., 2023, Cardoso et al., 2025].

Conformal Prediction is a framework that can be applied on top of any base model to produce *prediction sets* (or *intervals* in regression) instead of point predictions [Shafer and Vovk, 2008, Vovk et al., 2005, 1999]. These sets are guaranteed to contain the true label with a user-specified *marginal coverage probability*. Inductive conformal prediction [Papadopoulos et al., 2002, Papadopoulos, 2008], also known as split conformal prediction, has become the standard approach in practice due to its computational efficiency. CP has been extensively studied for classification [Sadinle et al., 2019, Romano et al., 2020] and regression [Romano et al., 2019], where it provides finite-sample, distribution-free coverage guarantees.

More recently, conformal prediction for ordinal classification has attracted increasing attention. Lu et al. [2022] and Zhang et al. [2026] construct contiguous prediction sets by expanding outward from the mode of the predictive distribution, performing a greedy search for a threshold that ensures marginal coverage while aiming to keep the sets as small as possible. Xu et al. [2023] formulate ordinal conformal prediction within the conformal risk control framework [Angelopoulos et al., 2024], pursuing essentially the same goals. A different approach is taken by Dey et al. [2023], who enforce unimodal predictive distributions, enabling the reuse of existing conformal methods such as least ambiguous set-valued classifiers (LAC) [Sadinle et al., 2019] and adaptive prediction sets (APS) [Romano et al., 2020], while guaranteeing contiguity. However, unimodality is a strong bias that is not always warranted and may negatively impact unbiased UQ in OC [Haas and Hüllermeier, 2025, 2026].

In contrast to these approaches, we propose an efficient, model-agnostic conformal method that leverages the full predictive distribution through a principled proper scoring rule. This method guarantees median-centered, contiguous prediction sets without relying on greedy or search-based procedures, mode-centered constructions, or unimodality assumptions, while faithfully respecting the ordinal structure of the label space and minimizing ordinal risk under oracle conditional coverage.

3 METHOD

3.1 PROBLEM FORMULATION

Consider a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, drawn from an underlying distribution $\mathcal{P}$ over the joint input-output space $\mathcal{X} \times \mathcal{Y}$. We focus on the ordinal classification setting, where the output space $\mathcal{Y} = \{y_1, \ldots, y_K\}$ consists of a finite set of class labels endowed with a natural (linear) order $y_1 \prec y_2 \prec \cdots \prec y_K$.

Let (X_{n+1}, Y_{n+1}) denote a test instance such that the augmented sample $\mathcal{D} \cup \{(X_{n+1}, Y_{n+1})\}$ is exchangeable. Assuming that the test label Y_{n+1} is unobserved, the goal of conformal prediction is to construct a prediction set $\mathcal{C}_\alpha(X_{n+1}) \subseteq \mathcal{Y}$ satisfying a marginal coverage guarantee.

Definition 3.1 (Marginal coverage)**.** *A conformal prediction procedure satisfies marginal coverage at level* $1 - \alpha$ *if the prediction set* $\mathcal{C}(X_{n+1})$ *it outputs satisfies*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1})\right) \geq 1 - \alpha, \tag{1}$$

where $\alpha \in (0, 1)$ *is a user-specified error rate and the probability is taken with respect to the joint distribution* $\mathcal{P}$ *and any randomness in the construction of* $\mathcal{C}$*.*

Distribution-free CP methods cannot generally guarantee instance-wise conditional coverage [Vovk, 2013], a strictly stronger requirement than marginal coverage.

Definition 3.2 (Conditional coverage)**.** *A conformal prediction procedure satisfies conditional coverage at level* $1 - \alpha$ *if for all* X_{n+1}*, the set* $\mathcal{C}(X_{n+1})$ *it outputs satisfies*

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}\right) \geq 1 - \alpha, \tag{2}$$

where $\alpha \in (0, 1)$ *is a user-specified error rate and the probability is taken with respect to the conditional distribution induced by* $\mathcal{P}$*.*

Conformal prediction constructs prediction sets through a *nonconformity score*, which quantifies how incompatible a candidate label is with a given input relative to a predictive model. Formally, a nonconformity score is a function

$s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where larger values indicate greater nonconformity between an input–label pair $(\boldsymbol{x}, y)$ and the model's predictive behavior. In this work, we consider probabilistic predictors $h : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$, which output a predictive probability vector $\boldsymbol{p} = (p(y_1), \ldots, p(y_K)) = (p_1, \ldots, p_K) \in \mathbb{P}(\mathcal{Y})$, where $p(y_k)$ denotes the predictive probability assigned to class y_k. Nonconformity scores are then derived from this predictive distribution to quantify the incompatibility of candidate labels with the model's predictions.

In this work, we adopt the inductive (or split) conformal prediction framework [Papadopoulos et al., 2002, Papadopoulos, 2008]. The dataset $\mathcal{D}$ is partitioned into a proper training set $\mathcal{D}_{\text{train}}$, used to train a predictive model h, and a calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$, used to compute nonconformity scores. Given a test input X_{n+1} and a candidate label $y_k \in \mathcal{Y}$, the conformal prediction set is defined as

$$\mathcal{C}_\alpha(X_{n+1}) = \{y \in \mathcal{Y} \ : \ s(X_{n+1}, y) \leq \hat{q}_{1-\alpha}\}, \quad (3)$$

where $\hat{q}_{1-\alpha}$ denotes the empirical $(1-\alpha)$-quantile of the nonconformity scores computed on the calibration set. This construction guarantees marginal coverage at level $1-\alpha$ under the exchangeability assumption on $\mathcal{D}_{\text{cal}} \cup \{(X_{n+1}, Y_{n+1})\}$ [Vovk et al., 2005, Shafer and Vovk, 2008]. Naturally, the choice of the nonconformity score plays a central role in conformal prediction, as it determines not only calibration but also how informative the resulting prediction sets are, particularly in structured output spaces such as ordinal classification.

3.2 AN ORDINAL ORACLE METHOD

Assume the true data-generating distribution $\mathcal{P}$ is known, which would in principle allow construction of prediction sets satisfying conditional coverage (Definition 3.2). In the ordinal classification setting, an additional common objective of conformal prediction is to construct the *smallest contiguous* prediction set achieving conditional coverage, thereby balancing coverage with efficiency [Lu et al., 2022, Xu et al., 2023, Zhang et al., 2026].

Formally, with contiguous intervals $[l, u] := \{l, l+1, \ldots, u\}$ and interval probabilities $p^*_{l,u}(X) := \sum_{j=l}^{u} p^*(y_j \mid X)$, the optimal contiguous prediction set for an instance X can be written as

$$\mathcal{C}^*_\alpha(X) = \underset{[l,u]:1\leq l\leq u\leq K}{\arg\min} \left\{ u - l \ : \ p^*_{l,u}(X) \geq 1-\alpha \right\} \quad (4)$$

Minimizing the length of the prediction set balances efficiency with conditional coverage, producing the most efficient contiguous sets possible. Contiguity respects the ordinal structure of $\mathcal{Y}$ and ensures that no gaps exist in the predicted set of labels. This oracle construction is not available in practice, as the true conditional distribution $p^*(y \mid \boldsymbol{x})$ is unknown. Notably, this oracle minimizes set length but does not explicitly account for ordinal risk, such as expected l_1 deviation from the true label.

3.3 CONFORMAL ORDINAL PREDICTION

To approximate the oracle construction (4) in practice, existing approaches [Lu et al., 2022, Zhang et al., 2026] aim to identify a threshold λ through greedy search that satisfies the marginal coverage guarantee (Definition 3.1) while producing efficient contiguous prediction sets.

$$\mathcal{C}_\lambda(X) = \{y_j \in \mathcal{Y} : l(X;\lambda) \leq j \leq u(X;\lambda)\},$$
$$\big(l(X;\lambda), u(X;\lambda)\big) = \underset{1\leq l\leq u\leq K}{\arg\min} \Big\{ u - l \ : \ p_{l,u}(X) \geq \lambda \Big\}. \quad (5)$$

To ensure marginal coverage, the threshold λ is selected using the calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$ as the smallest value satisfying

$$\sum_{i=1}^{n} \mathbb{1}\{Y_i \in \mathcal{C}_\lambda(X_i)\} \ \geq \ \lceil (1-\alpha)(n+1) \rceil. \quad (6)$$

While the procedure guarantees marginal coverage, relying on a single global threshold λ ignores the geometric structure of the ordinal predictive distribution and instead primarily controls set size, i.e., efficiency. As a consequence, the resulting prediction sets may be efficient in terms of cardinality, yet limited in faithfulness with respect to uncertainty quantification and ordinal risk.

Similarly, the approach of [Dey et al., 2023] constructs prediction sets by growing them outward from the mode of a unimodal predictive distribution.

Definition 3.3 (Unimodality [Keilson and Gerber, 1971])**.** *A discrete probability distribution $\boldsymbol{p}$ is* unimodal *if there exists at least one index m, called the mode, such that*

$$p_k \geq p_{k-1}, \quad \textit{for all } k \leq m,$$
$$p_{k+1} \leq p_k, \quad \textit{for all } k \geq m.$$

The assumption of unimodal predictive distributions is a common inductive bias in ordinal classification [da Costa et al., 2008, Beckham and Pal, 2017]. Restricting the output distributions of a predictor to be unimodal allows one to obtain contiguous prediction sets using standard nominal nonconformity scores, such as LAC or APS [Dey et al., 2023]. Nonetheless, both methods are driven by probability magnitude rather than ordinal geometry, and thus may insufficiently reflect uncertainty in low-probability tails of unimodal predictive distributions.

3.4 THE RANKED PROBABILITY SCORE (RPS)

To address these limitations, we propose the use of the *Ranked Probability Score (RPS)* [Epstein, 1969] as a non-

conformity score for conformal prediction in ordinal classification. RPS is a *proper scoring rule* [Gneiting and Raftery, 2007, Murphy, 1969] defined as follows.

Definition 3.4 (Proper scoring rule)**.** *A scoring rule* $s : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \to \mathbb{R}$ *is* proper *if the expected score is minimized when the predicted distribution* $\boldsymbol{p}$ *equals the true distribution* $\boldsymbol{p}^*$*, that is,*

$$\mathbb{E}_{Y\sim\boldsymbol{p}^*}\big[s(Y,\boldsymbol{p}^*)\big] \;\le\; \mathbb{E}_{Y\sim\boldsymbol{p}^*}\big[s(Y,\boldsymbol{p})\big] \quad \forall\, \boldsymbol{p} \in \mathbb{P}(\mathcal{Y}).$$

It is strictly proper *if equality holds if and only if* $\boldsymbol{p} = \boldsymbol{p}^*$.

The RPS can be viewed as the Brier score [Brier, 1950] applied to cumulative predictive probabilities, making it sensitive to the ordinal distances between classes [Jose et al., 2009]. For $K = 2$, it reduces exactly to the standard Brier score for binary outcomes.

Let $F_X(k) = \sum_{j=1}^{k} p(y_j \mid X)$ denote the predicted cumulative distribution function (CDF), and let $\mathbb{1}\{Y \le y_k\}$ be the corresponding cumulative indicator of the true label. The RPS is then defined as

$$\mathrm{RPS}(X,Y) = \frac{1}{K-1}\sum_{k=1}^{K-1}\left(F_X(k) - \mathbb{1}\{Y \le y_k\}\right)^2, \tag{7}$$

where the factor $\frac{1}{K-1}$ normalizes the score to lie in the interval $[0,1]$, making it independent of the number of classes K.

Unlike previous ordinal conformal methods, RPS leverages the *entire predictive distribution* rather than mode-centric summaries, yielding prediction sets that better adapt to input-dependent uncertainty. By construction, it provides a natural measure of nonconformity for ordinal candidate labels with respect to the full predictive distribution: it attains its maximum when the predicted mass is concentrated at the opposite end of the ordinal scale from the true label, and its minimum when the predicted mass is concentrated on the true class. As a strictly proper scoring rule for ordinal outcomes [Murphy, 1969], RPS is theoretically grounded and encourages calibrated probability forecasts. Its properness has recently attracted renewed attention for evaluating probabilistic ordinal classifiers [Galdran, 2023], despite its long-standing popularity in forecasting [Murphy, 1970, 1971].

3.5 VALIDITY OF RPS FOR ORDINAL CP

In this section, we establish the validity of the RPS as a nonconformity measure within the conformal prediction framework. Let $s_{\mathrm{RPS}}(X,Y) = \mathrm{RPS}(X,Y)$ denote the nonconformity score derived from the RPS. A key requirement for using RPS in conformal prediction is that the resulting sets satisfy marginal coverage (Definition 3.1).

Algorithm 1 Prediction Sets via RPS

Input: $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$, significance level α, new instance X_{n+1}
Output: Prediction set $\mathcal{C}_\alpha(X_{n+1})$
$s_i \leftarrow s_{\mathrm{RPS}}(X_i, Y_i), \quad i = 1, \dots, n$
$\hat{q}_{1-\alpha} \leftarrow \mathrm{quantile}_{(1-\alpha)(1+\frac{1}{n})}\big(\{s_i\}_{i=1}^{n}\big)$
$\mathcal{C}_\alpha(X_{n+1}) \leftarrow \{y \in \mathcal{Y} : s_{\mathrm{RPS}}(X_{n+1}, y) \le \hat{q}_{1-\alpha}\}$
return $\mathcal{C}_\alpha(X_{n+1})$

Proposition 3.1 (Marginal coverage guarantee of RPS-based sets)**.** *Under the exchangeability assumption on* $\mathcal{D}_{\mathrm{cal}} \cup \{(X_{n+1}, Y_{n+1})\}$*, the RPS-based conformal procedure satisfies*

$$\mathbb{P}\big(Y_{n+1} \in \mathcal{C}(X_{n+1})\big) \ge 1 - \alpha,$$

regardless of the underlying predictive model.

Proof Sketch. This result follows directly from the general theory of inductive conformal prediction [Shafer and Vovk, 2008, Vovk et al., 2005]. Since s_{RPS} is a real-valued nonconformity score and the calibration scores are exchangeable with the test score, the standard quantile-based construction (3) guarantees marginal coverage [Vovk et al., 2005, Shafer and Vovk, 2008]. □

Another desirable property for ordinal prediction sets is *nesting* with respect to α [Lu et al., 2022, Xu et al., 2023, Zhang et al., 2026], e.g., $\mathcal{C}_{\alpha_2}(X) = \{\texttt{baby}, \texttt{child}\} \subseteq \mathcal{C}_{\alpha_1}(X) = \{\texttt{baby}, \texttt{child}, \texttt{teenager}\}$ for $0 < \alpha_1 \le \alpha_2 < 1$. While RPS sets are always nested regardless of the predictive distribution, mode-based sets (5) are nested only when the predictive density is radially monotone, i.e., when it decreases monotonically from the mode along every direction [Zhang et al., 2026].

Proposition 3.2 (Nestedness of RPS-based sets in α)**.** *For any input* X *and any* $0 < \alpha_1 \le \alpha_2 < 1$*, the RPS-based conformal prediction sets satisfy* $\mathcal{C}_{\alpha_2}(X) \subseteq \mathcal{C}_{\alpha_1}(X)$.

Proof. If $\alpha_1 \le \alpha_2$, then $1 - \alpha_1 \ge 1 - \alpha_2$, which implies $\hat{q}_{1-\alpha_1} \ge \hat{q}_{1-\alpha_2}$. Hence, any label y such that $s_{\mathrm{RPS}}(X,y) \le \hat{q}_{1-\alpha_2}$ also satisfies $s_{\mathrm{RPS}}(X,y) \le \hat{q}_{1-\alpha_1}$, and therefore $\mathcal{C}_{\alpha_2}(X) \subseteq \mathcal{C}_{\alpha_1}(X)$. □

Furthermore, ordinal prediction sets must be contiguous along the label ordering, consistent with the oracle definition in (4) [Xu et al., 2023, Dey et al., 2023].

Theorem 3.1 (Contiguity of RPS-based sets)**.** *Let* $\mathcal{Y} = \{y_1 \prec y_2 \prec \cdots \prec y_K\}$ *be a set of ordered labels, and let* s_{RPS} *denote the RPS-based nonconformity score. Then, for any input* X *and any miscoverage level* $\alpha \in (0,1)$*, the conformal prediction set* $\mathcal{C}_\alpha^{\mathrm{RPS}}(X) = \{y \in \mathcal{Y} : s_{\mathrm{RPS}}(X,y) \le \hat{q}_{1-\alpha}\}$ *forms a contiguous interval of labels centered at the*

median m, i.e., there exist integers $1 \leq l \leq m \leq u \leq K$ such that $\mathcal{C}_\alpha(X) = \{y_l, \ldots, y_m, \ldots, y_u\}$.

Unlike existing nonconformity scores for nominal classification such as LAC [Sadinle et al., 2019] or APS [Romano et al., 2020], RPS-based prediction sets are contiguous by design. This contiguity follows directly from the ordinal structure and the monotonicity of the cumulative distribution function, and does not require imposing unimodality assumptions on the predictive probability mass function [Dey et al., 2023], which may be unwarranted for ordinal targets. A formal proof is provided in Appendix A. Together with marginal validity and nesting in α, these properties ensure that RPS-based conformal prediction produces prediction sets well suited for ordinal tasks.

Moreover, unlike *mode*-centered approaches, RPS-based prediction sets are *median*-centered (see Appendix A), thereby balancing the cumulative probability mass above and below the center. This property promotes robustness under skewed or heavy-tailed distributions; sublevel sets expand by minimizing the imbalance between the lower and upper predictive tails, rather than expanding outwards from a single high-probability mode. While this characteristic does not guarantee global risk optimality for the full conformal procedure, defined in terms of the set-based l_1 error (8), it does imply instance-level risk optimality at the level of the nonconformity scores. Furthermore, this instance-level optimality scales to full risk optimality within a conditional coverage oracle setting, where the true conditional distribution is known, causing pointwise risk minimization to aggregate directly into prediction sets that globally minimize the expected l_1 error.

Theorem 3.2 (Ordinal risk optimality of RPS-based median-grown prediction sets under oracle conditional coverage)**.** *For a fixed input X, let $\mathcal{C}_{\mathrm{RPS}}(X) = \{y_l, \ldots, y_m, \ldots, y_u\}$ denote the contiguous RPS-based prediction set, which is grown from a median index m as in Theorem 3.1. Define the* ordinal risk *of a set $\mathcal{C}(X)$ as the expected l_1-distance of true label from this set:*

$$R(\mathcal{C}(X)) := \sum_{y=1}^{K} p(y \mid X) \min_{c \in \mathcal{C}(X)} |y - c|. \tag{8}$$

Let $\mathcal{C}(X)$ be any other contiguous *conformal set of minimal cardinality satisfying the same coverage constraint as $\mathcal{C}_{\mathrm{RPS}}(X)$. Then*

$$R(\mathcal{C}_{\mathrm{RPS}}(X)) \leq R(\mathcal{C}(X)). \tag{9}$$

A formal proof is provided in Appendix B. RPS-based sets, directly target ordinal risk reduction under oracle conditional coverage by sequentially adding adjacent labels that minimize risk, starting from the singleton risk-minimizing set containing only the median. This contrasts with mode-centered procedures, which prioritize set-size efficiency and may overlook the ordinal structure of the labels (Section 3.2).

As illustrative examples, consider the unimodal distribution $\boldsymbol{p}_{\mathrm{umod}} = (0.06, 0.24, 0.32, 0.20, 0.18)$ and the multimodal distribution $\boldsymbol{p}_{\mathrm{multimod}} = (0.09, 0.12, 0.40, 0.04, 0.35)$, both with median $m = 3$, coinciding with the mode. Tables 1 and 2 compare step-wise set expansions based on RPS with greedy mode-based expansion (e.g., min-CPS). The greedy strategy expands from the mode by iteratively adding the class with the largest remaining probability, thereby maximizing local probability mass for a given set size. In contrast, the RPS-based expansion accounts for cumulative distance-weighted deviations and directly minimizes ordinal risk (8). While the differences are moderate in the unimodal case, they become substantial for the multimodal distribution, where heavy tail mass causes the greedy procedure to underestimate ordinal risk.

#	RPS-based	Risk	Mode-grown	Risk
1	{3}	0.92	{3}	0.92
2	{3, 4}	**0.54**	{2, 3}	0.62
3	{2, 3, 4}	0.24	{2, 3, 4}	0.24
4	{2, 3, 4, 5}	0.06	{2, 3, 4, 5}	0.06
5	{1, 2, 3, 4, 5}	0.00	{1, 2, 3, 4, 5}	0.00

Table 1: Step-wise comparison of ordinal risk for median RPS-based versus greedy mode-based set expansions for an exemplary unimodal distribution $\boldsymbol{p}_{\mathrm{umod}} = (0.06, 0.24, 0.32, 0.20, 0.18)$. Lower ordinal risk (8) is highlighted.

#	RPS-based	Risk	Mode-grown	Risk
1	{3}	1.04	{3}	1.04
2	{3, 4}	**0.65**	{2, 3}	0.83
3	{3, 4, 5}	**0.30**	{1, 2, 3}	0.74
4	{2, 3, 4, 5}	**0.09**	{1, 2, 3, 4}	0.35
5	{1, 2, 3, 4, 5}	0.00	{1, 2, 3, 4, 5}	0.00

Table 2: Step-wise comparison of ordinal risk for median RPS-based versus greedy mode-based set expansions for a multimodal distribution $\boldsymbol{p}_{\mathrm{multimod}} = (0.09, 0.12, 0.40, 0.04, 0.35)$. Lower ordinal risk (8) is highlighted.

3.6 COMPUTATIONAL COMPLEXITY OF RPS

During conformal **calibration**, we only need to compute the RPS score $s_{\mathrm{RPS}}(X_i, Y_i)$ for the true label Y_i of each calibration point X_i. This requires a single RPS evaluation per point, each taking $\mathcal{O}(K)$ time to compute the cumulative distribution $F_X(k)$ and the score. Hence, the total cost for the calibration dataset $\mathcal{D}_{\mathrm{cal}}$ is $\mathcal{O}(nK)$.

During conformal **inference**, computing a prediction set for a new input requires evaluating $s_{\mathrm{RPS}}(X, y_\ell)$ for all K candidate labels y_ℓ. Naively, this requires $\mathcal{O}(K^2)$ time. However, after computing $s_{\mathrm{RPS}}(X, y_1)$ once, we can exploit the exact

recurrence

$$s_{\mathrm{RPS}}(X, y_{\ell+1}) = s_{\mathrm{RPS}}(X, y_\ell) + \frac{2F(\ell) - 1}{K - 1},$$

to compute all K scores in linear time $\mathcal{O}(K)$ (see Appendix A). Specifically, the initial score $s_{\mathrm{RPS}}(X, y_1)$ is computed in $\mathcal{O}(K)$ time, and each subsequent score is obtained with a constant-time update, i.e., $\mathcal{O}(1)$ per label.

Overall, RPS-based conformal prediction is highly efficient: linear in the number of calibration points n and the number of labels K. By contrast, prior methods [Lu et al., 2022, Xu et al., 2023, Zhang et al., 2026] typically require an additional multiplicative cost of $\mathcal{O}(\log(1/\epsilon))$ to identify an ϵ-optimal contiguous probability mass threshold λ in (5) via binary search to ensure marginal coverage.

4 EXPERIMENTS

To evaluate the quality of RPS-based prediction sets, we conduct experiments on several ordinal image and tabular datasets, comparing against established ordinal conformal baselines. All experiments use neural networks as the underlying model class (we provide additional experiments using gradient boosted trees (GBTs) in Appendix E). The source code of the experiments is made publicly available [1].

4.1 BASELINE NONCONFORMITY SCORES

As nominal baseline nonconformity measures, we include the Least ambiguous set-valued classifier (LAC) score [Sadinle et al., 2019], as well as the adaptive prediction set (APS) score [Romano et al., 2020] (see Appendix C for details). Both LAC and APS treat class labels as unstructured, providing natural baselines for assessing the benefits of incorporating ordinal structure. As ordinal conformal prediction baselines, we consider conformal prediction sets for ordinal classification (COPOC) [Dey et al., 2023] combined with LAC (COPOCL) and APS (COPOCA), as well as the min-CPS approach [Zhang et al., 2026]. The latter is a greedy search–based algorithm that produces small contiguous mode-centered prediction sets and improves upon OrdinalAPS [Lu et al., 2022] in both computational efficiency and empirical performance. As a naive ordinal baseline method, we also include the ordinal CDF (OCDF) [Lu et al., 2022], which constructs prediction intervals from cumulative probabilities and is therefore an interesting baseline for RPS.

4.2 PERFORMANCE METRICS

To evaluate the performance of the different conformal methods, we compute the empirical coverage (COV), as well as the average prediction set size (PS). Additionally, we include the mean interval width (MW) (see Appendix C for details). Another important aspect is the contiguity of the produced prediction sets, which we measure following [Dey et al., 2023] via the contiguity violation (CV) metric, where 0 indicates no contiguity violations on $\mathcal{D}_{\mathrm{test}}$ and 1 corresponds to maximal contiguity violation.

Since a central objective in ordinal classification is to minimize error distances, we prioritize ordinal-specific metrics. First, with $M = \{i : Y_i \notin \mathcal{C}(X_i)\}$ and

$$d(Y_i, \mathcal{C}(X_i)) = \begin{cases} l_i - Y_i & \text{if } Y_i < l_i \\ Y_i - u_i & \text{if } Y_i > u_i \end{cases}$$

for a contiguous interval $\mathcal{C}(X_i) = [l_i, u_i]$, the mean absolute miscoverage magnitude (MAMM) is defined as

$$\mathrm{MAMM} := \frac{1}{|M|} \sum_{i \in M} d(Y_i, \mathcal{C}(X_i)) \,.$$

The worst-case absolute miscoverage magnitude (WAMM) is

$$\mathrm{WAMM} := \max_{i \in M} d(Y_i, \mathcal{C}(X_i)).$$

These metrics quantify ordinal risk under miscoverage, measuring how far the true label lies from the prediction set rather than merely whether it is excluded.

Another highly relevant metric, which evaluates the trade-off between interval efficiency and distance-based error, is the average interval score loss (AISL) [Gneiting and Raftery, 2007], recently applied in conformal prediction for regression [Cabezas et al., 2025]:

$$\mathrm{AISL} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{i \in \mathcal{D}_{\mathrm{test}}} \Bigg[(u_i - l_i) + \frac{2}{\alpha}(l_i - Y_i)\, \mathbb{1}\{Y_i < l_i\} + \frac{2}{\alpha}(Y_i - u_i)\, \mathbb{1}\{Y_i > u_i\} \Bigg].$$

AISL simultaneously accounts for interval width and miscoverage magnitude: the first term $(u_i - l_i)$ penalizes wide intervals, encouraging efficiency, while the second and third terms penalize distance-based errors outside the interval. The penalties are scaled by $2/\alpha$, so stricter coverage requirements (smaller α) amplify the cost of miscoverage. By combining these factors, AISL provides an interpretable metric capturing both compactness and ordinal risk, making it particularly suitable for evaluating conformal prediction methods in ordinal settings.

4.3 EXPERIMENTS ON ORDINAL DATASETS

We evaluate RPS-based prediction sets alongside several baseline methods on two medical image datasets

[1] https://github.com/stefanahaas41/rps-ordinal-conformal-prediction

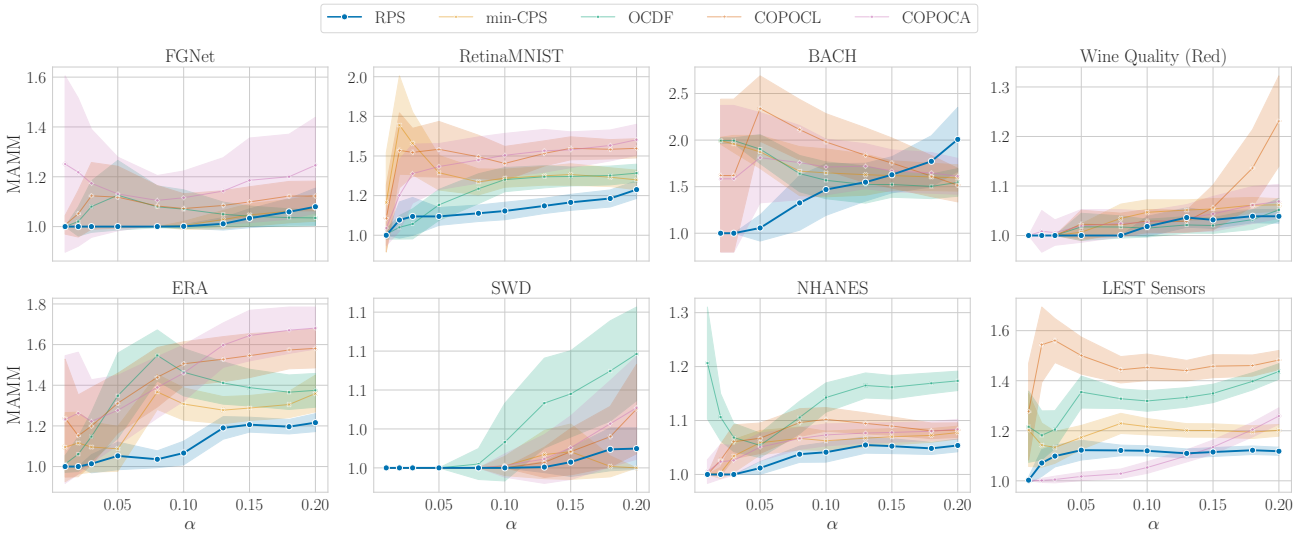


Figure 2: Comparison of prediction sets at $\alpha = \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$ across methods and datasets using the MAMM metric. Shaded regions indicate standard deviation over 50 trials.

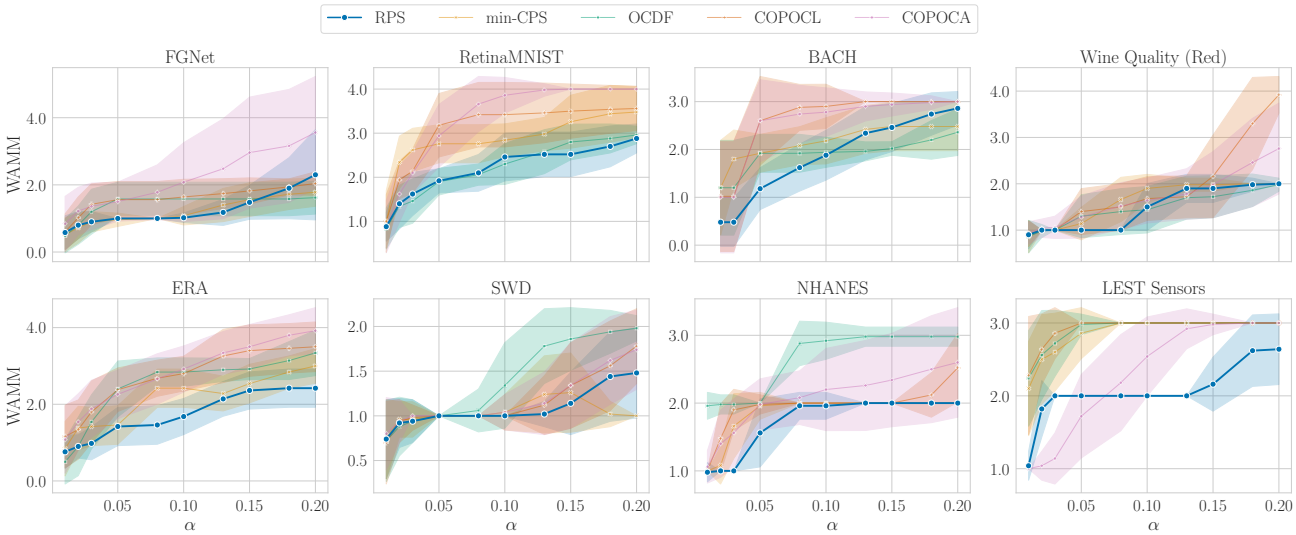


Figure 3: Comparison of prediction sets at $\alpha = \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$ across methods and datasets using the WAMM metric. Shaded regions indicate standard deviation over 50 trials.

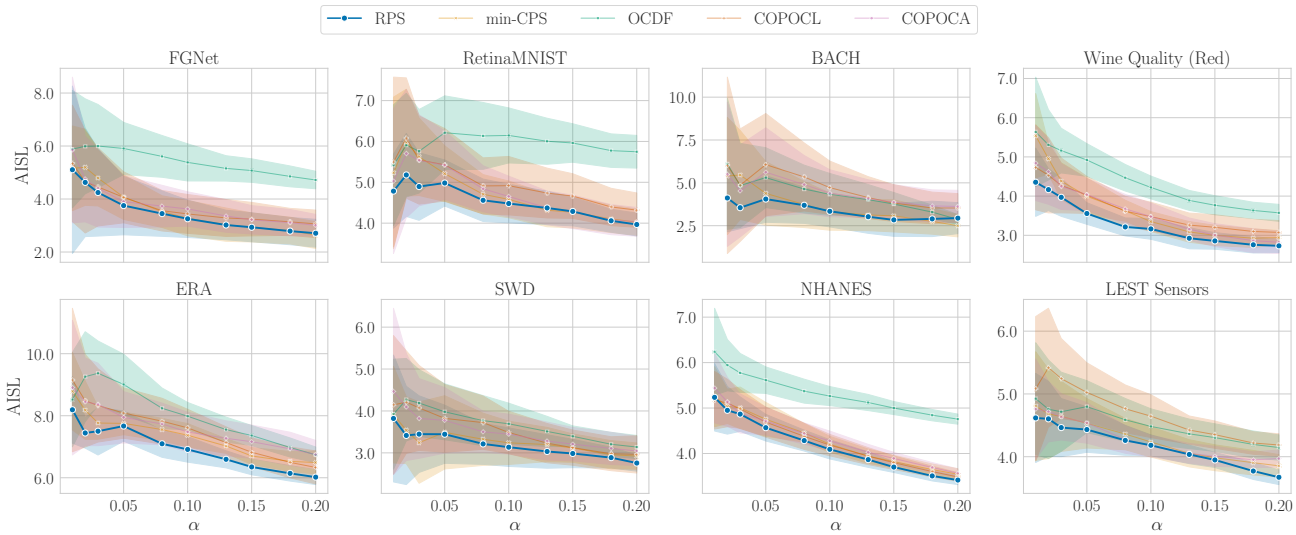


Figure 4: Comparison of prediction sets at $\alpha = \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$ across methods and datasets using the AISL metric. Shaded regions indicate standard deviation over 50 trials.

(BACH [Aresta et al., 2019] and RetinaMNIST [Yang et al., 2023]) and an age-estimation dataset (FGNet [Lanitis et al., 2002, Panis et al., 2016]). Additionally, we include multiple ordinal tabular benchmark datasets [Ayllón-Gavilán et al., 2026].

Due to space constraints, we focus on metrics that reflect ordinal miscoverage directly, consistent with our argument that evaluation in COP should account for the severity of missed predictions rather than treating all miscoverage events equally. Specifically, MAMM and WAMM quantify miscoverage magnitude, while AISL jointly captures prediction set size and ordinal miscoverage in a single score. LAC and APS violate the contiguity requirement for ordinal prediction sets and are therefore excluded from metrics that assume contiguous sets, such as MAMM, WAMM, and AISL. Additional experimental details and results over all datasets and metrics are provided in Appendix C and D.

While RPS-based sets do not always achieve the highest prediction set efficiency, with min-CPS [Zhang et al., 2026] showing particularly strong efficiency on this metric (see Appendix C and D), they tend to achieve lower ordinal miscoverage, as reflected in MAMM (Figure 2) and WAMM (Figure 3). In contrast, mode-centered methods tend to underestimate ordinal risk relative to RPS-based sets. These results provide empirical support for our theoretical claim that RPS-based sets optimize for ordinal risk rather than set size (Theorem 3.2).

Furthermore, when considering the trade-off between efficiency and miscoverage magnitude via AISL (Figure 4), RPS-based sets achieve a favorable balance, demonstrating their practical advantage for real-world ordinal prediction tasks.

5 CONCLUSION & DISCUSSION

We have demonstrated that ranked probability score (RPS)-based conformal sets are, by construction, contiguous and median-centered, providing robust prediction sets for ordinal classification that minimize ordinal risk under oracle conditional coverage, defined as the set-based l_1 error. These sets effectively balance efficiency and error reduction, a critical consideration in high-stakes applications. Specifically, RPS-based sets achieve a favorable trade-off between interval width and miscoverage magnitude while guaranteeing marginal coverage. They are highly competitive with existing ordinal conformal prediction methods, do not depend on specific model architectures, and can be applied to any underlying predictive model. Importantly, RPS-based sets produce meaningful contiguous intervals regardless of the data distribution, making them a practical and versatile solution for ordinal prediction tasks.

Building on this RPS-based framework for COP, a natural direction for future work is to incorporate a delineation of uncertainty into epistemic and aleatoric components [Hüllermeier and Waegeman, 2021]. This line of research has recently attracted attention in both OC [Haas and Hüllermeier, 2026] and CP [Sale et al., 2025, Javanmardi et al., 2025].

Acknowledgements

Alireza Javanmardi gratefully acknowledges funding by the Klaus Tschira Stiftung (project 00.019.2024).

References

Tomé Albuquerque, Ricardo P. M. Cruz, and Jaime S. Cardoso. Ordinal losses for classification of cervical cancer risk. *PeerJ Comput. Sci.*, 7:e457, 2021. doi: 10.7717/PEERJ-CS.457.

Tomé Albuquerque, Ricardo Cruz, and Jaime S Cardoso. Quasi-unimodal distributions for ordinal classification. *Mathematics*, 10(6):980, 2022.

John A Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(1):1–22, 1984.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56: 122–139, 2019.

Rafael Ayllón-Gavilán, David Guijo-Rubio, Antonio Manuel Gómez-Orellana, Francisco Bérchez-Moreno, Víctor Manuel Vargas, and Pedro A. Gutiérrez. Toc-uco: a comprehensive repository of tabular ordinal classification datasets. *Neurocomputing*, 684:133528, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.133528.

Christopher Beckham and Christopher J. Pal. Unimodal probability distributions for deep ordinal classification. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 411–419. PMLR, 2017.

Francisco Bérchez-Moreno, Rafael Ayllón-Gavilán, Víctor Manuel Vargas Yun, David Guijo-Rubio, César Hervás-Martínez, Juan Carlos Fernández, and Pedro Antonio Gutiérrez. dlordinal: A python package for deep ordinal classification. *Neurocomputing*, 622:129305, 2025. doi: 10.1016/J.NEUCOM.2024.129305.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Luben M. C. Cabezas, Vagner S. Santos, Thiago Ramos, and Rafael Izbicki. Epistemic uncertainty in conformal scores: A unified approach. In *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, volume 286 of *Proceedings of Machine Learning Research*, pages 443–470. PMLR, 2025.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit. Lett.*, 140: 325–331, 2020. doi: 10.1016/J.PATREC.2020.11.008.

Jaime S. Cardoso, Ricardo P. M. Cruz, and Tomé Albuquerque. Unimodal distributions for ordinal regression. *IEEE Trans. Artif. Intell.*, 6(9):2498–2509, 2025. doi: 10.1109/TAI.2025.3549740.

François Castagnos, Martin Mihelich, and Charles 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*, pages 4604–4609. International Committee on Computational Linguistics, 2022.

Tianqi Chen and Carlos 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*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Thibault Cordier, Vincent Blot, Louis Lacombe, Thomas Morzadec, Arnaud Capitaine, and Nicolas Brunel. Flexible and Systematic Uncertainty Estimation with Conformal Prediction via the MAPIE library. In *Conformal and Probabilistic Prediction with Applications*, 2023.

Joaquim F. Pinto da Costa, Hugo Alonso, and Jaime S. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008. doi: 10.1016/J.NEUNET.2007.10.003.

Jordi de La Torre, Domenec Puig, and Aïda Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognit. Lett.*, 105: 144–154, 2018. doi: 10.1016/J.PATREC.2017.05.018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan R. Kaveri. Conformal prediction sets for ordinal classification. In *Advances in Neural Information Processing*

Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4738–4747. Computer Vision Foundation / IEEE, 2019.

Manuel Dorado-Moreno, María Pérez-Ortiz, Pedro Antonio Gutiérrez, Rubén Ciria, Javier Briceño, and César Hervás-Martínez. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. *Artif. Intell. Medicine*, 77:1–11, 2017. doi: 10.1016/J.ARTMED.2017.02.004.

Edward S Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969.

Adrian Galdran. Performance metrics for probabilistic ordinal classifiers. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part III*, volume 14222 of *Lecture Notes in Computer Science*, pages 357–366. Springer, 2023. doi: 10.1007/978-3-031-43898-1_35.

Lisa Gaudette and Nathalie Japkowicz. Evaluation methods for ordinal classification. In *Advances in Artificial Intelligence, 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009, Kelowna, Canada, May 25-27, 2009, Proceedings*, volume 5549 of *Lecture Notes in Computer Science*, pages 207–210. Springer, 2009. doi: 10.1007/978-3-642-01818-3_25.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Stefan Haas and Eyke Hüllermeier. Rectifying bias in ordinal observational data using unimodal label smoothing. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part VI*, volume 14174 of *Lecture Notes in Computer Science*, pages 3–18. Springer, 2023. doi: 10.1007/978-3-031-43427-3_1.

Stefan Haas and Eyke Hüllermeier. Uncertainty quantification in ordinal classification: A comparison of measures. *Int. J. Approx. Reason.*, 186:109479, 2025. doi: 10.1016/J.IJAR.2025.109479.

Stefan Haas and Eyke Hüllermeier. Aleatoric and epistemic uncertainty measures for ordinal classification through binary reduction. *Mach. Learn.*, 115(3):63, 2026. doi: 10.1007/S10994-025-06960-5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Rainer Hirk, Kurt Hornik, and Laura Vana. Multivariate ordinal regression models: an analysis of corporate credit ratings. *Stat. Methods Appl.*, 28(3):507–539, 2019. doi: 10.1007/S10260-018-00437-7.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *CoRR*, abs/1611.05916, 2016.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3): 457–506, 2021. doi: 10.1007/S10994-021-05946-3.

Alireza Javanmardi, Soroush H Zargarbashi, Santo MAR Thies, Willem Waegeman, Aleksandar Bojchevski, and Eyke Hüllermeier. Optimal conformal prediction under epistemic uncertainty. *arXiv preprint arXiv:2505.19033*, 2025.

Victor Richmond R. Jose, Robert F. Nau, and Robert L. Winkler. Sensitivity to distance and baseline distributions in forecast evaluation. *Manag. Sci.*, 55(4):582–590, 2009. doi: 10.1287/MNSC.1080.0955.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan 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*, pages 3146–3154, 2017.

Julian Keilson and Hans Gerber. Some results for discrete unimodality. *Journal of the American Statistical Association*, 66(334):386–389, 1971.

Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on pattern Analysis and machine Intelligence*, 24(4):442–455, 2002.

Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu, and Jane You. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*, 388:34–44, 2020. doi: 10.1016/J.NEUCOM.2020.01.025.

Charles Lu, Anastasios N. Angelopoulos, and Stuart R. Pomerantz. Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and*

Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VIII, volume 13438 of *Lecture Notes in Computer Science*, pages 545–554. Springer, 2022. doi: 10.1007/978-3-031-16452-1_52.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Allan H Murphy. On the "ranked probability score". *Journal of Applied Meteorology and Climatology*, 8(6):988–989, 1969.

Allan H Murphy. The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98(12):917–924, 1970.

Allan H. Murphy. A note on the ranked probability score. *Journal of Applied Meteorology*, 10:155–155, 1971.

Inbar Nachmani, Bar Genossar, Coral Scharf, Roee Shraga, and Avigdor Gal. SLACE: A monotone and balance-sensitive loss function for ordinal regression. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 19598–19606. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34158.

National Cancer Institute. Melanoma on a patient's skin., 1985. AV Number: AV-8500-3850.

Gabriel Panis, Andreas Lanitis, Nicholas Tsapatsoulis, and Timothy F Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2): 37–46, 2016.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002, 13th European Conference on Machine Learning, Helsinki, Finland, August 19-23, 2002, Proceedings*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356. Springer, 2002. doi: 10.1007/3-540-36755-1_29.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gorkem Polat, Ümit Mert Çaglar, and Alptekin Temizel. Class distance weighted cross entropy loss for classification of disease severity. *Expert Syst. Appl.*, 269:126372, 2025. doi: 10.1016/J.ESWA.2024.126372.

Mathieu Prodeau, Elodie Drumez, Alain Duhamel, Eric Vibert, Olivier Farges, Guillaume Lassailly, Jean-Yves Mabrut, Jean Hardwigsen, Jean-Marc Régimbeau, Olivier Soubrane, et al. An ordinal model to predict the risk of symptomatic liver failure in patients with cirrhosis undergoing hepatectomy. *Journal of Hepatology*, 71(5): 920–929, 2019.

Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6639–6649, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3538–3548, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

Yusuf Sale, Alireza Javanmardi, and Eyke Hüllermeier. Aleatoric and epistemic uncertainty in conformal prediction. In *Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications, COPA 2025, 10-12 September 2025, London, UK*, volume 266 of *Proceedings of Machine Learning Research*, pages 784–786. PMLR, 2025.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, 2008. doi: 10.5555/1390681.1390693.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression

based on conditional probabilities. *Pattern Anal. Appl.*, 26 (3):941–955, 2023. doi: 10.1007/S10044-023-01181-9.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information fusion*, 81:84–90, 2022.

Tayyab Bin Tariq, Zobia Suhail, and Zubair Nawaz. Ordinal classification for knee osteoarthritis x-rays using vision transformers. *Multim. Tools Appl.*, 84(33):41353–41380, 2025. doi: 10.1007/S11042-025-20804-3.

Marian Tietz, Thomas J. Fan, Daniel Nouri, Benjamin Bossan, and skorch Developers. *skorch: A scikit-learn compatible neural network library that wraps PyTorch*, July 2017. URL `https://skorch.readthedocs.io/en/stable/`.

Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Unimodal regularisation based on beta distribution for deep ordinal regression. *Pattern Recognit.*, 122:108310, 2022. doi: 10.1016/J.PATCOG.2021.108310.

Víctor Manuel Vargas, Antonio Manuel Durán-Rosal, David Guijo-Rubio, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Generalised triangular distributions for ordinal deep learning: Novel proposal and optimisation. *Inf. Sci.*, 648:119606, 2023a. doi: 10.1016/J.INS.2023.119606.

Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Javier Barbero-Gómez, and César Hervás-Martínez. Soft labelling based on triangular distributions for ordinal classification. *Inf. Fusion*, 93:258–267, 2023b. doi: 10.1016/J.INFFUS.2023.01.003.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2-3):349–376, 2013. doi: 10.1007/S10994-013-5355-6.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, pages 444–453. Morgan Kaufmann, 1999.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Conformal risk control for ordinal classification. In *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, pages 2346–2355. PMLR, 2023.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Víctor Manuel Vargas Yun, Antonio M. Gómez-Orellana, David Guijo-Rubio, Francisco Bérchez-Moreno, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Age estimation using soft labelling ordinal classification approaches. In *Advances in Artificial Intelligence - 20th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2024, A Coruña, Spain, June 19-21, 2024, Proceedings*, volume 14640 of *Lecture Notes in Computer Science*, pages 40–49. Springer, 2024. doi: 10.1007/978-3-031-62799-6_5.

Zijian Zhang, Xinyu Chen, Yuanjie Shi, Liyuan Lillian Ma, Zifan Xu, and Yan Yan. Minimum-length conformal prediction sets for ordinal classification. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 28662–28670. AAAI Press, 2026. doi: 10.1609/AAAI.V40I34.40098.

A PROOF OF THEOREM 3.1

Theorem (Contiguity of RPS-based sets)**.** *Let $\mathcal{Y} = \{y_1 \prec y_2 \prec \cdots \prec y_K\}$ be a set of ordered labels, and let s_{RPS} denote the RPS-based nonconformity score. Then, for any input X and any miscoverage level $\alpha \in (0,1)$, the conformal prediction set $\mathcal{C}_\alpha^{\mathrm{RPS}}(X) = \{y \in \mathcal{Y} : s_{\mathrm{RPS}}(X,y) \le \hat{q}_{1-\alpha}\}$ forms a contiguous interval of labels centered at the median m, i.e., there exist integers $1 \le l \le m \le u \le K$ such that $\mathcal{C}_\alpha(X) = \{y_l, \ldots, y_m, \ldots, y_u\}$.*

Proof. Fix an input X. For a candidate label y_ℓ, the RPS nonconformity score is

$$s_{\mathrm{RPS}}(X, y_\ell) = \frac{1}{K-1} \sum_{k=1}^{K-1} \big(F_X(k) - \mathbb{1}\{k \ge \ell\}\big)^2,$$

where k indexes the cumulative sums $F_X(k)$ and ℓ indexes the candidate labels.

(1) Difference between consecutive labels. Consider

$$\Delta_\ell := s_{\mathrm{RPS}}(X, y_{\ell+1}) - s_{\mathrm{RPS}}(X, y_\ell).$$

The step function $\mathbb{1}\{k \ge \ell\}$ changes only at index $k = \ell$ when moving from y_ℓ to $y_{\ell+1}$; for all other k, the indicator remains the same. Therefore, the difference Δ_ℓ between two adjacent labels (y_ℓ and $y_{\ell+1}$) reduces to

$$\Delta_\ell = \frac{1}{K-1}\left[(F(\ell) - 0)^2 - (F(\ell) - 1)^2\right] = \frac{2F(\ell) - 1}{K-1}.$$

(2) Single minimum. Since the cumulative distribution $F(\ell)$ is non-decreasing in ℓ and satisfies $0 \le F(\ell) \le 1$, the consecutive differences

$$\Delta_\ell := s_{\mathrm{RPS}}(X, y_{\ell+1}) - s_{\mathrm{RPS}}(X, y_\ell) = \frac{2F(\ell) - 1}{K-1}$$

form a non-decreasing sequence in ℓ. Hence, there exists an index

$$m := \min\{\ell : F(\ell) \ge 1/2\},$$

corresponding to a (discrete) median of the predictive distribution, such that

$$\Delta_\ell \le 0 \quad \text{for } \ell < m, \qquad \Delta_\ell \ge 0 \quad \text{for } \ell \ge m.$$

It follows that the sequence $\ell \mapsto s_{\mathrm{RPS}}(X, y_\ell)$, which satisfies the recurrence

$$s_{\mathrm{RPS}}(X, y_{\ell+1}) = s_{\mathrm{RPS}}(X, y_\ell) + \frac{2F(\ell) - 1}{K-1},$$

is non-increasing for $\ell < m$ and non-decreasing for $\ell \ge m$. Therefore, the RPS score attains a single minimum at the median index m.

Consequently, starting from the minimum at m, extending the candidate set to the right ($\ell \ge m$) increases the RPS score by increments $(2F(\ell) - 1)/(K-1) \ge 0$, while extending it to the left ($\ell < m$) increases the score by increments $(1 - 2F(\ell))/(K-1) \ge 0$.

Thus, the RPS score is V-shaped around the median of the predictive distribution, being non-increasing to the left of the median and non-decreasing to the right.

(3) Contiguity of conformal sets. Since the mapping $\ell \mapsto s_{\mathrm{RPS}}(X, y_\ell)$ is V-shaped, its sublevel set

$$\mathcal{C}_\alpha^{\mathrm{RPS}}(X) = \{y_\ell : s_{\mathrm{RPS}}(X, y_\ell) \le \hat{q}_{1-\alpha}\}$$

forms a contiguous interval along the ordinal axis, expanding from the median index m toward the tails. Hence, the RPS-based conformal prediction set is contiguous.

□

B PROOF OF THEOREM 3.2

Theorem (Ordinal risk optimality of RPS-based median-grown prediction sets under oracle conditional coverage)**.** *For a fixed input X, let $\mathcal{C}_{\mathrm{RPS}}(X) = \{y_l, \ldots, y_m, \ldots, y_u\}$ denote the contiguous RPS-based prediction set, which is grown from a median index m as in Theorem 3.1. Define the* ordinal risk *of a set $\mathcal{C}(X)$ as the expected l_1-distance of true label from this set:*

$$R(\mathcal{C}(X)) := \sum_{y=1}^{K} p(y \mid X) \min_{c \in \mathcal{C}(X)} |y - c|. \tag{10}$$

Let $\mathcal{C}(X)$ be any other contiguous *conformal set of minimal cardinality satisfying the same coverage constraint as $\mathcal{C}_{\mathrm{RPS}}(X)$. Then*

$$R(\mathcal{C}_{\mathrm{RPS}}(X)) \leq R(\mathcal{C}(X)). \tag{11}$$

Proof. **(1) Singleton case.** For a singleton set $\{c\}$, the ordinal risk is

$$R(\{c\}) = \sum_{y=1}^{K} p(y \mid X)\, |y - c|.$$

It is well known that this expected absolute deviation is minimized at a (discrete) median m of $p(\cdot \mid X)$. Hence, starting with $\{m\}$, as done by RPS-based sets (see Theorem 3.1), is optimal among all sets of size 1. Any discrete median suffices.

(2) Adding adjacent labels. We initialize the contiguous set $\mathcal{C}$ at the singleton median of the predictive distribution,

$$m := \min\{\ell : F(\ell) \geq 1/2\}, \qquad \mathcal{C} = \{m\},$$

and then expand it by adding adjacent labels toward the side with larger tail probability.

Let $\mathcal{C} = \{l, \ldots, u\}$ be a contiguous set of labels, and let

$$F(y) := \sum_{j \leq y} p(y_j \mid X)$$

be the cumulative probability function.

Ordinal risk reduction. For any label $y \leq l - 1$, we have

$$\min_{c \in \mathcal{C}} |y - c| = l - y, \qquad \min_{c \in \mathcal{C} \cup \{l-1\}} |y - c| = (l - 1) - y,$$

so the distance decreases by 1, while for $y \geq l$ the distance is unchanged. Hence, the reduction in ordinal risk from adding $l - 1$ is

$$R(\mathcal{C}) - R(\mathcal{C} \cup \{l - 1\}) = \sum_{y \leq l-1} p(y \mid X) = F(l - 1).$$

Similarly, if $u < K$, adding $u + 1$ reduces the ordinal risk by

$$R(\mathcal{C}) - R(\mathcal{C} \cup \{u + 1\}) = \sum_{y \geq u+1} p(y \mid X) = 1 - F(u).$$

Thus, among the two possible single-step contiguous extensions of $\mathcal{C}$, the one yielding the largest reduction in ordinal risk is toward the side with larger tail probability outside $\mathcal{C}$:

$$\text{add left if } F(l - 1) \geq 1 - F(u), \quad \text{otherwise add right.} \tag{12}$$

Connection to RPS sublevel sets. By Theorem 3.1, extending the candidate set to the right $(u+1)$ increases the RPS score by increments $(2F(u)-1)/(K-1) \geq 0$, while extending it to the left $(l-1)$ increases the score by increments $(1-2F(l-1))/(K-1) \geq 0$.

The conformal prediction procedure selects the next label that yields the *smallest increase in the RPS score* (3), so the greedy expansion chooses

$$\text{add left if } \frac{1-2F(l-1)}{K-1} \leq \frac{2F(u)-1}{K-1}, \quad \text{otherwise add right,}$$

which is algebraically equivalent to (12):

$$\begin{aligned} 1-2F(l-1) \leq 2F(u)-1 &\iff 2 \leq 2F(u)+2F(l-1) \\ &\iff 1 \leq F(u)+F(l-1) \\ &\iff F(l-1) \geq 1-F(u). \end{aligned}$$

(3) Risk optimality among contiguous intervals of fixed length. Fix a length $s \geq 1$ and consider all contiguous sets $\mathcal{C} = \{l, \ldots, l+s-1\}$ of size s. For such sets, the ordinal risk

$$R(\mathcal{C}(X)) = \sum_{y=1}^{K} p(y \mid X) \min_{c \in \mathcal{C}} |y-c|$$

is the expected l_1 distance from Y to the set $\mathcal{C}$. For fixed s, this risk is minimized when the interval is (in a discrete sense) centered at a median index m of $p(\cdot \mid X)$: moving the interval one step away from m increases the expected absolute distance. Consequently, among all contiguous sets of size s, those whose center is closest to m have minimal ordinal risk.

By Theorem 3.1, for each cardinality s, the RPS sublevel set expands around m in a way that keeps the tail imbalance $|F(\ell)-1/2|$ as small as possible. Consequently, for each s, the RPS-based set is (discretely) centered at the median, which minimizes the ordinal risk among all contiguous sets of size s.

(4) Risk dominance over minimal-cardinality coverage sets. Let $\mathcal{C}(X)$ be any contiguous minimal-cardinality set satisfying coverage $\sum_{y \in \mathcal{C}} p(y \mid X) \geq 1-\alpha$, with size $s^\star$. Since $\mathcal{C}_{\text{RPS}}(X)$ also satisfies coverage (Proposition 3.1), $|\mathcal{C}_{\text{RPS}}(X)| \geq s^\star$, so the size-$s^\star$ sublevel set $\mathcal{C}^{s^\star}_{\text{RPS}}(X) \subseteq \mathcal{C}_{\text{RPS}}(X)$ exists (Proposition 3.2). By (3) it minimizes ordinal risk among contiguous size-$s^\star$ sets, and by monotonicity of R under inclusion,

$$R(\mathcal{C}_{\text{RPS}}(X)) \leq R(\mathcal{C}^{s^\star}_{\text{RPS}}(X)) \leq R(\mathcal{C}(X)).$$

□

C ADDITIONAL DETAILS FOR EXPERIMENTS ON ORDINAL IMAGE DATASETS

This section provides additional details for the experiments on ordinal image datasets.

Baseline Nonconformity Scores. Least ambiguous set-valued classifier (LAC) score [Sadinle et al., 2019],

$$s_{\text{LAC}}(X, y) := 1 - p(y \mid X).$$

Adaptive prediction set (APS) score [Romano et al., 2020],

$$s_{\text{APS}}(X, y) := \sum_{y' :\, p(y' \mid X) \geq p(y \mid X)} p(y' \mid X).$$

Performance Metrics. Empirical coverage (COV),

$$\mathrm{COV} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{i \in \mathcal{D}_{\mathrm{test}}} \mathbb{1}\{Y_i \in \mathcal{C}(X_i)\}.$$

Average prediction set size (PS),

$$\mathrm{PS} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{i \in \mathcal{D}_{\mathrm{test}}} |\mathcal{C}(X_i)|.$$

Mean interval width (MW),

$$\mathrm{MW} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{i \in \mathcal{D}_{\mathrm{test}}} (u_i - l_i),$$

where $\mathcal{C}(X_i) = [l_i, u_i]$ denotes the contiguous prediction interval for the i-th test point.

To measure ordinal error distance over the entire test set $\mathcal{D}_{\mathrm{test}}$, we also report the mean absolute interval error (MAIE),

$$\mathrm{MAIE} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{i \in \mathcal{D}_{\mathrm{test}}} \begin{cases} l_i - Y_i & \text{if } Y_i < l_i \\ Y_i - u_i & \text{if } Y_i > u_i \\ 0 & \text{otherwise} \end{cases}.$$

Model implementation. All models are implemented using `skorch` [Tietz et al., 2017], a `scikit-learn` [Pedregosa et al., 2011]-compatible wrapper for `PyTorch` [Paszke et al., 2019], and are conformalized with `MAPIE` [Cordier et al., 2023]. The `dlordinal` [Bérchez-Moreno et al., 2025] package is used to implement ordinal-specific methodologies, such as COPOC [Dey et al., 2023].

For image datasets, we employ a computationally efficient `ResNet-18` [He et al., 2016] model pretrained on ImageNet [Deng et al., 2009], as our primary objective is not to maximize predictive performance but to evaluate the proposed methodology. Nonetheless, ResNet-18 is widely used as a backbone in image-based ordinal classification research, where it achieves reasonable performance [Bérchez-Moreno et al., 2025]. See Table 3 for the neural network training hyperparameters used.

Table 3: Neural network training hyperparameters for the image experiments. Both models (cross-entropy and COPOC) use a ResNet-18 pretrained on ImageNet with the final layer replaced.

	BACH / RetinaMNIST	FG-NET
Optimizer	Adam	AdamW
Learning rate	10^{-3}	$2 \cdot 10^{-4}$
Weight decay	0	10^{-4}
Batch size	128	32
Max. epochs	100	
Early stopping	patience 40 (validation loss)	
LR schedule	ReduceLROnPlateau on training loss (patience 10, factor 0.5, min. 10^{-6})	
Validation split	10%	

Loss functions. We train all models using the standard cross-entropy (CE) loss, which is a proper scoring rule and has been shown to yield unbiased predictive probability distributions, including in ordinal classification [Haas and Hüllermeier, 2025, 2026].

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

where $\boldsymbol{p} = (p_1, \ldots, p_K)$ is the predicted probability distribution over the K classes, and y is the true label. In addition, we consider the non-parametric conformal prediction sets for ordinal classification (COPOC) approach proposed by [Dey et al., 2023], which enforces unimodality in the predictive probability distribution. This also serves as an exemplary ordinal-specific loss, encouraging predictions to respect the natural ordering of the labels.

BACH Dataset The BACH (BreAst Cancer Histology) dataset [Aresta et al., 2019] is a benchmark for breast cancer histopathological image classification, originally introduced as part of the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images. It consists of hematoxylin and eosin (H&E) stained microscopy images of breast tissue, annotated by expert pathologists. The dataset contains 400 high-resolution images (2048×1536 pixels) with 100 images per class, representing four ordinal classes that follow the natural progression of breast cancer:

1. **Normal** – healthy breast tissue
2. **Benign** – non-cancerous abnormal tissue
3. **In situ carcinoma** – cancerous cells that have not invaded surrounding tissue
4. **Invasive carcinoma** – cancerous cells that have spread to surrounding tissue

The ordinal structure of these classes reflects increasing disease severity, making BACH particularly well-suited for evaluating ordinal classification methods in medical imaging. The dataset is perfectly balanced with an imbalance ratio (IR) of 1.0. For our experiments, we resize all images to 224×224 pixels and apply standard normalization with mean and standard deviation of 0.5 across all channels. During training, we augment the data with random rotations (up to 10 degrees) and random horizontal flips to improve model generalization. We use a ResNet-18 model pretrained on ImageNet and fine-tune it for our task. We use two thirds of the data to fine-tune a ResNet-18 model and split the remaining one third equally into calibration and test sets. This random split is repeated over 50 trials. BACH is widely used in medical imaging research and provides an important testbed for uncertainty quantification methods, as reliable predictions with well-calibrated confidence are critical in clinical decision-making for cancer diagnosis.

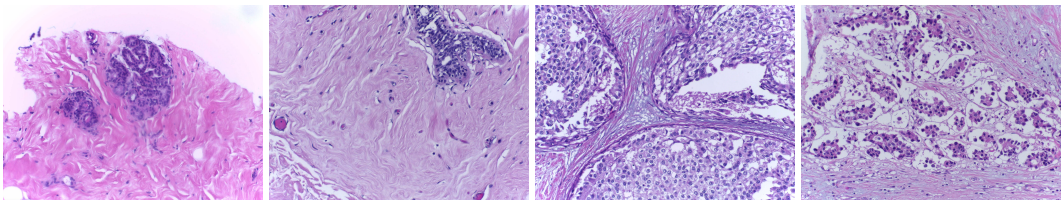


Figure 5: Example images from the BACH dataset [Aresta et al., 2019].

RetinaMNIST Dataset RetinaMNIST is a benchmark dataset of retinal fundus images from the MedMNIST [Yang et al., 2023] collection, consisting of 1,600 28×28 grayscale images labeled with diabetic retinopathy severity levels. The labels form an ordered set of five classes (*No DR*, *Mild*, *Moderate*, *Severe*, *Proliferative DR*) reflecting increasing disease severity, which makes RetinaMNIST a common choice in image-based ordinal classification research, e.g., [Dey et al., 2023]. This ordered structure makes RetinaMNIST well-suited for evaluating ordinal classification and uncertainty estimation methods in medical image analysis. In our experiments, we use a ResNet-18 backbone pretrained on ImageNet and adapted to the RetinaMNIST resolution. ResNet-18 provides a good trade-off between performance and computational efficiency for this task. We use the dedicated training set to train the model, while the original validation and test sets are merged and then randomly split equally into calibration and test sets, again repeated over 50 trials.

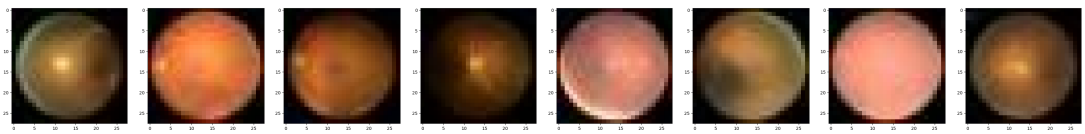


Figure 6: Example images from the RetinaMNIST dataset [Yang et al., 2023]

FGNet Dataset FGNet [Lanitis et al., 2002] is a widely used benchmark for age estimation from facial images, containing 1,002 images of 82 subjects spanning ages 0–69 years. For our experiments, we group ages into six ordinal classes to evaluate coarse age prediction performance, which aligns with the natural ordering of ages and allows ordinal classification evaluation. We preprocess images by resizing them to 256 pixels on the smaller side, followed by a random resized crop to 224×224 (scale 0.85–1.0). Data augmentation includes random horizontal flips (50% probability), color jitter (brightness, contrast, saturation, and hue variations), and random rotations up to 10 degrees. Images are converted to tensors and then normalized using the ImageNet channel-wise mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

We use a ResNet-18 backbone pretrained on ImageNet and fine-tune it on FGNet for the six-class age classification task, leveraging the ordinal structure of the labels to evaluate our ordinal classification methods. We use the dedicated training set to train the model, and split the test set into calibration and test sets, also repeated over 50 trials.

Figure 7: Example images from the FGNet dataset [Lanitis et al., 2002]

Predictive Performance. Table 4 depicts the predictive performance obtained using COPOC and cross-entropy (CE) loss across the different image datasets. Overall, CE achieves the best performance on FGNet and BACH, consistently improving ACC, MAE, MSE, and QWK. On RetinaMNIST, CE attains higher ACC and lower error metrics (MAE and MSE), while COPOC slightly outperforms in terms of 1-OFF accuracy and QWK. These results indicate that although CE generally provides stronger overall classification accuracy, COPOC remains competitive, particularly with respect to ordinal consistency metrics. Furthermore, COPOC ensures unimodality across all datasets, as indicated by its degree of unimodality (UMOD) being equal to one.

Table 4: Predictive performance on the image datasets. Results are reported as mean $\pm$ stddev. Metrics: ACC (Accuracy), 1-OFF (1-Off Accuracy), MAE (Mean Absolute Error), MSE (Mean Squared Error), QWK (Quadratic Weighted Kappa), UMOD (Degree of Unimodality). For ACC, 1-OFF, and QWK, higher is better; for MAE and MSE, lower is better. Bold values indicate the best performance for each metric within each dataset.

Dataset	**Loss**	**ACC** ($\uparrow$)	**1-OFF** ($\uparrow$)	**MAE** ($\downarrow$)	**MSE** ($\downarrow$)	**QWK** ($\uparrow$)	**UMOD**
BACH	COPOC	0.662 ± 0.041	0.839 ± 0.030	0.549 ± 0.070	1.073 ± 0.166	0.549 ± 0.072	1.000 ± 0.000
	CE	**0.839** ± 0.033	**0.912** ± 0.026	**0.256** ± 0.057	**0.462** ± 0.118	**0.806** ± 0.051	0.325 ± 0.031
RetinaMNIST	COPOC	0.481 ± 0.018	**0.759** ± 0.016	0.867 ± 0.040	1.803 ± 0.123	**0.504** ± 0.033	1.000 ± 0.000
	CE	**0.519** ± 0.019	0.757 ± 0.017	**0.826** ± 0.038	**1.741** ± 0.108	0.460 ± 0.032	0.208 ± 0.018
FGNet	COPOC	0.502 ± 0.032	0.945 ± 0.013	0.563 ± 0.040	0.717 ± 0.080	0.821 ± 0.021	1.000 ± 0.000
	CE	**0.579** ± 0.035	**0.952** ± 0.013	**0.474** ± 0.043	**0.591** ± 0.072	**0.860** ± 0.018	0.131 ± 0.023

Detailed Experimental Results. Results over 50 trials for $\alpha = \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$ on the BACH, RetinaMNIST, and FGNet datasets, covering the considered CP methods and metrics, are shown in Figures 8, 9, and 10, respectively, and summarized in Table 5 for $\alpha = \{0.02, 0.05, 0.1\}$. In all cases, as expected, LAC and APS violate the contiguity of prediction sets (CV) and are therefore excluded from measures that require contiguous sets, i.e., MW, MAMM, WAMM, AISL, and MAIE. All ordinal methods (min-CPS, COPOCL, COPOCA, OCDF, and RPS), however, output purely contiguous prediction sets. Considering the trade-off between efficiency, as indicated by MW, and ordinal errors, as indicated by MAMM, WAMM, and MAIE, RPS-based sets strike a favorable balance, which is also reflected by the AISL metric that combines both aspects into a single score. Compared to the other ordinal methods (min-CPS, COPOCL, COPOCA, and OCDF), RPS-based sets are reliable in the sense that risk is neither underestimated nor overestimated. In contrast, min-CPS, COPOCL, and COPOCA tend to underestimate risk, as indicated by higher MAMM, WAMM, and MAIE values induced by overly small PS and MW, whereas OCDF tends to produce very large PS and MW values, particularly for the RetinaMNIST and FGNet datasets (see Figure 9). The claim that RPS-based sets strike a favorable balance between ordinal error awareness and efficient, small intervals is further supported by the tabular experiments reported in Appendix D. These experiments also support our earlier claim that mode-centered set construction does not accurately capture uncertainty, whereas RPS-based sets faithfully account for the full ordinal structure and the associated risk.

D ADDITIONAL EXPERIMENTS ON TABULAR ORDINAL DATASETS

Model implementation. We conduct additional experiments on tabular datasets using multilayer perceptrons (MLPs), which allow straightforward integration of the COPOC architecture [Dey et al., 2023]. We use a simple MLP with a single

Table 5: Comparison of the different non-conformity measures over the BACH, RetinaMNIST and FGNet datasets at $\alpha = 0.02$, $\alpha = 0.05$ and $\alpha = 0.1$.

Dataset	α	Method	COV	PS (↓)	CV (↓)	MW (↓)	MAMM (↓)	WAMM (↓)	MAIE (↓)	AISL (↓)
BACH	0.02	APS	0.982 ± 0.021	2.475 ± 0.603	0.240 ± 0.060	-	-	-	-	-
		COPOCA	0.985 ± 0.021	3.593 ± 0.380	0.000 ± 0.000	2.593 ± 0.380	1.584 ± 0.788	1.000 ± 1.161	0.029 ± 0.046	5.502 ± 4.206
		COPOCL	0.983 ± 0.022	3.539 ± 0.459	0.000 ± 0.000	2.539 ± 0.459	1.620 ± 0.820	1.020 ± 1.152	0.035 ± 0.056	6.024 ± 5.145
		LAC	0.981 ± 0.020	**2.417** ± 0.658	0.210 ± 0.059	-	-	-	-	-
		OCDF	0.981 ± 0.021	3.319 ± 0.320	0.000 ± 0.000	2.319 ± 0.320	1.994 ± 0.030	1.200 ± 0.990	0.038 ± 0.041	6.107 ± 3.818
		RPS	0.983 ± 0.022	3.419 ± 0.286	0.000 ± 0.000	2.419 ± 0.286	**1.000** ± 0.000	**0.480** ± 0.505	**0.017** ± 0.022	**4.116** ± 1.971
		min-CPS	0.981 ± 0.020	2.671 ± 0.594	0.000 ± 0.000	**1.671** ± 0.594	1.993 ± 0.037	1.200 ± 0.990	0.037 ± 0.039	5.398 ± 3.416
	0.05	APS	0.949 ± 0.030	1.773 ± 0.188	0.235 ± 0.052	-	-	-	-	-
	–	COPOCA	0.943 ± 0.041	2.612 ± 0.310	0.000 ± 0.000	1.612 ± 0.310	1.811 ± 0.485	2.600 ± 0.857	0.100 ± 0.071	5.624 ± 2.586
	–	COPOCL	0.949 ± 0.038	2.505 ± 0.340	0.000 ± 0.000	1.505 ± 0.340	2.339 ± 0.351	2.600 ± 0.926	0.114 ± 0.083	6.062 ± 2.983
	0.05	LAC	0.955 ± 0.029	**1.767** ± 0.148	0.220 ± 0.049	-	-	-	-	-
	–	OCDF	0.954 ± 0.029	2.910 ± 0.164	0.000 ± 0.000	1.910 ± 0.164	1.903 ± 0.155	1.920 ± 0.396	0.085 ± 0.047	5.304 ± 1.748
	–	RPS	0.951 ± 0.029	2.860 ± 0.320	0.000 ± 0.000	1.860 ± 0.320	**1.057** ± 0.140	**1.180** ± 0.438	**0.055** ± 0.041	**4.054** ± 1.356
	–	min-CPS	0.952 ± 0.033	1.979 ± 0.198	0.000 ± 0.000	**0.979** ± 0.198	1.873 ± 0.177	1.920 ± 0.396	0.085 ± 0.050	4.385 ± 1.854
	0.1	APS	0.913 ± 0.043	1.468 ± 0.205	0.165 ± 0.056	-	-	-	-	-
	–	COPOCA	0.905 ± 0.043	2.158 ± 0.230	0.000 ± 0.000	1.158 ± 0.230	1.712 ± 0.299	2.780 ± 0.507	0.162 ± 0.078	4.394 ± 1.369
	–	COPOCL	0.900 ± 0.056	1.927 ± 0.205	0.000 ± 0.000	0.927 ± 0.205	1.980 ± 0.304	2.900 ± 0.463	0.190 ± 0.086	4.721 ± 1.550
	0.1	LAC	0.916 ± 0.040	**1.451** ± 0.223	0.153 ± 0.059	-	-	-	-	-
	–	OCDF	0.912 ± 0.049	2.592 ± 0.226	0.000 ± 0.000	1.592 ± 0.226	1.568 ± 0.198	1.940 ± 0.314	0.137 ± 0.071	4.331 ± 1.220
	–	RPS	0.917 ± 0.042	1.788 ± 0.614	0.000 ± 0.000	0.788 ± 0.614	**1.469** ± 0.277	**1.880** ± 0.521	**0.128** ± 0.071	3.340 ± 0.911
	–	min-CPS	0.917 ± 0.040	1.613 ± 0.293	0.000 ± 0.000	**0.613** ± 0.293	1.648 ± 0.257	2.180 ± 0.482	0.135 ± 0.067	**3.322** ± 1.088
RetinaMNIST	0.02	APS	0.984 ± 0.010	4.080 ± 0.156	0.057 ± 0.020	-	-	-	-	-
		COPOCA	0.982 ± 0.011	4.266 ± 0.190	0.000 ± 0.000	3.266 ± 0.190	1.252 ± 0.227	1.620 ± 0.635	0.024 ± 0.017	5.705 ± 1.570
		COPOCL	0.981 ± 0.011	4.222 ± 0.156	0.000 ± 0.000	3.222 ± 0.156	1.535 ± 0.236	1.940 ± 0.424	0.029 ± 0.016	6.083 ± 1.473
		LAC	0.982 ± 0.010	**4.025** ± 0.086	0.046 ± 0.013	-	-	-	-	-
		OCDF	0.981 ± 0.012	4.848 ± 0.047	0.000 ± 0.000	3.848 ± 0.047	**1.050** ± 0.074	**1.340** ± 0.479	0.021 ± 0.013	5.910 ± 1.267
		RPS	0.982 ± 0.010	4.172 ± 0.100	0.000 ± 0.000	3.172 ± 0.100	1.096 ± 0.114	1.400 ± 0.571	**0.020** ± 0.011	**5.180** ± 1.016
		min-CPS	0.983 ± 0.008	4.032 ± 0.090	0.000 ± 0.000	**3.032** ± 0.090	1.695 ± 0.310	2.340 ± 0.593	0.029 ± 0.014	5.971 ± 1.306
	0.05	APS	0.951 ± 0.014	3.600 ± 0.146	0.121 ± 0.025	-	-	-	-	-
	–	COPOCA	0.950 ± 0.016	3.620 ± 0.149	0.000 ± 0.000	2.620 ± 0.149	1.433 ± 0.145	2.940 ± 0.712	0.072 ± 0.026	5.420 ± 0.840
	–	COPOCL	0.953 ± 0.016	3.527 ± 0.201	0.000 ± 0.000	2.527 ± 0.201	1.542 ± 0.176	3.180 ± 0.720	0.073 ± 0.027	5.428 ± 0.887
	0.05	LAC	0.952 ± 0.015	3.639 ± 0.216	0.095 ± 0.030	-	-	-	-	-
	–	OCDF	0.947 ± 0.019	4.660 ± 0.075	0.000 ± 0.000	3.660 ± 0.075	1.192 ± 0.098	**1.900** ± 0.303	0.064 ± 0.024	6.214 ± 0.903
	–	RPS	0.951 ± 0.015	3.804 ± 0.137	0.000 ± 0.000	2.804 ± 0.137	**1.119** ± 0.056	1.920 ± 0.274	**0.054** ± 0.017	**4.983** ± 0.564
	–	min-CPS	0.950 ± 0.016	**3.476** ± 0.143	0.000 ± 0.000	**2.476** ± 0.143	1.394 ± 0.109	2.760 ± 0.431	0.068 ± 0.020	5.212 ± 0.681
	0.1	APS	0.901 ± 0.024	2.958 ± 0.144	0.249 ± 0.034	-	-	-	-	-
	–	COPOCA	0.901 ± 0.025	3.135 ± 0.137	0.000 ± 0.000	2.141 ± 0.135	1.504 ± 0.138	3.860 ± 0.405	0.149 ± 0.040	4.653 ± 0.497
	–	COPOCL	0.899 ± 0.027	2.990 ± 0.134	0.000 ± 0.000	1.990 ± 0.134	1.453 ± 0.107	3.420 ± 0.731	0.146 ± 0.041	4.917 ± 0.715
	0.1	LAC	0.902 ± 0.018	**2.783** ± 0.127	0.203 ± 0.024	-	-	-	-	-
	–	OCDF	0.899 ± 0.026	4.413 ± 0.100	0.000 ± 0.000	3.413 ± 0.100	1.351 ± 0.076	**2.300** ± 0.463	0.137 ± 0.038	6.147 ± 0.674
	–	RPS	0.898 ± 0.025	3.128 ± 0.097	0.000 ± 0.000	2.128 ± 0.097	**1.153** ± 0.055	2.460 ± 0.542	**0.118** ± 0.028	**4.480** ± 0.472
	–	min-CPS	0.900 ± 0.020	2.849 ± 0.119	0.000 ± 0.000	**1.849** ± 0.119	1.360 ± 0.057	2.840 ± 0.370	0.137 ± 0.028	4.582 ± 0.449
FGNet	0.02	APS	0.984 ± 0.019	5.393 ± 0.450	0.315 ± 0.137	-	-	-	-	-
		COPOCA	0.980 ± 0.016	3.417 ± 0.306	0.000 ± 0.000	2.417 ± 0.306	1.218 ± 0.297	1.220 ± 0.708	0.023 ± 0.018	4.757 ± 1.617
		COPOCL	0.978 ± 0.018	3.227 ± 0.269	0.000 ± 0.000	2.227 ± 0.269	1.051 ± 0.093	1.020 ± 0.622	0.024 ± 0.021	4.647 ± 1.930
		LAC	0.984 ± 0.017	5.175 ± 0.321	0.214 ± 0.041	-	-	-	-	-
		OCDF	0.981 ± 0.017	5.050 ± 0.143	0.000 ± 0.000	4.050 ± 0.143	1.020 ± 0.059	0.800 ± 0.571	**0.019** ± 0.019	5.990 ± 1.792
		RPS	0.975 ± 0.023	**3.148** ± 0.274	0.000 ± 0.000	**2.148** ± 0.274	**1.000** ± 0.000	0.800 ± 0.404	0.025 ± 0.023	**4.628** ± 2.046
		min-CPS	0.980 ± 0.018	4.236 ± 0.400	0.000 ± 0.000	3.236 ± 0.400	**1.000** ± 0.000	**0.720** ± 0.454	0.020 ± 0.018	5.196 ± 1.434
	0.05	APS	0.950 ± 0.030	3.875 ± 0.661	0.446 ± 0.094	-	-	-	-	-
	–	COPOCA	0.957 ± 0.026	2.964 ± 0.216	0.000 ± 0.000	1.964 ± 0.216	1.131 ± 0.148	1.500 ± 0.580	0.048 ± 0.029	3.900 ± 0.972
	–	COPOCL	0.948 ± 0.027	2.759 ± 0.205	0.000 ± 0.000	1.759 ± 0.205	1.118 ± 0.124	1.560 ± 0.541	0.058 ± 0.029	4.063 ± 0.962
	0.05	LAC	0.952 ± 0.032	3.914 ± 0.601	0.247 ± 0.045	-	-	-	-	-
	–	OCDF	0.951 ± 0.026	4.732 ± 0.186	0.000 ± 0.000	3.732 ± 0.186	1.126 ± 0.140	1.580 ± 0.499	0.054 ± 0.029	5.908 ± 0.985
	–	RPS	0.946 ± 0.032	**2.584** ± 0.198	0.000 ± 0.000	**1.584** ± 0.198	**1.000** ± 0.000	1.000 ± 0.000	0.054 ± 0.032	**3.752** ± 1.102
	–	min-CPS	0.953 ± 0.030	3.190 ± 0.421	0.000 ± 0.000	2.190 ± 0.421	**1.000** ± 0.000	**0.960** ± 0.198	**0.047** ± 0.030	4.070 ± 0.839
	0.1	APS	0.902 ± 0.040	2.562 ± 0.343	0.230 ± 0.066	-	-	-	-	-
	–	COPOCA	0.898 ± 0.037	2.428 ± 0.182	0.000 ± 0.000	1.429 ± 0.181	1.116 ± 0.107	2.080 ± 1.175	0.115 ± 0.047	3.625 ± 0.647
	–	COPOCL	0.900 ± 0.040	2.294 ± 0.147	0.000 ± 0.000	1.294 ± 0.147	1.072 ± 0.060	1.640 ± 0.525	0.107 ± 0.043	3.438 ± 0.733
	0.1	LAC	0.908 ± 0.035	2.558 ± 0.332	0.170 ± 0.039	-	-	-	-	-
	–	OCDF	0.905 ± 0.042	4.359 ± 0.181	0.000 ± 0.000	3.359 ± 0.181	1.070 ± 0.078	1.580 ± 0.499	0.101 ± 0.044	5.383 ± 0.701
	–	RPS	0.897 ± 0.043	**2.182** ± 0.172	0.000 ± 0.000	**1.182** ± 0.172	**1.001** ± 0.007	**1.020** ± 0.141	0.104 ± 0.043	**3.254** ± 0.705
	–	min-CPS	0.902 ± 0.044	2.303 ± 0.293	0.000 ± 0.000	1.303 ± 0.293	1.005 ± 0.016	1.080 ± 0.274	**0.099** ± 0.046	3.275 ± 0.665

hidden layer of 64 units to maintain consistency across datasets and isolate the effect of the conformal prediction method. Our focus is on evaluating conformal prediction performance rather than optimizing base model accuracy; accordingly, we do not perform hyperparameter tuning. Nonetheless, the model achieves reasonably close-to-standard predictive performance on the datasets and successfully captures the ordinal structure of the labels. See Table 6 for model settings and Table 8 for the predictive performance of COPOC and CE loss on the considered datasets.

Tabular Datasets. The tabular ordinal datasets are obtained from the TOC-UCO repository [Ayllón-Gavilán et al., 2026] (see Table 7 for details). All features are already numeric and are then standardized using standard scaling. For evaluation, approximately 60% of each dataset is used for training, with the remaining 40% split evenly between calibration and test sets. The results are averaged over 50 random splits between calibration and test sets, with the training data held constant. We focus on several larger datasets exhibiting diverse class distributions, which induce predictive distributions ranging from bimodal to unimodal, as reflected by the mean predicted probabilities (MP) (see Figure 21).

Experimental Results. Again, we exclude LAC and APS, which may produce non-contiguous prediction sets, from metrics other than COV, CV, and PS. See the figures below for CP results across several α values ($\alpha =$

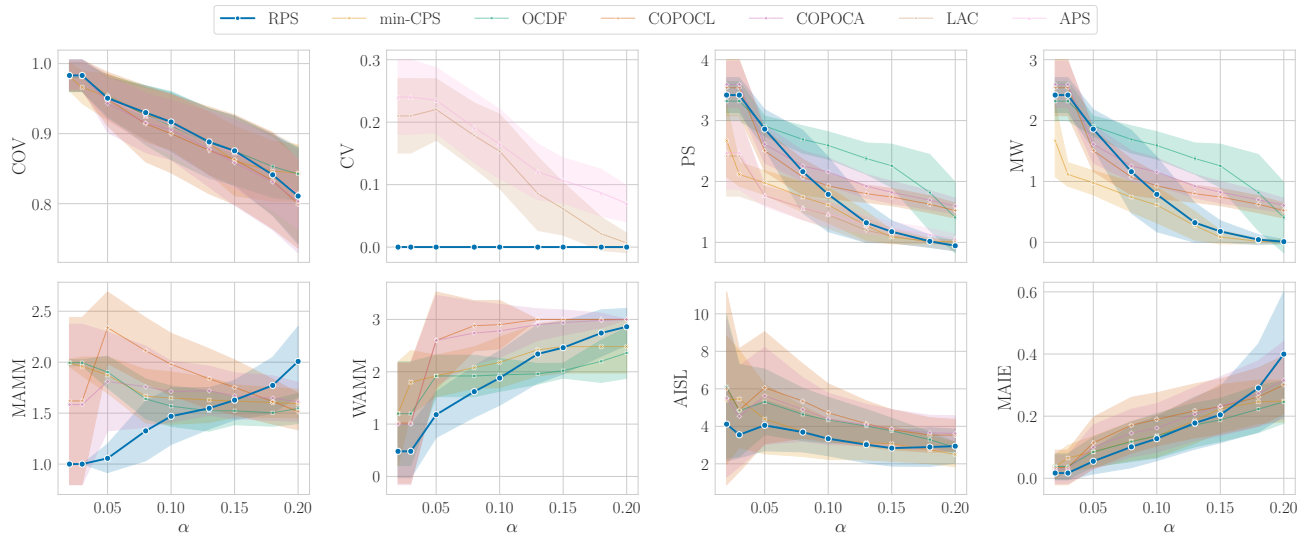


Figure 8: Comparison of prediction sets at $\alpha = \{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$ across methods on the BACH dataset [Aresta et al., 2019]. Shaded regions indicate standard deviation over 50 trials.

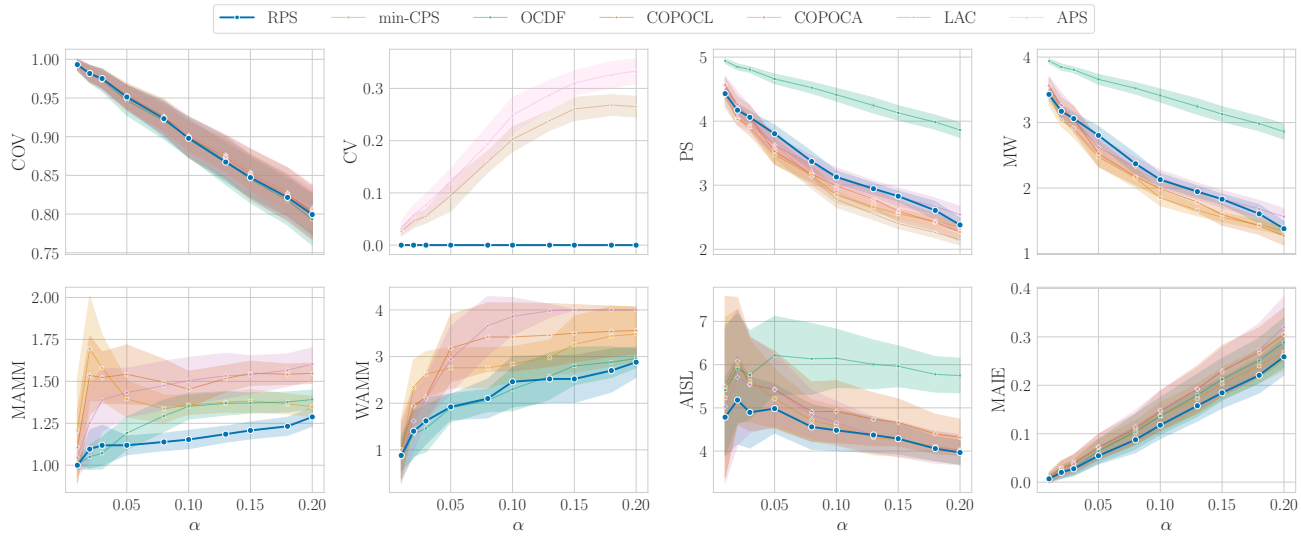


Figure 9: Comparison of prediction sets across methods on the RetinaMNIST dataset [Yang et al., 2023]. Shaded regions indicate standard deviation.

$\{0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.18, 0.2\}$) and Table 9 for detailed results at $\alpha = 0.1$. Consistent with previous findings, RPS-based conformal prediction tends to reduce ordinal error magnitudes while often striking a favorable balance between interval width and miscoverage magnitude, as indicated by AISL. COPOCA remains a strong competitor, particularly on the LESTSensors and LEVXSensors datasets, which exhibit bimodal predictive distributions; however, this advantage comes at the cost of larger prediction sets and wider intervals compared to RPS. Because COPOCA enforces unimodal predictive distributions, bimodal distributions are effectively centered between the extreme classes, thereby reducing ordinal risk in a manner similar to RPS and improving over min-CPS. In general, LAC and its unimodal COPOCL variant tend to produce the most efficient sets, as indicated by PS and MW. However, this often comes at the cost of underestimating the ordinal risk indicated by MAMM, WAMM, and MAIE.

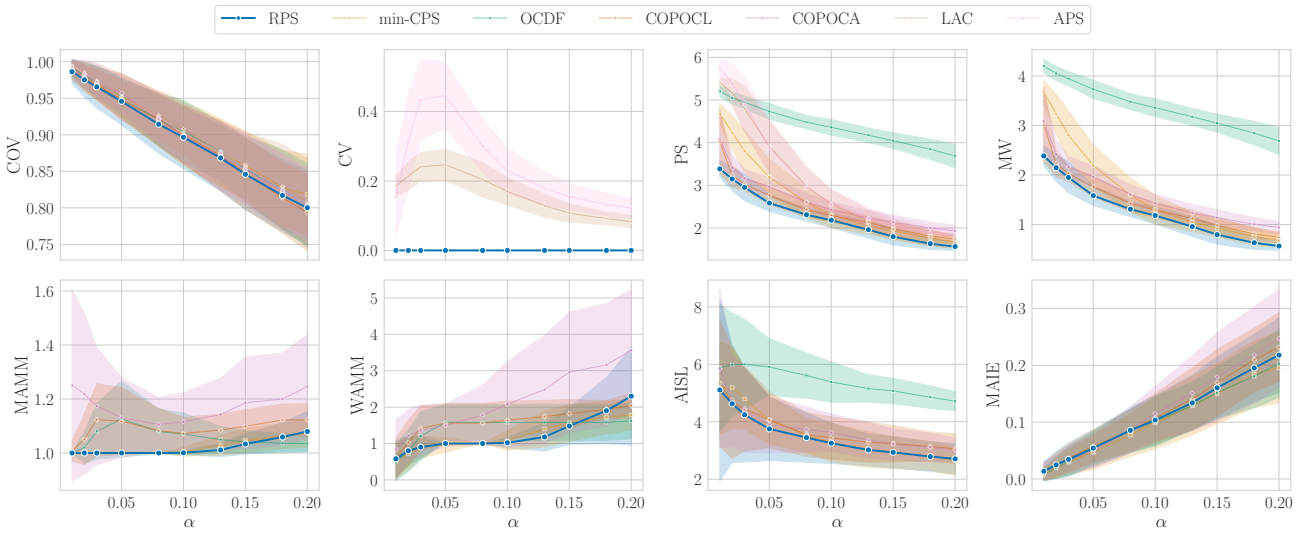


Figure 10: Comparison of prediction sets across methods on the FGNet dataset [Lanitis et al., 2002]. Shaded regions indicate standard deviation.

Table 6: Model settings for the tabular experiments (TOC-UCO datasets).

MLP	
Architecture	1 hidden layer $\times$ 64 units, ReLU
Optimizer	Adam, learning rate 10^{-3}
Batch size	$\min(200, n_{\text{train}})$
Epochs	20 (no early stopping)
Input scaling	standardization (zero mean, unit variance)

E ADDITIONAL EXPERIMENTS USING GRADIENT BOOSTED TREES

In this section, we compare the different nonconformity scores using gradient boosted trees (GBTs), a model class of high practical relevance for tabular data [Shwartz-Ziv and Armon, 2022]. We select LightGBM [Ke et al., 2017] with library defaults as a representative GBT instantiation; other popular implementations such as CatBoost [Prokhorenkova et al., 2018] or XGBoost [Chen and Guestrin, 2016] typically yield similar results. Among the nonconformity measures, we exclude COPOCA and COPOCL from this comparison, as COPOC [Dey et al., 2023] relies on a specific neural network architecture that is not trivial to replicate with GBTs. For our experiments, we use six tabular datasets from the TOC-UCO repository [Ayllón-Gavilán et al., 2026] (Table 10). The results are reported in Table 11 and, broken down per dataset across all metrics, in the figures below. They corroborate our earlier findings: RPS remains competitive with min-CPS [Zhang et al., 2026] and strikes a favorable balance between reducing ordinal miscoverage and maintaining small prediction set sizes.

Table 7: Summary of tabular ordinal datasets used in our experiments [Ayllón-Gavilán et al., 2026].

Dataset	#Samples	#Features	#Classes	Class distribution	IR
LESTSensors	5,112	6	4	(0.16, 0.14, 0.23, 0.46)	0.98
LEVXSensors	5,112	6	4	(0.33, 0.08, 0.10, 0.48)	1.45
nhanes	5,223	30	5	(0.11, 0.28, 0.40, 0.18, 0.04)	1.72
swd	1,000	10	4	(0.03, 0.35, 0.40, 0.22)	2.33
winequalityRed	1,599	11	5	(0.04, 0.43, 0.40, 0.12, 0.01)	4.88
insurance	1,338	9	9	(0.28, 0.19, 0.16, 0.12, 0.06, 0.05, 0.03, 0.03, 0.08)	1.55
melbourneAirbnb	20,036	48	10	(0.08, 0.09, 0.10, 0.09, 0.05, 0.12, 0.11, 0.09, 0.09, 0.18)	1.01
cancerDeathRate	3,047	29	10	(0.10, 0.09, 0.10, 0.11, 0.12, 0.13, 0.11, 0.08, 0.06, 0.10)	0.95
era	1,000	4	9	(0.09, 0.14, 0.18, 0.17, 0.16, 0.12, 0.09, 0.03, 0.02)	1.66
lev	1,000	4	5	(0.09, 0.28, 0.40, 0.20, 0.03)	2.16

Table 8: Predictive performance on the datasets. Results are reported as mean ± stddev. Metrics: ACC (Accuracy), 1-OFF (1-Off Accuracy), MAE (Mean Absolute Error), MSE (Mean Squared Error), QWK (Quadratic Weighted Kappa), UMOD (Degree of Unimodality). For ACC, 1-OFF, and QWK, higher is better; for MAE and MSE, lower is better. Bold values indicate the best performance for each metric within each dataset.

Dataset	**Loss**	**ACC** (↑)	**1-OFF** (↑)	**MAE** (↓)	**MSE** (↓)	**QWK** (↑)	**UMOD**
LESTSensors	COPOC	0.4851 ± 0.0077	$\mathbf{0.7589} \pm 0.0043$	$\mathbf{0.8590} \pm 0.0104$	$\mathbf{1.7536} \pm 0.0265$	0.1119 ± 0.0113	1.000 ± 0.000
	CE	$\mathbf{0.4928} \pm 0.0091$	0.7552 ± 0.0082	0.8597 ± 0.0191	1.7799 ± 0.0534	$\mathbf{0.2579} \pm 0.0207$	0.471 ± 0.013
LEVXSensors	COPOC	0.5220 ± 0.0078	0.6740 ± 0.0085	1.0011 ± 0.0215	2.4413 ± 0.0656	0.3253 ± 0.0187	1.000 ± 0.000
	CE	$\mathbf{0.6091} \pm 0.0091$	$\mathbf{0.7106} \pm 0.0101$	$\mathbf{0.8894} \pm 0.0278$	$\mathbf{2.3046} \pm 0.0835$	$\mathbf{0.4368} \pm 0.0204$	0.0278 ± 0.004
nhanes	COPOC	0.4098 ± 0.0093	0.8884 ± 0.0056	0.7073 ± 0.0121	0.9526 ± 0.0224	$\mathbf{0.3198} \pm 0.0168$	1.000 ± 0.000
	CE	$\mathbf{0.4225} \pm 0.0078$	$\mathbf{0.8944} \pm 0.0054$	$\mathbf{0.6870} \pm 0.0097$	$\mathbf{0.9140} \pm 0.0206$	0.2944 ± 0.0179	0.941 ± 0.005
SWD	COPOC	0.4730 ± 0.0267	0.8827 ± 0.0141	0.6443 ± 0.0341	0.8789 ± 0.0571	0.4249 ± 0.0364	1.000 ± 0.000
	CE	$\mathbf{0.5596} \pm 0.0208$	$\mathbf{0.9648} \pm 0.0091$	$\mathbf{0.4756} \pm 0.0236$	$\mathbf{0.5460} \pm 0.0361$	$\mathbf{0.5031} \pm 0.0327$	1.000 ± 0.000
winequalityRed	COPOC	0.5298 ± 0.0235	0.9672 ± 0.0060	0.5030 ± 0.0248	0.5686 ± 0.0309	$\mathbf{0.4927} \pm 0.0304$	1.000 ± 0.000
	CE	$\mathbf{0.5693} \pm 0.0156$	$\mathbf{0.9724} \pm 0.0057$	$\mathbf{0.4583} \pm 0.0152$	$\mathbf{0.5136} \pm 0.0201$	0.4379 ± 0.0204	1.000 ± 0.000
insurance	COPOC	$\mathbf{0.5605} \pm 0.0163$	$\mathbf{0.7562} \pm 0.0144$	$\mathbf{1.1058} \pm 0.0507$	$\mathbf{4.2700} \pm 0.3290$	$\mathbf{0.6236} \pm 0.0320$	1.000 ± 0.000
	CE	0.4943 ± 0.0151	0.7527 ± 0.0105	1.2496 ± 0.0541	5.0466 ± 0.3768	0.6021 ± 0.0345	0.000 ± 0.000
melbourneAirbnb	COPOC	0.3447 ± 0.0054	$\mathbf{0.6779} \pm 0.0058$	$\mathbf{1.2867} \pm 0.0168$	$\mathbf{3.7262} \pm 0.0873$	$\mathbf{0.8007} \pm 0.0046$	1.000 ± 0.000
	CE	$\mathbf{0.3493} \pm 0.0046$	0.6483 ± 0.0056	1.3835 ± 0.0184	4.4412 ± 0.1079	0.7740 ± 0.0054	0.197 ± 0.005
cancerDeathRate	COPOC	0.2675 ± 0.0127	$\mathbf{0.6067} \pm 0.0146$	$\mathbf{1.5839} \pm 0.0532$	$\mathbf{5.0903} \pm 0.3254$	$\mathbf{0.6973} \pm 0.0192$	1.000 ± 0.000
	CE	$\mathbf{0.2685} \pm 0.0123$	0.5339 ± 0.0146	1.8201 ± 0.0569	6.6766 ± 0.3864	0.6432 ± 0.0202	0.083 ± 0.008
era	COPOC	0.2171 ± 0.0223	$\mathbf{0.6362} \pm 0.0258$	$\mathbf{1.3670} \pm 0.0615$	3.1862 ± 0.2774	$\mathbf{0.5692} \pm 0.0374$	1.000 ± 0.000
	CE	$\mathbf{0.2312} \pm 0.0184$	0.5846 ± 0.0214	1.3874 ± 0.0389	$\mathbf{3.1610} \pm 0.1495$	0.3379 ± 0.0306	0.166 ± 0.015
lev	COPOC	$\mathbf{0.5737} \pm 0.0223$	$\mathbf{0.9573} \pm 0.0095$	$\mathbf{0.4690} \pm 0.0249$	$\mathbf{0.5544} \pm 0.0377$	$\mathbf{0.6668} \pm 0.0245$	1.000 ± 0.000
	CE	0.4633 ± 0.0186	0.9500 ± 0.0082	0.5867 ± 0.0193	0.6867 ± 0.0289	0.4009 ± 0.0272	0.777 ± 0.012

Table 9: Performance comparison of the different CP methods on the tabular datasets at $\alpha = 0.1$.

Dataset	Method	COV	PS (↓)	CV (↓)	MW (↓)	MAMM (↓)	WAMM (↓)	MAIE (↓)	AISL (↓)
LESTSensors	APS	0.904 ± 0.012	2.932 ± 0.063	0.343 ± 0.015	–	–	–	–	–
	COPOCA	0.901 ± 0.009	3.08 ± 0.049	0.0 ± 0.0	2.081 ± 0.049	**1.053** ± 0.023	2.54 ± 0.542	**0.104** ± 0.01	**4.17** ± 0.171
	COPOCL	0.902 ± 0.012	2.803 ± 0.065	0.0 ± 0.0	**1.803** ± 0.065	1.453 ± 0.053	3.0 ± 0.0	0.142 ± 0.02	4.646 ± 0.341
	LAC	0.901 ± 0.01	**2.755** ± 0.049	0.232 ± 0.016	–	–	–	–	–
	OCDF	0.901 ± 0.012	2.871 ± 0.048	0.0 ± 0.0	1.871 ± 0.048	1.319 ± 0.041	3.0 ± 0.0	0.131 ± 0.016	4.482 ± 0.281
	RPS	0.901 ± 0.008	2.966 ± 0.028	0.0 ± 0.0	1.966 ± 0.028	1.12 ± 0.02	**2.0** ± 0.0	0.111 ± 0.009	4.182 ± 0.166
	min-CPS	0.901 ± 0.01	2.823 ± 0.036	0.0 ± 0.0	1.823 ± 0.036	1.216 ± 0.032	3.0 ± 0.0	0.121 ± 0.014	4.235 ± 0.241
LEVXSensors	APS	0.897 ± 0.009	2.487 ± 0.051	0.78 ± 0.01	–	–	–	–	–
	COPOCA	0.897 ± 0.012	3.414 ± 0.051	0.0 ± 0.0	2.414 ± 0.051	**1.027** ± 0.016	2.06 ± 0.424	**0.106** ± 0.012	**4.525** ± 0.208
	COPOCL	0.899 ± 0.014	2.927 ± 0.053	0.0 ± 0.0	1.927 ± 0.053	1.843 ± 0.066	3.0 ± 0.0	0.186 ± 0.026	5.638 ± 0.474
	LAC	0.899 ± 0.011	**2.405** ± 0.055	0.626 ± 0.016	–	–	–	–	–
	OCDF	0.898 ± 0.012	2.808 ± 0.057	0.0 ± 0.0	**1.808** ± 0.057	1.876 ± 0.059	3.0 ± 0.0	0.192 ± 0.027	5.65 ± 0.483
	RPS	0.896 ± 0.014	3.104 ± 0.042	0.0 ± 0.0	2.104 ± 0.042	1.174 ± 0.029	**2.0** ± 0.0	0.122 ± 0.017	4.538 ± 0.295
	min-CPS	0.896 ± 0.015	2.874 ± 0.052	0.0 ± 0.0	1.874 ± 0.052	1.458 ± 0.059	3.0 ± 0.0	0.152 ± 0.022	4.911 ± 0.394
nhanes	APS	0.899 ± 0.013	3.129 ± 0.066	0.03 ± 0.005	–	–	–	–	–
	COPOCA	0.899 ± 0.012	3.107 ± 0.062	0.0 ± 0.0	2.107 ± 0.062	1.073 ± 0.02	2.2 ± 0.606	0.109 ± 0.013	4.279 ± 0.214
	COPOCL	0.9 ± 0.013	2.974 ± 0.055	0.0 ± 0.0	**1.974** ± 0.055	1.102 ± 0.022	2.0 ± 0.0	0.111 ± 0.014	4.185 ± 0.235
	LAC	0.899 ± 0.012	**2.971** ± 0.052	0.011 ± 0.002	–	–	–	–	–
	OCDF	0.902 ± 0.01	4.018 ± 0.045	0.0 ± 0.0	3.018 ± 0.045	1.143 ± 0.027	2.92 ± 0.274	0.112 ± 0.013	5.264 ± 0.209
	RPS	0.9 ± 0.012	3.005 ± 0.052	0.0 ± 0.0	2.005 ± 0.052	**1.041** ± 0.018	**1.96** ± 0.198	**0.104** ± 0.013	**4.092** ± 0.212
	min-CPS	0.899 ± 0.01	3.064 ± 0.046	0.0 ± 0.0	2.064 ± 0.046	1.062 ± 0.018	2.0 ± 0.0	0.107 ± 0.011	4.203 ± 0.173
swd	APS	0.906 ± 0.025	2.221 ± 0.086	0.0 ± 0.0	–	–	–	–	–
	COPOCA	0.901 ± 0.027	2.459 ± 0.098	0.0 ± 0.0	1.459 ± 0.098	1.002 ± 0.007	1.04 ± 0.198	0.099 ± 0.028	3.447 ± 0.476
	COPOCL	0.899 ± 0.033	2.464 ± 0.168	0.0 ± 0.0	1.464 ± 0.168	**1.0** ± 0.0	**1.0** ± 0.0	0.101 ± 0.033	3.486 ± 0.506
	LAC	0.903 ± 0.026	**2.131** ± 0.074	0.0 ± 0.0	–	–	–	–	–
	OCDF	0.902 ± 0.027	2.686 ± 0.102	0.0 ± 0.0	1.686 ± 0.102	1.017 ± 0.025	1.34 ± 0.479	0.1 ± 0.029	3.694 ± 0.492
	RPS	0.901 ± 0.025	2.15 ± 0.051	0.0 ± 0.0	**1.15** ± 0.051	**1.0** ± 0.0	**1.0** ± 0.0	0.099 ± 0.025	**3.136** ± 0.46
	min-CPS	0.906 ± 0.028	2.356 ± 0.07	0.0 ± 0.0	1.356 ± 0.07	1.001 ± 0.006	1.04 ± 0.198	**0.094** ± 0.029	3.24 ± 0.516
winequalityRed	APS	0.905 ± 0.014	2.383 ± 0.101	0.0 ± 0.0	–	–	–	–	–
	COPOCA	0.903 ± 0.023	2.481 ± 0.132	0.0 ± 0.0	1.481 ± 0.132	1.026 ± 0.023	1.64 ± 0.485	0.1 ± 0.023	3.478 ± 0.344
	COPOCL	0.903 ± 0.021	2.471 ± 0.084	0.0 ± 0.0	1.471 ± 0.084	1.029 ± 0.023	1.68 ± 0.471	0.1 ± 0.022	3.476 ± 0.369
	LAC	0.901 ± 0.015	2.138 ± 0.126	0.0 ± 0.0	–	–	–	–	–
	OCDF	0.903 ± 0.021	3.25 ± 0.14	0.0 ± 0.0	2.25 ± 0.14	**1.015** ± 0.018	**1.44** ± 0.501	**0.099** ± 0.021	4.221 ± 0.291
	RPS	0.901 ± 0.016	**2.135** ± 0.092	0.0 ± 0.0	**1.135** ± 0.092	1.018 ± 0.02	1.5 ± 0.505	0.101 ± 0.017	**3.163** ± 0.264
	min-CPS	0.902 ± 0.017	2.296 ± 0.103	0.0 ± 0.0	1.296 ± 0.103	1.047 ± 0.025	1.9 ± 0.303	0.103 ± 0.017	3.35 ± 0.255
insurance	APS	0.9 ± 0.018	4.284 ± 0.277	0.495 ± 0.072	–	–	–	–	–
	COPOCA	0.9 ± 0.019	4.599 ± 0.38	0.0 ± 0.0	3.599 ± 0.38	1.914 ± 0.247	4.22 ± 0.79	0.196 ± 0.061	7.51 ± 0.852
	COPOCL	0.901 ± 0.02	4.464 ± 0.241	0.0 ± 0.0	**3.464** ± 0.241	2.163 ± 0.177	4.48 ± 0.646	0.216 ± 0.056	7.793 ± 0.891
	LAC	0.899 ± 0.018	**3.922** ± 0.241	0.342 ± 0.043	–	–	–	–	–
	OCDF	0.901 ± 0.021	5.631 ± 0.257	0.0 ± 0.0	4.631 ± 0.257	2.259 ± 0.199	4.9 ± 0.303	0.223 ± 0.052	9.088 ± 0.818
	RPS	0.901 ± 0.017	5.954 ± 0.138	0.0 ± 0.0	4.954 ± 0.138	**1.231** ± 0.083	**2.58** ± 0.538	**0.123** ± 0.027	**7.416** ± 0.404
	min-CPS	0.898 ± 0.021	4.75 ± 0.364	0.0 ± 0.0	3.75 ± 0.364	1.963 ± 0.229	4.04 ± 0.402	0.204 ± 0.063	7.829 ± 0.911
melbourneAirbnb	APS	0.899 ± 0.008	4.851 ± 0.069	0.233 ± 0.005	–	–	–	–	–
	COPOCA	0.899 ± 0.007	4.905 ± 0.066	0.0 ± 0.0	3.907 ± 0.066	1.879 ± 0.06	8.98 ± 0.141	0.191 ± 0.015	7.716 ± 0.248
	COPOCL	0.898 ± 0.007	4.741 ± 0.052	0.0 ± 0.0	**3.741** ± 0.052	1.924 ± 0.044	7.82 ± 0.438	0.196 ± 0.014	7.653 ± 0.229
	LAC	0.899 ± 0.007	**4.725** ± 0.055	0.163 ± 0.004	–	–	–	–	–
	OCDF	0.9 ± 0.006	6.76 ± 0.056	0.0 ± 0.0	5.76 ± 0.056	1.942 ± 0.04	7.78 ± 0.418	0.194 ± 0.012	9.632 ± 0.187
	RPS	0.899 ± 0.007	4.915 ± 0.052	0.0 ± 0.0	3.915 ± 0.052	**1.768** ± 0.041	**6.56** ± 0.501	**0.178** ± 0.013	**7.47** ± 0.204
	min-CPS	0.899 ± 0.009	4.813 ± 0.061	0.0 ± 0.0	3.813 ± 0.061	1.822 ± 0.039	7.76 ± 0.431	0.185 ± 0.016	7.51 ± 0.255
cancerDeathRate	APS	0.899 ± 0.018	6.025 ± 0.186	0.179 ± 0.018	–	–	–	–	–
	COPOCA	0.902 ± 0.021	5.952 ± 0.218	0.0 ± 0.0	**4.952** ± 0.218	1.91 ± 0.128	5.72 ± 0.784	0.187 ± 0.045	8.695 ± 0.686
	COPOCL	0.904 ± 0.021	6.002 ± 0.221	0.0 ± 0.0	5.002 ± 0.221	1.92 ± 0.106	5.2 ± 0.404	0.185 ± 0.043	8.693 ± 0.659
	LAC	0.899 ± 0.016	**6.04** ± 0.168	0.112 ± 0.016	–	–	–	–	–
	OCDF	0.902 ± 0.017	7.213 ± 0.159	0.0 ± 0.0	6.213 ± 0.159	2.278 ± 0.143	6.52 ± 0.735	0.223 ± 0.037	10.677 ± 0.618
	RPS	0.9 ± 0.019	6.093 ± 0.184	0.0 ± 0.0	5.093 ± 0.184	**1.638** ± 0.102	**4.88** ± 0.328	**0.163** ± 0.033	**8.354** ± 0.483
	min-CPS	0.899 ± 0.017	5.987 ± 0.155	0.0 ± 0.0	4.987 ± 0.155	1.882 ± 0.135	5.58 ± 0.499	0.191 ± 0.038	8.812 ± 0.62
era	APS	0.9 ± 0.024	5.706 ± 0.178	0.19 ± 0.035	–	–	–	–	–
	COPOCA	0.905 ± 0.027	5.679 ± 0.276	0.0 ± 0.0	4.679 ± 0.276	1.46 ± 0.138	2.92 ± 0.601	0.14 ± 0.045	7.477 ± 0.634
	COPOCL	0.902 ± 0.024	**5.639** ± 0.218	0.0 ± 0.0	**4.639** ± 0.218	1.506 ± 0.108	2.8 ± 0.452	0.148 ± 0.04	7.609 ± 0.598
	LAC	0.903 ± 0.022	5.842 ± 0.234	0.116 ± 0.044	–	–	–	–	–
	OCDF	0.908 ± 0.018	6.289 ± 0.144	0.0 ± 0.0	5.289 ± 0.144	1.464 ± 0.117	2.84 ± 0.37	0.135 ± 0.028	7.985 ± 0.443
	RPS	0.905 ± 0.026	5.877 ± 0.244	0.0 ± 0.0	4.877 ± 0.244	**1.066** ± 0.058	**1.68** ± 0.471	**0.102** ± 0.032	**6.913** ± 0.41
	min-CPS	0.909 ± 0.019	6.013 ± 0.036	0.0 ± 0.0	5.013 ± 0.036	1.308 ± 0.079	2.42 ± 0.499	0.118 ± 0.02	7.371 ± 0.376
lev	APS	0.902 ± 0.021	2.824 ± 0.111	0.014 ± 0.01	–	–	–	–	–
	COPOCA	0.902 ± 0.025	2.675 ± 0.148	0.0 ± 0.0	1.675 ± 0.148	1.066 ± 0.041	1.88 ± 0.328	0.105 ± 0.027	3.771 ± 0.429
	COPOCL	0.905 ± 0.024	**2.45** ± 0.146	0.0 ± 0.0	**1.45** ± 0.146	1.059 ± 0.038	1.82 ± 0.388	0.101 ± 0.026	**3.462** ± 0.386
	LAC	0.897 ± 0.022	2.628 ± 0.115	0.0 ± 0.0	–	–	–	–	–
	OCDF	0.901 ± 0.026	3.515 ± 0.106	0.0 ± 0.0	2.515 ± 0.106	1.106 ± 0.052	1.94 ± 0.24	0.11 ± 0.029	4.709 ± 0.49
	RPS	0.904 ± 0.014	2.857 ± 0.073	0.0 ± 0.0	1.857 ± 0.073	**1.0** ± 0.0	**1.0** ± 0.0	**0.096** ± 0.014	3.781 ± 0.216
	min-CPS	0.902 ± 0.026	2.812 ± 0.106	0.0 ± 0.0	1.812 ± 0.106	1.057 ± 0.045	1.78 ± 0.418	0.102 ± 0.026	3.862 ± 0.426

Table 10: Summary of tabular ordinal datasets used in our LightGBM experiments [Ayllón-Gavilán et al., 2026].

Dataset	#Samples	#Features	#Classes	Class distribution	IR
heartDisease	294	13	4	(0.64, 0.13, 0.09, 0.15)	1.48
mammoexp	412	5	3	(0.57, 0.25, 0.18)	0.92
support	730	19	4	(0.39, 0.13, 0.08, 0.40)	1.34
winequalityRed	1,599	11	5	(0.04, 0.43, 0.40, 0.12, 0.01)	4.88
nhanes	5,223	30	5	(0.11, 0.28, 0.40, 0.18, 0.04)	1.72
LEVXSensors	5,112	6	4	(0.33, 0.08, 0.10, 0.48)	1.45

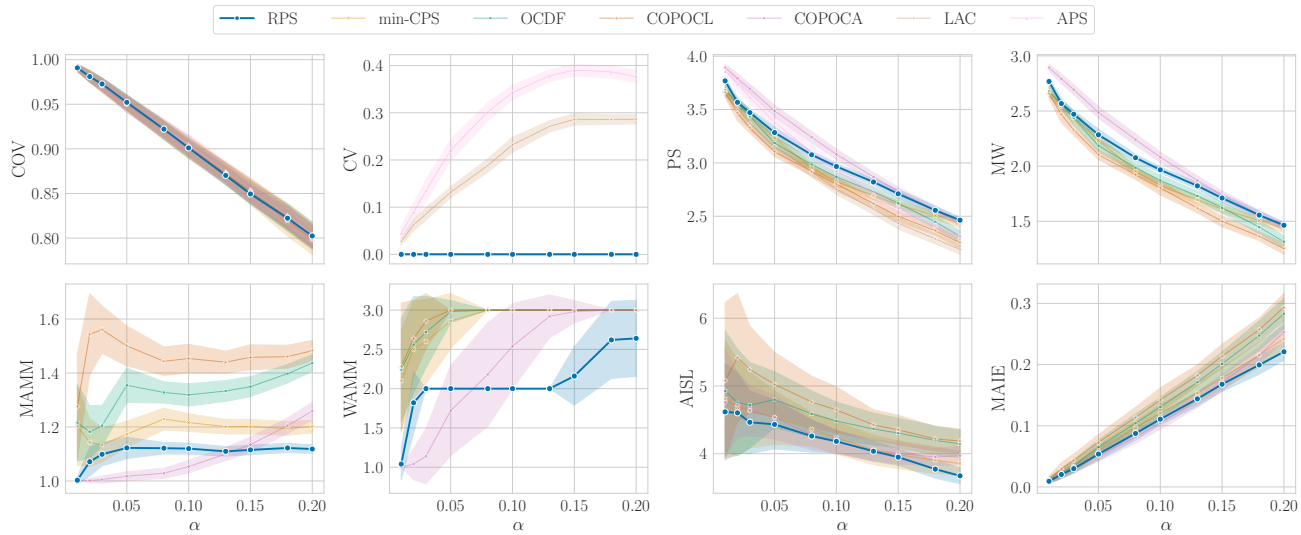


Figure 11: Comparison of prediction sets across methods on LESTSensors dataset. Shaded regions indicate standard deviation.

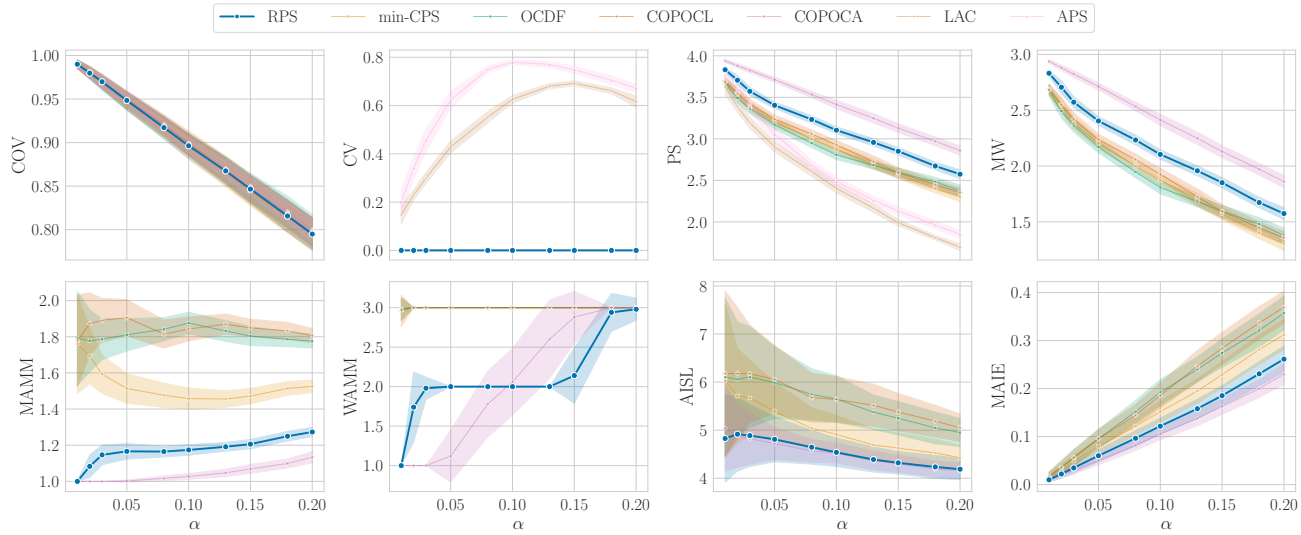


Figure 12: Comparison of prediction sets across methods on LEVXSensors dataset. Shaded regions indicate standard deviation.

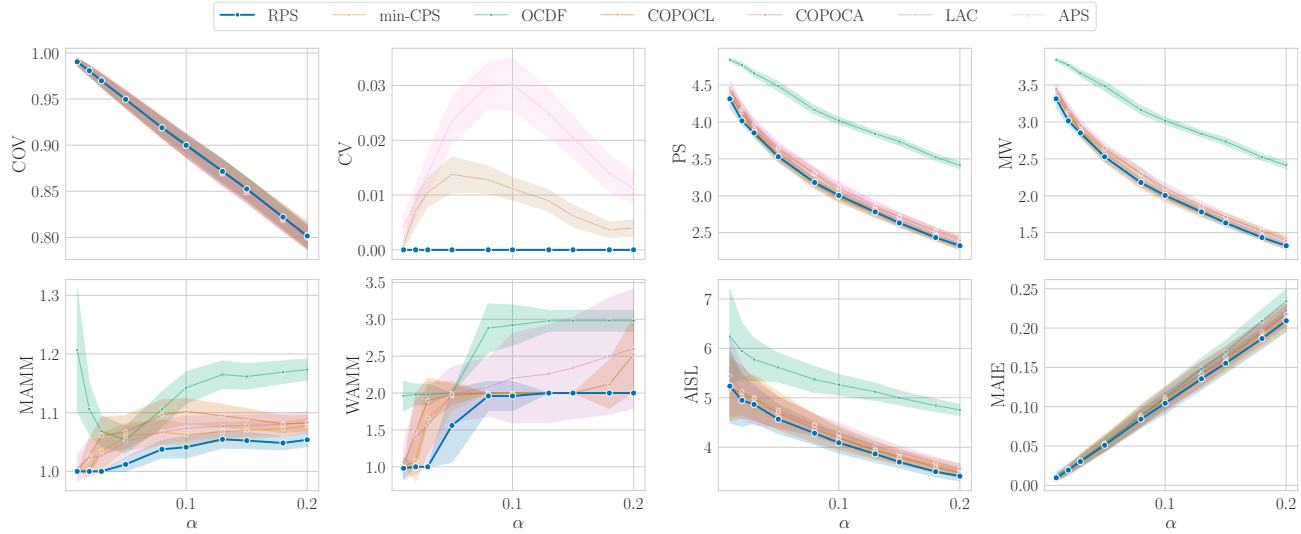


Figure 13: Comparison of prediction sets across methods on Planes dataset. Shaded regions indicate standard deviation.

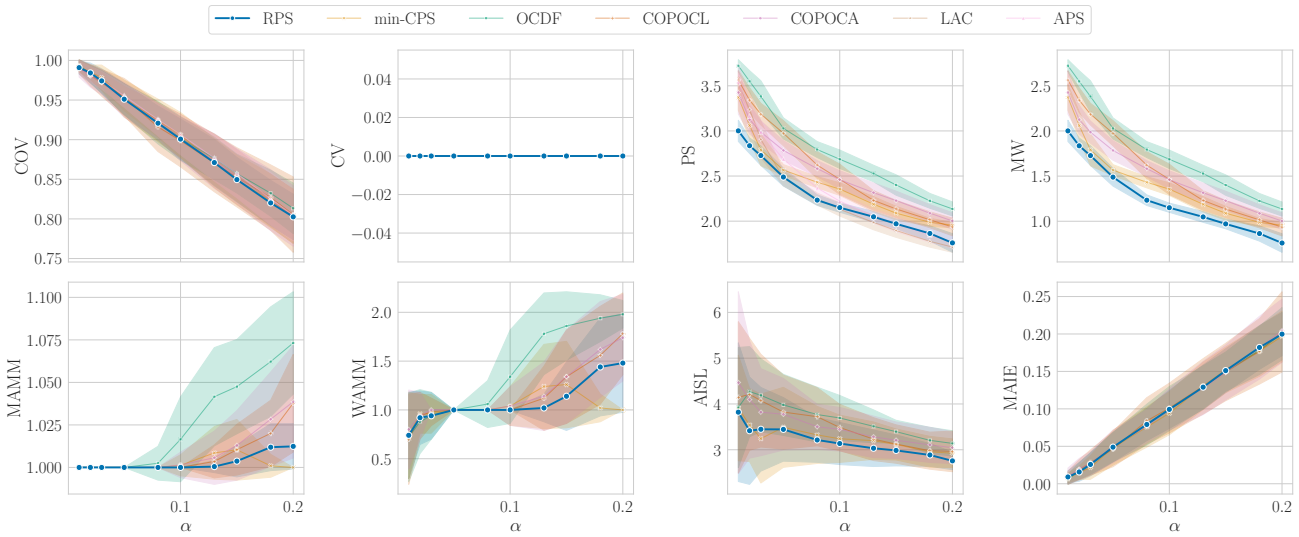


Figure 14: Comparison of prediction sets across methods on swd dataset. Shaded regions indicate standard deviation.

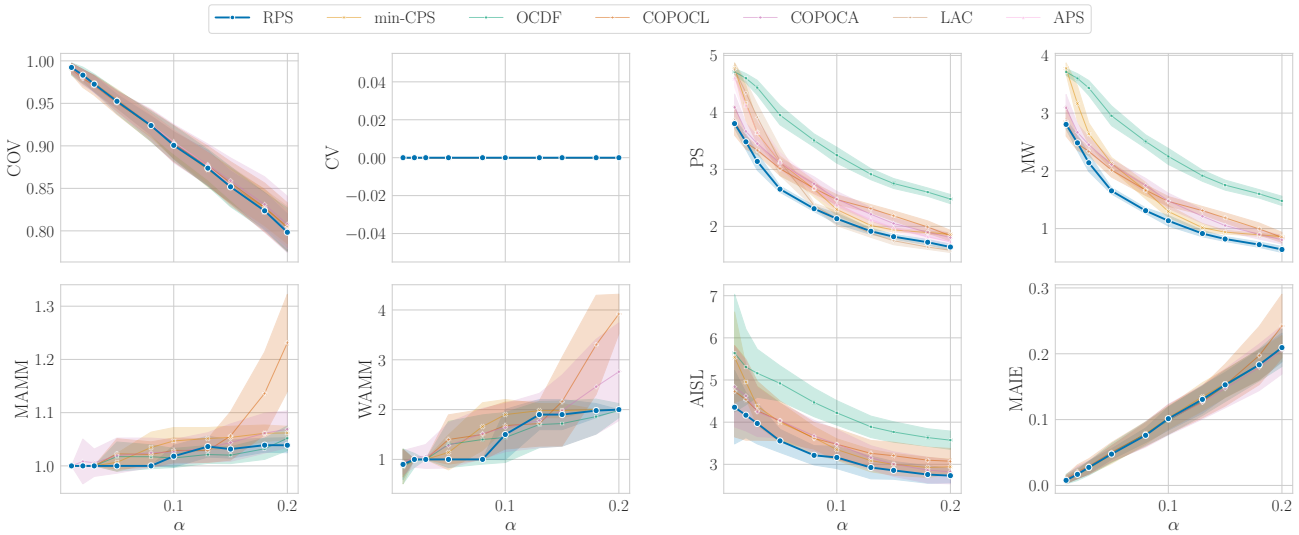


Figure 15: Comparison of prediction sets across methods on winequalityRed dataset. Shaded regions indicate standard deviation.

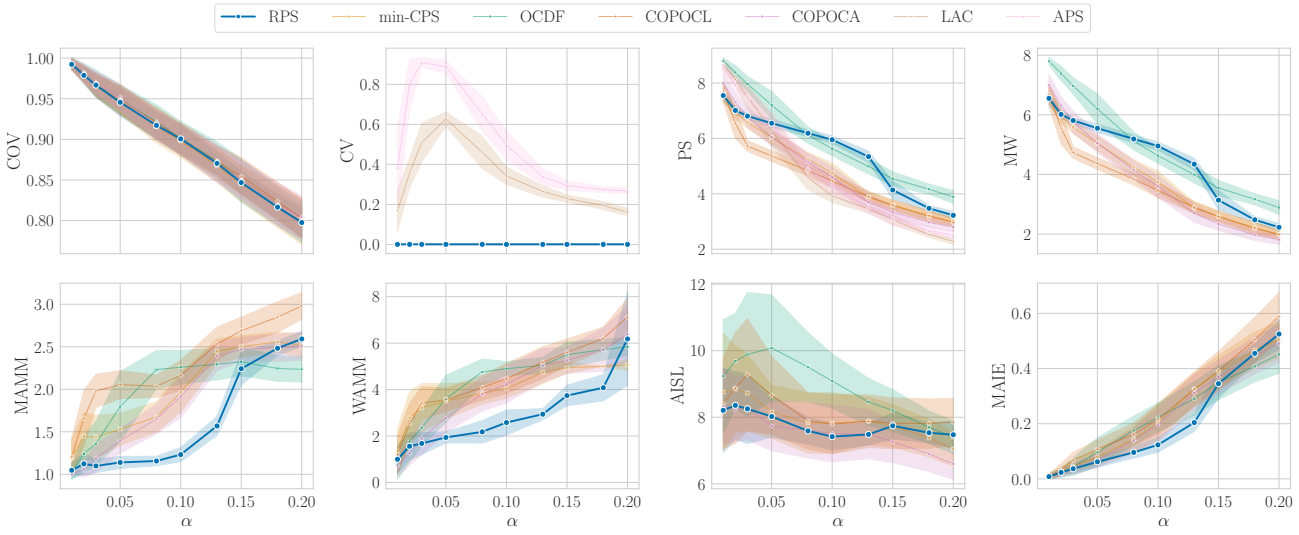


Figure 16: Comparison of prediction sets across methods on insurance dataset. Shaded regions indicate standard deviation.

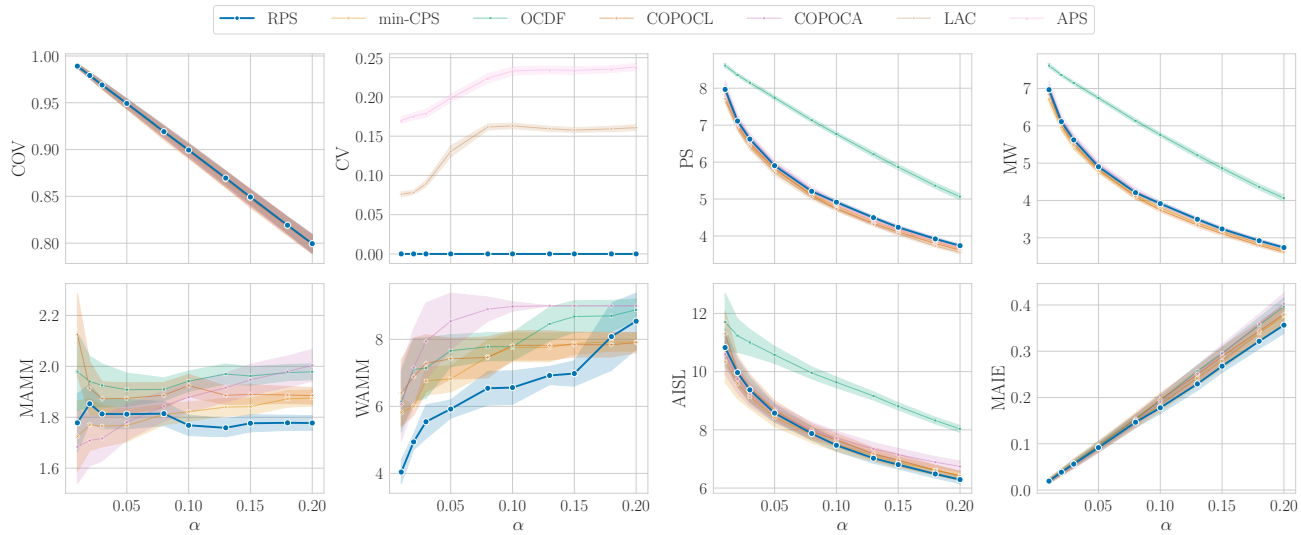


Figure 17: Comparison of prediction sets across methods on melbourneAirbnb dataset. Shaded regions indicate standard deviation.

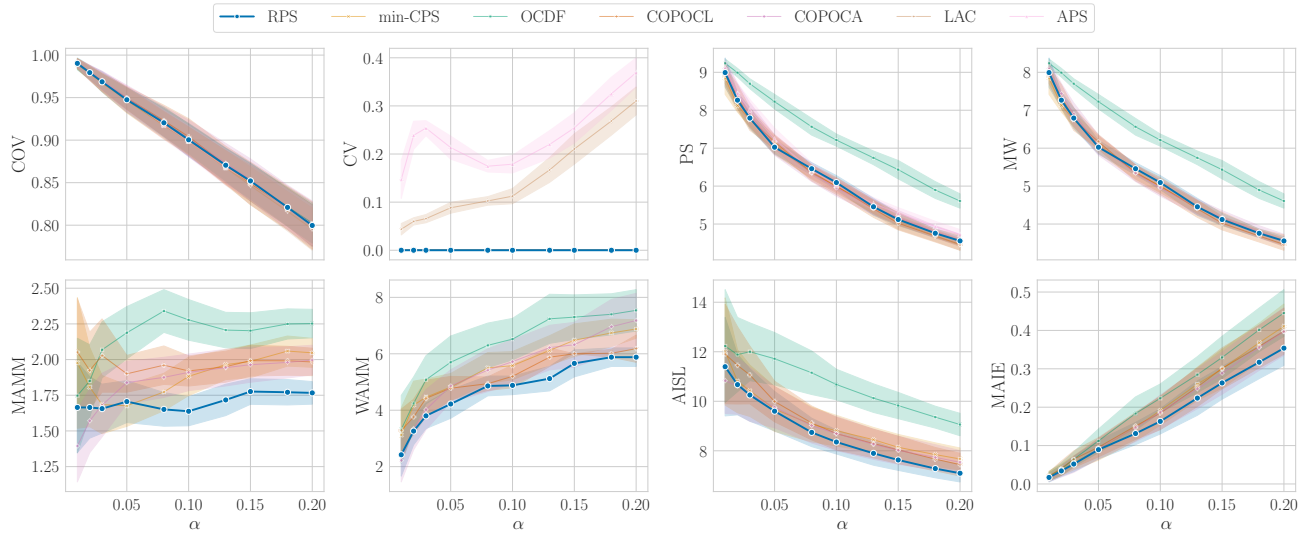


Figure 18: Comparison of prediction sets across methods on cancerDeathRate dataset. Shaded regions indicate standard deviation.

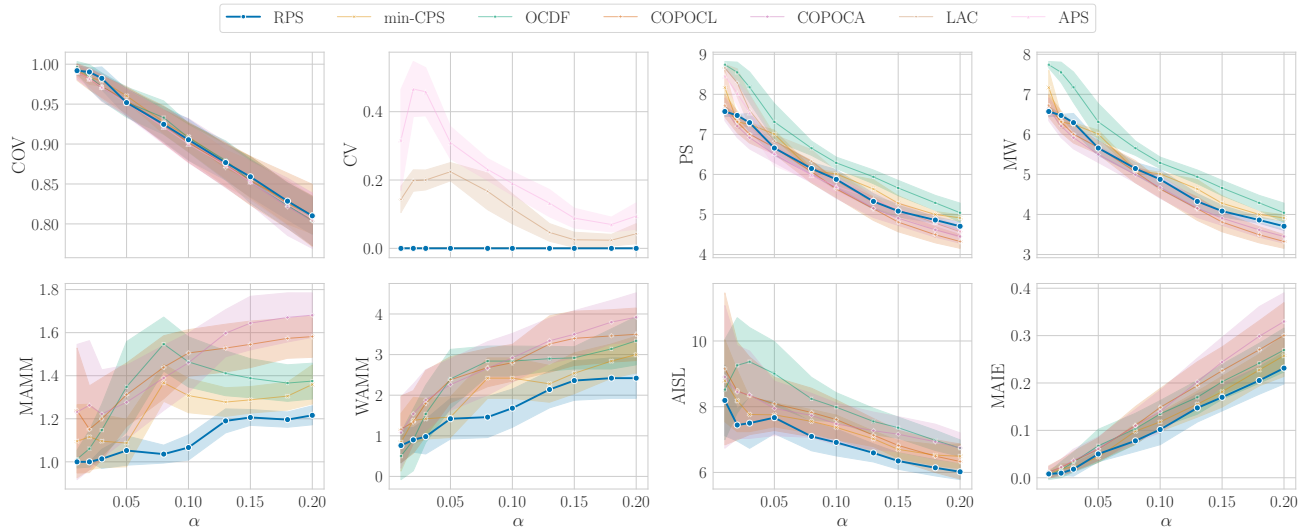


Figure 19: Comparison of prediction sets across methods on era dataset. Shaded regions indicate standard deviation.

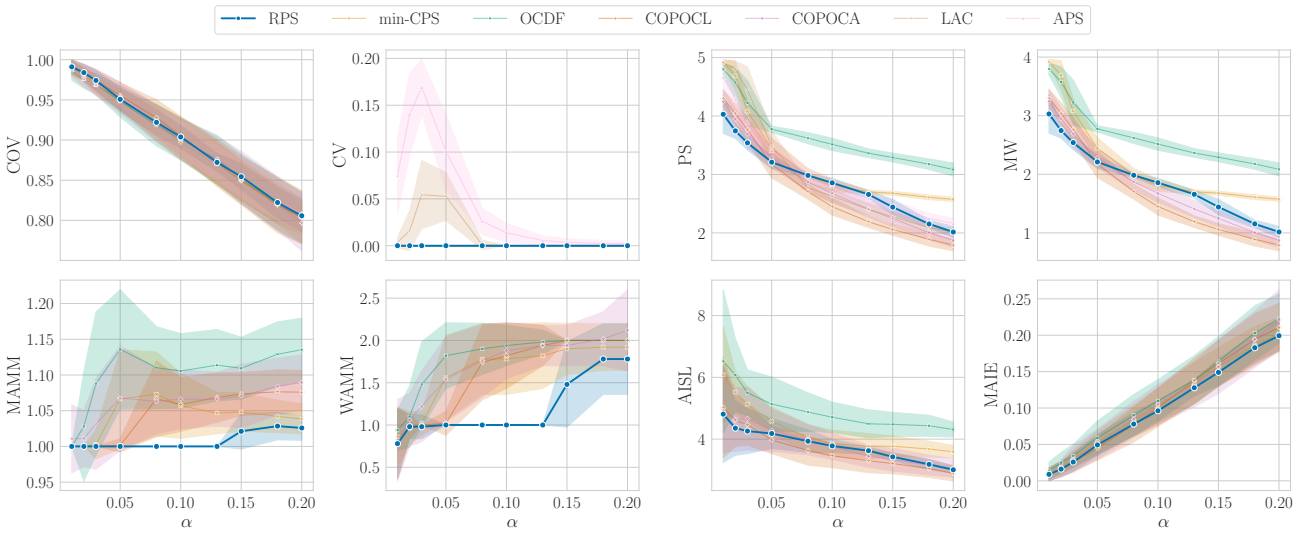
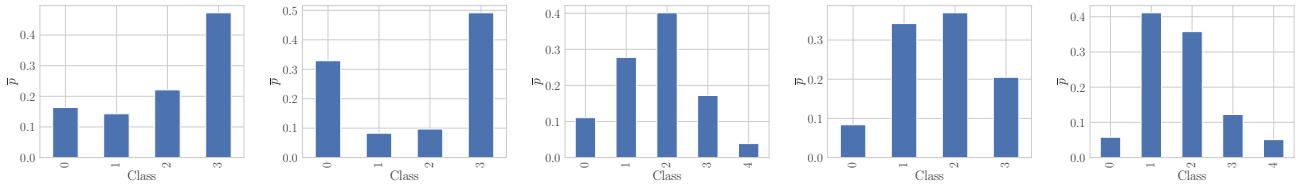
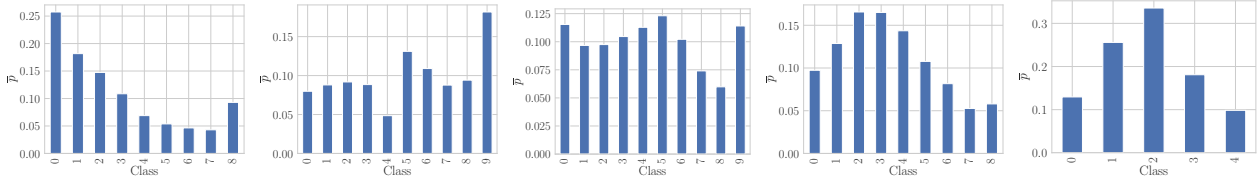


Figure 20: Comparison of prediction sets across methods on lev dataset. Shaded regions indicate standard deviation.



(a) LESTSensors (b) LEVXSensors (c) nhanes (d) swd (e) winequalityRed



(f) insurance (g) melbourneAirbnb (h) cancerDeathRate (i) era (j) lev

Figure 21: Mean predictive probabilities (MP) for the different datasets with the CE loss.

Table 11: Comparison of the different non-conformity measures using LightGBM over tabular datasets at $\alpha = 0.02$, $\alpha = 0.05$ and $\alpha = 0.1$.

Dataset	α	Method	COV	PS (↓)	CV (↓)	MW (↓)	MAMM (↓)	WAMM (↓)	MAIE (↓)	AISL (↓)
heartDisease	0.02	APS	0.985 ± 0.019	3.625 ± 0.388	0.072 ± 0.071	-	-	-	-	-
		LAC	0.984 ± 0.023	3.634 ± 0.427	0.042 ± 0.049	-	-	-	-	-
		OCDF	0.981 ± 0.023	3.872 ± 0.109	0.000 ± 0.000	2.872 ± 0.109	0.940 ± 0.874	1.120 ± 1.003	0.028 ± 0.029	5.686 ± 2.836
		RPS	0.985 ± 0.019	3.718 ± 0.240	0.000 ± 0.000	2.718 ± 0.240	**0.480** ± 0.505	**0.480** ± 0.505	**0.015** ± 0.019	**4.176** ± 1.687
		min-CPS	0.984 ± 0.022	**3.592** ± 0.489	0.000 ± 0.000	**2.592** ± 0.489	1.028 ± 1.097	1.220 ± 1.329	0.033 ± 0.043	5.880 ± 3.808
	0.05	APS	0.967 ± 0.031	3.033 ± 0.407	0.149 ± 0.056	-	-	-	-	-
		LAC	0.973 ± 0.026	3.081 ± 0.402	0.096 ± 0.033	-	-	-	-	-
		OCDF	0.966 ± 0.034	3.774 ± 0.109	0.000 ± 0.000	2.774 ± 0.109	1.101 ± 0.696	1.320 ± 0.844	0.048 ± 0.049	4.686 ± 1.871
		RPS	0.967 ± 0.027	3.363 ± 0.307	0.000 ± 0.000	2.363 ± 0.307	**0.800** ± 0.437	**0.840** ± 0.510	**0.035** ± 0.031	**3.760** ± 0.946
		min-CPS	0.957 ± 0.030	**2.693** ± 0.204	0.000 ± 0.000	**1.693** ± 0.204	1.633 ± 0.796	2.260 ± 1.065	0.075 ± 0.050	4.703 ± 1.826
	0.1	APS	0.926 ± 0.048	2.275 ± 0.271	0.196 ± 0.040	-	-	-	-	-
		LAC	0.918 ± 0.051	2.226 ± 0.275	0.152 ± 0.036	-	-	-	-	-
		OCDF	0.905 ± 0.047	3.576 ± 0.135	0.000 ± 0.000	2.576 ± 0.135	1.411 ± 0.275	2.080 ± 0.601	0.139 ± 0.072	5.356 ± 1.303
		RPS	0.910 ± 0.038	2.418 ± 0.239	0.000 ± 0.000	1.418 ± 0.239	**1.345** ± 0.172	**1.920** ± 0.340	**0.124** ± 0.058	**3.893** ± 0.955
		min-CPS	0.905 ± 0.052	**2.068** ± 0.299	0.000 ± 0.000	**1.068** ± 0.299	1.542 ± 0.513	2.640 ± 0.802	0.149 ± 0.073	4.051 ± 1.195
mammoexp	0.02	APS	0.987 ± 0.016	2.914 ± 0.072	0.049 ± 0.042	-	-	-	-	-
		LAC	0.988 ± 0.016	2.920 ± 0.078	0.050 ± 0.035	-	-	-	-	-
		OCDF	0.986 ± 0.017	2.909 ± 0.053	0.000 ± 0.000	1.909 ± 0.053	0.584 ± 0.503	0.600 ± 0.535	0.014 ± 0.018	3.331 ± 1.753
		RPS	0.990 ± 0.013	2.940 ± 0.068	0.000 ± 0.000	1.940 ± 0.068	**0.500** ± 0.505	**0.500** ± 0.505	**0.010** ± 0.013	**2.976** ± 1.227
		min-CPS	0.990 ± 0.012	**2.893** ± 0.155	0.000 ± 0.000	**1.893** ± 0.155	0.670 ± 0.733	0.740 ± 0.828	0.015 ± 0.023	3.436 ± 2.122
	0.05	APS	0.948 ± 0.032	2.667 ± 0.143	0.147 ± 0.045	-	-	-	-	-
		LAC	0.958 ± 0.031	2.617 ± 0.190	0.107 ± 0.035	-	-	-	-	-
		OCDF	0.961 ± 0.029	2.801 ± 0.069	0.000 ± 0.000	1.801 ± 0.069	1.120 ± 0.347	1.480 ± 0.614	0.046 ± 0.032	3.633 ± 1.206
		RPS	0.957 ± 0.028	2.739 ± 0.092	0.000 ± 0.000	1.739 ± 0.092	**0.962** ± 0.199	**0.980** ± 0.247	**0.044** ± 0.028	**3.483** ± 1.058
		min-CPS	0.957 ± 0.031	**2.551** ± 0.081	0.000 ± 0.000	**1.551** ± 0.081	1.278 ± 0.535	1.740 ± 0.664	0.059 ± 0.039	3.903 ± 1.507
	0.1	APS	0.883 ± 0.047	2.308 ± 0.154	0.194 ± 0.038	-	-	-	-	-
		LAC	0.905 ± 0.047	**2.276** ± 0.136	0.149 ± 0.033	-	-	-	-	-
		OCDF	0.904 ± 0.046	2.651 ± 0.081	0.000 ± 0.000	1.651 ± 0.081	**1.062** ± 0.179	**1.560** ± 0.541	**0.103** ± 0.048	3.714 ± 0.883
		RPS	0.901 ± 0.031	2.323 ± 0.173	0.000 ± 0.000	**1.323** ± 0.173	1.095 ± 0.095	1.600 ± 0.495	0.110 ± 0.039	3.516 ± 0.633
		min-CPS	0.910 ± 0.040	2.389 ± 0.145	0.000 ± 0.000	1.389 ± 0.145	1.198 ± 0.152	1.860 ± 0.351	0.106 ± 0.045	**3.510** ± 0.770
support	0.02	APS	0.984 ± 0.016	**2.288** ± 0.335	0.130 ± 0.066	-	-	-	-	-
		LAC	0.983 ± 0.016	2.292 ± 0.355	0.105 ± 0.060	-	-	-	-	-
		OCDF	0.987 ± 0.013	2.758 ± 0.046	0.000 ± 0.000	1.758 ± 0.046	1.423 ± 0.777	1.440 ± 0.787	0.024 ± 0.025	4.197 ± 2.443
		RPS	0.988 ± 0.011	2.628 ± 0.175	0.000 ± 0.000	1.628 ± 0.175	**0.760** ± 0.431	**0.760** ± 0.431	**0.012** ± 0.011	**2.875** ± 0.958
		min-CPS	0.985 ± 0.014	2.383 ± 0.510	0.000 ± 0.000	**1.383** ± 0.510	1.239 ± 0.746	1.520 ± 0.863	0.023 ± 0.019	3.684 ± 1.529
	0.05	APS	0.953 ± 0.022	1.656 ± 0.128	0.060 ± 0.017	-	-	-	-	-
		LAC	0.954 ± 0.023	**1.651** ± 0.124	0.051 ± 0.016	-	-	-	-	-
		OCDF	0.952 ± 0.024	2.627 ± 0.066	0.000 ± 0.000	1.627 ± 0.066	1.654 ± 0.168	2.000 ± 0.000	0.078 ± 0.037	4.739 ± 1.409
		RPS	0.956 ± 0.022	2.327 ± 0.077	0.000 ± 0.000	1.327 ± 0.077	**1.000** ± 0.000	**1.000** ± 0.000	**0.044** ± 0.022	**3.080** ± 0.803
		min-CPS	0.955 ± 0.021	1.661 ± 0.102	0.000 ± 0.000	**0.661** ± 0.102	1.363 ± 0.163	1.960 ± 0.198	0.061 ± 0.030	3.088 ± 1.100
	0.1	APS	0.904 ± 0.035	1.372 ± 0.094	0.044 ± 0.017	-	-	-	-	-
		LAC	0.907 ± 0.033	**1.346** ± 0.117	0.043 ± 0.015	-	-	-	-	-
		OCDF	0.902 ± 0.030	2.370 ± 0.167	0.000 ± 0.000	1.370 ± 0.167	1.507 ± 0.113	2.000 ± 0.000	0.148 ± 0.044	4.323 ± 0.717
		RPS	0.909 ± 0.033	1.676 ± 0.346	0.000 ± 0.000	0.676 ± 0.346	**1.120** ± 0.126	**1.620** ± 0.490	**0.105** ± 0.050	**2.786** ± 0.709
		min-CPS	0.908 ± 0.031	1.356 ± 0.102	0.000 ± 0.000	**0.356** ± 0.102	1.416 ± 0.109	2.000 ± 0.000	0.131 ± 0.047	2.978 ± 0.845
winequalityRed	0.02	APS	0.981 ± 0.008	3.090 ± 0.185	0.002 ± 0.003	-	-	-	-	-
		LAC	0.981 ± 0.009	3.151 ± 0.279	0.001 ± 0.002	-	-	-	-	-
		OCDF	0.981 ± 0.012	4.297 ± 0.147	0.000 ± 0.000	3.297 ± 0.147	1.022 ± 0.043	1.220 ± 0.418	0.020 ± 0.013	5.303 ± 1.141
		RPS	0.981 ± 0.009	**2.820** ± 0.127	0.000 ± 0.000	**1.820** ± 0.127	**1.000** ± 0.000	**1.000** ± 0.000	0.019 ± 0.009	**3.720** ± 0.815
		min-CPS	0.982 ± 0.008	3.120 ± 0.308	0.000 ± 0.000	2.120 ± 0.308	1.000 ± 0.000	1.000 ± 0.000	**0.019** ± 0.008	3.970 ± 0.566
	0.05	APS	0.950 ± 0.015	2.374 ± 0.127	0.006 ± 0.004	-	-	-	-	-
		LAC	0.949 ± 0.015	2.347 ± 0.137	0.006 ± 0.004	-	-	-	-	-
		OCDF	0.948 ± 0.015	3.951 ± 0.061	0.000 ± 0.000	2.951 ± 0.061	1.027 ± 0.032	1.460 ± 0.503	0.054 ± 0.016	5.104 ± 0.571
		RPS	0.949 ± 0.014	**2.198** ± 0.062	0.000 ± 0.000	**1.198** ± 0.062	**1.000** ± 0.000	**1.000** ± 0.000	0.051 ± 0.014	**3.233** ± 0.511
		min-CPS	0.950 ± 0.017	2.354 ± 0.142	0.000 ± 0.000	1.354 ± 0.142	1.000 ± 0.000	1.000 ± 0.000	**0.050** ± 0.017	3.374 ± 0.549
	0.1	APS	0.897 ± 0.023	1.879 ± 0.086	0.011 ± 0.005	-	-	-	-	-
		LAC	0.897 ± 0.024	**1.824** ± 0.111	0.008 ± 0.004	-	-	-	-	-
		OCDF	0.897 ± 0.024	3.664 ± 0.087	0.000 ± 0.000	2.664 ± 0.087	1.018 ± 0.019	1.540 ± 0.503	0.105 ± 0.026	4.766 ± 0.439
		RPS	0.896 ± 0.022	1.831 ± 0.091	0.000 ± 0.000	0.831 ± 0.091	**1.007** ± 0.014	**1.220** ± 0.418	**0.104** ± 0.023	**2.921** ± 0.376
		min-CPS	0.896 ± 0.024	1.825 ± 0.113	0.000 ± 0.000	**0.825** ± 0.113	1.016 ± 0.019	1.480 ± 0.505	0.106 ± 0.026	2.951 ± 0.411
nhanes	0.02	APS	0.982 ± 0.006	4.143 ± 0.066	0.010 ± 0.003	-	-	-	-	-
		LAC	0.980 ± 0.006	**4.077** ± 0.073	0.008 ± 0.003	-	-	-	-	-
		OCDF	0.980 ± 0.005	4.705 ± 0.029	0.000 ± 0.000	3.705 ± 0.029	1.090 ± 0.049	1.880 ± 0.328	0.022 ± 0.005	5.858 ± 0.482
		RPS	0.981 ± 0.006	4.094 ± 0.071	0.000 ± 0.000	3.094 ± 0.071	**1.000** ± 0.000	**1.000** ± 0.000	**0.019** ± 0.006	**4.952** ± 0.563
		min-CPS	0.981 ± 0.006	4.091 ± 0.065	0.000 ± 0.000	**3.091** ± 0.065	1.026 ± 0.026	1.540 ± 0.503	0.020 ± 0.006	5.087 ± 0.566
	0.05	APS	0.950 ± 0.010	3.628 ± 0.073	0.034 ± 0.005	-	-	-	-	-
		LAC	0.950 ± 0.008	3.540 ± 0.058	0.025 ± 0.003	-	-	-	-	-
		OCDF	0.949 ± 0.009	4.476 ± 0.045	0.000 ± 0.000	3.476 ± 0.045	1.106 ± 0.030	2.000 ± 0.000	0.056 ± 0.010	5.730 ± 0.357
		RPS	0.949 ± 0.008	**3.536** ± 0.041	0.000 ± 0.000	**2.536** ± 0.041	**1.014** ± 0.014	**1.560** ± 0.501	**0.051** ± 0.009	**4.594** ± 0.315
		min-CPS	0.950 ± 0.008	3.552 ± 0.039	0.000 ± 0.000	2.552 ± 0.039	1.030 ± 0.015	1.920 ± 0.274	0.052 ± 0.008	4.622 ± 0.299
	0.1	APS	0.899 ± 0.012	3.129 ± 0.056	0.050 ± 0.005	-	-	-	-	-
		LAC	0.903 ± 0.012	3.091 ± 0.056	0.029 ± 0.003	-	-	-	-	-
		OCDF	0.900 ± 0.010	4.174 ± 0.036	0.000 ± 0.000	3.174 ± 0.036	1.138 ± 0.025	2.200 ± 0.404	0.114 ± 0.012	5.450 ± 0.209
		RPS	0.900 ± 0.011	**3.001** ± 0.047	0.000 ± 0.000	**2.001** ± 0.047	**1.048** ± 0.017	**2.000** ± 0.000	**0.104** ± 0.012	**4.087** ± 0.201
		min-CPS	0.901 ± 0.013	3.094 ± 0.061	0.000 ± 0.000	2.094 ± 0.061	1.052 ± 0.017	2.000 ± 0.000	0.104 ± 0.014	4.183 ± 0.219
LEVXSensors	0.02	APS	0.980 ± 0.006	3.552 ± 0.066	0.296 ± 0.037	-	-	-	-	-
		LAC	0.981 ± 0.007	3.464 ± 0.069	0.252 ± 0.029	-	-	-	-	-
		OCDF	0.980 ± 0.006	**3.411** ± 0.055	0.000 ± 0.000	**2.411** ± 0.055	1.800 ± 0.121	3.000 ± 0.000	0.036 ± 0.012	6.039 ± 1.179
		RPS	0.980 ± 0.006	3.734 ± 0.021	0.000 ± 0.000	2.734 ± 0.021	**1.000** ± 0.000	**1.000** ± 0.000	**0.020** ± 0.006	**4.739** ± 0.571
		min-CPS	0.981 ± 0.005	3.482 ± 0.052	0.000 ± 0.000	2.482 ± 0.052	1.520 ± 0.153	2.960 ± 0.198	0.029 ± 0.010	5.356 ± 0.927
	0.05	APS	0.948 ± 0.010	3.107 ± 0.073	0.501 ± 0.028	-	-	-	-	-
		LAC	0.949 ± 0.010	**3.001** ± 0.070	0.426 ± 0.026	-	-	-	-	-
		OCDF	0.948 ± 0.010	3.149 ± 0.057	0.000 ± 0.000	**2.149** ± 0.057	1.970 ± 0.089	3.000 ± 0.000	0.102 ± 0.022	6.242 ± 0.824
		RPS	0.948 ± 0.010	3.552 ± 0.060	0.000 ± 0.000	2.552 ± 0.060	**1.008** ± 0.013	**1.320** ± 0.471	**0.052** ± 0.010	**4.643** ± 0.362
		min-CPS	0.952 ± 0.008	3.208 ± 0.035	0.000 ± 0.000	2.208 ± 0.035	1.585 ± 0.086	3.000 ± 0.000	0.076 ± 0.015	5.265 ± 0.557
	0.1	APS	0.900 ± 0.012	2.590 ± 0.059	0.638 ± 0.015	-	-	-	-	-
		LAC	0.898 ± 0.012	**2.498** ± 0.067	0.562 ± 0.017	-	-	-	-	-
		OCDF	0.897 ± 0.012	2.805 ± 0.049	0.000 ± 0.000	**1.805** ± 0.049	2.002 ± 0.049	3.000 ± 0.000	0.205 ± 0.023	5.907 ± 0.422
		RPS	0.900 ± 0.014	3.114 ± 0.038	0.000 ± 0.000	2.114 ± 0.038	**1.224** ± 0.035	**2.000** ± 0.000	**0.123** ± 0.019	**4.579** ± 0.346
		min-CPS	0.899 ± 0.013	2.839 ± 0.053	0.000 ± 0.000	1.839 ± 0.053	1.738 ± 0.058	3.000 ± 0.000	0.175 ± 0.025	5.343 ± 0.463

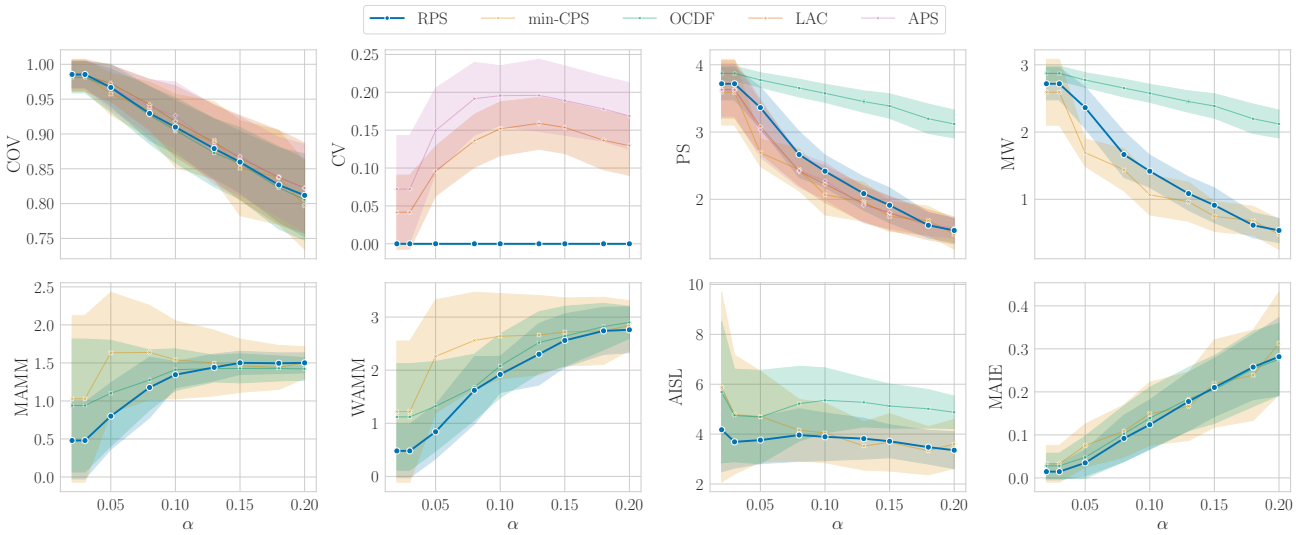


Figure 22: Comparison of prediction sets across methods on heartDisease dataset using LightGBM. Shaded regions indicate standard deviation.

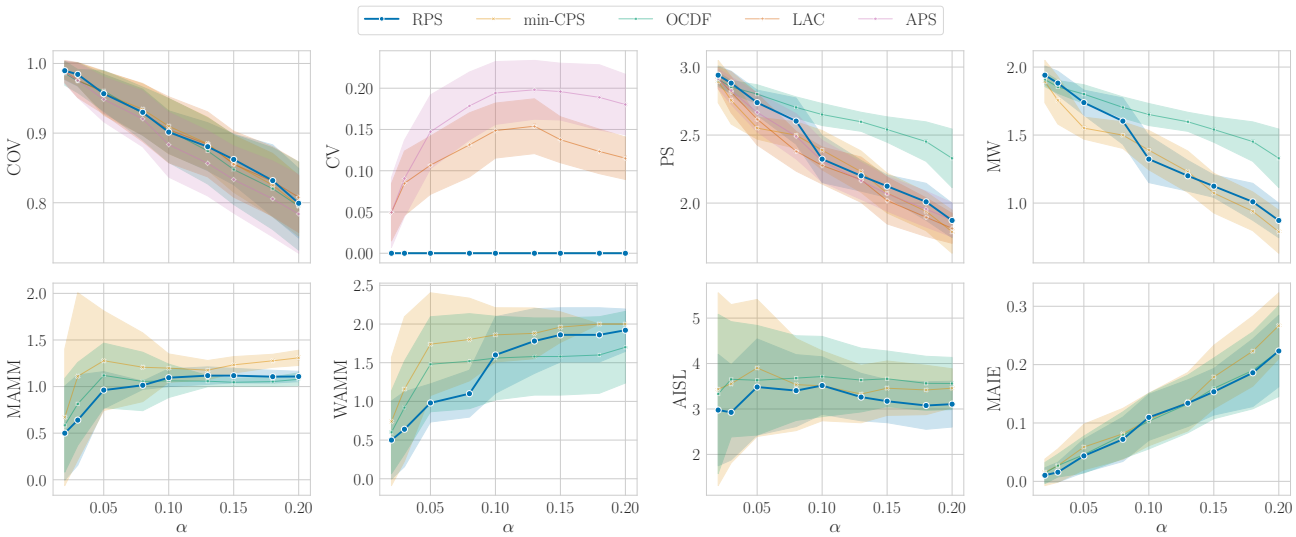


Figure 23: Comparison of prediction sets across methods on mammoexp dataset using LightGBM. Shaded regions indicate standard deviation.

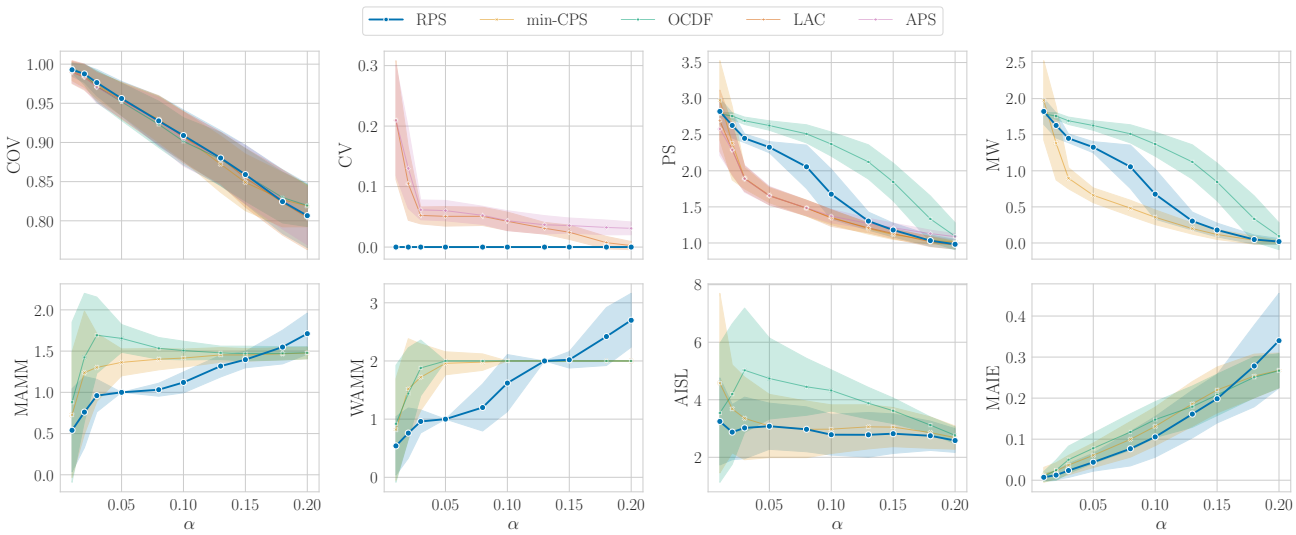


Figure 24: Comparison of prediction sets across methods on support dataset using LightGBM. Shaded regions indicate standard deviation.

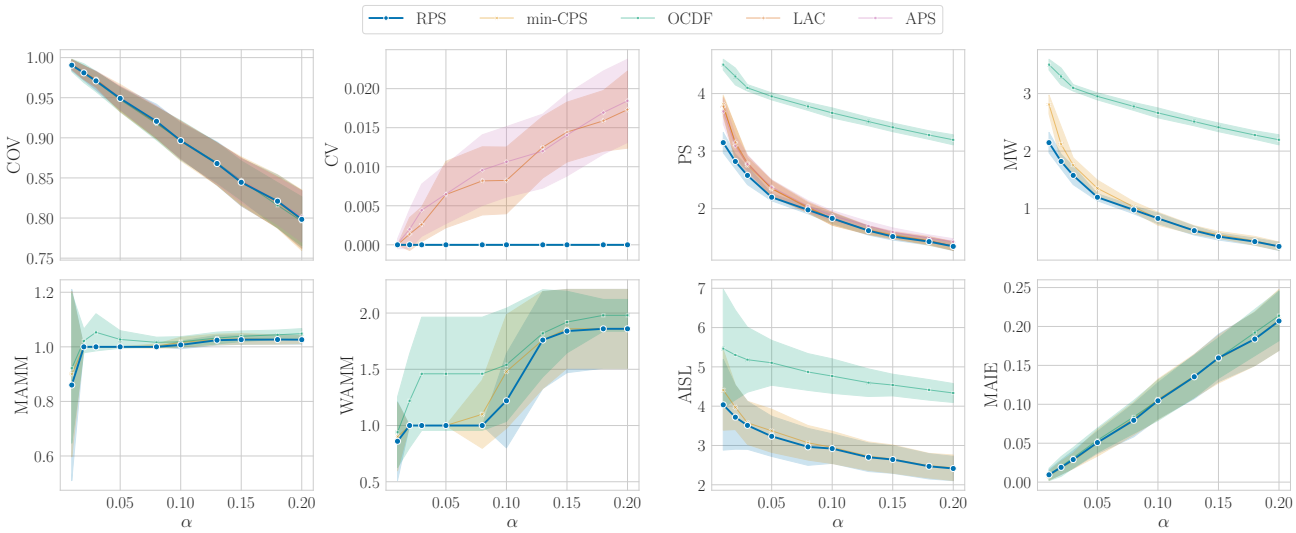


Figure 25: Comparison of prediction sets across methods on winequalityRed dataset using LightGBM. Shaded regions indicate standard deviation.

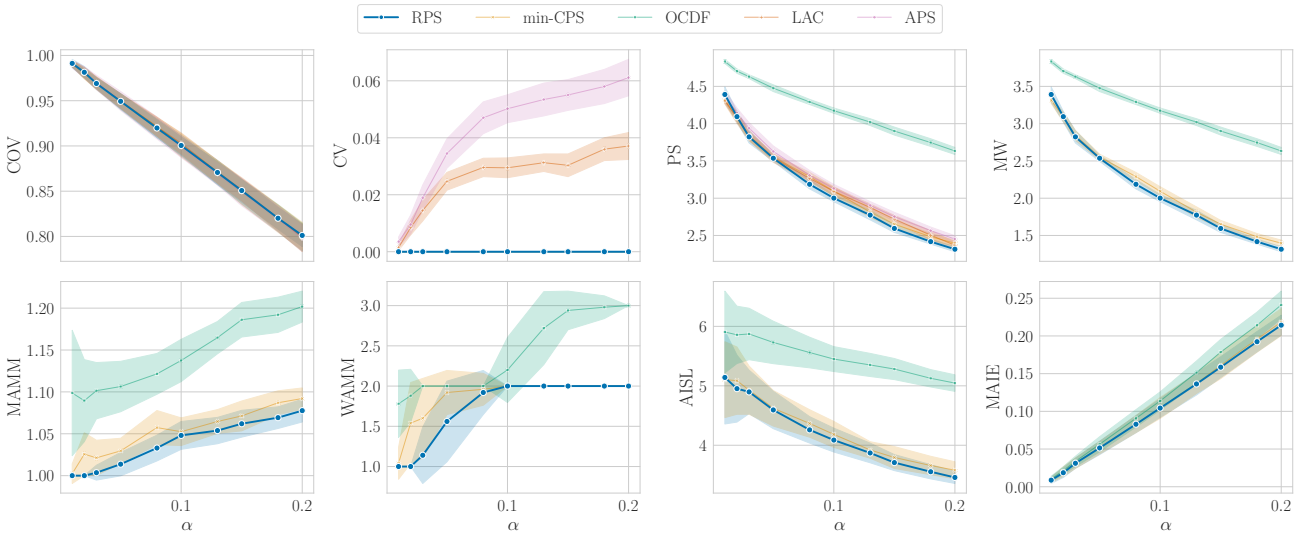


Figure 26: Comparison of prediction sets across methods on nhanes dataset using LightGBM. Shaded regions indicate standard deviation.

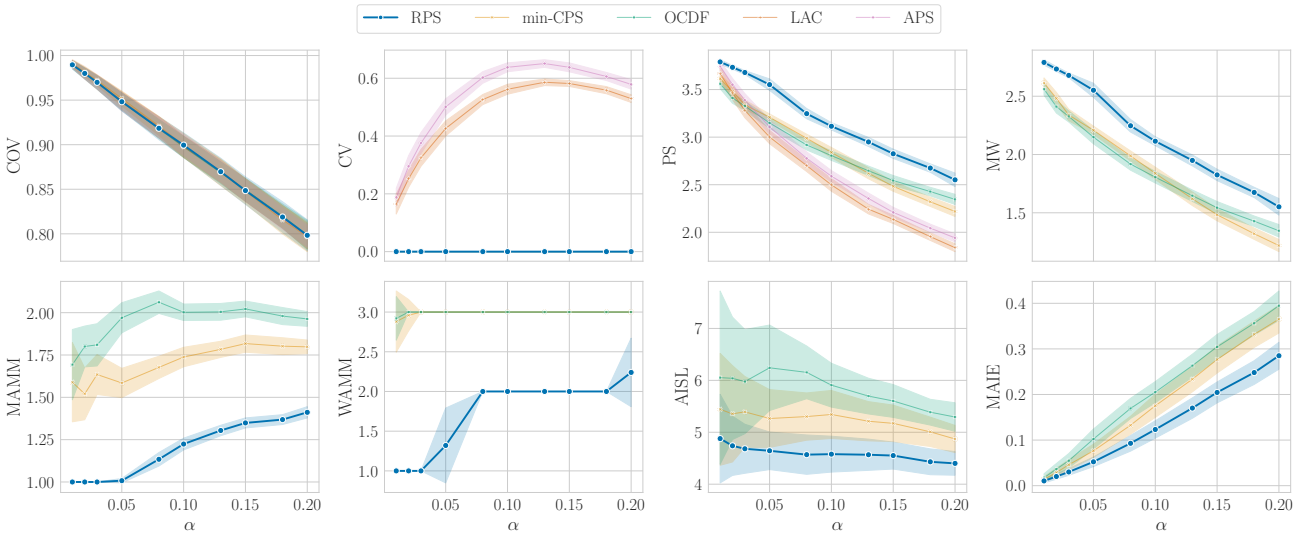


Figure 27: Comparison of prediction sets across methods on LEVXSensors dataset using LightGBM. Shaded regions indicate standard deviation.