Machine Learning-Based Prescriptive Analytics for Automating Expert Ratings

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
Doctor rerum naturalium (Dr. rer. nat.)

eingereicht von

Stefan Anthony Haas

am 08.09.2025

Stefan Anthony Haas
Machine Learning-Based Prescriptive Analytics for Automating Expert Ratings

1. Gutachter/in: Prof. Dr. Eyke Hüllermeier
Ludwig-Maximilians-Universität München

2. Gutachter/in: Prof. Dr. Oliver Müller
Universität Paderborn

3. Gutachter/in: Prof. Dr. Stefan Feuerriegel
Ludwig-Maximilians-Universität München

Tag der Einreichung: 08.09.2025
Tag der Disputation: 11.12.2025

Acknowledgement

I would like to express my deepest gratitude to Prof. Dr. Hüllermeier for accepting me as a PhD student and providing me with the invaluable opportunity to pursue my doctoral studies under his guidance.

I am also very grateful to Prof. Dr. Müller and Prof. Dr. Feuerriegel for serving on my PhD committee, as well as to Prof. Dr. Seidl for chairing the committee.

My sincere thanks go to my managers and team at BMW, particularly Mr. Thomas Aichberger, for their support and trust throughout this endeavor.

Special thanks are also due to the team at the Chair of Artificial Intelligence and Machine Learning (AIML) at LMU Munich, for always making me feel welcome as a member of the group during conferences, workshops, seminars, and visits to the Christmas market.

Finally, I want to express my heartfelt gratitude for the unwavering support and understanding of my family.

Abstract

Expert ratings play an important role in many business domains such as credit scoring, risk assessment, and performance appraisals. However, the growing volume of data and the shortage of qualified human experts necessitate the automation of these ratings. Machine learning-based prescriptive analytics, which leverages historical data to develop predictive models and recommend actionable insights, offers a promising solution to this challenge.

Despite its potential, automating expert ratings through *behavioral cloning*, in which a supervised machine learning model is trained on historical expert decisions, presents unique challenges. The data used in such settings is often biased and inconsistent, including overly used extreme ratings that create data imbalances, midpoint rounding biases that under-utilize the rating scale, and high uncertainty stemming from the stochastic nature of human judgments. These challenges are further exacerbated by the absence of information about the true outcomes of the observed actions, complicating the development of reliable decision models.

This thesis addresses these challenges through the real-world use case of automotive goodwill claim assessment, which necessitates balancing customer satisfaction with financial considerations, particularly in the context of a dynamic environment and safety implications. The proposed contributions explore approaches such as ordinal cost-sensitive hierarchical learning, unimodal soft-label methods, explainable artificial intelligence (XAI), and rating-scale-aware uncertainty quantification. These methods aim to reduce decision biases, improve model alignment with domain-specific objectives, and enhance the interpretability and reliability of machine learning models in prescriptive analytics for expert ratings.

Overall, the thesis contributes to the fields of weakly supervised learning, ordinal classification and uncertainty quantification by underscoring the complexities of human rating data, which often violate standard assumptions in ordinal classification, notably the existence of ground-truth outcome labels and the expectation of unimodal predictive probability distributions. It further explores the underrepresented area of uncertainty quantification for ordinal or cardinal discrete targets, an area often overlooked compared to the more extensively studied nominal classification and regression tasks. By introducing novel uncertainty measures for these target types, this work addresses a critical gap in the field and advances the development of robust machine learning systems for automating expert ratings.

Zusammenfassung

Expertenbewertungen sind in vielen Anwendungsfeldern wie Kreditbewertung, Risikoklassifizierung oder Leistungsbeurteilungen zentral. Angesichts wachsender Datenmengen und fehlender Fachkräfte wird deren Automatisierung immer wichtiger. *Prescriptive Analytics* auf Basis von *Machine Learning*, welche historische Daten für Vorhersagen und Handlungsempfehlungen nutzt, bietet hierfür eine vielversprechende Lösung.

Die Automatisierung mittels *Behavioral Cloning*, dem Training überwachter Modelle auf Basis historischer Experten-Entscheidungen, ist jedoch mit spezifischen Herausforderungen verbunden. Typische Probleme sind verzerrte und inkonsistente Daten, etwa durch häufige Extrembewertungen, Rundungstendenzen zur Skalenmitte oder die intrinsische Unsicherheit menschlicher Urteile. Verstärkt wird dies durch das Fehlen von Informationen über die tatsächlichen Resultate der Entscheidungen, was die Entwicklung verlässlicher Modelle erschwert.

Diese Arbeit adressiert die genannten Probleme am Beispiel der Beurteilung von Kulanzanträgen in der Automobilindustrie, wo Kundenzufriedenheit und finanzielle Erwägungen des Herstellers ausbalanciert werden müssen. Vorgestellt und untersucht werden Ansätze wie ordinale, kosten-sensitive hierarchische Lernverfahren, unimodale *Soft-Label*-Methoden, erklärbare Künstliche Intelligenz (XAI), sowie Unsicherheitsquantifizierung, welche die Struktur ordinaler und kardinaler Expertenbewertungen berücksichtigt. Ziel ist es, Verzerrungen zu reduzieren, Modelle stärker an domänenspezifischen Zielen auszurichten und deren Nachvollziehbarkeit sowie Verlässlichkeit zu verbessern.

Die Arbeit leistet hierbei Beiträge zu den Forschungsfeldern *Weakly Supervised Learning*, *Ordinal Classification* und *Uncertainty Quantification*, indem sie die Besonderheiten menschlicher Bewertungsdaten aufzeigt, die oftmals zentrale Annahmen ordinaler Klassifikation verletzen, insbesondere die Existenz von Zielgrößen als verlässliche Grundwahrheit und die Erwartung unimodaler Vorhersagewahrscheinlichkeiten. Zudem wird das bislang kaum untersuchte Feld der Unsicherheitsquantifizierung für ordinale und kardinale Zielgrößen betrachtet. Durch die Einführung neuer Unsicherheitsmaße wird eine wesentliche Forschungslücke geschlossen und die Grundlage für robuste *Machine Learning*-Systeme zur Automatisierung von Expertenbewertungen geschaffen.

Contents

1 Introduction

In the contemporary business landscape, the ability to make informed and timely decisions is paramount. The advent of big data and advanced analytics has revolutionized the way organizations operate, providing unprecedented insights into various aspects of business performance. Among the various analytical approaches, *prescriptive analytics* stands out as a powerful tool that not only predicts future outcomes but also recommends actions to achieve desired results [Lep+20].

Expert ratings play a crucial role in various business domains, such as credit scoring, risk assessment, product evaluations, stock valuations, and performance appraisals. Traditionally, these ratings are based on the subjective judgments of human experts. While these judgments are valuable, they are often time-consuming, costly, and susceptible to personal biases. Alternatively, rule-based systems are frequently employed; however, these systems also require significant manual maintenance efforts and tend to become large and difficult to manage [Ben08]. The increasing complexity and volume of data in modern business environments necessitate more efficient and consistent methods for generating these ratings. *Machine Learning* (ML) [BN06; Mur22], a subset of *Artificial Intelligence* (AI), offers a promising solution to this challenge. By leveraging historical decision data, ML models can identify decision patterns and make predictions on previously unseen inputs. When combined with prescriptive analytics, these models can not only forecast outcomes but also provide actionable recommendations, either through *Automated Decision Making* (ADM), where actions are taken autonomously, or *Decision Support Systems* (DSS), which assist human decision-makers.

However, such prescriptive models trained solely on historical decisions face a fundamental limitation: the absence of fully supervised data, as historical decisions or actions cannot serve as ground-truth labels for optimal decisions, especially given the lack of actual consequential outcome data. Consequently, this thesis situates itself within the domain of *weakly supervised learning* [Zho18], where models must learn from supervision that, while grounded in trusted expert judgments, is inherently weak and noisy due to variability and occasional misjudgments. Within this broader context, the specific learning paradigm most aligned with this thesis is *behavioral cloning* [Sam11; BUS95; Kum+22], a subfield of *imitation learning* [Zar+24; Hus+17], in which a supervised machine learning model is trained on expert demon-

strations to replicate the decision-making policies of experts. This paradigm operates without access to ground-truth outcomes, a condition that stands in contrast to the core assumption of supervised machine learning [Hül21]. While this approach enables the automation of expert ratings, it also introduces fundamental challenges in model development.

One major concern is the risk of human biases, such as tendencies toward extreme decisions or the underutilization of certain ratings, being directly transferred to ML models, thereby perpetuating existing human decision biases [Akt+21; Nto+20]. To address this, this thesis explores data modeling techniques to mitigate these biases, drawing from the fields of *data-centric AI* [Zha+25] and *weakly supervised learning* [Zho18]. Moreover, the observational nature of the training data introduces inherent noise and variability, which manifests as substantial irreducible *aleatoric* uncertainty due to the stochasticity in human judgments [HW21]. To tackle this challenge, the thesis explores methods for *uncertainty quantification*, aiming to capture the confidence of model predictions to better understand and manage the risks associated with prescribed ratings. Furthermore, information about uncertainty (or inversely confidence) related to a rating request can serve as a proxy signal, helping to partially compensate for the lack of actual outcome data. For instance, it could be assumed that high confidence in a model prediction for a given case reflects a consensus among historically observed expert decisions regarding similar cases, which, given the experts' expertise, is likely to result in fair and reasonable outcomes [Ash85; KW89; CW99]. When empirical validation of a judgment is difficult, delayed, or even impossible, utilizing expert consensus as a surrogate for optimal decisions is a common practice, for instance in auditing and medicine [Ash85; Eva97; Giu+14; Hoh+18]. Nonetheless, confidence is only a proxy and needs to be treated carefully, especially if important criteria like independence of experts and diverse expertise among them are not present [Sur05; Jor15].

An added layer of complexity stems from the structured nature of the prediction task. Expert ratings are typically expressed on ordinal or cardinal rating scales, which require careful modeling of the relative severity of prediction errors across categories. This thesis considers multi-category rating scales with K discrete, ordered categories $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, where the categories follow a natural order, $c_1 \prec c_2 \prec \ldots \prec c_K$. For example, a risk assessment scenario may use ordinal categories such as $\mathcal{C} = \{\texttt{none}, \texttt{low}, \texttt{moderate}, \texttt{high}, \texttt{very high}\}$. Within this broader ordinal framework, the specific focus of this thesis is on cardinal rating scales, where categories are evenly spaced, such as $\mathcal{C} = \{1, 2, \ldots, K\}$.

All in all, the main objective of this thesis is to explore and develop machine learning methods for automating expert ratings in a reliable manner, with a particular focus

on the real-world use case of automotive goodwill claim assessments [MB05]. In this setting, human experts decide whether, and to what extent, customers should be compensated for repair costs when issues arise outside the legal warranty period. These compensation decisions are made on a cardinal rating scale ranging from 0% to 100%, in 10% steps. The task requires balancing the manufacturer's financial interests with customer satisfaction, and it reflects the challenges discussed earlier.

This work has relevance for both academic and industrial contexts. From a research point of view, it extends the use of machine learning and prescriptive analytics to real-world decision-making problems where uncertainty and lack of outcome data play a major role, especially in the area of mimicking expert ratings. In doing so, the thesis contributes to the broader field of weakly supervised learning and introduces novel strategies for modeling and quantifying uncertainty for ordinal and cardinal data. On the industry side, the methods investigated here could provide scalable, consistent alternatives to manual expert ratings, with the potential to save time and costs while maintaining decision reliability and transparency.

The remainder of this thesis is organized as follows:

- **Chapter 2** provides the methodological and general background of this thesis, introducing fundamental concepts of prescriptive analytics (Section 2.1), supervised machine learning, and ordinal classification, which is highly relevant for ordinal and cardinal scale rating data (Section 2.2). Another critical component of modern machine learning is the representation and quantification of uncertainty, which is discussed in Section 2.3. The chapter closes with a discussion of *Explainable Artificial Intelligence* (XAI) and its relevance for the interpretability of machine learning models in Section 2.4.
- **Chapter 3** introduces the specific challenges related to observational rating data using the exemplary use case of automotive goodwill claim assessment. It highlights how these challenges undermine common assumptions in supervised machine learning, particularly in ordinal classification, such as the absence of ground-truth labels and the assumption of unimodal predictive probability distributions. Additionally, the chapter proposes a conceptual framework for automating expert ratings through behavioral cloning, by enhancing a supervised machine learning model with components for bias mitigation, uncertainty representation and quantification, selective classification, and explainability, while addressing cross-cutting concerns such as ordinal target awareness and human oversight. All of this aims to model valuable decisions despite the lack of outcome data.

- **Chapter 4** presents the concrete contributions to the conceptual framework outlined in Chapter 3 of this thesis, detailing the methodologies developed, the experiments conducted, and the insights gained in the realm of machine learning-based prescriptive analytics for automotive goodwill claim assessment and beyond.

  List of publications and contributions to the literature:

  1. **Stefan Haas** and Eyke Hüllermeier. "A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part VI*. vol. 13718. Lecture Notes in Computer Science. Springer, 2022, pp. 170–184
  2. **Stefan Haas** and Eyke Hüllermeier. "Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests". In: *International Journal of Data Science and Analytics* (2024)
  3. **Stefan Haas**, Konstantin Hegestweiler, Michael Rapp, Maximilian Muschalik, and Eyke Hüllermeier. "Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments". In: *Frontiers in Artificial Intelligence - AI in Business* 7 (2024)
  4. **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*. vol. 14174. Lecture Notes in Computer Science. Springer, 2023, pp. 3–18
  5. **Stefan Haas** and Eyke Hüllermeier. "Uncertainty quantification in ordinal classification: A comparison of measures". In: *Int. J. Approx. Reason.* 186 (2025), p. 109479
  6. **Stefan Haas** and Eyke Hüllermeier. "Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction". In: *Machine Learning* (2026)

- **Chapter 5** concludes the thesis by providing a summary, discussing its limitations, and offering an outlook for future work.

2 Foundations

This chapter provides an overview of foundational methods for automating expert ratings using machine learning-based prescriptive analytics, specifically in the form of behavioral cloning. It begins with an exploration of prescriptive analytics, followed by a discussion of supervised learning, with a particular emphasis on ordinal classification due to its applicability and relevance to both ordinal and cardinal rating scale data. Additionally, this chapter addresses uncertainty representation and quantification in machine learning, as well as explainable AI, which are essential components for ensuring the reliable automation of expert ratings.

2.1 Prescriptive Analytics

With the ever-increasing amount of data produced in today's business world, the demand for gaining insights from this data and automating business processes based on it remains strong. A crucial role in this regard is played by *business analytics*, which aims to enable organizations to make quicker, better, and more intelligent decisions with the goal of creating business value [Fra+19; Lep+20]. Specifically, it refers to the extensive use of data created by business operations to support better decision-making through statistical and quantitative analysis, explanatory and predictive models, and fact-based management decisions. Business analytics is categorized into four main stages characterized by different levels of complexity, business value, and intelligence [ŠP18] (Figure 2.1):

- **Descriptive Analytics**: The first stage focuses on collecting, categorizing, and classifying data, as well as identifying and visualizing relevant patterns. The goal is to answer the question "What has happened?" This is commonly achieved through visualization, dashboards, statistical analysis, and data mining, for example, sales or employee performance reports shown in dashboards.
- **Diagnostic Analytics**: The second stage aims to provide further insights into past events by answering the question "Why did it happen?" Examples include root cause analysis of quality issues in manufacturing or employee turnover analysis.

- **Predictive Analytics**: This stage addresses the question "What will happen?" by predicting future events. Typically, this is done using large volumes of historical data and techniques such as machine learning, data mining, and statistics, for instance, sales forecasting, customer churn prediction, healthcare outcome prediction, or predictive maintenance.
- **Prescriptive Analytics**: The final and most advanced stage is prescriptive analytics, which provides actionable outcomes or suggestions aimed at answering "What should I do?" It seeks to recommend optimal decisions that maximize business value or other specific criteria set by the organization. Examples include supply chain optimization, healthcare treatment optimization, or financial portfolio management.

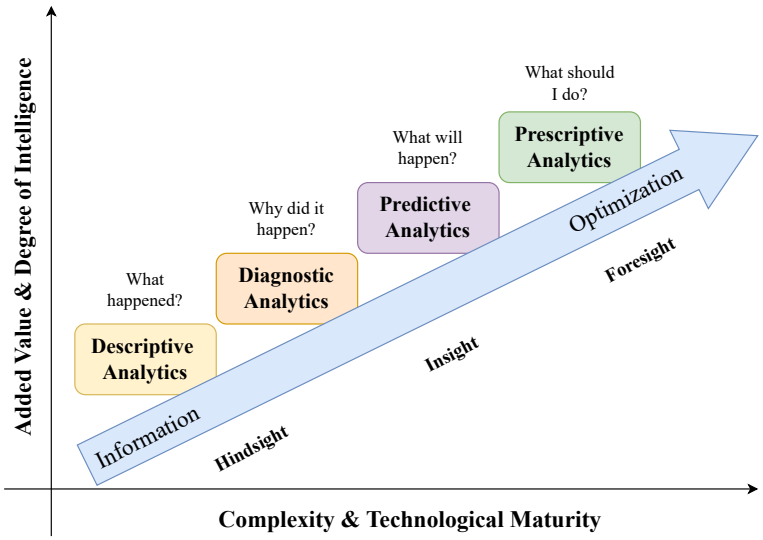


Fig. 2.1: Added business value and complexity of different stages of business analytics [ŠP18].

To date, the major focus of industry and academia has been on descriptive, diagnostic, and predictive analytics [Lep+20]. Nevertheless, there has been an increasing interest and shift towards prescriptive analytics in recent years [BK20; PSM20; Pes+20; Lat+24], as it possesses the highest potential for delivering increased business value, either through decision support or automation. While machine learning is popular for predictive analytics, the dominant techniques underlying prescriptive analytics applications are *mathematical programming* and *logic-based models* [Lep+20]. This focus appears reasonable, as mathematical programming, implemented in the form of objective functions and constraints, and logic-based models, implemented through `if-then` decision rules [Gro+11], allow for precise, transparent, and effective definition of objectives and constraints. Nevertheless, a significant drawback of these approaches is their reliance on domain expert

knowledge to manually define objectives, constraints, or decision rules upfront. Due to the limited availability and high costs of domain experts, this situation is not ideal.

Nowadays, given the availability of big data, there is a significant opportunity to decrease dependency on domain expertise by utilizing these data sets through machine learning to "learn knowledge" instead of requiring human experts to specify it manually upfront. The role of the human thereby shifts from direct input to a more supervisory role in the lifecycle. Shifting the application of machine learning from predictive to prescriptive analytics has been termed *prescriptive machine learning* and introduces entirely new challenges, particularly when the data is observational and human decision-makers act as teachers [Hül21], such as in mimicking expert judgments using supervised learning. For instance, subjective personal preferences and views of decision-makers can influence prescription choices in medicine or loan approvals in finance. Training models on such biased data is likely to result in biased models, which can lead to severe unintended consequences when deployed in real-world scenarios [Ang+22b]. Consequently, the core assumption of supervised machine learning, that the training data accurately represents ground truth outcomes (Section 2.2), is often violated. Moreover, there is even the question of whether something like ground truth exists in prescriptive ML, as the goal in prescriptive ML shifts from approximating ground-truth data, as in standard predictive supervised ML, to learning "practicable decisions" that, for instance, maximize business outcomes [Hül21].

Causal machine learning is increasingly preferred over classical supervised machine learning for prescriptive analytics in general, and automated decision making in particular, as it enables more nuanced estimation of treatment effects based on inputs X, treatments T, and outcomes Y, rather than merely predicting outcomes or mimicking expert decisions [Ker+25; Yao+21], for instance, in medicine [Feu+24] or the public sector [Fis+24]. However, obtaining such data is often challenging in practice due to confounding factors, incomplete observations, and the high cost of controlled experiments, and in many cases, collecting this data is not even feasible or applicable. Concretely, causal ML enables the estimation of treatment effects such as the *Average Treatment Effect* (ATE), defined as

$$\text{ATE} = \mathbb{E}[Y(T=1) - Y(T=0)],$$

the *Conditional Average Treatment Effect* (CATE), defined as

$$\text{CATE} = \mathbb{E}[Y(T=1) - Y(T=0) \mid X = \boldsymbol{x}],$$

or the *Individual Treatment Effect* (ITE), defined as

$$\text{ITE} = Y(T = 1) - Y(T = 0).$$

Here, $Y(T = 1)$ and $Y(T = 0)$ are potential outcomes under treatment and control, respectively, such as improved medical conditions or not. $T \in \{0, 1\}$ represents the presence or absence of treatment, such as medication or no medication [Yao+21]. While ATE and CATE can be estimated under certain assumptions [CBM22]: *Unconfoundedness* $(Y(T = 0), Y(T = 1) \perp T|X)$ requires that outcomes Y are independent of treatments T given covariates X, *Positivity* assumes that any subject $\boldsymbol{x} \in \mathcal{X}$ has a non-zero probability of receiving the treatment $0 < P(T = 1 \mid X = \boldsymbol{x})$, and *Stable Unit Treatment Value Assumption* (SUTVA) assumes that individuals do not interfere with each other, meaning a treatment applied to one individual does not affect the others. The ITE itself is fundamentally unobservable for any individual instance, since only one of the two potential outcomes, under treatment or control, can be observed. This limitation is known as the *fundamental problem of causal inference* [Hol86]. Furthermore, using causal machine learning in a decision-making task is a two-step process: First, treatment effects are estimated, which are then used to determine an optimal downstream decision considering other external factors and constraints. This is similar to the usage of supervised machine learning in prescriptive analytics, where an outcome is first predicted (e.g., whether the customer will churn) and subsequently used to make a decision. Hence, the estimated target is only indirectly linked to the optimal decision policy to be discovered from data.

A related paradigm also highly relevant for prescriptive machine learning is *offline contextual bandit learning* [Sak23], or more broadly *off-policy learning* [Hül21; Fis+24], where an algorithm selects an action a (e.g., show an ad) for a given input $\boldsymbol{x}$ (e.g., a user profile) and receives partial feedback, only for the action performed, in the form of a reward r (e.g., user clicks on the ad) [SJ15a; SJ15b; JSD18]. The goal is to learn a policy $\pi(a \mid \boldsymbol{x})$ that maximizes the expected reward R based on the chosen actions a and given contexts $\boldsymbol{x}$:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} R(\pi) = \arg\max_{\pi \in \Pi} \mathbb{E}_{(\boldsymbol{x},a)\sim\pi} \left[r(a \mid \boldsymbol{x}) \right],$$

where $r(a \mid \boldsymbol{x})$ denotes the reward received when taking action a in context $\boldsymbol{x}$. This approach offers a more direct method for discovering an optimal decision policy compared to deriving it from estimated treatment effects or predictions. In the offline setting, learning must occur from previously logged interactions rather than through active exploration, which is typically unsuitable in high-stakes environments [Fis+24]. Unlike *reinforcement learning* [SB+98], which involves

sequential decision-making over multiple time steps and evolving environments, bandit learning focuses on a single decision point with a static environment. Though, in particular, *offline reinforcement learning* [Lev+20] may also be a good fit for prescriptive sequential decision-making tasks, the focus here is on static, one-time decisions due to their natural applicability to rating applications in general and the considered use case of automotive goodwill claim assessment in particular.

Table 2.1 summarizes common learning settings in prescriptive machine learning, categorized by the types of observed data. As already mentioned, the setting most aligned with this thesis is *Behavioral Cloning* (BC), specifically behavioral cloning from observational data [TWS18]. This approach uses supervised learning (Section 2.2) to capture the decision policy π of experts from historical observed context-action pairs $\mathcal{O} = \{(\boldsymbol{x}_1, a_1), \ldots, (\boldsymbol{x}_N, a_N)\}$ by minimizing a loss function l:

$$\hat{\pi} = \arg\min_{\pi \in \Pi} \mathbb{E}_{(\boldsymbol{x},a)\sim\pi_E} \left[l(\pi(\boldsymbol{x}), a) \right] = \arg\min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} l(\pi(\boldsymbol{x}_i), a_i),$$

where π_E denotes the underlying expert policy distribution generating the observed context-action pairs $(\boldsymbol{x}, a)$. Although BC originates from domains such as robotics and autonomous driving [Cod+19; LA20], it has also been applied in areas like algorithmic trading [Liu+20b; Xu+24; SXS25] or medicine [Mat+25]. In the reinforcement learning literature, BC typically refers to supervised policy learning from expert demonstrations over sequential decision-making trajectories. Here, the term is used in a broader sense, encompassing the special case of *single-step* decision problems in which each expert decision is independent (e.g., an isolated expert rating for a specific case). Although no temporal dependencies or state transitions exist in this setting, the underlying methodological principle, training a model to imitate expert-chosen actions from historical (context, action) pairs, remains the same. Hence, BC is used within this thesis as a generic label to clearly distinguish learning from observed one-step expert actions or decisions, without ground-truth outcome data, from standard supervised learning settings in which such outcomes are available.

BC provides recommendations aligned with what experts would likely decide in similar scenarios ("What should be done according to experts?"). It is the least informed among prescriptive learning paradigms, as it lacks access to actual outcomes Y and only provides information about inputs X and expert decisions or actions A. Nonetheless, BC can achieve several prescriptive goals valued by organizations, such as ensuring consistency in decision-making, since most traditional supervised machine learning models produce deterministic results given identical inputs, and enabling (near) real-time decisions. In particular, results from BC intuitively cor-

respond to the decisions taken by the majority of experts in the observed data for similar contexts, thereby leveraging collective expertise [Mat+25]. Thus, BC standardizes frequent decision patterns, reducing unwarranted variations. Combined with proper uncertainty quantification (Section 2.3) and XAI (Section 2.4), BC can enhance uncertainty awareness and transparency for generated judgments, which are often obscured in manual decision processes. Transparency and uncertainty awareness, in turn, support feedback loops that improve decision quality over time. Furthermore, automation can free human experts for complex or exceptional cases, increasing operational efficiency. Overall, exploring proxy signals such as explainability and uncertainty to improve the reliability of BC for expert ratings is central to this thesis, alongside addressing potential biases inherited from human experts.

Tab. 2.1: Overview of prescriptive machine learning settings characterized by the type of observed data. The setting studied in this work is shown in bold.

Setting	**Observed Data**	**Goal**
Supervised Machine Learning	(X, Y)	Predict outcomes
Causal Machine Learning	(X, T, Y)	Estimate treatment effects
Offline Contextual Bandits	(X, A, R)	Maximize expected rewards
Behavioral Cloning	(X, A)	Mimic expert decisions

2.2 Supervised Machine Learning

This section provides an introduction to supervised machine learning and its two most common problem types: *classification* and *regression*. Furthermore, it presents the problem of *ordinal classification* (or ordinal regression), which lies somewhat between the two and is highly relevant for both ordinal and cardinal rating data.

2.2.1 Problem Formulation

In the context of supervised machine learning, it is assumed that there is some ground-truth mapping $g : \mathcal{X} \to \mathcal{Y}$ from a feature space $\mathcal{X} \subseteq \mathbb{R}^m$ to a target space $\mathcal{Y} \subseteq \mathbb{R}$. The goal of supervised learning is then to learn a hypothesis h from a hypothesis space $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}} = \{h : \mathcal{X} \to \mathcal{Y}\}$, also called *model* or *predictor*, that approximates the ground-truth mapping g as accurate as possible. The identification of the model h is hereby based on $N \in \mathbb{N}$ data samples of the form

$$\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}. \tag{2.1}$$

Commonly, $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$ is referred to as a *feature vector*, consisting of m *features*, and $y \in \mathcal{Y} \subseteq \mathbb{R}$ is referred to as a *label*.

Another standard assumption in supervised machine learning is that the samples $(\boldsymbol{x}, y) \in \mathcal{D}$ are *independent and identically distributed* (i.i.d.) and originate from an (unknown) *data-generating process* that defines a joint probability distribution $P(\boldsymbol{x}, y)$ on $\mathcal{X} \times \mathcal{Y}$. How well h approximates $P(\boldsymbol{x}, y)$ is typically measured using a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, which compares the ground-truth labels y with the predictions of the model $\hat{y} = h(\boldsymbol{x})$. The overall expected loss over $\mathcal{X} \times \mathcal{Y}$ is captured by the so-called true risk $R(h)$ of h, also referred to as the *population risk* [Mur22]:

$$R(h) := \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[l(h(\boldsymbol{x}), y)\right] = \int_{\mathcal{X}\times\mathcal{Y}} l(h(\boldsymbol{x}), y)\, dP(\boldsymbol{x}, y).$$

The true risk-minimizing hypothesis h^*, which achieves the minimum expected loss $\mathbb{E}[l(\cdot)]$ with respect to the loss function l across the entire joint probability distribution $P(\boldsymbol{x}, y)$, is called the *Bayes predictor*. It is defined as:

$$h^* := \arg\min_{h\in\mathcal{H}} R(h). \tag{2.2}$$

However, in practice, the learner does not observe $P(\boldsymbol{x}, y)$ and can only estimate the true risk $R(h)$ using the empirical risk $R_{\text{emp}}(h)$, which is computed based on the data $\mathcal{D}$:

$$R_{\text{emp}}(h) := \frac{1}{N}\sum_{i=1}^{N} l(h(\boldsymbol{x}_i), y_i). \tag{2.3}$$

The *empirical risk minimization* procedure yields the empirical risk-minimizing hypothesis $\hat{h}$:

$$\hat{h} := \arg\min_{h\in\mathcal{H}} R_{\text{emp}}(h).$$

To prevent overfitting and ensure the generalization of the trained model $\hat{h}$ to new, unseen data instances, the dataset $\mathcal{D}$ is typically split into training, $\mathcal{D}_{\text{train}}$, and test, $\mathcal{D}_{\text{test}}$, datasets [Mur22]. Whereas the training dataset, $\mathcal{D}_{\text{train}}$, is used to fit the model, the test dataset, $\mathcal{D}_{\text{test}}$, is subsequently employed to evaluate the model's generalization performance on previously unseen data.

$$\hat{h} := \arg\min_{h\in\mathcal{H}} R_{\text{emp}}(h) = \arg\min_{h\in\mathcal{H}} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{train}}} l(h(\boldsymbol{x}), y).$$

See Figure 2.2 for a graphical illustration of this process.

2.2.2 Nominal Classification

Nominal classification problems involve a target space $\mathcal{Y}$ consisting of $K \geq 2$ unordered, mutually exclusive labels, known as *classes*, defined as $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$. These classes unambiguously identify semantically coherent subsets of the input space $\mathcal{X}$. In contrast, when labels are not mutually exclusive and inputs $\boldsymbol{x} \in \mathcal{X}$ can be associated with multiple $y \in \mathcal{Y}$, this setting is referred to as *multi-label classification* [Her+16].

Classification problems can generally be categorized into two distinct types based on the number of classes involved:

1. Binary classification: This occurs when the number of classes is exactly 2 ($K = 2$). For example, emails can be classified as either `spam` or `non-spam` based on their content and metadata, with $\mathcal{Y} = \{\texttt{spam}, \texttt{non-spam}\}$.
2. Multi-class classification: This occurs when an instance may be classified into one of three or more categories ($K > 2$). For instance, images can be classified into categories such as animals or vehicles, with $\mathcal{Y} = \{\texttt{airplane}, \texttt{automobile}, \texttt{bird}, \texttt{cat}, \texttt{deer}, \texttt{dog}, \texttt{frog}, \texttt{horse}, \texttt{ship}, \texttt{truck}\}$ [KH+09].

A common loss function in classification is the *zero-one* (01) loss

$$l_{01}(y, \hat{y}) = [\![y \neq \hat{y}]\!],$$

where $\hat{y} = h(\boldsymbol{x})$ and $[\![\cdot]\!]$ denotes the indicator function returning 1 if the argument is true and 0 otherwise.

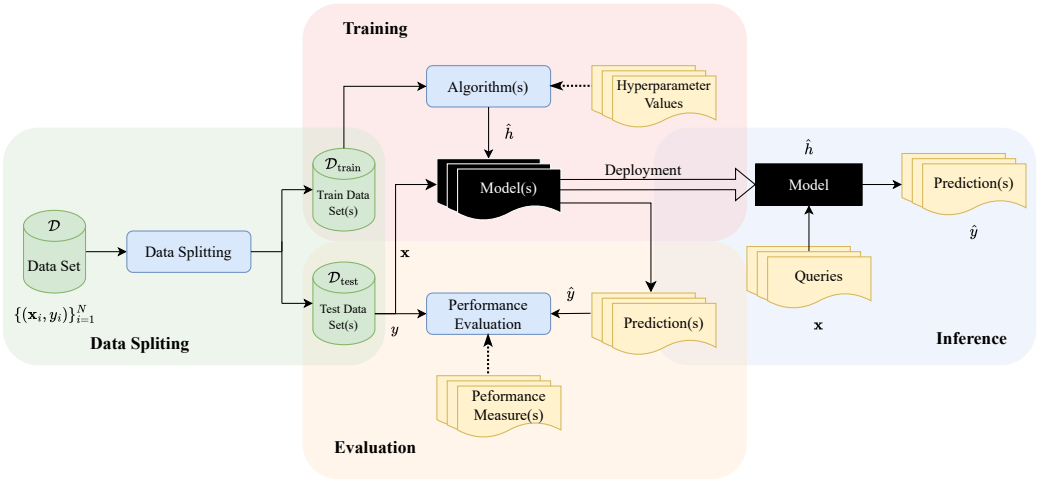


Fig. 2.2: Overview of the supervised machine learning process, from data splitting, training, and evaluation to deployment and inference.

2.2.2.1 Probabilistic Classification

In probabilistic classification, instead of committing fully to a single class or label $\hat{y}$, the model or predictor outputs a conditional probability distribution $p(y \mid \boldsymbol{x})$ over all possible classes $y \in \mathcal{Y}$ for an input instance $\boldsymbol{x} \in \mathcal{X}$, with $p(y_k \mid \boldsymbol{x}) \in [0, 1]$ denoting the conditional probability of the k-th class. In general, the training process of the predictor does not differ from the one outlined in Section 2.2.1. The only difference is that, instead of comparing a true class y with a deterministically predicted class $\hat{y}$, the loss function l compares a predicted probability distribution

$$\boldsymbol{p} = \hat{h}(\boldsymbol{x}_q) = (p(y_1), \ldots, p(y_K)) = (p_1, \ldots, p_K) \in \mathbb{P}(\mathcal{Y}), \tag{2.4}$$

with y, where p_k represents the predicted probability for the k-th class y_k. Hence, the signature of the loss function changes to $l : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \to \mathbb{R}$, where $\mathbb{P}(\mathcal{Y})$ denotes the set of *Probability Mass Functions* (PMFs) on $\mathcal{Y}$.

Obviously, a deterministic prediction $\hat{y}$ can also be obtained from a probabilistic prediction, usually by selecting the class with the highest posterior probability:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}).$$

This decision rule minimizes the expected l_{01}-loss and, under this loss function, is also known as the *Bayes estimator*.

A commonly used loss function for probabilistic multi-class classification is the *Cross-Entropy* (CE) loss:

$$l_{\text{CE}}(\boldsymbol{y}, \boldsymbol{p}) = -\sum_{k=1}^{K} y_k \log(p_k), \tag{2.5}$$

where $\boldsymbol{p}$ is the predicted probability distribution, and $\boldsymbol{y}$ is the one-hot (0/1) encoded true label vector, with $y_k = 1$ for the correct class and $y_k = 0$ for all other classes. For example, in the case of five classes and $y = 2$, the one-hot (0/1) encoded vector for y would be $\boldsymbol{y} = (0, 1, 0, 0, 0)$, where the second entry corresponds to the correct class.

Moreover, the CE loss is a strictly proper scoring rule, which incentivizes the true probability distribution $\boldsymbol{p}^*$ as the optimal prediction [GR07; MW70]. Formally, a loss is proper if it is minimized when the predicted probability distribution $\boldsymbol{p}$ matches the true probability distribution $\boldsymbol{p}^*$ in expectation:

$$\mathbb{E}_{y \sim \boldsymbol{p}^*}[l(y, \boldsymbol{p}^*)] \leq \mathbb{E}_{y \sim \boldsymbol{p}^*}[l(y, \boldsymbol{p})].$$

A loss is strictly proper if the equality holds only when the predicted distribution $\boldsymbol{p}$ exactly matches the true distribution $\boldsymbol{p}^*$.

Although, in theory, minimizing the empirical loss using a proper scoring rule should lead to accurate predictive probabilities, in practice, this may not always be the case. For instance, it has been shown that neural networks are poorly calibrated despite commonly utilizing the CE loss and tend to be overconfident [Guo+17; Sze+16]. This means that the predicted probabilities are often higher than the true probabilities. To address such biased predictive probability distributions, so-called *calibration* methods have been developed [Men+23]. These methods leverage a calibration dataset to learn mapping functions that adjust biased predictive probabilities to more accurate ones [NC05]. A predictor is considered *well-calibrated* when the predicted probabilities $\boldsymbol{p}$ align with the actual observed frequencies of the true labels. There are at least three ways to define calibration [Men+23]:

- *Confidence calibration* considers only the highest predicted probability [Guo+17].
- *Classwise calibration* considers marginal predicted probabilities for each class separately in a *one-vs-rest* fashion [ZE02].
- *Multi-class calibration* considers the entire vector of predicted probabilities [WLZ19].

Formally, a probabilistic multi-class classifier is considered multi-class calibrated if the predicted probability for each class p_k matches the true probability of the class $k \in \mathcal{Y} = \{1, \ldots, K\}$ given the predicted distribution:

$$\mathbb{P}(Y = k \mid h(X) = \boldsymbol{p}) = p_k.$$

Calibration is of fundamental importance when the predictor is used for *cost-sensitive learning* [Elk01; LS10; Fer+18a] and human decision-making. In such cases, an accurate quantification of the level of uncertainty is of utmost importance.

A common metric for assessing calibration is the *Expected Calibration Error* (ECE) [Guo+17]:

$$\text{ECE} = \sum_{j=1}^{B} \frac{|B_j|}{N} \left| o(B_j) - p(B_j) \right|,$$

where the predicted probability space is divided into B equal-sized bins (commonly 5, 10, or up to 15). For each bin B_j, the observed accuracy of the instances within the bin is calculated as:

$$o(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} [\![y_i = \hat{y}_i]\!].$$

This observed accuracy is then compared to the average predicted probability of the instances in bin B_j, which is given by:

$$p(B_j) = \frac{1}{|B_j|} \sum_{i \in B_j} p(y_i \mid \boldsymbol{x}_i).$$

The closer these two values $o(B_j)$ and $p(B_j)$ are, the lower the ECE score, indicating better calibration of the classifier.

2.2.2.2 Measuring Performance

Arguably, the most popular performance evaluation metrics for classification, given a vector of ground-truth labels $\boldsymbol{y} = (y_1, \ldots, y_N)$ from a given test data set $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ and corresponding point predictions denoted by a vector $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$, are *Accuracy* (ACC) and its inverse, the *Misclassification Rate* (MCR) (sometimes also referred to as *Mean Zero-One Error* (MZE)). Whereas accuracy is defined as the proportion of correct predictions

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^{N} [\![y_i = \hat{y}_i]\!]$$

and misclassification rate as the proportion of incorrect predictions

$$\text{MCR} = \frac{1}{N} \sum_{i=1}^{N} [\![y_i \neq \hat{y}_i]\!] = 1 - \text{ACC}.$$

The *Negative Log Likelihood* (NLL) is a performance metric commonly used to evaluate probabilistic classifiers. Unlike metrics that only assess the correctness of predictions, NLL also evaluates the confidence of the predicted probabilities.

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i), \tag{2.6}$$

where y_i is the true label for the i-th sample, and $p(y_i \mid \boldsymbol{x}_i)$ is the predicted probability assigned to the true label y_i given the input $\boldsymbol{x}_i$. NLL is equivalent to the CE or log loss (2.5) when used as a performance metric, as it measures how well the predicted probability distribution aligns with the true labels.

Another option for evaluating probabilistic predictions in nominal classification is the *Brier Score* (BS) [Bri50], which is also a proper scoring rule [GR07]. Unlike the NLL, the Brier Score considers all probabilities in the predictive distribution and

measures the mean squared difference between the true one-hot (0/1) encoded label vector $\boldsymbol{y}$ and the predicted probability distribution vector $\boldsymbol{p}$.

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(y_{i,k} - p_{i,k})^2, \tag{2.7}$$

where $y_{i,k}$ is the k-th element of the one-hot (0/1) encoded label vector for the i-th instance, and $p_{i,k}$ is the predicted probability for class k for the same instance.

2.2.3 Regression

The second prevalent problem in supervised learning is regression, where the target space $\mathcal{Y}$ is continuous, i.e., $\mathcal{Y} = \mathbb{R}$. Unlike classification, the goal in regression is to predict a real-valued quantity $y \in \mathbb{R}$. For example, one might predict house prices based on features such as property size, number of rooms, neighborhood characteristics, and other relevant attributes.

Arguably, the most common loss function for regression is the quadratic or l_2 loss:

$$l_2(y, \hat{y}) = (y - \hat{y})^2,$$

which penalizes the squared residuals, $(y - \hat{y})^2$. While the l_2 loss is widely used due to its mathematical properties (e.g., differentiability and convexity), it penalizes large residuals more heavily than small ones, making it sensitive to outliers. A more robust alternative is the l_1 loss, which measures the absolute difference between y and $\hat{y}$:

$$l_1(y, \hat{y}) = |y - \hat{y}|.$$

The l_1 loss is less sensitive to outliers because it grows linearly with the residuals rather than quadratically, making it a preferred choice in some applications.

2.2.3.1 Measuring Performance

Typically, the performance of a regression problem, given a vector of ground-truth labels $\boldsymbol{y} = (y_1, \ldots, y_N)$ from a test dataset $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ and corresponding predictions denoted by a vector $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$, is measured using the *Mean Absolute Error* (MAE) or the *Mean Squared Error* (MSE).

The MAE represents the average magnitude of the errors between y and $\hat{y}$:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

In contrast, the MSE squares the difference between y and $\hat{y}$ before summing them, thus giving more weight to larger errors:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

Taking the square root of the MSE yields the *Root Mean Squared Error* (RMSE), which is often considered easier to interpret than the MSE, as it is measured in the same units as the target variable:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}.$$

Another measure often used for evaluating regression models is the *Mean Absolute Percentage Error* (MAPE), which quantifies the performance of a regression model as a percentage. It is defined as the average of the absolute percentage errors between the actual values and the predicted values:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \cdot 100.$$

Since MAPE provides relative percentage values, it is scale-independent, making it suitable for comparisons across different datasets or models with varying scales. Furthermore, it is often more interpretable for stakeholders, as it provides a clear percentage that indicates how much the predictions deviate from the actual values on average.

2.2.4 Ordinal Classification

Ordinal classification (also called *ordinal regression* in statistics [McC80]) is a special form of supervised learning that lies between classification and regression, where the set of class labels exhibits a natural (linear) order:

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

As an example, consider credit scoring, where a customer's creditworthiness is rated on the categorical ordinal scale $\mathcal{Y} = \{\texttt{poor}, \texttt{fair}, \texttt{good}, \texttt{very good}, \texttt{excellent}\}$. Many additional examples of ordinal classification problems can, for instance, be found in medicine [Bar+21; Riv+23; Lin+22; Lei+22; Yon+22; Liu+18] or finance [Sol+22; Man+20; KA12]. This highlights the practical relevance of ordinal

classification as a subproblem of supervised learning, particularly in high-stakes settings such as medical diagnosis or financial risk assessment.

As it is common practice to encode ordinal labels by their rank in the ordinal scale, thereby turning the ordinal scale into a cardinal (interval) scale,

$$\begin{aligned}\mathcal{Y} &= \{y_1, y_2, y_3, \ldots, y_k, \ldots, y_K\} \\ &= \{1, 2, 3, \ldots, k, \ldots, K\},\end{aligned}$$

ordinal classification problems can thus be naively treated not only as classification problems but also as regression problems [Gut+16]. However, this ordinal encoding is theoretically disputable, as it assumes equal distances between class labels, a property that may not always hold. Nevertheless, given that this encoding is a common practice for ordinal data and aligns well with the cardinal nature of the goodwill assessment use case under consideration, as well as rating data in general, we will adopt it throughout this thesis.

One significant difference between ordinal classification and nominal classification is that misclassification costs are not uniform. For example, misclassifying an actual `poor` creditworthiness as `good` or even `excellent` may have a dramatic impact and should be penalized more heavily than misclassifying it as `fair`. In nominal classification, using losses such as the CE loss, these varying misclassification costs are not accounted for during model training, which can result in more severe and distant errors when applied to ordinal classification problems. In terms of regression, a significant difference is that the labels are not continuous real values ($y \in \mathbb{R}$) but discrete positive integers ($y \in \mathbb{N}$). This distinction necessitates rounding continuous-valued predictions to the nearest natural number in the set of integer-encoded class labels as a post-processing step, which can lead to information loss due to rounding errors [Kra+01]. Kramer et al. [Kra+01] further highlight the inherent trade-off in ordinal classification between achieving optimal categorical classification accuracy (hit rate) and minimizing distance-based error, which distinguishes it as a unique problem.

In recent decades, numerous dedicated ordinal classification (or regression) methods have been developed, aiming to leverage the ordering information to construct more accurate models. Gutiérrez et al. [Gut+16] provide a taxonomy of ordinal classification methods, grouping them based on their construction into *Naive Approaches*, *Ordinal Binary Decompositions*, and *Threshold Models* (Figure 2.3). We extend this taxonomy by incorporating more recently proposed *Ordinal Losses*, which have primarily been developed in the context of *Deep Learning* (DL) and *Unimodal Model* approaches. The latter can be further categorized into *Soft Labeling* and *Con-*

straints. Exploiting the ordinal structure has been shown to lead to improvements in model performance, as demonstrated by Hühn and Hüllermeier [HH08], as well as Gutiérrez et al. [Gut+16]. However, Gutiérrez et al. [Gut+16] also note that naive approaches can achieve competitive performance and are difficult to surpass for certain datasets. This observation is further supported, among others, by Ben-David et al. [BST09] and Kasa et al. [Kas+24], who acknowledge the competitiveness of naive approaches, particularly in terms of accuracy.

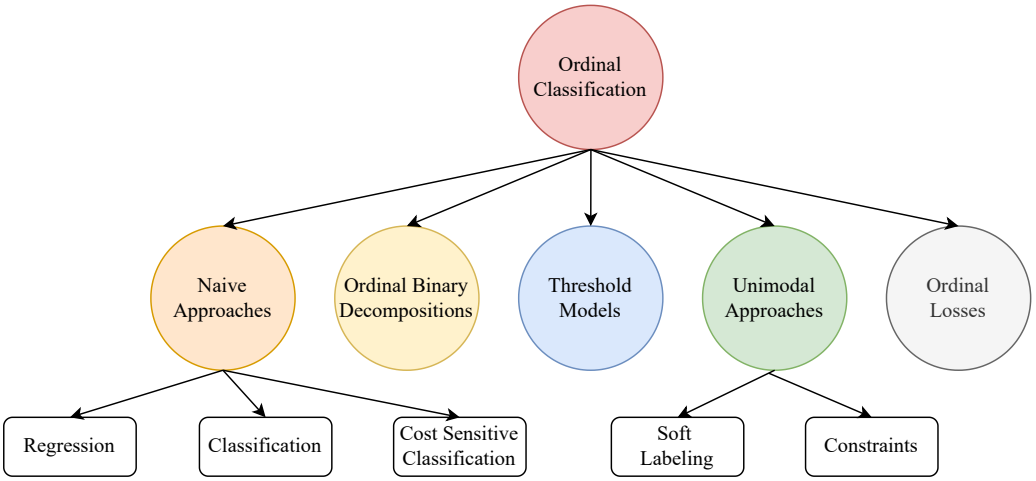


Fig. 2.3: Enhanced taxonomy of ordinal classification methods based on Gutiérrez et al. [Gut+16]

The following sections present a non-exhaustive exemplary overview of several ordinal classification methods, organized according to the taxonomy shown in Figure 2.3, with a focus on popular and more recent approaches.

2.2.4.1 Naive Approaches

As previously mentioned, ordinal classification problems can naively be treated as standard classification or regression problems. A more advanced method within the naive approaches is *cost-sensitive learning* [Elk01; Fer+18b], which assigns different costs to different misclassification errors [Gut+16]. Kotsiantis and Pintelas [KP04] propose a simple cost-sensitive post-processing method for ordinal classification, where the risk of predicting a certain class is evaluated based on the output of any probabilistic classification model. This approach has the advantage that it does not require any modification of the underlying learning algorithm. In general, their approach aligns with Bayesian decision theory, where decisions are made to minimize the expected loss or Bayes risk R given a posterior predictive probability distribution $p(y \mid \boldsymbol{x})$. However, instead of employing a loss function, Kotsiantis and

Pintelas propose the use of fixed $K \times K$ cost matrices C, which represent the costs of misclassification between the different ordinally encoded labels.

Concretely, they calculate the conditional risk $R(y_i \mid \boldsymbol{x})$ of selecting class $y_i \in \mathcal{Y}$, with $i \in \{1, \ldots, K\}$, using an ordinal cost matrix C and the given predictive posterior probabilities $p(y_j \mid \boldsymbol{x})$ with $y_j \in \mathcal{Y}$ and $j \in \{1, \ldots, K\}$. The entry $C_{i,j}$ represents the cost of predicting class y_i when the true class is y_j:

$$R(y_i \mid \boldsymbol{x}) = \sum_{j=1}^{K} C_{i,j} \cdot p(y_j \mid \boldsymbol{x}) \,.$$

The final prediction $\hat{y}$ is determined by selecting the class y_i that minimizes the conditional risk:

$$\hat{y} = \underset{y_i \in \mathcal{Y}}{\arg\min}\ R(y_i \mid \boldsymbol{x}).$$

Table 2.2 presents different exemplary ordinal cost matrices for a five-class ordinal classification problem. Common choices include absolute costs (left), $C_{i,j} = |i - j|$, and quadratic costs (middle), $C_{i,j} = (i-j)^2$, which correspond to the l_1 and l_2 losses, respectively, when applied to ordinally encoded labels. However, some problems may require an asymmetric cost matrix (right) in cases where misclassification costs are not symmetric. For example, in medical disease severity rating, failing to detect a higher disease severity may have more severe consequences than initially overestimating it.

Tab. 2.2: Different cost matrices for a five class ordinal classification problem $\mathcal{Y} = \{y_1, y_2, y_3, y_4, y_5\}$.

Absolute Costs	Quadratic Costs	Asymmetric Costs
$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 4 & 9 & 16 \\ 1 & 0 & 1 & 4 & 9 \\ 4 & 1 & 0 & 1 & 4 \\ 9 & 4 & 1 & 0 & 1 \\ 16 & 9 & 4 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 4 & 9 & 16 \\ 1 & 0 & 1 & 4 & 9 \\ 2 & 1 & 0 & 1 & 4 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$

In general, Kotsiantis and Pintelas note that on discretized regression datasets, their method effectively minimizes the distances between actual and predicted classes, as indicated by MAE, even with slight improvements in accuracy, which is highly valuable in an ordinal classification setting.

More formally, as previously mentioned, their method is well aligned with Bayesian decision theory, where a decision $\hat{y}$ is made to minimize the expected loss (Bayes

risk). The optimal policy that minimizes the risk, given a loss function, is also referred to as the Bayes estimator:

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

For absolute and quadratic costs, the Bayes risk corresponds to the l_1 and l_2 loss, respectively:

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

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

It is important to note that the above approach relies on accurate predictive probabilities and may require the use of appropriate scoring rules as a loss function, as well as proper calibration.

2.2.4.2 Ordinal Binary Decompositions

Ordinal binary decomposition methods are a popular and natural approach to dealing with ordinal classification problems. In their original taxonomy, Gutiérrez et al. further subdivide ordinal binary decomposition methods into *Multiple Model* and *Multiple-Output Single Model* approaches [Gut+16].

A simple and popular multiple model approach that exploits ordinal labels is to decompose the problem into $K-1$ sequential binary problems that essentially answer the following question for a given ordinal class k: "Is the label of $\boldsymbol{x}$ greater than k?" This splits $\mathcal{Y}$ into two meta-classes, $\{y_1, \ldots, y_k\}$ and $\{y_{k+1}, \ldots, y_K\}$, which serve as the negative and positive classes in a binary problem, respectively (Figure 2.4).

This approach results in several ordered binary problems that can be combined into a final probabilistic multiclass prediction, as described by Frank and Hall [FH01]:

$$\begin{aligned} p(y_1 \mid \boldsymbol{x}) &= 1 - p(y \succ y_1 \mid \boldsymbol{x}) \\ p(y_k \mid \boldsymbol{x}) &= \max\{p(y \succ y_{k-1} \mid \boldsymbol{x}) - p(y \succ y_k \mid \boldsymbol{x}),\, 0\} \\ p(y_K \mid \boldsymbol{x}) &= p(y \succ y_{K-1} \mid \boldsymbol{x}) \end{aligned} \tag{2.8}$$

Moreover, Li and Lin [LL06] introduced another binary reduction framework based on *extended examples*, which are derived from the original examples, and a specified

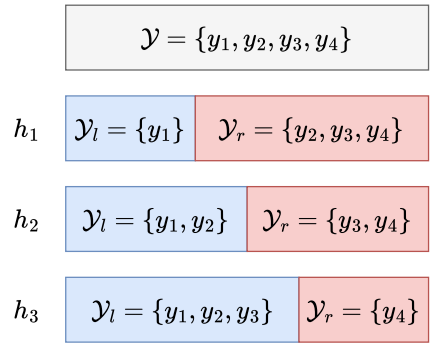


Fig. 2.4: Illustration of transforming an ordinal classification problem with $K = 4$ into $K - 1 = 3$ sequentially ordered binary subproblems, on which binary learners h_k (for $k = 1, 2, 3$) are trained.

mislabeling cost matrix $\mathcal{C}$ with V-shaped rows. In this matrix, the costs increase as the ordinal class k moves further away from the true label y. For instance, an absolute cost matrix is defined as $\mathcal{C}_{y,k} = |y - k|$. Unlike the approach by Frank and Hall [FH01], which solves each binary classification problem independently, this framework addresses all binary classification problems simultaneously using a single binary classifier. It achieves this by encoding the necessary ordinal information within the extended examples $\boldsymbol{x}^{(k)}$ and their corresponding binary labels $y^{(k)}$, while incorporating different misclassification cost weights $w_{y,k}$:

$$\boldsymbol{x}^{(k)} = (\boldsymbol{x}, k), \qquad y^{(k)} = 2[\![k < y]\!] - 1, \qquad w_{y,k} = |\mathcal{C}_{y,k} - \mathcal{C}_{y,k+1}|$$

The final aggregated result can then be obtained using a simple ranking rule r with a single predictor h, which sums the positive results across all thresholds to determine the ordinal class:

$$r(\boldsymbol{x}) = 1 + \sum_{k=1}^{K-1} [\![\, h(\boldsymbol{x}, k) > 0 \,]\!].$$

Additionally, the framework unifies many existing ordinal regression algorithms.

Cheng et al. [CWP08] and Niu et al. [Niu+16] proposed multiple-output single model adaptations of this method, utilizing multi-output neural networks. However, Cao et al. [CMR20] highlighted a key limitation of these binary decomposition neural network implementations: they do not ensure rank consistency, which refers to a monotonic relationship between the binary predictors and the ordinal outcome, where the probabilities of being in a higher category should decrease monotonically.

Another flexible binary decomposition method that can be applied to ordinal classification is *Nested Dichotomies* (NDs). A nested dichotomy is a binary tree that recursively partitions the label space $\mathcal{Y}$ into pairs of disjoint, nonempty subsets

$(\mathcal{Y}_l, \mathcal{Y}_r)$ (Figure 2.5). To transform a nested dichotomy into a multi-class classifier, a binary classifier $h_{\mathcal{Y}_l,\mathcal{Y}_r}$ is assigned to each inner node of the tree. This classifier is tasked with separating the set of classes $\mathcal{Y}_l$ associated with its left successor node from the set of classes $\mathcal{Y}_r$ associated with its right successor node, using a suitable *base learner* [MH18]. Usually, the same base learner is used for all binary classification problems of the nested dichotomy. Obviously, there are many ways to partition $\mathcal{Y}$ recursively, and the choice of partitioning can significantly impact the performance of the nested dichotomy. To address this, nested dichotomies were originally combined into an *Ensemble of Nested Dichotomies* (END), which has demonstrated excellent classification accuracy [FK04]. An END consists of a set of randomly generated nested dichotomies, and their predictions are combined by averaging the respective probability distributions. Initially, ENDs were applied to nominal classification problems, where no restrictions were imposed on how the label space was split into subsets. However, for ordinal classification, it is more appropriate to restrict the splits to ordered splits that respect and leverage the ordinal structure of the labels. See Figure 2.5 for an example of an ordinal dichotomy, where splits are performed in a way that respects the ordering of $\mathcal{Y}$. Restricting the set of potential dichotomies to ordered dichotomies significantly reduces the number of possible dichotomies from $(3^K - (2^{K+1} - 1))/2$ to $(K^3 - K)/6$ [HH08]. Depending on the number of classes K, this reduced number may still be computationally infeasible.

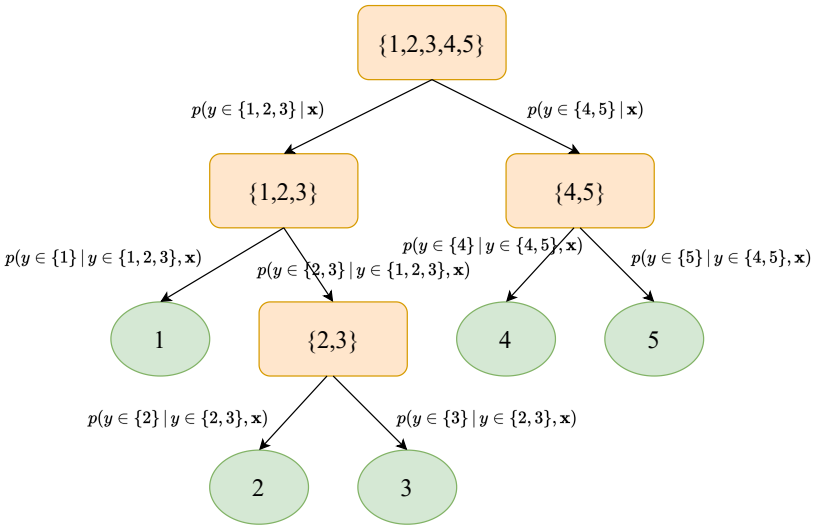


Fig. 2.5: Nested dichotomy in which a five class ordinal classification problem is decomposed into binary decision problems maintaining the ordinal structure.

Several heuristics for constructing nested dichotomies have been proposed for nominal classification tasks [MH18]. In the case of ordinal classification, nested dichotomies have so far been utilized only in specific contexts: either within an ensem-

ble of 20 randomly generated ordinal dichotomies restricted to ordinal splits [HH08], which, according to Frank and Kramer, is sufficient to achieve near-optimal performance [FK04], or in conjunction with imprecise probabilities, where splits were determined based on a risk-averse utility function [DY14; YDM17].

Once a hierarchy of binary classifiers has been constructed and trained, a new instance $\boldsymbol{x}$ can be classified probabilistically by multiplying the predicted probabilities of the inner binary classifiers, $h_{\mathcal{Y}_l,\mathcal{Y}_r}$, along the path from the root to the leaf nodes. The probability p_k of a specific class $y_k \in \mathcal{Y}$ is then computed using the chain rule of probability, given the chain of subsets from the root to the leaf, $\mathcal{Y} = \mathcal{Y}_0 \supset \mathcal{Y}_1 \supset \ldots \supset \mathcal{Y}_{K-1} = \{y_k\}$:

$$p(y_k \mid \boldsymbol{x}) = \prod_{i=1}^{K-1} p(y_k \in \mathcal{Y}_{i+1} \mid y_k \in \mathcal{Y}_i, \boldsymbol{x}),$$

where $p(\mathcal{Y}_{i+1} \mid \mathcal{Y}_i, \boldsymbol{x})$ is given by $h_{\mathcal{Y}_{i+1},\mathcal{Y}_i \setminus \mathcal{Y}_{i+1}}(\boldsymbol{x})$ if $\mathcal{Y}_{i+1}$ is the left successor of $\mathcal{Y}_i$ and $1 - h_{\mathcal{Y}_{i+1},\mathcal{Y}_i \setminus \mathcal{Y}_{i+1}}(\boldsymbol{x})$ if $\mathcal{Y}_{i+1}$ is the right successor.

2.2.4.3 Threshold Models

Originating from statistics, a common approach to addressing ordinal classification problems is to assume the existence of an underlying continuous variable, often referred to as a latent variable, along with a set of thresholds that partition the real line into contiguous intervals corresponding to the ordinal categories. Such models are classified as *Threshold Models* by Gutiérrez et al. [Gut+16].

Classic statistical models, such as the *Proportional Odds Model* (POM) [McC80], a specific type of *Cumulative Link Model* (CLM) that employs the logit link function, belong to this group of models. In general, CLMs, as the name suggests, estimate cumulative probabilities for an ordinal target variable y with K ordered categories up to category k, rather than estimating class probabilities for each individual category k:

$$p(y \preceq y_k \mid \boldsymbol{x}) = p(y_1 \mid \boldsymbol{x}) + \ldots + p(y_k \mid \boldsymbol{x}),$$

which can then be related to class probabilities as follows:

$$p(y_k \mid \boldsymbol{x}) = p(y \preceq y_k \mid \boldsymbol{x}) - p(y \preceq y_{k-1} \mid \boldsymbol{x}), \tag{2.9}$$

with $p(y_1 \mid \boldsymbol{x}) = p(y \preceq y_1 \mid \boldsymbol{x})$ and $p(y \preceq y_K \mid \boldsymbol{x}) = 1$ by definition.

Concretely, a CLM is typically formulated as a *Generalized Linear Model* (GLM):

$$g\Big(p(y \preceq y_k \mid \boldsymbol{x})\Big) = \log\left(\frac{p(y \preceq y_k \mid \boldsymbol{x})}{1 - p(y \preceq y_k \mid \boldsymbol{x})}\right) = \theta_k - \boldsymbol{w}^\top \boldsymbol{x},$$

where g is a link function, in the case of the POM, the logit function, which transforms the log-odds of cumulative probabilities (i.e., the probability of being in a category less than or equal to k versus being in a category higher than k) into a real value to match the linear predictor ($\theta_k - \boldsymbol{w}^\top \boldsymbol{x}$). Here, $\boldsymbol{w}$ is the coefficient vector representing the effect of the predictors, and θ represents the set of thresholds $\theta_1, \theta_2, \ldots, \theta_{K-1}$ with the property $\theta_1 < \theta_2 < \ldots < \theta_{K-1}$, which define contiguous intervals corresponding to the ordinal categories. The proportional odds assumption implies that the coefficients $\boldsymbol{w}$ are the same for all categories, meaning that the effect of the predictors on the odds ratio is constant across all thresholds. While this assumption simplifies the model, it may be violated in practice. To address this, various extensions of the POM have been proposed to relax the proportional odds assumption [PH90].

To obtain cumulative probabilities from the linear predictor, the inverse of the logit link function, also known as the logistic or sigmoid function, is applied:

$$p(y \preceq y_k \mid \boldsymbol{x}) = g^{-1}\Big(\theta_k - \boldsymbol{w}^T \boldsymbol{x}\Big) = \frac{1}{1 + e^{-(\theta_k - \boldsymbol{w}^T \boldsymbol{x})}}. \tag{2.10}$$

where g^{-1} denotes the inverse of the logit function. This function transforms the linear predictor $(\theta_k - \boldsymbol{w}^\top \boldsymbol{x})$ into a cumulative probability, ensuring it lies within the range $[0, 1]$.

More recently, CLMs were also integrated into deep learning architectures [VGH19; VGH20; Ros+22]. The only requirement for the *Deep Neural Network* (DNN) is that the last layer of the model should be a linear layer with only one unit $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$, which serves as the latent variable that the CLM takes as input (Figure 2.6). The thresholds $\theta_1, \theta_2, \ldots, \theta_{K-1}$ hereby become learnable parameters of the CLM layer and are constraint in the following way to fulfill the property $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{K-1}$:

$$\theta_k = \theta_1 + \sum_{k=1}^{K-1} \alpha_k^2,$$

where θ_1 and α_k are the learnable threshold parameters of the network. Through the inverse link function and the learned thresholds (2.10), the network can then output probabilities for each class according to (2.9). Alternatives to the logit link function are the *probit* and *clog-log* functions [VGH20]. Essentially, the CLM

layer replaces the standard softmax layer commonly used in DNNs to convert logits $\boldsymbol{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$ (unnormalized predictions) into probabilities:

$$p(y_k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \tag{2.11}$$

where $p(y_k \mid x)$ denotes the predicted probability of class y_k, and z_k is the logit (or unnormalized score) for class k. As a loss function, the authors either make use of the *Quadratic Weighted Kappa* (QWK) loss (2.14) [VGH19; VGH20] as proposed by de la Torre et al. [LPV18] (Sections 2.2.4.5 and 2.2.4.6), or a standard CE loss [Ros+22]. Interestingly, they note that the experimental findings in [Ros+22] suggest that a standard CE loss together with the CLM output layer is sufficient to model the ordinal structure of the label, without requiring the minimization of an ordinal loss (e.g., QWK).

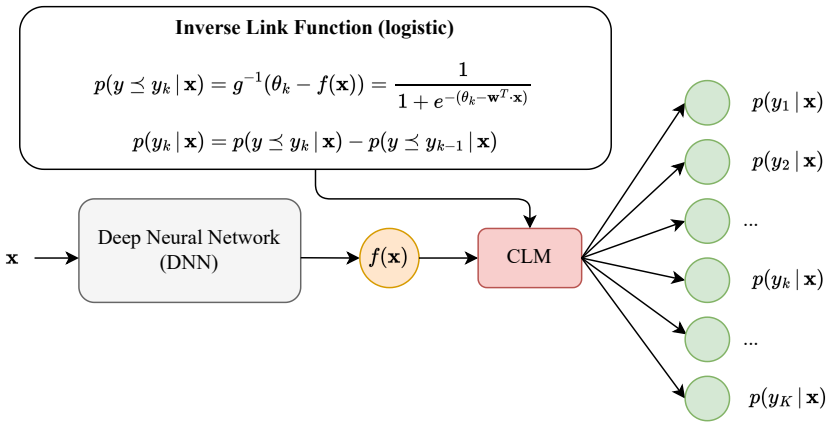


Fig. 2.6: Cumulative Link Model (CLM) integrated into a Deep Neural Network (DNN) as the output layer [VGH19].

2.2.4.4 Unimodal Approaches

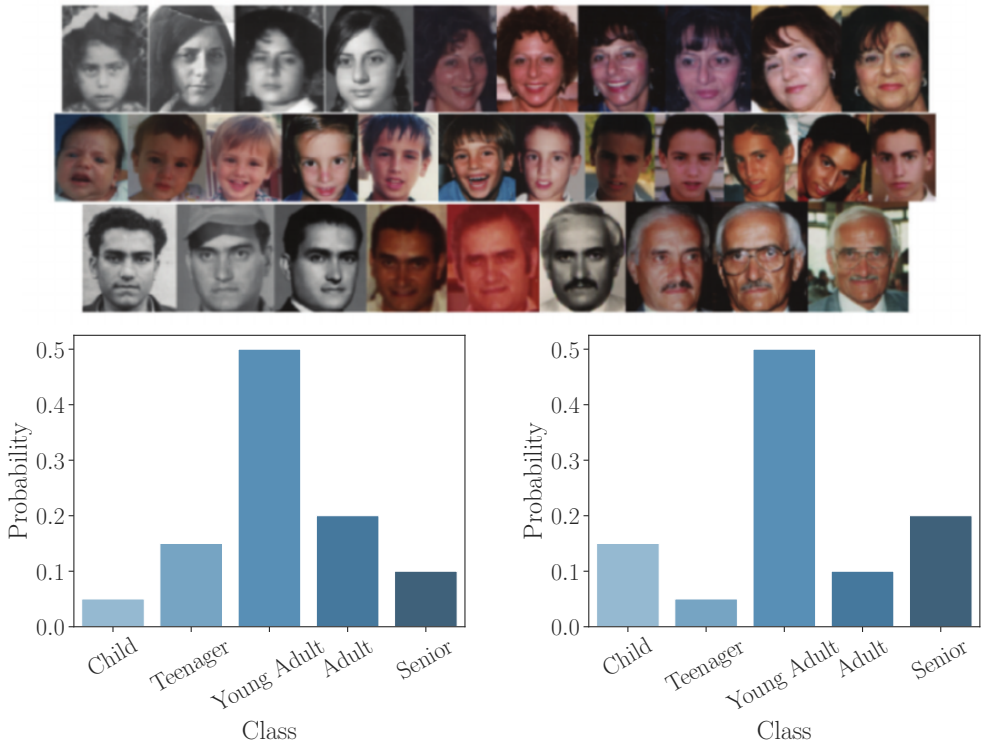
An inductive bias that has recently gained significant popularity in ordinal classification is the assumption that predictive probabilities should exhibit unimodality [Liu+19; Li+22; DMK23]. This means that the probabilities decrease monotonically on either side of the mode, which corresponds to the class with the highest probability in the distribution. This assumption reflects the ordinal nature of the problem, suggesting that adjacent classes are more likely than those further away. Figure 2.7 illustrates this concept with a graphical comparison of unimodal and multimodal probability distributions, using age estimation from images as an example [Lan08]. While a unimodal distribution has a single peak or mode, a multimodal distribution is characterized by multiple peaks or local maxima. More formally, a

probability distribution $\boldsymbol{p}$ is defined as unimodal if there exists at least one index m, the location of the mode, such that [KG71]:

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

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

The assumption of unimodality in ordinal classification appears reasonable, as classes closer to the mode are expected to be more likely than those further away, given the natural order of the classes. For instance, consider the distributions shown in Figure 2.7. The distribution in Figure 2.7a is far more plausible because it reflects the ordinal structure of the problem, where probabilities decrease as the distance from the mode increases. If the highest probability is assigned to the class `Young Adult`, other plausible classes would include `Teenager` or `Adult`, but not `Senior` or `Child`, as observed in the multimodal distribution in Figure 2.7b.



(a) Unimodal probability distribution. **(b)** Multimodal probability distribution.

Fig. 2.7: Exemplary unimodal and multimodal probability distributions for age estimation from images [Lan08] with $\mathcal{Y} = \{\texttt{Child}, \texttt{Teenager}, \texttt{Young Adult}, \texttt{Adult}, \texttt{Senior}\}$.

Due to the numerous unimodal approaches that have been proposed in recent years for ordinal classification, particularly in the deep learning domain, the taxonomy of Gutiérrez et al. needs to be extended (Figure 2.3) to incorporate this class of

models. Specifically, one can distinguish between unimodal *soft labeling* approaches, where targets are modeled not as deterministic one-hot (0/1) encoded vectors but as unimodal probability distributions, and unimodal *constraints*, where the predictive output probabilities of a model are explicitly constrained to exhibit unimodality. Unimodal constraints are typically implemented as specific loss functions that enforce the unimodal property during training. In contrast, unimodal soft labels serve as a regularization technique, inspired by *label smoothing* [Sze+16] and, more broadly, by concepts from *weakly supervised learning* [Zho18].

Constraints As previously mentioned, a naive approach to ordinal classification is to treat it as nominal classification and employ the CE loss. However, this approach may result in multimodal predictive probabilities [BP17; LPV18], as the CE loss focuses solely on maximizing the probability of the true class during training while disregarding the ordinal relationships between classes. As already discussed, non-unimodal predictive probability distributions are widely regarded as inappropriate for ordinal classification tasks, as they fail to reflect the ordinal structure of the classes [CC05; CAC08; BP17]. To enforce unimodality in the predictive output probabilities of neural networks, da Costa and Cardoso [CC05; CAC08], as well as Beckham and Pal, proposed utilizing the PMFs of the Poisson and binomial distributions [BP17]. Figure 2.8 illustrates the PMFs of the binomial and Poisson distributions, highlighting how these distributions inherently exhibit unimodality. The binomial distribution is defined as:

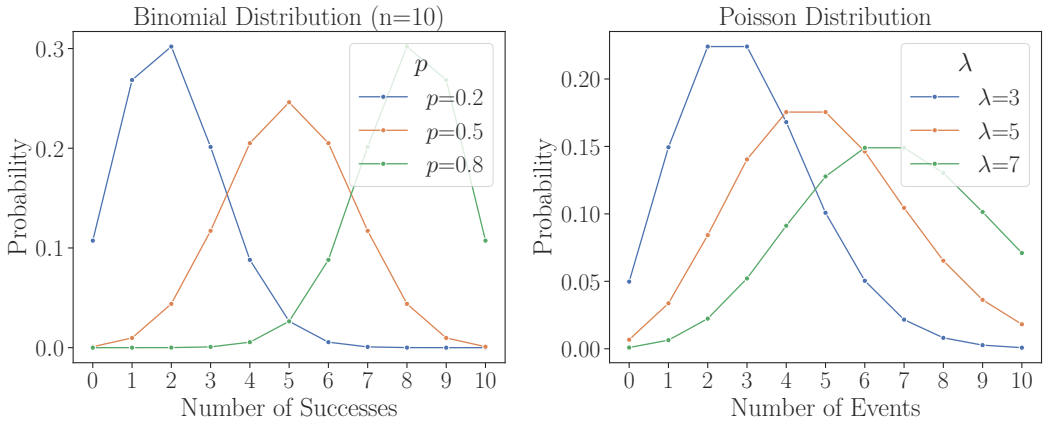


Fig. 2.8: Exemplary unimodal distributions produced by the binomial and Poisson distributions.

$$p(k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k},$$

where n is the number of trials, p is the probability of success, and k is the number of successes. This distribution is commonly used to model the number of successes in a fixed number of independent trials. The Poisson distribution is defined as:

$$p(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where λ is the average rate of occurrence of an event, and k is the number of occurrences of the event. The Poisson distribution is often used to model the number of events occurring in a fixed interval of time or space.

In the implementation by Beckham and Pal, the deep neural network outputs a single scalar value that serves either as the rate parameter $\lambda \in \mathbb{R}^+$ for the Poisson distribution or as the probability parameter $p \in [0, 1]$ for the binomial distribution (Figures 2.9 and 2.10). In the context of enforcing unimodal predictive probabilities, k hereby denotes the index of the class y_k to be predicted, and K represent the total number of classes. In both cases, Beckham and Pal apply the logarithm to the PMFs (omitted here) to address numerical instabilities and use a softmax (2.11) layer to normalize the probabilities. This approach also facilitates truncating the Poisson distribution, as it has infinite support, unlike the binomial distribution, which has finite support. Additionally, they introduce a temperature parameter τ to control the magnitudes of the probabilities output by the Poisson or binomial distribution when passed through the final softmax layer. This allows the distribution to become more uniform ($\tau \to \infty$) or more concentrated ($\tau \to 0$) around the class with the largest pre-softmax value. For the loss function, they employ the CE loss. Experimentally,

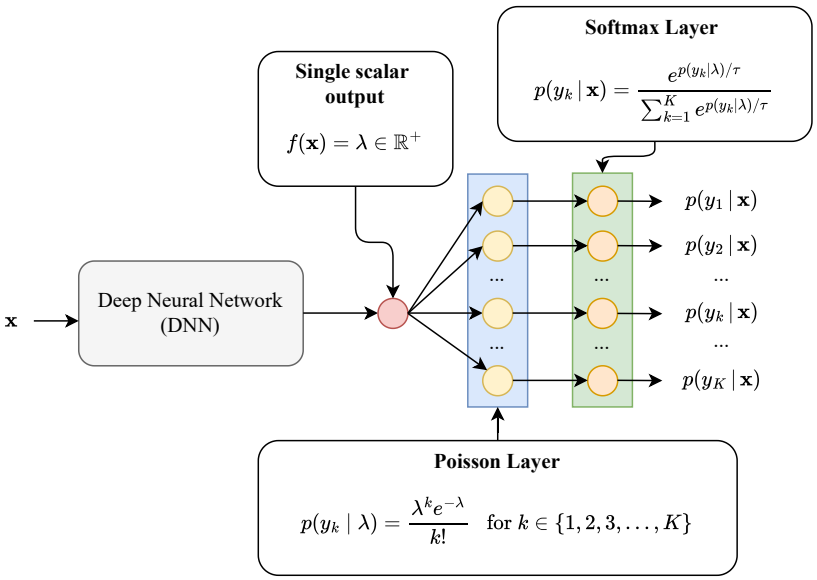


Fig. 2.9: Unimodal constrainment using the Poisson distribution [BP17].

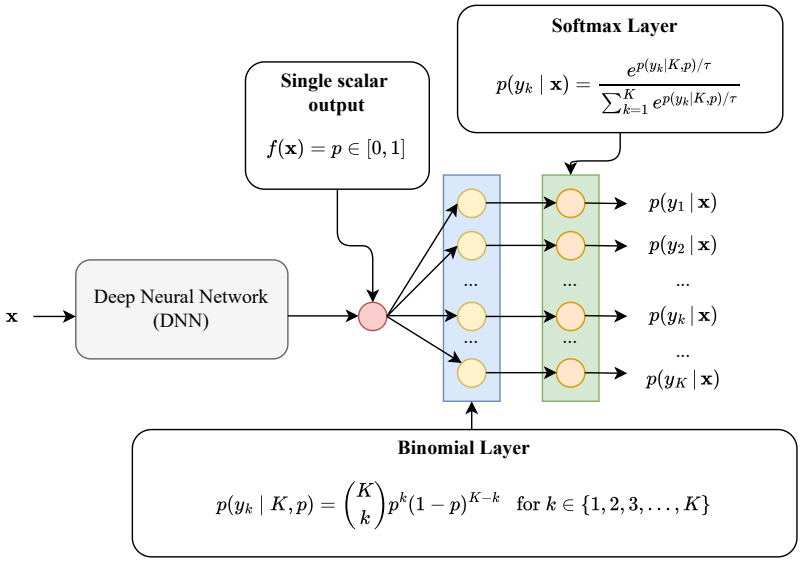


Fig. 2.10: Unimodal constrainment using the Binomial distribution [BP17].

they demonstrate that their approaches effectively produce unimodal predictive probability distributions and perform either superiorly or at least competitively compared to a cross-entropy baseline in terms of QWK (2.17) and top-k accuracy.

Instead of relying on parametric distributions, Albuquerque et al. propose various non-parametric approaches to enforce unimodality in predictive probability distributions [ACC21]. Specifically, they achieve this by introducing a regularization term to the CE loss, which penalizes deviations from unimodality. Additionally, they propose a less restrictive regularization to enforce what they term as "quasi-unimodal" distributions [ACC22]. In this approach, unimodality is incentivized only for the top three probabilities, while the remaining probabilities are constrained to not exceed the top three probabilities. Another non-parametric approach is proposed by Dey et al. [DMK23], who adapt the softmax function to enforce unimodality.

Soft Labeling An inverse approach compared to the previously described constraint-based methods is taken by unimodal soft labeling approaches. Instead of explicitly enforcing unimodality at the predictive probabilistic output of the deep neural network, it is implicitly encouraged through the label representation at the input. The true labels y of the training data are no longer encoded in a deterministic one-hot (0/1) format, e.g., $\boldsymbol{y} = (0, 0, 1, 0, 0)$ for a five-class problem with $y = 3$, but are instead represented as a *unimodal label-smoothed* (ULS) probability distribution $\boldsymbol{p}^{\text{ULS}}$, e.g., $\boldsymbol{p}^{\text{ULS}} = (0.1, 0.2, 0.4, 0.2, 0.1)$. This unimodal surrogate distribuption has two significant effects: First, it serves as a form of regularization by reducing the model's overconfidence in the true class, as the probability mass is not entirely

concentrated on a single class. Second, it effectively transforms the model into an ordinal classifier. The unimodal distribution implies that classes adjacent to the true class are more likely than those further away, thereby helping to minimize distant errors, which is advantageous in ordinal classification.

The corresponding technique widely used in nominal classification and deep learning to mitigate overconfidence in one-hot (0/1) encoded deterministic labels is known as *Label Smoothing* (LS) [Sze+16; MKH19]:

$$p_k^{\text{LS}} = (1 - \alpha)y_k + \alpha\frac{1}{K},$$

where y_k represents the value for the k-th class in the one-hot (0/1) encoded vector $\boldsymbol{y}$, with $y_k = 1$ for the correct class and $y_k = 0$ for all other classes. The parameter $\alpha \in [0, 1]$ is a smoothing factor hyperparameter that must be tuned. This factor is subtracted from the deterministic true class probability and uniformly redistributed across all classes. As a result, it does not convey any ordering information and treats all classes as equally likely.

In contrast, with *Unimodal Label Smoothing* (ULS), the probability of a specific class k is expressed as:

$$p_k^{\text{ULS}} = (1 - \alpha)y_k + \alpha \cdot p^U(k \mid y),$$

where p^U denotes a unimodal probability distribution, and $p^U(k \mid y)$ represents the probability of class k given the true class y. As previously mentioned, reallocating some of the probability mass from the true label to adjacent classes appears reasonable in ordinal classification, as they are the next logical candidates when the target space exhibits a natural order.

In contrast to the CE loss with one-hot (0/1) encoded labels:

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

the CE loss with the unimodal surrogate probability distribution $\boldsymbol{p}^{\text{ULS}}$ does not degenerate to the log loss by ignoring all classes other than the true one y:

$$l_{\text{CE}}(\boldsymbol{p}^{\text{ULS}}, \boldsymbol{p}) = -\sum_{k=1}^{K} p_k^{\text{ULS}} \log(p_k).$$

Instead, the unimodal surrogate probability distribution $\boldsymbol{p}^{\text{ULS}}$ arguably provides a more realistic representation by suggesting adjacent classes as plausible candidates,

simultaneously reducing overconfidence in potentially *noisy labels* [Son+23]. In general, soft labels address the problem of noisy labels by representing the labels in a less confident manner, which results in less overconfident models and helps prevent the learner from overfitting to noisy labels. In the case of ordinal classification, predictors trained using unimodal soft labels have demonstrated increased predictive performance and robustness to noisy labels [VGH22; Var+24]. For nominal multi-class classification using neural networks, it has been shown that employing soft targets significantly improves the generalization, learning speed, and calibration of neural networks [MKH19].

To represent targets in ordinal classification as unimodal probability distributions $\boldsymbol{p}^U$ rather than deterministic one-hot (0/1) encoded labels $\boldsymbol{y}$, several approaches have been proposed. One of the earliest approaches is by Díaz and Marathe [DM19], who adapt the softmax function (2.11) to transform the deterministic true label y into a unimodal soft label:

$$p^{\text{SORD}}(k \mid y) = \frac{e^{-\phi(k,y)}}{\sum_{j=1}^{K} e^{-\phi(j,y)}} = \frac{e^{-|k-y|}}{\sum_{j=1}^{K} e^{-|j-y|}}, \tag{2.12}$$

where y is the ordinally encoded true class and k is the index of the current class. In general, ϕ can be any distance measure, but the authors propose using the absolute distance, as MAE is a common evaluation metric in ordinal classification. Note that the negative exponent essentially inverts the standard softmax function, resulting in smaller distances producing higher values and larger distances yielding smaller values, which correspond to higher and lower probabilities, respectively. These soft labels are then used in conjunction with the CE loss to train the model. This approach is referred to as *Soft Ordinal Regression* (SORD) by the authors. The soft labels are unimodal, as they decrease monotonically on either side of the mode, which corresponds to the true class y.

Another approach is by Liu et al. [Liu+20a], who propose using the PMFs of the binomial and Poisson distributions (Figure 2.11). In the case of the Poisson distribution, the true class y is mapped one-to-one to the λ parameter of the Poisson distribution to model y as a probabilistic soft label (e.g., $y = 1 \rightarrow \lambda_y = 1, y = 2 \rightarrow \lambda_y = 2$, etc.). The probability for a certain class k, given λ_y, is then expressed as:

$$p^{\text{Poiss}}(k \mid \lambda_y) = \frac{\lambda_y^k e^{-\lambda_y}}{k!}.$$

To truncate the distribution, a softmax normalization is applied, as already proposed by Beckham and Pal [BP17].

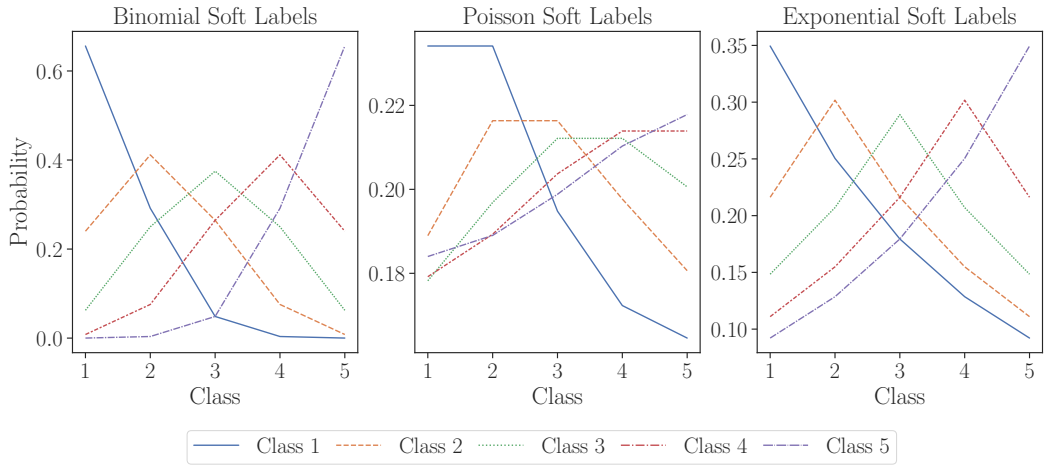


Fig. 2.11: Unimodal soft labeling approaches [Bér+25]: Soft labels based on the Binomial distribution [Liu+20a] (left), the Poisson distribution [Liu+20a] (center), and the Exponential function [Liu+20a; Var+23c] (right).

The parameter $p \in [0, 1]$ required for the Binomial distribution is determined by dividing the interval $[0, 1]$ into equal-sized segments based on the number of classes K. The true class-specific probability parameter p_y for the Binomial distribution, corresponding to the given true class y, is then calculated as follows:

$$p_y = 0.1 + \left(\frac{0.9 - 0.1}{K - 1} \right) \cdot y.$$

Since the Binomial distribution has finite support, no softmax normalization is required. The probability for a certain class k, given p_y, is then expressed as:

$$p^{\text{Bin}}(k \mid K, p_y) = \binom{K}{k} p_y^k (1 - p_y)^{K-k}.$$

However, Liu et al. report that both approaches are relatively static, making it challenging to adjust their shape. To address this, they propose an alternative method based on the exponential function (Figure 2.11):

$$p^{\text{Exp}}(k \mid y, \tau) = e^{-\frac{|k-y|}{\tau}},$$

where k is the current class, y again the ordinally encoded true label, and τ is a scaling factor that controls the influence of the absolute distance between k and y, thereby determining the strength of the smoothing. Smaller values of τ result in a more peaked smoothed probability distribution, while larger values of τ produce more uniform distributions. To obtain probabilities, a softmax normalization is required once again. Since the exponential function does not inherently produce

probabilities, it can be somewhat opaque with respect to tuning the hyperparameter τ and may require significant experimental effort. Nonetheless, according to Liu et al., this unimodal label regularization improves performance on several medical ordinal image classification tasks.

To address the inflexibility of the Poisson and binomial distributions, as well as the lack of interpretability of the exponential function, Vargas et al. propose a unimodal regularization method based on the beta distribution [VGH22] (Figure 2.12). They argue that the beta distribution is particularly well-suited for unimodal regularization because it is defined over the range $[0, 1]$, eliminating the need for softmax normalization. Additionally, it avoids high variance by allowing the true class to be modeled as a relatively sharp peak. Unlike the previously described soft label methods, the beta distribution is a continuous distribution and must be discretized to align with the discrete nature of ordinal classification tasks. Its *Probability Density Function* (PDF) is given by:

$$f(x, p, q) = \frac{1}{B(p, q)} x^{p-1}(1 - x)^{q-1},$$

where $0 < x < 1$, $p > 0$, $q > 0$, and $\frac{1}{B(p,q)}$ is a normalizing factor that ensures the total probability is 1:

$$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p + q)},$$

with Γ denoting the *Gamma function*.

To discretize the continuous beta distribution for use with discrete ordinal labels, the interval $[0, 1]$ is divided into equal-sized subintervals based on the number of classes K. For example, for $K = 5$, the intervals are defined as follows: Class 1 corresponds to $[0, 0.2]$, Class 2 to $[0.2, 0.4]$, Class 3 to $[0.4, 0.6]$, Class 4 to $[0.6, 0.8]$, and Class 5 to $[0.8, 1.0]$. The probability associated with a specific class k is then computed as the integral:

$$p^{\text{Beta}}(k \mid p_y, q_y) = \int_{(k-1)/K}^{k/K} f(x, p_y, q_y), dx,$$

where the parameters p_y and q_y, which control the shape of the beta distribution, are determined by the true class y. Vargas et al. propose determining the parameters p_y and q_y based on the expected value of the beta distribution, which is defined as

$$\mathbb{E}(x) = \frac{p}{p + q}.$$

The assumption is that for a certain ordinally encoded true class y, the expected value should lie at the center of the respective interval. For instance, $\mathbb{E}(x) = \frac{1}{2K}$ for $y = 1$, $\mathbb{E}(x) = \frac{3}{2K}$ for $y = 2$, etc. In general, the mean of the interval for

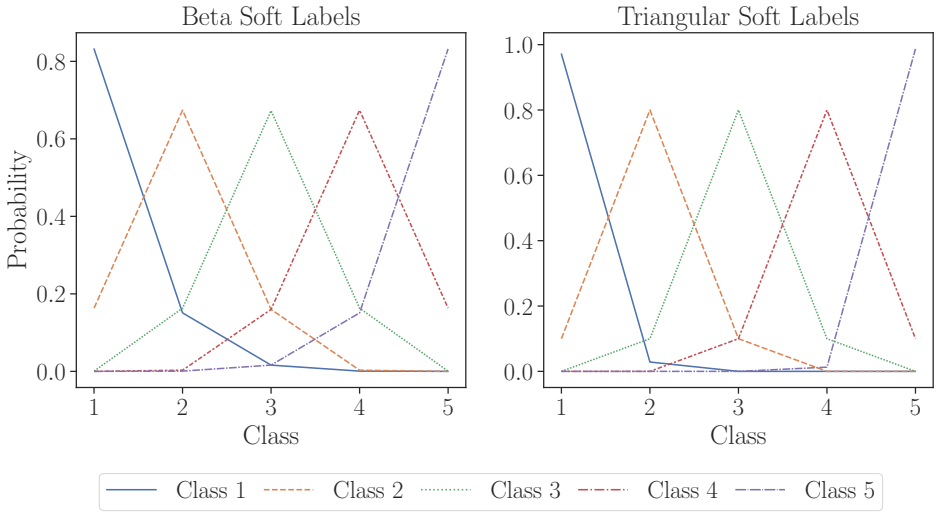


Fig. 2.12: Discretized continuous unimodal soft labeling approaches [Bér+25]: Soft labels based on the Beta distribution [VGH22] (left) and the Triangular distribution (with $\alpha = 0.2$) [Var+23b; Var+23a] (right).

a certain true class y is determined as: $\mathbb{E}(x) = \frac{2y-1}{2K}$. Based on this formula, the parameters p_y and q_y can be analytically derived. Experimentally, Vargas et al. report improvements in performance on several ordinal image classification tasks through the usage of beta distribution-based soft labels. There appears to be an increase in the robustness of the model in the presence of noisy targets by shifting probability mass to neighboring classes.

One issue Vargas et al. notice with the beta distribution-based soft labeling is that, while the mean can be centered in the middle of the interval using the analytical method, the variance tends to decrease as the number of classes increases. This can cause the soft labels to become crisp labels in problems with many classes, as the probability mass allocated to the adjacent classes becomes too small. To deal with this issue, Vargas et al. propose using triangular distributions [Var+23b] (Figure 2.12). The triangular distribution is also continuous like the beta distribution and requires three parameters to control its shape: a, b, and c, with a as the lower limit, b as the upper limit, and c as the mode, where $a < b$ and $a \leq c \leq b$. Its PDF is given by

$$f(x, a, b, c) = \begin{cases} 0 & \text{for } x < a, \\ \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \leq x < c, \\ \frac{2}{b-a} & \text{for } x = c, \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \leq b, \\ 0 & \text{for } b < x \end{cases} \tag{2.13}$$

To discretize the triangular distribution, the interval $[0, 1]$ is again split into K evenly spaced sub-intervals, just like for the beta distribution. The probability associated with a certain class k given an ordinally encoded true class y is then given by:

$$p^{\text{Tri}}(k \mid a_y, b_y, c_y) = \int_{(k-1)/K}^{k/K} f(x, a_y, b_y, c_y)\, dx$$

where the parameters a_y, b_y and c_y, which control the shape of the Triangular distribution, are determined by the true class y.

In general, three cases need to be distinguished: For the first class, for the last class, and for the intermediate classes. In the case of the first class, with $y = 1$, $a_1 = c_1 = 0$ and only the upper limit b_1 needs to be estimated. In the case of the last class, with $y = K$, $b_K = c_K = 1$ and only the lower limit a_K needs to be estimated. In case of a middle class, the mode is placed in the center of the respective interval with $c_k = \frac{2y-1}{2K}$ and both a_k and b_k need to be estimated. Vargas et al. also introduce a parameter α that controls the reallocation of probability mass to adjacent classes for middle classes ($1 < k < K$) in a symmetric manner (refer to [Var+23b] for the detailed derivations of the parameters a_y, b_y and c_y).

In [Var+23a], this approach is extended to allow for asymmetric class-wise reallocation of probability mass to adjacent classes using two parameters, α_{2k-1} and α_{2k}. Here, α_{2k-1} specifies the probability mass allocated to the left adjacent class, while α_{2k} specifies the probability mass allocated to the right adjacent class. For the extreme classes, only one parameter is used: α_2 for the first class, which defines the probability mass allocated to the right, and α_{2K-1} for the last class, which defines the probability mass allocated to the left. The class-wise asymmetric smoothing, achieved through generalized Triangular distributions [Var+23a], significantly enhances the flexibility of soft labeling compared to earlier methods [Var+23b; VGH22]. This approach allows for more precise control over the probability mass reallocation to adjacent classes, addressing limitations of prior symmetric methods. To estimate the various α parameters, Vargas et al. employ particle swarm optimization, which enables them to achieve superior performance compared to previous soft labeling techniques across eight ordinal datasets [Var+23a].

Table 2.3 lists and compares the unimodal soft label approaches discussed earlier. The comparison criteria include flexibility in controlling the shape of the unimodal soft label distribution (shape control), support for class-wise smoothing (class-wise), and the ability to redistribute probability mass asymmetrically (asymmetric).

Tab. 2.3: Comparison of different unimodal soft label approaches.

Method	Shape control	Class-wise	Asymmetric
SORD [DM19]	✓	✗	✗
Poisson [Liu+20a]	✗	✗	✗
Binomial [Liu+20a]	✗	✗	✗
Exponential [Liu+20a]	✓	✗	✗
Beta [VGH22]	✗	✗	✗
Triangular [Var+23b]	✓	✗	✗
Generalised Triangular [Var+23a]	✓	✓	✓

2.2.4.5 Ordinal Losses

In past years, several specific loss functions for deep learning-based ordinal classification have been proposed, which take the ordinal structure into account at the loss level, particularly focusing on reducing error distances between classes. For instance, De la Torre et al. propose directly optimizing the *Quadratic Weighted Kappa* (QWK) metric [LPV18], which is considered an important metric in the evaluation of ordinal classification predictors (Section 2.2.4.6). QWK measures the agreement between two raters (e.g., predictions and ground truth) from -1 to 1, where 1 indicates full agreement. It accounts for the degree of disagreement, by penalizing differences in ordinal ratings quadratically. To reformulate QWK as a loss function suitable for minimization, consistent with the standard approach in training predictive models, De la Torre et al. suggest taking the logarithm of the complement of QWK as follows:

$$l_{\text{QWK}} = \log\left(1 - QWK\right), \tag{2.14}$$

with $l_{\text{QWK}} \in (-\infty, \log(2)]$ for $QWK \in [-1, 1]$. Furthermore, they adapt QWK to work with probability distributions $\boldsymbol{p}$ over classes, rather than deterministic class labels and a confusion matrix (2.17). The QWK loss can then be computed using the following equation, with the numerator representing observed frequencies and the denominator representing expected frequencies:

$$\text{QWK} = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{i,j} \sum_{k=1}^{N} p_{k,i}\, y_{k,j}}{\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{K} w_{i,j} (\sum_{k=1}^{N} p_{k,i})(\sum_{k=1}^{N} y_{k,j})}, \tag{2.15}$$

where w is a $K \times K$ quadratic weight matrix that quadratically penalizes discrepancies between predicted and true class distributions (2.18), $\boldsymbol{y}$ represents the one-hot (0/1) encoded true labels, and $\boldsymbol{p}$ denotes the predicted probability vector over the ordinal classes. This formulation makes the QWK continuous and, therefore, differentiable, which enables the training of deep neural networks on datasets with the

explicit objective of optimizing the QWK metric [Bér+25]. However, it is important to note that QWK cannot be applied to a single instance and requires a larger batch size for calculation, as it relies on a complete confusion matrix.

Hou et al. propose using the squared *Earth Mover's Distance* (EMD) as an ordinal loss [HYS16]:

$$l_{\text{EMD}}(\boldsymbol{y}, \boldsymbol{p}) = \sum_{k=1}^{K-1} \Big(F_k(\boldsymbol{y}) - F_k(\boldsymbol{p}) \Big)^2, \tag{2.16}$$

where probabilistic predictions $\boldsymbol{p}$ are evaluated against one-hot (0/1) encoded deterministic observations or decisions $\boldsymbol{y}$. Here, $F_k(\boldsymbol{p}) = \sum_{j=1}^{k} p_j$ denotes the cumulative sum of predicted probabilities over the set $\mathcal{Y}$ up to the k-th class, and $F_k(\boldsymbol{y}) = \sum_{j=1}^{k} y_j$ denotes the cumulative sum of the one-hot (0/1) encoded true outcomes over $\mathcal{Y}$ up to the k-th class. Notably, this loss is equivalent to the *Ranked Probability Score* (RPS) [Eps69; Mur70], which is a proper scoring rule [GR07] for ordinal outcomes. It encourages truthful probability reporting in the ordinal case by incentivizing unimodal predictive probability distributions. In this context, the probability smoothly decreases as the distance from the true class increases, with the loss penalizing probability mass farther from the true class more strongly.

Another recently proposed ordinal loss loss is the *Class Distance Weighted Cross-Entropy* (CDW-CE) loss proposed by Polat et al. [Pol+22; PÇT25], which is a variant of the CE loss that penalizes errors based on the distance between the predicted class and the true class. The loss is defined as:

$$l_{\text{CDW-CE}}(y, \boldsymbol{p}) = -\sum_{k=1}^{K} \log(1 - p_k) \cdot |k - y|^{\alpha},$$

where y is the ordinally encoded ground-truth class, and the power term α is a hyperparameter that determines the strength of the coefficient. A similar loss called *Ordinal Log Loss* (OLL) was proposed by Castagnos et al. [CMD22], with the only difference being the use of configurable $K \times K$ distance matrices D instead of fixed absolute distances:

$$l_{\text{OLL}}(y, \boldsymbol{p}) = -\sum_{k=1}^{K} \log(1 - p_k) \cdot d(k, y)^{\alpha},$$

Here, $d(k, y)$ is the distance between the predicted class k and the ordinally encoded true class y, and it can be read from the $K \times K$ matrix as $d(k, y) = D_{k,y}$.

Kasa et al. [Kas+24] notice that ordinal losses excel primarily in ordinal distance-based metrics while compromising performance on nominal metrics. That's why

they introduce a loss which combines CE loss with the OLL loss [CMD22], which they call *Multi-task Log Loss* (MLL):

$$l_{\text{MLL}} = \lambda \cdot l_{\text{CE}} + (1 - \lambda) \cdot l_{\text{OLL}}$$

Here, $\lambda \in [0, 1]$ is a hyperparameter that balances the trade-off between the two losses, reflecting the inherent compromise in ordinal classification between achieving a higher exact hit-rate and minimizing error distances. Moreover, the authors argue that this loss can enhance the performance of ordinal classifiers across both ordinal and nominal metrics by leveraging the strengths of both loss functions.

In summary, ordinal losses are characterized by penalizing larger or more distant errors more heavily than those in adjacent or nearby classes. Consequently, they encourage the formation of more compressed unimodal predictive probability distributions, as demonstrated by De la Torre et al. [LPV18] and Liu et al. [Liu+20a]. However, similar to unimodal soft label approaches, they do not explicitly enforce this behavior, in contrast to unimodal constraint methods.

2.2.4.6 Measuring Performance

Since ordinal classification overlaps with nominal classification and regression, metrics such as ACC (or its complement, MCR/MZE), MAE, and (R)MSE are commonly used to evaluate the performance of ordinal predictors [GJ09; BES09; JTB16].

Furthermore, the *Quadratic Weighted Kappa* (QWK) is a very popular measure for evaluating ordinal classifiers [Coh60; Coh68; YD23]. It measures the level of agreement between two raters (or between ground-truth ratings and a predictor) who classify items into ordered categories. It is particularly useful in ordinal classification tasks because it takes into account the ordering of the categories and penalizes larger discrepancies more heavily. QWK values range from -1 to 1, where

- $QWK = 1$ indicates perfect agreement, meaning the observed ratings of both raters perfectly align.
- $QWK = 0$ indicates no agreement beyond what would be expected by chance, reflecting that the level of agreement is what would be expected randomly.
- $QWK < 0$ indicates worse than chance agreement, meaning the raters are in disagreement more often than would be expected by random chance.

Values between 0.81 and 1.0 are considered very good, whereas values between 0.61 and 0.8 are considered good (Table 2.4) [LPV18].

Formally, QWK is defined as follows:

$$\text{QWK} = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{i,j}O_{i,j}}{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{i,j}E_{i,j}}, \tag{2.17}$$

where $w_{i,j}$ is the weight assigned to the disagreement between the categories $i \in \{1,\dots,K\}$ and $j \in \{1,\dots,K\}$, usually quadratic in the context of ordinal classification:

$$w_{i,j} = \frac{(i-j)^2}{(K-1)^2}, \tag{2.18}$$

$O_{i,j}$ is the observed frequency of true category i being predicted as j, derived from the confusion matrix O, and $E_{i,j}$ is the expected frequency of the true category i being predicted as j derived from the marginal probabilities and under the assumption of independence:

$$E_{i,j} = \frac{\sum_{k=1}^{K} O_{i,k}\sum_{k=1}^{K} O_{k,j}}{N},$$

with N as the total number of observations.

Tab. 2.4: Interpretation of QWK values [LK77; LPV18].

QWK	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

Another common metric to evaluate ordinal classifiers is the k-off metric, where a prediction "close" to the true label is considered partially correct. Normally, it is used in its 1-off instantiation [HYS16; VGH22; CMD22; Bér+25].

$$\text{1-off Accuracy} = \frac{1}{N}\sum_{i=1}^{N} [\![\, |y_i - \hat{y}_i| \leq 1 \,]\!].$$

Additional metrics for ordinal classification sets specifically targeted at imbalanced dataset problems are the *Average MAE* (AMAE) [BES09] and the *Maximum MAE* (MMAE) [Cru+14]. Both metrics evaluate the performance of an ordinal classifier separately for each specific true class j, accounting for each class equally. The class-specific MAE for class j can be written as:

$$\text{MAE}_j = \frac{1}{N_j}\sum_{k=1}^{K} |y_j - \hat{y}_k| N_{j,k},$$

where $|y_j - \hat{y}_k|$ is the absolute difference between the ordinally encoded true class y_j and the ordinally encoded predicted class $\hat{y}_k$, N_j is the number of instances in the true class j, and $N_{j,k}$ is the count of instances with true class y_j and predicted class $\hat{y}_k$. The AMAE, which evaluates the performance of an ordinal classifier across all classes, is defined as:

$$\text{AMAE} = \frac{1}{K} \sum_{j=1}^{K} \text{MAE}_j$$

The MMAE then just reports the max MAE among the class wise MAEs, highlightng the worst performing class, which is particularly useful in imbalanced datasets where some classes may have significantly fewer samples:

$$\text{MMAE} = \max\{\text{MAE}_j : j = 1, \ldots, K\}$$

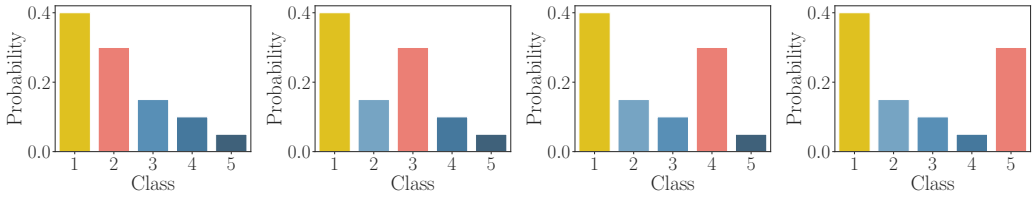
The *Ranked Probability Score* (RPS), originally introduced by Epstein [Eps69], is a popular measure for the evaluation of probabilistic forecasts of ordinal variables. For instance, it is used in weather forecasts where the possible outcomes might be `no rain`, `light rain`, `moderate rain`, and `heavy rain` [Mur70]. The RPS assumes unimodal distributions for ordinal outcomes and penalizes unimodally concentrated distributions less than multimodal distributions (Figure 2.13). Hence, unlike NLL (2.6) and BS (2.7), it is not invariant to probability mass re-distribution. From a mathematical formulation point of view, it is the Brier score on cumulative probability vectors. Naturally, it can also be applied to evaluate probabilistic predictions of an ordinal classifier [JTB16; Gal23]. It is defined as follows:

$$\text{RPS} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{K-1} \sum_{k=1}^{K-1} \Big(F_k(\boldsymbol{p}_i) - F_k(\boldsymbol{y}_i) \Big)^2 \right), \tag{2.19}$$

where $F_k(\boldsymbol{p}_i)$ is the cumulative probability of the predictive outcome k for the i-th test instance, $F_k(\boldsymbol{y}_i)$ is the cumulative probability of the true outcome k for the i-th one-hot (0/1) encoded target $\boldsymbol{y}_i$, and the term $\frac{1}{K-1}$ normalizes the RPS to lie between 0 (perfect score) and 1 (worst score). Essentially, RPS is the corresponding metric to the EMD loss (2.16).

Notably, RPS is also considered a proper scoring rule because it is only minimized when the predicted probability distribution matches the true distribution. It is particularly useful for ordered outcomes, as it assigns better scores to probabilistic forecasts that place high probabilities on the correct class and its adjacent classes. As mentioned previously, this aligns with the expected proper behavior for ordinal forecasting and classification. Nonetheless, QWK is generally more popular than RPS for evaluating ordinal classifiers. Galdran [Gal23] traces this back to the fact

that RPS increases penalties linearly for increasing error distances, unlike QWK, which penalizes errors quadratically with increasing distance from the true class or label. Another reason is that the focus is commonly on predicting deterministic labels rather than evaluating probabilistic predictions. However, with the increasing importance of uncertainty awareness and proper calibrated predictive probability distributions in predictors, RPS appears to be gaining more popularity in machine learning, particularly in ordinal classification [JTB16; Gal23; Bér+25].



(a) RPS $= 0.475$, BS $= 0.485$, NLL $= 0.9163$.

(b) RPS $= 0.5875$, BS $= 0.485$, NLL $= 0.9163$.

(c) RPS $= 0.6875$, BS $= 0.485$, NLL $= 0.9163$.

(d) RPS $= 0.775$, BS $= 0.485$, NLL $= 0.9163$.

Fig. 2.13: Illustration of the RPS in comparison to NLL and BS for four different probability distributions $\boldsymbol{p}_1 = (0.4, 0.3, 0.1, 0.15, 0.05)$ (2.13a), $\boldsymbol{p}_2 = (0.4, 0.1, 0.3, 0.15, 0.05)$ (2.13b), $\boldsymbol{p}_3 = (0.4, 0.1, 0.15, 0.3, 0.05)$ (2.13c), and $\boldsymbol{p}_4 = (0.4, 0.1, 0.15, 0.05, 0.3)$ (2.13d) given that the true label is $y = 1$. As shown, successively redistributing probability mass from class 2 to higher classes only affects the RPS, while BS and NLL remain unchanged.

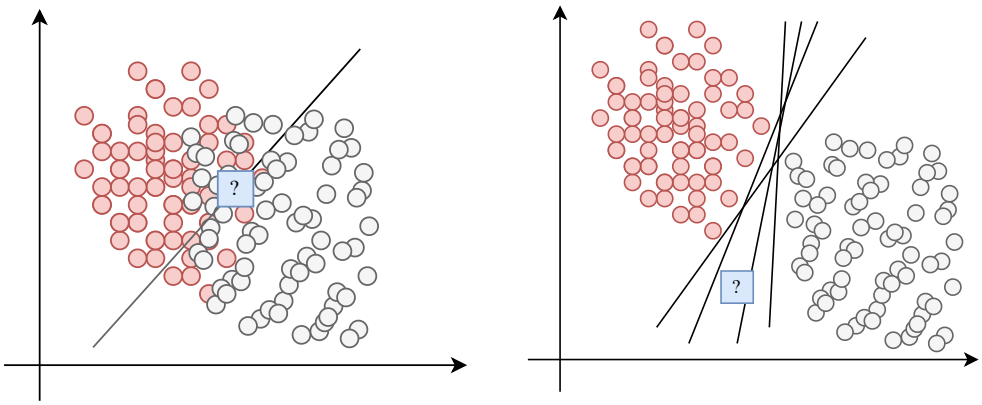
2.3 Uncertainty in Machine Learning

With machine learning models increasingly being deployed in high-stakes domains like finance or healthcare, the notion of uncertainty is likewise becoming more and more important. Usually, one is interested in the uncertainty related to a prediction $\hat{y}$ for a particular query $\boldsymbol{x}_q$, which is commonly referred to as *predictive uncertainty* [HW21]. Nowadays, it is also considered crucial to distinguish between two types of uncertainties in machine learning, namely *Aleatoric Uncertainty* (AU) and *Epistemic Uncertainty* (EU) [Sen+14]. See Figure 2.14 for an illustration. EU hereby refers to a lack of knowledge and is reducible on the basis of additional information or data. Since the learned predictor $\hat{h} \in \mathcal{H}$ is only an estimate of the Bayes predictor h^* (2.2) based on empirical risk minimization (2.3) and strongly depends on the amount and quality of the training data $\mathcal{D}$ (2.1), there naturally remains uncertainty about how well $\hat{h}$ approximates h^*. This is also referred to as *approximation uncertainty*. Another uncertainty, which is also of epistemic nature, is *model uncertainty*, which arises due to the choice of the hypothesis space $\mathcal{H}$. Concretely, which model to be fit to the data, e.g., neural networks or tree-based

models, or the particular neural network architecture. In contrast, AU is irreducible and refers to inherent stochasticity of the data-generating process. So even if the learner had perfect knowledge about $P(\boldsymbol{x}, y)$, it cannot determine a single outcome $y \in \mathcal{Y}$, but only the conditional probability

$$p(y \mid \boldsymbol{x}_q) = \frac{p(\boldsymbol{x}_q, y)}{p(\boldsymbol{x}_q)}.$$

When the training data is of observational nature stemming from expert decisions, this could be due to different opinions of experts for particular cases. For instance, think of doctors who are not in agreement with regards to their disease severity rating for a patient, or financial analysts, who are not in accordance with regards to an investment and its associated risk.



(a) Under the assumption of an optimal linear hypothesis, the prediction for the query point (indicated by a question mark) remains aleatorically uncertain, as the two classes overlap in this region.

(b) The query point (indicated by a question mark) lies outside the training data region, resulting in epistemic uncertainty due to a lack of knowledge about the correct hypothesis.

Fig. 2.14: An illustration of aleatoric and epistemic uncertainty in the context of a binary classification problem.

Information about predictive uncertainty is highly valuable in downstream tasks such as *selective classification* [GE17]. In this context, queries with high predictive uncertainty are not processed by the model; instead, the model abstains [Hen+24] and, for example, forwards them to human experts for manual inspection. This approach enhances the model's predictive accuracy and overall reliability. Another important application is *Out-Of-Distribution* (OOD) detection, which aims to determine whether a query originates from the same distribution as the training data or from a different, unknown distribution [HG17]. The underlying assumption is that epistemic uncertainty is particularly elevated when a sample is drawn from an unknown distribution. Detecting OOD samples is critical in high-stakes applications,

such as medical diagnosis and security systems. Additionally, uncertainty sampling-based *active learning* presents another potential downstream task. In this scenario, instances with high epistemic uncertainty are deemed more informative than those with high aleatoric uncertainty. Therefore, these instances should be prioritized for labeling and added to the training data first, facilitating a rapid improvement in model performance [NSH22].

The following two sections give an overview of popular uncertainty representation and quantification methods in machine learning, separated into *first-order* methods only capable of quantifying aleatoric uncertainty and *second-order* methods that are capable of quantifying both aleatoric and epistemic uncertainty.

2.3.1 First-Order Uncertainty Representation

Common approaches to representing first-order predictive uncertainty in machine learning include probabilities and sets, which will be briefly introduced in the following two sections. Our focus here will be on the classification setting.

2.3.1.1 Representation based on probability distributions

The common practice in machine learning is to represent uncertainty as a probability distribution over classes, in which a single predictor $\hat{h} \in \mathcal{H}$ outputs a probability distribution given a query instance $\boldsymbol{x}_q \in \mathcal{X}$ as input.

$$\boldsymbol{p} = \hat{h}(\boldsymbol{x}_q) = (p(y_1), \ldots, p(y_K)) = (p_1, \ldots, p_K) \in \mathbb{P}(\mathcal{Y}), \tag{2.20}$$

where p_k is the predicted probability for the k-th class y_k. This prediction is considered an estimate of the true conditional probability $p(y \mid \boldsymbol{x}_q)$.

A common uncertainty measure for nominal classification based on such first-order probability distributions is the (Shannon) entropy [Sha48],

$$\mathbb{H}(\boldsymbol{p}) = -\sum_{k=1}^{K} p_k \log p_k, \tag{2.21}$$

which essentially quantifies the non-uniformity or "peakedness" of a probability distribution, with $0 \log 0 = 0$ by definition.

Other uncertainty measures include *confidence*,

$$C(\boldsymbol{p}) = 1 - \max_{k \in K} p_k, \tag{2.22}$$

which measures the confidence of the predictor with regards to the class with the highest probability, or *margin*,

$$M(\boldsymbol{p}) = p_m - p_n, \tag{2.23}$$

where $m = \underset{k \in K}{\arg\max}\ p_k$ and $n = \underset{k \in K \setminus m}{\arg\max}\ p_k$, which measures the difference between the two highest probabilities in the distribution to quantify predictive uncertainty.

However, all of these measures on first-order probabilities have in common that they do not express any uncertainty related to the hypothesis $h \in \mathcal{H}$ and hence are only capable of expressing aleatoric uncertainty.

2.3.1.2 Representation based on sets

Another way to express predictive uncertainty is to make set-valued predictions instead of point predictions covering the true outcome with high probability. The key idea hereby is to construct a set $\mathcal{C}(\boldsymbol{x})$ that contains the true outcome y with a pre-specified (high) probability $1 - \alpha$, where $\alpha \in [0, 1]$ is a user-specified desired level of significance or risk. Formally, this can be expressed as:

$$P\Big(y \in \mathcal{C}(\boldsymbol{x})\Big) \geq 1 - \alpha. \tag{2.24}$$

A popular framework to construct such prediction sets which comes with strong *marginal coverage* guarantees and only assumes exchangeability of the data is *Conformal Prediction* (CP) [VGS05; BHV14; AB23]. Note that exchangeability, where the order of the data points is irrelevant, represents a weaker assumption than i.i.d. (Section 2.2.1). Essentially, conformal prediction guarantees that the true outcome y will be contained in the prediction set $\mathcal{C}(\boldsymbol{x})$ with probability at least $1 - \alpha$, regardless of the underlying data distribution and quality of the predictive model. The other way round, the probability of an invalid prediction in which $y \notin \mathcal{C}(\boldsymbol{x})$ is (asymptotically) bounded by $\alpha > 0$. A stronger notion of coverage than marginal coverage (2.24) is *conditional coverage* (2.25) [Vov13], which guarantees that the prediction set $\mathcal{C}(\boldsymbol{x})$ contains the true outcome y with high probability $(1 - \alpha)$ for each query instance $\boldsymbol{x}$, thus providing per-instance guarantees instead of average guarantees:

$$P\Big(y \in \mathcal{C}(\boldsymbol{x}) \;\Big|\; \boldsymbol{x}\Big) \geq 1 - \alpha. \tag{2.25}$$

However, conformal procedures are not guaranteed to satisfy (2.25).

Arguably the most popular instantiation of confromal prediction is *Inductive Conformal Prediction* (ICP) (or *split conformal prediction*) [Pap+02; Pap08], in which a

calibration data set is set aside on top of training and test set, to estimate so-called nonconformity scores. A widely used and straightforward nonconformity score in classification is the softmax-based score. This score is defined using the probability assigned to the true label from the calibration dataset by a probabilistic classifier, denoted as $p(y \mid \boldsymbol{x})$:

$$s(\boldsymbol{x}, y) = 1 - p(y \mid \boldsymbol{x}).$$

In this formulation, $s(\boldsymbol{x}, y) \in \mathbb{R}$ quantifies the degree of nonconformity of the instance $\boldsymbol{x}$ with respect to the true label y. A higher score indicates greater nonconformity, suggesting that the model is less confident in its prediction for that particular instance. After the nonconformity scores are calculated for all examples in the calibration set, the nonconformity scores are sorted in ascending order. The $(1-\alpha)$-quantile, computed as $\lceil (n+1)(1-\alpha) \rceil / n$, is selected as the critical value and is denoted by $\hat{q}$. Prediction sets can then be constructed using the nonconformity scores across all labels for a new query $\boldsymbol{x}$ and the critical value threshold $\hat{q}$:

$$\mathcal{C}(\boldsymbol{x}) = \left\{ y \in \mathcal{Y} \;\middle|\; s(\boldsymbol{x}, y) \leq \hat{q} \right\}$$

The above explained CP method is called the *Least Ambiguous Set-Valued Classifier* (LAC) [MW19] and has been shown to produce prediction sets with the smallest average size. See Figure 2.15 for an illustration of the LAC method. An improvement over LAC, which considers cumulative probabilities instead, is the *Adaptive Prediction Set* (APS) [RSC20] method, which leads to more adaptive prediction sets depending on the difficulty of the query.

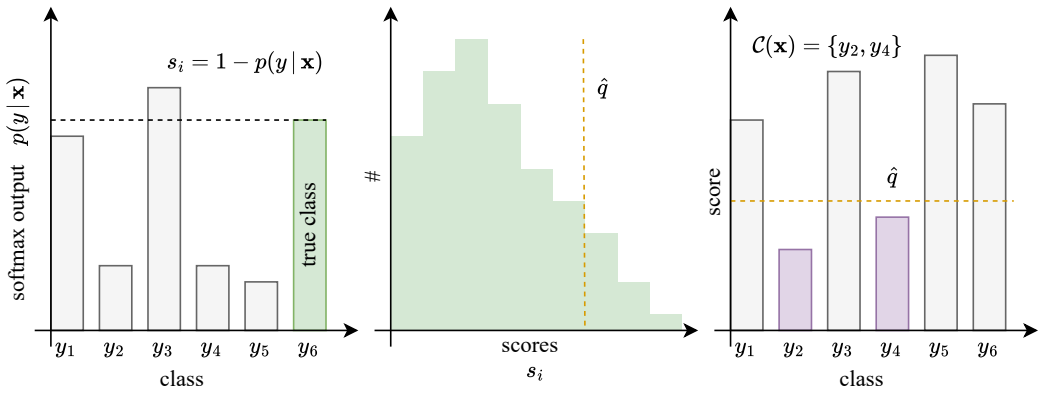


Fig. 2.15: Illustration of the *Least Ambiguous Set-Valued Classifier* (LAC) conformal prediction method, as proposed by Sadinle and Wasserman [MW19].

There have also been several works adapting conformal prediction for ordinal classification, where the common theme is to enforce contiguous prediction sets, e.g., $\{y_1, y_2, y_3\}$ in contrast to $\{y_1, y_2, y_4\}$, either by enforcing unimodal probabilities

in the underlying model [DMK23] or by adjusting the score function to grow a potential prediction set only from the mode outwards [LAP22]. Moreover, Xu et al. [XGW23] propose a conformal risk control method [Ang+22a] for ordinal classification, which aims to control the expected risk of the prediction set instead of marginal coverage. They define risk either as a weight or as a divergence-based loss, depending on the distance between the true label and the prediction set. This, in turn, also leads to more centripetal and hence more contiguous prediction sets.

The *validity* of a conformal prediction approach is usually assessed through the *Empirical Coverage* (EC),

$$\text{EC} = \frac{1}{N} \sum_{i=1}^{N} [\![\, y_i \in \mathcal{C}(\boldsymbol{x}_i) \,]\!],$$

which ensures that the marginal coverage condition (2.24), holds in practice. Additionally, the *Mean Prediction Set Size* (MPS) is evaluated as

$$\text{MPS} = \frac{1}{N} \sum_{i=1}^{N} |\, \mathcal{C}(\boldsymbol{x}_i) \,|,$$

where smaller prediction sets are considered more *efficient* [AB23; KTL23]. In the case of ordinal classification, Dey et al. [DMK23] introduce the fraction of non-contiguous prediction sets as an additional performance metric, where smaller values are preferable in ordinal classification.

Though the coverage guarantee of conformal prediction always holds, the usefulness of the prediction sets is, in addition to the underlying predictor itself, primarily determined by the score function. Consequently, the choice of the score function is an important engineering decision to be made [AB23]. Just like first-order probabilities obtained through a single probabilistic model, conformal prediction only quantifies aleatoric uncertainty by capturing the inherent randomness in the data generating process. It does not quantify any uncertainty about the model itself.

2.3.2 Second-Order Uncertainty Representation

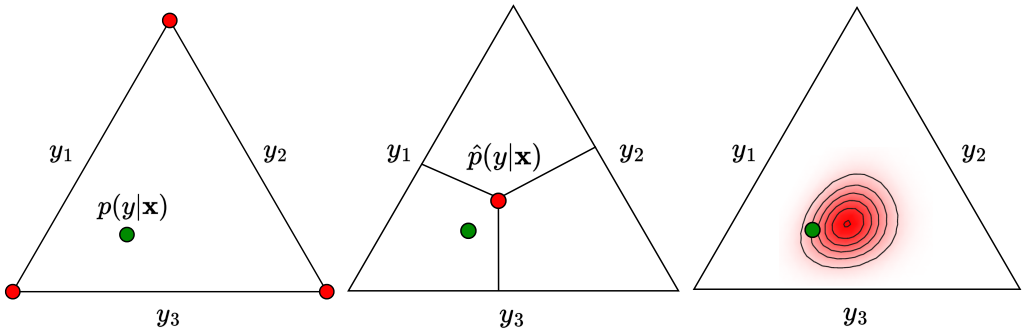
As already stated, it is crucial in machine learning to distinguish between two types of uncertainty: aleatoric and epistemic uncertainty. A principled way to represent epistemic uncertainty, which is related to the model and its approximation quality, is Bayesian inference [HW21]. Instead of committing to a single hypothesis $h \in \mathcal{H}$, as in the first-order setting, a hypothesis space $\mathcal{H}$ consisting of probabilistic predictors is

considered and a prior $p(h)$ is placed over each candidate $h \in \mathcal{H}$. Learning essentially consists of updating the prior distribution $p(h)$ to the posterior distribution $p(h \mid \mathcal{D})$ in light of the training data $\mathcal{D}$:

$$p(h \mid \mathcal{D}) = \frac{p(h) \cdot p(\mathcal{D} \mid h)}{p(\mathcal{D})} \propto p(h) \cdot p(\mathcal{D} \mid h),$$

where $p(\mathcal{D} \mid h)$ is the likelihood of the hypothesis h, (i.e., the probability of the data $\mathcal{D}$ given h) and $p(\mathcal{D})$ is the marginal probability of the data, which serves as a normalization factor. Intuitively, $p(h \mid \mathcal{D})$ captures the state of knowledge of the learner and hence its epistemic uncertainty. The more concentrated (or "peaked") the probability mass in a small region of $\mathcal{H}$ is, the less uncertain the learner is. Since every $h \in \mathcal{H}$ produces a probabilistic prediction, the belief about the outcome y is represented by a second-order probability: a probability distribution of probability distributions [SH21]. See Figure 2.16 for an illustration of different degrees of uncertainty awareness. The predictive posterior distribution specifies the posterior probability of each outcome $y \in \mathcal{Y}$. It is defined in terms of the *expected* probability $p(y \mid \boldsymbol{x}, h)$, where the expectation is taken with respect to the posterior distribution $p(h \mid \mathcal{D})$:

$$p(y \mid \boldsymbol{x}) = \mathbb{E}_{p(h \mid \mathcal{D})}[\, p(y \mid \boldsymbol{x}, h)\,] = \int_{\mathcal{H}} p(y \mid \boldsymbol{x}, h)\, d\, p(h \mid \mathcal{D})$$



(a) Deterministic predictor $h : \mathcal{X} \to \mathcal{Y}$.

(b) Probabilistic predictor $h : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$.

(c) Second-order predictor $h : \mathcal{X} \to \mathbb{P}(\mathbb{P}(\mathcal{Y}))$.

Fig. 2.16: Uncertainty awareness illustrated on the probability simplex for $\mathcal{Y} = \{y_1, y_2, y_3\}$. Left: A deterministic predictor is only capable of predicting deterministic labels. Middle: A probabilistic predictor is able to quantify aleatoric uncertainty as a probability distribution on $\mathcal{Y}$. Right: A second-order predictor provides a probability distribution over probability distributions and is thereby able to also quantify epistemic uncertainty.

An instantiation of Bayesian inference in machine learning are Bayesian neural networks [Nea12], where real-valued weights $\boldsymbol{w}$ in the network are replaced by probability distributions (typically Gaussians). Learning then comes down to Bayesian inference, i.e., computing the posterior over the weights given the data $p(\boldsymbol{w} \mid \mathcal{D})$.

As extact posterior inference is difficult to compute [Abd+21], several approximations have been proposed. For instance, Gal and Ghahramani [GG16] propose *Dropout* [Sri+14], which was originally intended to prevent overfitting of deep neural networks, to approximate Bayesian inference. Similarly, Mobiny et al. show that *DropConnect* [Wan+13] can also be seen as a way to approximate the posterior distribution over the network weights [Mob+19]. Both Dropout and DropConnect can capture epistemic uncertainty by performing multiple stochastic forward passes with different random masks during inference. The variability of these predictions then constitutes the epistemic uncertainty. Aleatoric uncertainty, on the other hand, can be directly captured from the different model's outputs. Another simple and flexible approach based on deep ensembles, which also allows for the separation of epistemic and aleatoric uncertainty by analyzing the variations in the ensemble member predictions, was proposed by Lakshminarayanan et al. [LPB17].

Explicit attempts to model and separate EU and AU in classification and regression are presented by Depeweg et al. [Dep+18]. Specifically, they propose to measure TU in terms of the entropy of the predictive posterior distribution, which in the case of discrete $\mathcal{Y}$ is given by

$$\mathbb{H}[p(y \mid \boldsymbol{x})] = \mathbb{E}_{p(y|\boldsymbol{x})}[-\log_2 p(y \mid \boldsymbol{x})] = -\sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \log_2 p(y \mid \boldsymbol{x}). \tag{2.26}$$

In turn, AU is given by the expectation over the entropies of the different probabilistic predictions:

$$\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\mathbb{H}[p(y \mid \boldsymbol{x}, \boldsymbol{w})] = -\int p(\boldsymbol{w} \mid \mathcal{D}) \left(\sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{w}, \boldsymbol{x}) \log_2 p(y \mid \boldsymbol{w}, \boldsymbol{x}) \right) d\boldsymbol{w} \tag{2.27}$$

Eventually, EU is then the *mutual information* between the weights $\boldsymbol{w}$ and the prediction y:

$$\underbrace{\mathbb{I}[y, \boldsymbol{w} | \boldsymbol{x}, \mathcal{D}]}_{\mathrm{EU}(\boldsymbol{x})} = \underbrace{\mathbb{H}[p(y \mid \boldsymbol{x})]}_{\mathrm{TU}(\boldsymbol{x})} - \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\mathbb{H}[p(y \mid \boldsymbol{x}, \boldsymbol{w})]}_{\mathrm{AU}(\boldsymbol{x})} \tag{2.28}$$

This is expressed as the difference between TU and AU.

In case of regression, uncertainty estimates can be derived via the *law of total variance* instead:

$$\underbrace{\mathbb{V}_{p(y|\boldsymbol{x},\mathcal{D})}[y \mid \boldsymbol{x}]}_{\mathrm{TU}(\boldsymbol{x})} = \underbrace{\mathbb{V}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathbb{E}_{p(y|\boldsymbol{x},\boldsymbol{w})}[y \mid \boldsymbol{x}]\right]}_{\mathrm{EU}(\boldsymbol{x})} + \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathbb{V}_{p(y|\boldsymbol{x},\boldsymbol{w})}[y \mid \boldsymbol{x}]\right]}_{\mathrm{AU}(\boldsymbol{x})} \tag{2.29}$$

Moreover, the entropy- and variance-based approaches do not necessarily need to be implemented using neural networks with different weight vectors $\boldsymbol{w}$. The measures (2.28) and (2.29) can also be approximated using a finite ensemble of arbitrary M models, $H = \{h_1, \ldots, h_M\}$, for example, through ensembles of deep neural networks [LPB17], trees within a *Random Forest* [SH20], or ensembles of *Gradient Boosted Trees* (GBT) [MPU21].

An approximation of (2.27) can be obtained by

$$\text{AU}(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{H}\big[\, p(y \,|\, \boldsymbol{x}, h_m) \,\big], \tag{2.30}$$

an approximation of (2.26) by

$$\text{TU}(\boldsymbol{x}) \approx \mathbb{H}\left[\frac{1}{M} \sum_{m=1}^{M} p(y \,|\, \boldsymbol{x}, h_m) \right], \tag{2.31}$$

and finally an approximation of (2.28) by

$$\text{EU}(\boldsymbol{x}) = \text{TU}(\boldsymbol{x}) - \text{AU}(\boldsymbol{x}). \tag{2.32}$$

In the case of variance (2.29), AU and EU can be approximated as follows:

$$\text{AU}(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \sigma_m^2 \,, \tag{2.33}$$

$$\text{EU}(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \big[\mu - \mu_m\big]^2 \,, \tag{2.34}$$

with

$$\mu = \frac{1}{M} \sum_{m=1}^{M} \mu_m,$$

and regression models parameterizing the normal distribution, yielding a mean and standard deviation:

$$h_m(\boldsymbol{x}) = \{\mu_m, \sigma_m\}.$$

Other popular approaches for disentangling aleatoric and epistemic uncertainty, apart from the outlined Bayesian approach, include evidential deep learning [SKK18; Ami+20], which explicitly models second-order uncertainty by using evidence to parameterize a Dirichlet distribution, thereby allowing the variance of the Dirichlet distribution to capture epistemic uncertainty, and credal sets, which represent a set of probability distributions [SH21; HDS22; SCH23].

2.4 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a rapidly growing field that encompasses a set of methods and techniques designed to make the internal behaviors and decisions of ML systems understandable to humans [Ali+23; Dwi+23; Has+24]. This need arose from the fact that many ML models are difficult to comprehend and trust due to their black-box nature. Understanding the reasoning behind an ML model's decisions is often essential and, in some cases, legally mandated by the EU [GF17; Kam21]. Figure 2.17 provides a concise taxonomy of explainable AI as outlined by Molnar [Mol25]. In general, XAI methods can be broadly categorized into two groups: those that are *interpretable by design* (also referred to as *glass-box* or *white-box* models), such as linear regression, decision trees, or rule-based models, and those that enable *post-hoc interpretability*, which generate explanations for a model after it has been trained. Post-hoc methods are typically aimed at providing insights into non-interpretable complex models, such as tree ensembles [Bre01; Fri02; CG16; Ke+17; Pro+18] or (deep) neural networks [LBH15; GBC16]. These post-hoc methods can further be divided into *model-agnostic* methods, which can be applied to any type of model, and *model-specific* methods, which are only applicable to a specific class of models, such as tree-based models [LEL18]. In terms of the scope of the explanation, a distinction is made between single-prediction explanations, called *local explanations* (Figure 2.18), and explanations that describe the overall behavior of the model, called *global explanations*.

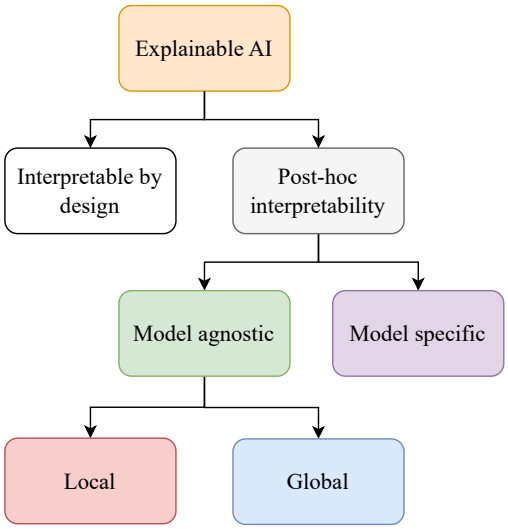


Fig. 2.17: Concise taxonomy of XAI methods according to Molnar [Mol25].

Furthermore, there are various types of explanations [Bod+23]. For tabular data [SAR21], which is the focus of this thesis and the considered use case, com-

mon explanation types include rule-based explanations [Gui+18; RSG18], feature attributions [LL17; RSG16], prototypes [KKK16], and counterfactuals [Dan+20; WMR17; MST20; Gui24].

Rule-based Explanations In rule-based explanations, the goal is to explain a prediction by providing an interpretable `if-then` rule for a black-box model's prediction, as such rules are considered human-interpretable. One example of a rule-based explanation method is *Anchors* [RSG18], which aims to find a decision rule that "anchors" the prediction sufficiently. Formally, an anchor A is defined as follows:

$$\mathbb{E}_{\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{z}|A)}[1_{f(\boldsymbol{x})=f(\boldsymbol{z})}] \geq \tau; A(\boldsymbol{x}) = 1, \tag{2.35}$$

where A is an interpretable *anchor* rule (a set of predicates) that returns 1 if all its feature predicates are true for the instance to be explained, $\boldsymbol{x}$. The term $\mathcal{D}_{\boldsymbol{x}}(z|A)$ represents the distribution of instances in the neighborhood of $\boldsymbol{x}$, and f denotes the black-box model (not h, as is common in machine learning; e.g., a neural network). The anchor rule A is considered a good explanation if it holds for a large proportion of instances in the neighborhood $\mathcal{D}_{\boldsymbol{x}}(z|A)$ of $\boldsymbol{x}$, where $\tau \in [0, 1]$ is a user-defined threshold. This threshold ensures that only rules achieving a certain level of local fidelity with respect to the black-box model's prediction are considered valid.

Feature Attribution Methods Feature attribution methods are arguably the most widely used and popular explanation methods in the tabular data context. They assign a score to each feature, indicating its contribution to the model's prediction. *Local Interpretable Model-agnostic Explanations* (LIME) [RSG16] is a well-known method for computing feature attributions, which approximates the model locally using an interpretable surrogate model, such as linear regression. Formally, a local surrogate model can be expressed as follows:

$$\xi(\boldsymbol{x}) = \underset{g \in G}{\arg\min}\ l(f, g, \pi_{\boldsymbol{x}}) + \Omega(g), \tag{2.36}$$

where $g \in G$ is the explanation model from the space of interpretable surrogate models G, such as linear models, decision trees, or decision rules. A *faithful* surrogate model g minimizes the loss function $l(f, g, \pi_{\boldsymbol{x}})$ (e.g., mean squared error), which quantifies the discrepancy between the black-box model f (e.g., XGBoost [CG16]) and the surrogate model g (e.g., linear regression) in the neighborhood of the instance $\boldsymbol{x}$, as defined by the proximity measure $\pi_{\boldsymbol{x}}$. Additionally, the surrogate model is subject to a complexity constraint $\Omega(g)$ that ensures interpretability, such as keeping the number of non-zero weights low in a linear model.

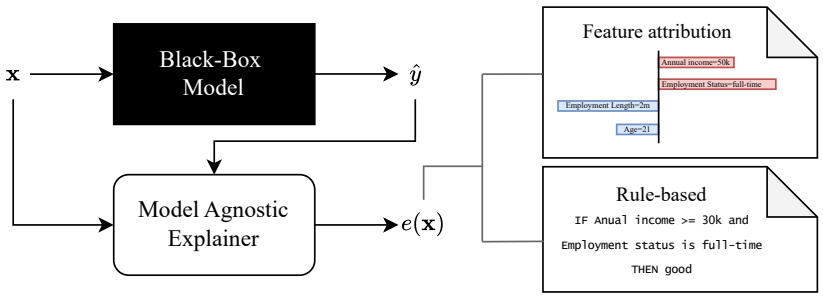


Fig. 2.18: An illustration of local, model-agnostic feature attribution methods and rule-based explanations for a black-box model in the context of credit scoring.

Another very popular feature attribution method is *SHapley Additive exPlanations* (SHAP) [LL17], which is based on Shapley values [Sha+53] from cooperative game theory. Formally, SHAP represents the contributions of each feature to the model's prediction as follows:

$$g(\boldsymbol{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i', \tag{2.37}$$

where g is a linear explanation model, $\boldsymbol{z}' \in \{0,1\}^M$ is a simplified binary feature vector with M features indicating the absence (0) or presence (1) of a feature, and $\phi_i \in \mathbb{R}$ is the SHAP value for feature i. SHAP adheres to three key theoretical properties that ensure reliable and interpretable feature attribution, namely *Local Accuracy*, *Missingness*, and *Consistency*. *Local Accuracy* ensures that the sum of SHAP values for all features equals the model's prediction minus the baseline (expected value) ϕ_0, which makes it *faithful* to the model's output:

$$f(\boldsymbol{x}) = g(\boldsymbol{x}') = \phi_0 + \sum_{i=1}^{M} \phi_i x' i = \mathbb{E}[f(\boldsymbol{x})] + \sum_{i=1}^{M} \phi_i.$$

Missingness guarantees that features which are absent or irrelevant to a prediction are assigned a SHAP value of zero.

$$x_i' = 0 \implies \phi_i = 0.$$

Consistency ensures that if a feature's marginal contribution increases in an updated model f', its SHAP value also increases or remains unchanged. Let $f_{\boldsymbol{x}}(\boldsymbol{z}') = f(h_{\boldsymbol{x}}(\boldsymbol{z}'))$, where $h_{\boldsymbol{x}}$ is a mapping function that maps simplified features to original inputs, and $\boldsymbol{z}' \setminus i$ denotes setting $z_i' = 0$. Then, for any two models f and f', the following holds:

$$f'_{\boldsymbol{x}}(\boldsymbol{z}') - f'_{\boldsymbol{x}}(\boldsymbol{z}' \setminus i) \geq f_{\boldsymbol{x}}(\boldsymbol{z}') - f_{\boldsymbol{x}}(\boldsymbol{z}' \setminus i) \quad \implies \quad \phi_i'(f', \boldsymbol{x}) \geq \phi_i(f, \boldsymbol{x}),$$

where $f'_{\boldsymbol{x}}$ is the updated model, and $\boldsymbol{z}' \setminus i$ is the feature vector with feature i removed. SHAP values can be estimated using different methods. The original model-agnostic method is *KernelSHAP* [LL17], which approximates the Shapley values using a linear model:

$$l(f, g, \pi_{\boldsymbol{x}}) = \sum_{\boldsymbol{z}' \in Z} \left[f(h_{\boldsymbol{x}}(\boldsymbol{z}')) - g(\boldsymbol{z}') \right]^2 \pi_{\boldsymbol{x}}(\boldsymbol{z}'),$$

and a Shapley kernel to achive Shapley compliant weighting of features:

$$\pi_{\boldsymbol{x}}(\boldsymbol{z}') = \frac{(M-1)}{\binom{M}{|\boldsymbol{z}'|} |\boldsymbol{z}'| (M - |\boldsymbol{z}'|)},$$

where M is the maximum number of features, and $|\boldsymbol{z}'|$ the number of present features in the simplified feature vector $\boldsymbol{z}'$. The formula assigns higher importance to subsets that are smaller or larger (close to extremes) and lower importance to subsets of intermediate sizes, balancing the contributions across all feature subsets. Moreover, this formulation establishes a connection between SHAP and LIME (2.36). However, this method is computationally expensive, as it requires 2^M evaluations of the model to compute the exact Shapley values for M features. To address this issue, TreeSHAP [LEL18] was introduced as a more efficient and model-specific algorithm for tree-based models, such as XGBoost [CG16], LightGBM [Ke+17], or CatBoost [Pro+18]. TreeSHAP computes SHAP values in polynomial time, $O(TLD^2)$, instead of exponential time, $O(TL2^M)$, where T is the number of trees, L is the maximum number of leaves, and D is the maximum depth of any tree. Unlike LIME, SHAP can also be used to obtain global explanations int the form of global feature importance scores for each fetaure I_i, which are computed by averaging the absolute SHAP values per feature ϕ_i over all N instances in the dataset:

$$I_i = \frac{1}{N} \sum_{j=1}^{N} |\phi_i^{(j)}|, \quad \text{with } i = 1, \ldots, M.$$

Prototypes Another type of explanation is to describe a dataset through representative examples, called *prototypes*, which represent a set of similar data records [KKK16; Gur+19]. In general, any clustering algorithm that returns actual data instances as cluster centers can be used to identify prototypes, such as k-medoids [RK87]. Kim et al. [KKK16] propose a method to identify prototypes based on the squared *Maximum Mean Discrepancy* (MMD) [Gre+12], which is a measure of the distance between two probability distributions. The squared MMD is defined as follows:

$$\text{MMD}^2 = \frac{1}{m^2} \sum_{i,j=1}^{m} k(\boldsymbol{z}_i, \boldsymbol{z}_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\boldsymbol{z}_i, \boldsymbol{x}_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad (2.38)$$

where m is the number of prototypes $\boldsymbol{z}$ that need to be defined upfront, and n is the number of actual instances $\boldsymbol{x}$. The goal is to minimize the squared MMD between the prototypes and the actual instances, using a kernel function as the distance measure, such as the radial basis function (RBF) kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \epsilon^{(\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|)},$$

where $\|x - \boldsymbol{x}'\|$ is the Euclidean distance between two points and γ is a scaling parameter. However, Kim et al. [KKK16] argue that prototypes alone are insufficient to explain a dataset and propose the use of *criticisms* in addition to prototypes to provide a more complete explanation of the data. Criticisms represent instances that are not well explained by the prototypes and are the points that deviate the most from both the prototypes and the dataset. They can be identified through a witness function, which is defined as follows:

$$f_{witness}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} k(\boldsymbol{x}, \boldsymbol{x}^{(i)}) - \frac{1}{m} \sum_{j=1}^{m} k(\boldsymbol{x}, \boldsymbol{z}^{(j)}),$$

Just as the number of prototypes needs to be defined upfront, so does the number of criticisms. In their original form, prototypes and criticisms serve as a global explanation method. However, they can also be employed as a local explanation technique by identifying the "nearest prototype" $s \in S$ for a new data instance $\boldsymbol{x}$ to be predicted:

$$f(\boldsymbol{x}) = \arg\max_{i \in S} k(\boldsymbol{x}, \boldsymbol{x}_i).$$

Counterfactuals Finally, *counterfactuals* [WMR17] are another type of explanation that can be used to clarify a model's predictions on tabular data. A counterfactual explanation of a prediction describes the minimal change necessary to alter the outcome of the prediction to a predefined, different output [WMR17; Dan+20]. To identify counterfactuals, Wachter et al. propose minimizing the following loss function [WMR17; Mol25]:

$$l(\boldsymbol{x}, \boldsymbol{x}', y', \lambda) = \lambda \cdot (f(\boldsymbol{x}') - y')^2 + d(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.39}$$

where the optimization problem is defined as:

$$\arg\min_{\boldsymbol{x}'} l(\boldsymbol{x}, \boldsymbol{x}', y', \lambda). \tag{2.40}$$

In 2.39, the first term represents the quadratic distance between the model's prediction, $f(\boldsymbol{x}')$, for the counterfactual instance, $\boldsymbol{x}$, and the desired prediction, y'. The

second term, $d(\boldsymbol{x}, \boldsymbol{x}')$, is a distance measure between the original instance $\boldsymbol{x}$ and the counterfactual instance $\boldsymbol{x}'$. The parameter λ is a trade-off parameter that balances the importance of the two terms. The distance measure $d(\boldsymbol{x}, \boldsymbol{x}')$ is defined as the Manhattan distance weighted by the inverse *Median Absolute Deviation* (MAD) of each feature, which enhances robustness to outliers:

$$d(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}, \tag{2.41}$$

where the *Median Absolute Deviation* for the j-th feature is given by:

$$MAD_j = \text{median}_{i \in \{1,\dots,n\}} \left(|x^{(i)}j - \text{median}_{l \in \{1,\dots,n\}}(x_j^{(l)})| \right), \tag{2.42}$$

and p is the number of features, x_j and x'_j are the values of the j-th feature for the original and counterfactual instances, respectively, and n is the number of data points. The ideal counterfactual is one that requires minimal changes to feature values while keeping the rest constant.

Although the use of post-hoc methods is appealing, as they do not restrict the choice of model class, it is also worth considering white-box models. In particular, *Generalized Additive Models* (GAMs) [HT87] are noteworthy, as they hardly exhibit any performance-interpretability trade-off for tabular data [Kru+25]. GAMs are capable of capturing complex, non-linear patterns while remaining fully interpretable [Lou+13].

3 Automating Expert Ratings

This chapter introduces the automotive goodwill claim assessment process as an exemplary prescriptive analytics use case and discusses the challenges associated with automating expert ratings through behavioral cloning. Subsequently, to address these challenges, a conceptual framework is proposed that integrates a supervised machine learning model with components for uncertainty representation and quantification, selective classification, explainability, and bias mitigation.

3.1 Exemplary Use Case

An example of a prescriptive analytics use case that will be explored throughout the contributing publications of this thesis (Chapter 4) is automotive goodwill claim assessment. At the car manufacturer under investigation, goodwill claim assessments are currently conducted either through automatic decision rules or manually by human experts. The latter method is predominantly used, as manually assembled decision rules often fail to cover all aspects due to the high maintenance effort required and the limited availability of human experts. Moreover, goodwill is not governed by clear laws like warranty, which is a legal obligation for manufacturers [MB05]. The patterns underlying goodwill are variable and dynamic, making it even more challenging to maintain them as decision rules in a rule-based system. Given that the process already employs a logic-based decision rules model to automatically assess certain goodwill claim cases, it can be considered a "classic" prescriptive analytics use case. However, this approach also reveals its limitations and suggests that transitioning to a machine learning model, where decision patterns are dynamically learned from data rather than pre-specified by experts, may be beneficial.

Figure 3.1 provides an overview of the goodwill process at the car manufacturer. When a customer encounters an issue with a vehicle outside the warranty period, they may still receive compensation through the manufacturer's goodwill. To determine whether the manufacturer is willing to contribute to repair costs, specifically for parts and labor, the dealer workshop contracted for the repair can submit a goodwill request to the manufacturer. The manufacturer then assesses whether, and to what extent, it is reasonable to contribute to the requested costs. These decisions

are based on a range of vehicle-specific and case-specific factors. In the case of the investigated car manufacturer, contributions are granted on a percentage rating scale 0%, 10%, 20%,. . ., 100%, binned in ten percent increments for the requested repair costs, separately for parts and labor. Unlike during the warranty period, there is no legal obligation for the manufacturer to contribute; rather, it is a commercial decision that must balance customer satisfaction against financial expenses. Given this context, the goodwill claim assessment process qualifies as a high-stakes financial process. Incorrect assessments that are perceived too low can negatively impact customer satisfaction, while assessments that are too high can adversely affect the manufacturer's financial bottom line. As already mentioned, current decisions are based either on predefined static rules defined by the manufacturer or on the experience and intuition of human experts, without immediate and direct customer feedback. Consequently, decisions are based on what the manufacturer and the experts perceive as fair and appropriate, rather than on the beneficiality of outcomes, such as future vehicle or service purchases triggered by valuable goodwill decisions. Nonetheless, since the manufacturer receives hundreds of thousands of goodwill requests each year, most of which require manual assessment by human experts, there is an urgent need for increased automation, faster decision-making, standardization, and reduced reliance on human expertise, all of which could potentially be achieved through the use of machine learning.

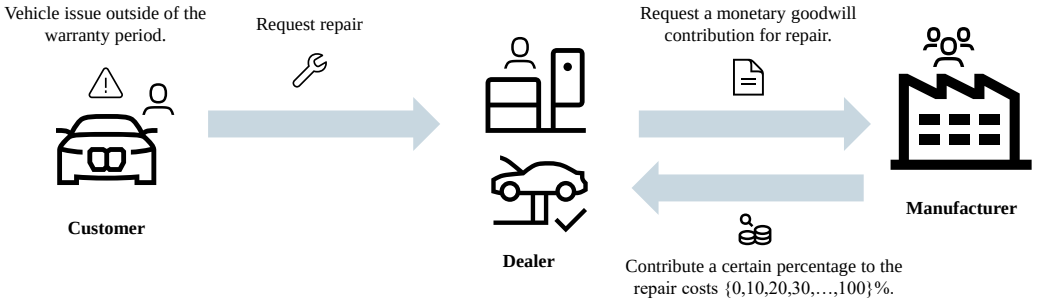


Fig. 3.1: Overview of the automotive goodwill claim assessment process.

3.2 Challenges

Given the abundance of historical goodwill claim assessment data, it is tempting to employ supervised machine learning to automate these expert ratings. By training decision models on historical data, the reliance on human domain knowledge and input, as compared to manual and rule-based decisions, can be reduced. This approach also allows for addressing more complex decision patterns. However,

this approach also presents several pitfalls and challenges, as illustrated by the current use case of goodwill, which are also transferable to observed expert ratings in general:

Lack of outcome data A strong foundational limitation is the lack of actual outcome data regarding rating decisions. As briefly discussed in Section 2.1, only behavioral cloning is within scope for the considered use case and this thesis from the set of potential learning paradigms, as only inputs and corresponding decisions are observed (X, A). While causal ML is increasingly regarded as the gold standard for prescriptive analytics [Ker+25], it requires access to treatment-outcome pairs (T, Y) to estimate treatment effects, such as the CATE, which are not available for the use case at hand. These estimated effects provide principled guidance for decision-making, e.g., determining whether a certain goodwill treatment would encourage the customer to remain loyal to the brand. However, acquiring such data is highly non-trivial, as counterfactual outcomes, that is, the outcomes of unchosen treatments (or actions), are never jointly observable for the same instance. Actively enforcing randomized treatment assignment in this setting would be ethically questionable and could result in individuals not receiving beneficial interventions, such as a goodwill contribution. Furthermore, it is often unclear what the appropriate outcome is: outcomes may be multi-dimensional, delayed, or only partially observable. Similarly, as in causal ML, the reward variable R assumed in offline contextual bandit learning is also not available in the current application. Real-world outcomes (e.g., future vehicle purchases) are often delayed or unobservable. All in all, the behavioral cloning setting is the only applicable learning paradigm and is a very limited one, as in its standard form, it will only mimic expert decisions instead of optimizing for improved business outcomes.

Human decision bias The available goodwill decision data is observational, with human decision-makers acting as "teachers". Hence, each expert's judgment is subject to individual variations, which can stem from a multitude of factors, including but not limited to personal experience, perception of brand loyalty and market situation, or even the time of the day. This leads to two obvious rating biases, visible in four exemplary sales market-specific goodwill data sets in Figure 3.2:

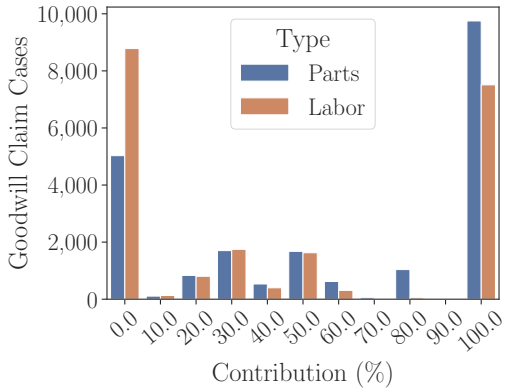
- **Absolute Judgement Bias:** The overall shape of the rating distributions resembles a U-curve, characterized by a pronounced preference for extreme ratings, such as 0 or 100 percent, which is a common bias in rating scales [Tse17; BGM22]. This pattern results in highly imbalanced datasets, which can pose significant challenges for statistical analysis and predictive modeling [FGH11;

Fer+18b]. The concentration of ratings at the extremes suggests that raters may exhibit a bias towards absolute judgments, potentially underutilizing the nuanced options available within the rating scale. This bias contrasts with another common bias in rating scales, the "central tendency bias", where raters cluster their ratings in the center of the scale [Alf+24; AEV24].

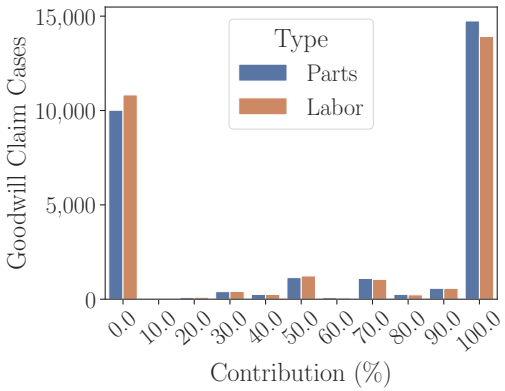
- **Midpoint Rounding Bias:** Decision-makers often exhibit a notable tendency to favor certain midpoint ratings, such as 50 percent, over more precise values like 40 or 60 percent. Similarly, they may prefer ratings like 30 or 70 over those closer to the extremes, such as 10, or 90. This behavior results in a phenomenon where the rating scale is not fully utilized, creating gaps or "holes" in the distribution. Such clustering around common ratings may reflect a psychological bias toward simplicity or a reluctance to commit to more granular assessments and can be attributed to effort reduction, as choosing midpoints requires less thought [All+16]. Moreover, a "central tendency bias" for partial contributions (0 to 90 percent) at 50 percent is also observable for certain sales markets (Figure 3.2c).

Concept drift Another challenge associated with human decisions in general, and specifically for the goodwill claim assessment use case, is that these decisions do not necessarily remain static over time (Figure 3.3). Depending on the context or market situation, these decisions may change. In goodwill claim assessments, decisions can be influenced by factors such as the available budget, the current sales situation, or marketing demands. In machine learning, this behavior is referred to as concept drift or shift, depending on how quickly or abruptly the statistical properties of the target variable in relation to the features change over time [Lu+18]. More formally, the conditional probabilities between inputs and outputs $p(y \mid \boldsymbol{x})$ change over time while the input distribution $p(\boldsymbol{x})$ stays the same.

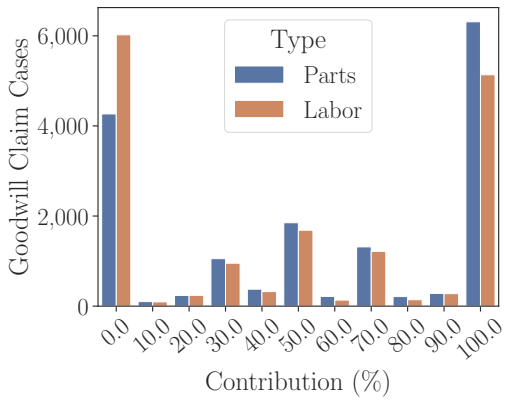
Data uncertainty Given that historical expert ratings are biased and subject to change over time, they cannot be considered the "gold standard" for goodwill claim decisions or for consistent decision-making in general. In extreme cases, expert decisions may even be *polarized* [ER94], lying at opposite extremes (0% vs. 100%) despite having the same input. In machine learning, such label issues are typically referred to as *label noise* [FV14; Son+23; Shi+24], which involves a lack of high-quality labels due to inaccuracies in labeling. A particular bias leading to noisy labels is *label bias* [JN20; DB20], which occurs in cases where there is a systematic issue, such as biases inherited from the views of the annotators due to cognitive bias, subjectivity or fatigue. This issue is also closely linked to the topic of *fairness*



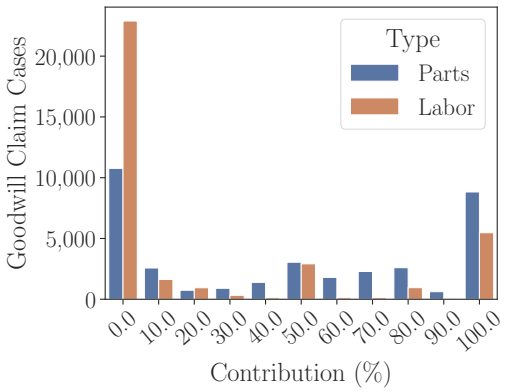
(a) Distribution of goodwill ratings for sales market A.



(b) Distribution of goodwill ratings for sales market B.

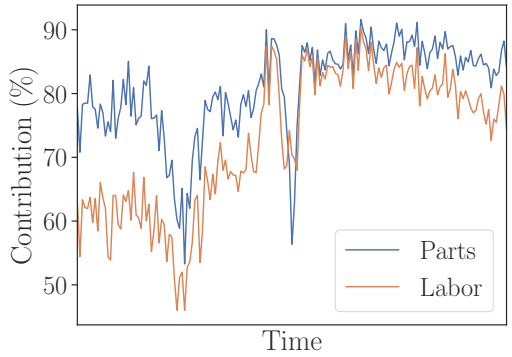


(c) Distribution of goodwill ratings for sales market C.

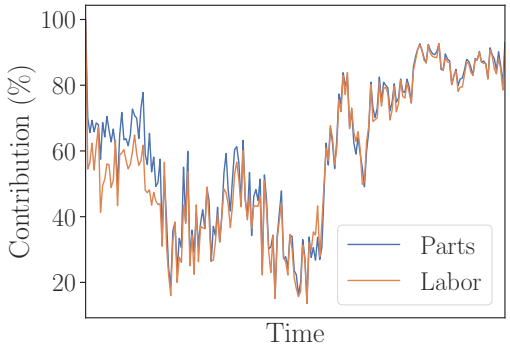


(d) Distribution of goodwill ratings for sales market D.

Fig. 3.2: Goodwill rating distributions for four exemplary sales markets.



(a) Bi-weekly average goodwill contributions for sales market A over time.



(b) Bi-weekly average goodwill contributions for sales market B over time.

Fig. 3.3: Concept drift of goodwill claim ratings for two exemplary sales markets.

in machine learning [Meh+21; Fav+23], as it may lead to unfair models that may discriminate against certain groups. *Selection bias* occurs when the selection of data instances is not representative of the underlying distribution [Fav+23], which, for instance, can occur when raters have varying levels of training or experience and use ratings inconsistently. Another form of label issue is *label ambiguity*, where multiple labels are provided for an instance by multiple annotators, and it is unclear which one to select as the ground truth due to the inherent complexity of the labeling task, overlapping categories, or varying expertise of the annotators [PK19; RPR13; Yan+14]. Due to the availability of only single deterministic labels, goodwill ratings will, at first glance, fall under the first label issue categories. However, the third issue is also highly relevant, as ambiguity is certainly present in controversial goodwill rating decisions, albeit concealed by the use of a single deterministic label. Needless to say, the quality of the labels is crucial for the performance of the machine learning model. If the data is used as is, it will likely result in a high degree of *aleatoric* uncertainty in model predictions, which arises from the randomness and ambiguity inherent in the human decision-making process. This type of uncertainty is irreducible and must be explicitly addressed [HW21].

Knowledge uncertainty Another type of uncertainty that may arise in machine learning-based goodwill claim assessment is epistemic uncertainty, also referred to as knowledge uncertainty. This type of uncertainty can be mitigated through the incorporation of additional training data [HW21]. Specifically, in the context of goodwill claim assessment, epistemic uncertainty may occur with new vehicle types or components, leading to requests that were not included in the model's training data and are therefore *Out-Of-Distribution* (OOD) [Yan+24]. Knowledge uncertainty can also be introduced over time by *data drift*, which occurs when the input data distribution changes. For example, the transition from combustion engines to electric motors introduces new and different relevant criteria for goodwill claim assessments. In this case, the input data distribution $p(\boldsymbol{x})$ changes, which may disrupt the relationship between inputs and outputs as represented by the conditional probability distribution $p(y \mid \boldsymbol{x})$.

Ordinal rating scale Another complexity of this use case is the ordinal rating scale. Naively, goodwill assessment could be treated as a nominal classification or regression problem, but neither of the two would capture the exact essence of the problem. Classification focuses too much on the exact hit rate, allowing larger errors, whereas regression prioritizes reducing the error distance at the cost of the exact hit rate and also requires rounding since it outputs continuous values, which may lead to

additional information loss. Ordinal classification or regression, as introduced in Section 2.2.4, seeks to find a good trade-off between the exact hit rate and reduced error distance and is therefore highly relevant for the use case at hand and this thesis.

Interpretability Moreover, a significant issue with high-performance machine learning models, such as deep neural networks [LBH15; GBC16] or decision tree ensembles like *Random Forests* (RF) [Bre01] and *Gradient Boosted Trees* (GBTs) [Fri02; Ke+17; Pro+18; CG16], is their lack of transparency compared to traditional interpretable prescriptive technologies such as rule-based systems and mathematical programming, although the interpretability of these technologies may also be debatable. This presumably also explains, to a certain extent, the reluctance to adopt machine learning for prescriptive analytics use cases [Lep+20], where interpretability and post-hoc auditability of decisions are crucial, especially in high-stakes settings such as goodwill claim assessment, where both financial outcomes and customer satisfaction are at stake.

3.3 A Conceptual Framework

To address the challenges outlined above, this section introduces a generic conceptual framework, applicable across domains, aimed at automating expert ratings, such as those utilized in goodwill claim assessments, through a machine learning-based prescriptive analytics approach based on behavioral cloning. Given the absence of observable outcomes or immediate rewards, the goal is not to predict outcomes or estimate causal effects, but rather to model expert decision-making in a manner that is **consistent**, **auditable**, **uncertainty-aware**, **unbiased**, and **aligned with domain-specific requirements**. To support this objective, promising auxiliary signals include tailored uncertainty quantification, particularly the distinction between *aleatoric uncertainty*, which captures inherent noise or randomness in the data, and *epistemic uncertainty*, which reflects uncertainty due to limited knowledge or model capacity [HW21], explainable AI techniques [Dwi+23], weakly supervised learning approaches [Zho18], and the integration of expert domain knowledge [Che+23]. As is standard in behavioral cloning, supervised learning serves as the methodological foundation for approximating the expert's decision policy $\pi(a \mid \boldsymbol{x})$ using a supervised machine learning model $h(\boldsymbol{x})$, where expert-observed decisions A are treated as outcome labels Y during training for inputs X. Figure 3.4 illustrates the various components of this framework. It integrates a *Machine Learning Model* that is enhanced with features for *Uncertainty Representation and Quantification*,

Selective Classification, *Explainability*, and *Bias Mitigation*, all designed to tackle the identified challenges. A key requirement is the awareness of ordinal targets, which is complemented by the principle of *Human-in-the-Loop* (HITL) [Wu+22].

As the machine learning model can only imitate the expert's observed behavior, which may not be aligned with organizational utility in every case, the bias mitigation component is essential to ensure that the model adheres to broader domain-specific requirements and ensures fair decision-making. Additionally, HITL and explainability play important roles in this regard, ensuring broadly aligned decision-making that includes all stakeholder interests. Uncertainty quantification and selective classification allow for risk management when decisions are not clear, enabling deferral and seeking additional information to arrive at valuable decisions. Consequently, the framework is embedded in classical statistical decision theory but emphasizes uncertainty, bias mitigation, explainability, and human involvement due to the unclear utilities of experts that need to be discovered and, if necessary, adjusted to align with broader organizational or societal utility. The subsequent paragraphs will provide a detailed explanation of these components and how they can be implemented.

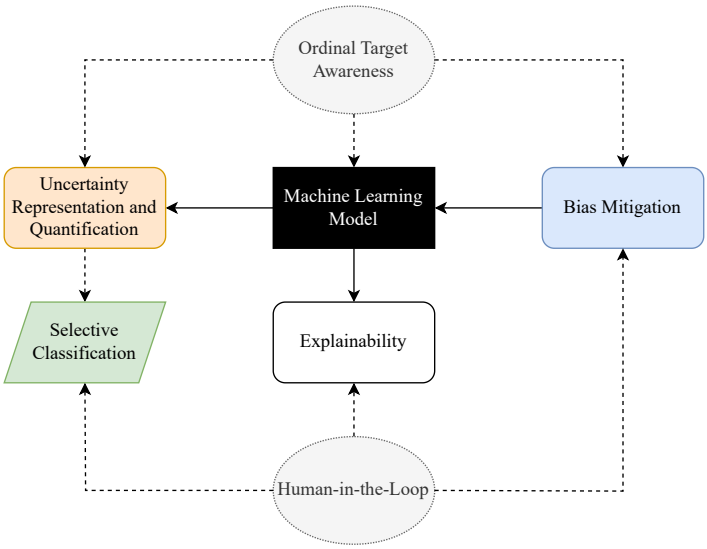


Fig. 3.4: A conceptual framework for machine learning-driven prescriptive analytics aimed at automating expert rating processes.

Bias Mitigation As outlined above, rating data derived from human decision-makers is susceptible to various biases and violates the common assumption in supervised learning that the data, specifically the relationship between inputs and observed decisions, represent ground-truth mappings. Consequently, it is advisable to handle such data with greater caution and reduced confidence, or to explicitly address the

inherent biases to prevent the direct propagation of these biases into the machine learning models being trained [Bol+16]. In other words, the objective should be to learn a prescriptive model capable of making "appropriate" or "practicable" decisions, rather than merely learning a predictive model that replicates the observed decisions [Hül21].

Data-centric AI [Zha+25] and *weakly supervised learning* [Zho18] are fields within machine learning that aim to address this issue from different perspectives. Data-centric AI focuses on improving the quality of the data itself before training a predictor, for example, by filtering out instances with labels considered erroneous [NJC21; VEJ21; Zha+24b] or by utilizing data more efficiently [Ben+09]. In contrast, weakly supervised learning seeks to learn from weakly labeled data, such as data with inaccurate supervision (commonly referred to as *noisy labels*) [Son+23]. Both approaches can generally be employed to mitigate bias in the goodwill claim assessment process or, more generally, to address biased rating data.

A simple approach to bias mitigation is to introduce a weighting scheme that assigns higher weights to underrepresented classes or samples. This approach is common when dealing with imbalanced datasets or in general *sample selection bias* [Hec79; Fav+23]. By assigning higher weights to underrepresented classes or samples, the model is encouraged to focus more on these instances during loss minimization in the training process, potentially reducing bias in the predictions. In goodwill claim assessments, this could be the instances with less frequently used ratings compared to the more commonly used extreme or rounded ratings. The general assumption here is the existence of underlying, unknown, and unbiased labels that can be recovered through an appropriate weighting scheme, ultimately resulting in an unbiased machine learning classifier [JN20]. The equation below shows the empirical risk minimizer (2.3) with weights w_i for each sample i in the training set $\mathcal{D}_{\text{train}}$ [Hua+06]:

$$R^{w}_{\text{emp}}(h) := \frac{1}{N} \sum_{i=1}^{N} w_i \cdot l(h(\boldsymbol{x}_i), y_i) \tag{3.1}$$

Instance weighting is also a way to implement *cost-sensitive* learning by assigning weights to instances according to their misclassification costs [ZLA03; Tin02; LL06]. Depending on the current context or market situation, higher goodwill claim ratings could entail higher misclassification costs than lower ones or vice versa. Other approaches to strengthen the influence of underrepresented classes in machine learning are oversampling [MRA20], undersampling, or more advanced techniques like SMOTE [Cha+02] or ADASYN [He+08], which use data augmentation for oversampling instead of copying existing data points. Another data-centric approach to bias mitigation is to filter out instances with potentially biased labels. This can be

achieved through outlier detection techniques, which identify instances that deviate significantly from the majority of the data [Smi20]. By removing these outliers, the model can focus on learning from more representative and consistent instances, potentially reducing bias in the predictions. Outlier detection can be performed using various unsupervised techniques [LTZ08; Est+96; KM22; AP23]. In general, clustering can be used to identify labels that may not correspond to natural data clusters [Est+96; HW79].

Approaches from weakly supervised learning include *soft label* methods [Sze+16; MKH19] and *superset learning* [LD14; HC15], where label information is imprecise. This imprecision is arguably more realistic than deterministic labels, particularly for observational data that involve human decision biases. For instance, in contrast to soft label approaches, where label information is modeled as a distribution on $\mathcal{Y}$ and which has already been discussed thoroughly in Section 2.2.4.4, superset learning assumes that the true (precise and possibly noisy) label $y \in \mathcal{Y}$ is contained in a set of possible candidate labels $Y \subseteq \mathcal{Y}$ [HC15]. Hence, the available training data for training a predictor changes to observational data of the form

$$\mathcal{O} = \{\boldsymbol{x}_i, Y_i\}_{i=1}^N \in (\mathcal{X} \times 2^{\mathcal{Y}})^N. \tag{3.2}$$

In the case of ordered targets, such as rating data, the assumption of a contiguous superset (e.g., $Y = \{y_2, y_3, y_4\}$) containing the actual label (e.g., $y = y_3$) appears quite natural [DMK23].

A corresponding loss function then needs to perform model identification and "data disambiguation" simultaneously. The extended loss functions hereby needs to compare precise predictions with set-valued observations: $L^* : 2^{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_+$. Hüllermeier and Cheng [HC15] propose the *Optimistic Superset Loss* (OSL) for this purpose:

$$L^*(Y, \hat{y}) = \min\{\, l(y, \hat{y}) \mid y \in Y \,\}, \tag{3.3}$$

which always considers the minimal loss over the set of possible candidate labels Y and the predicted label $\hat{y}$ when performing empirical risk minimization. Moreover, the framework of superset learning is not limited to inherently imprecise or ambiguous data; there is also the option to deliberately turn precise data into imprecise set-valued data [LH21b]. This process of "imprecisation" offers a means to control the influence of individual observations, similar to instance weighting, the more imprecise an observation is made, the less influence it will have on the final model.

More formally, the larger the (super)set Y is, the smaller its corresponding OSL and hence its influence during empirical risk minimization:

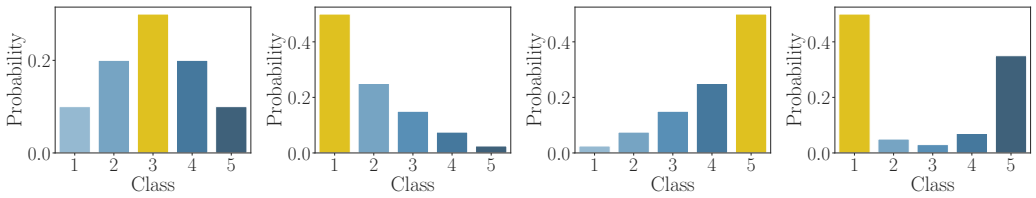
$$Y \supseteq Y' \Rightarrow \forall \hat{y} \in \mathcal{Y} : \; L^*(Y, \hat{y}) \leq L^*(Y', \hat{y}).$$

A further extension of superset learning, inspired by label smoothing [Sze+16; MKH19] and superset learning [HC15], was introduced by Lienen and Hüllermeier [LH21a]. In this approach, labels are modeled as credal sets, which are sets of probability distributions, enabling even greater flexibility.

Returning to goodwill claim assessment, or expert ratings in general, transforming overly used ratings into imprecise probability distributions or sets, rather than overconfident crisp labels, may help mitigate their influence on the final model and ultimately reduce bias in predictions. This approach could, in general, recover the realistic ambiguity lost due to single deterministic decisions, potentially leading to improved predictions in expectation. See Figure 3.5 for examples of labels made imprecise through soft labeling, in which some probability mass is removed from the observed class and distributed to other classes. In the case of rating data, unimodal soft labels, as commonly assumed in ordinal classification, are illustrated in Figures 3.5a, 3.5b, and 3.5c. However, the correct imprecisation for an observed rating might also exhibit a bimodal distribution when the decision is controversial (Figure 3.5d).

Concretely, in the case of goodwill claim assessment, mitigating bias could involve addressing the tendency to favor extreme decisions while underutilizing the intermediate scale. Furthermore, guiding the model toward a specific strategic objective, whether customer-friendly, manufacturer-advantageous, or a balanced compromise between the two, is a critical requirement. As previously discussed, the goal of a prescriptive model is not necessarily to replicate past decisions, particularly when they exhibit evident biases. Instead, the aim is to model the data or pre-select it, ideally in an automated manner, in a way that facilitates the generation of desired prescriptions.

Uncertainty Representation and Quantification To address the uncertainty associated with human expert decisions, machine learning models should be capable of representing and quantifying this uncertainty. Furthermore, the learned machine learning models should be aware of their own limitations and incompetence [Hül21]. This can be achieved through the various approaches outlined in Section 2.3. Particularly, in automating expert ratings through behavioral cloning, predictive uncertainty can serve as a proxy signal to reflect the degree of agreement among experts. This



(a) Unimodal imprecisation for observed rating $y = 3$.

(b) Unimodal imprecisation for observed rating $y = 1$.

(c) Unimodal imprecisation for observed rating $y = 5$.

(d) Bimodal imprecisation for a controversial or polarized observed rating $y = 1$.

Fig. 3.5: Soft label imprecisation examples for observed ratings in which probability mass is deducted from the observed rating and redistributed to other classes.

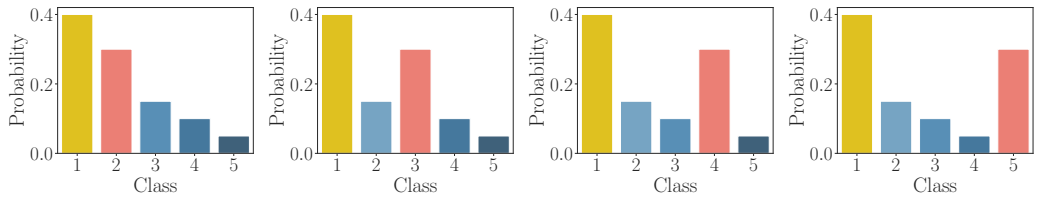
agreement, derived from historical decision data, is typically captured by a machine learning model in the form of a single predictive probability distribution. Achieving such a representation is also a common objective in decision science, as highlighted in various studies [JC96; CW99; MBD24; McA+21; CW93; MMM99]. Moreover, predictive uncertainty can then serve as an indicator of both the prediction's quality and the potential value of the resulting outcome. Using consensus among experts as a surrogate for optimal outcomes is common practice and has been proven effective [Ash85; KW89; CW99; SH89], though it is not warranted in every case [Hoc87; DKM00]. Nonetheless, under the assumption of diverse and independent ratings, and given the expertise of well-intentioned decision-makers, confident decisions, as indicated by a consensus among experts, are likely to lead to good decisions with beneficial outcomes [Jor15].

However, there is a notable research gap in uncertainty quantification for ordered discrete data, such as rating data, as most existing work focuses on nominal classification and regression [HW21; Abd+21]. In particular, common measures for quantifying uncertainty based on probabilities (Section 2.3.1.1), such as entropy or confidence, may not be well-suited for rating data because they fail to account for the inherent distances between ordinal categories. See Figure 3.6 for an illustration. In this figure, entropy (2.21), confidence (2.22), and margin (2.23) fail to distinguish between four different probability distributions because they are invariant to the re-distribution of probability mass. Variance (3.4), a commonly used uncertainty measure in regression tasks [MPU21], appears more suitable for ordered outcomes, as it effectively captures the increased dispersion of the probability distributions.

$$\mathbb{V}(\boldsymbol{p}) = \sum_{k=1}^{K} p_k \cdot (k - \mu), \text{ with } \mu = \sum_{k=1}^{K} p_k \cdot k. \tag{3.4}$$

Variance, however, requires ratings on a cardinal scale, assuming equal distances between categories, which may not always be appropriate. Nevertheless, as already discussed, converting ordinal targets to a cardinal scale is a common practice in ordinal classification [Gut+16].

Similarly, perfect *confidence* calibration [Guo+17], the calibration of the maximum predicted class probability, which is often used as a measure of confidence in a prediction (or inversely, uncertainty) [LHA24], does not provide complete information in this context. For example, even if the distributions depicted in Figures 3.6a and 3.6d are perfectly confidence-calibrated and their confidence levels are identical, there remains a significant difference in their shapes and the potential uncertainty associated with them. This highlights the necessity of additional measures to capture uncertainty beyond simple confidence calibration for ordinal outcomes, leveraging well-calibrated multiclass probabilities [Kul+19].



(a) $H(\boldsymbol{p}_1) = 1.3923$, $C(\boldsymbol{p}_1) = 0.6$, $M(\boldsymbol{p}_1) = 0.1$, $\mathbb{V}(\boldsymbol{p}_1) = 1.39$.

(b) $\mathbb{H}(\boldsymbol{p}_2) = 1.3923$, $C(\boldsymbol{p}_2) = 0.6$, $M(\boldsymbol{p}_2) = 0.1$, $\mathbb{V}(\boldsymbol{p}_2) = 1.4875$.

(c) $\mathbb{H}(\boldsymbol{p}_3) = 1.3923$, $C(\boldsymbol{p}_3) = 0.6$, $M(\boldsymbol{p}_3) = 0.1$, $\mathbb{V}(\boldsymbol{p}_3) = 1.9475$.

(d) $\mathbb{H}(\boldsymbol{p}_4) = 1.3923$, $C(\boldsymbol{p}_4) = 0.6$, $M(\boldsymbol{p}_4) = 0.1$, $\mathbb{V}(\boldsymbol{p}_4) = 2.91$.

Fig. 3.6: Illustration of four different uncertainty measures (entropy $\mathbb{H}(\boldsymbol{p})$ (2.21), confidence $C(\boldsymbol{p})$ (2.22), margin $M(\boldsymbol{p})$ (2.23), and variance $\mathbb{V}(\boldsymbol{p})$ (3.4)) on four different probability distributions $\boldsymbol{p}_1 = (0.4, 0.3, 0.1, 0.15, 0.05)$ (3.6a), $\boldsymbol{p}_2 = (0.4, 0.1, 0.3, 0.15, 0.05)$ (3.6b), $\boldsymbol{p}_3 = (0.4, 0.1, 0.15, 0.3, 0.05)$ (3.6c), and $\boldsymbol{p}_4 = (0.4, 0.1, 0.15, 0.05, 0.3)$ (3.6d) given that the true label is $y = 1$.

On the other hand, if the assumption of unimodal probability distributions in ordinal classification holds, accounting for dispersion may not be particularly relevant, as distributions like those shown in Figures 3.6b, 3.6c, and 3.6d ideally should not occur. Nonetheless, in cases of potentially polarized rating data, such as in the goodwill claim assessment use case (Figure 3.2), this assumption cannot be guaranteed unless strong unimodal constraints are explicitly enforced (Figure 3.7). Unimodality represents a strong inductive bias that does not appear to be justified in the context of goodwill claim assessment and expert ratings in general, given the possibility of polarized ratings. In general, Anderson [And84] distinguishes between two types of ordinal variables. The first type, referred to as "grouped continuous" ordered categorical variables, involves an inherently continuous variable, such as age, being discretized into groups. This type is the primary focus of the ordinal

classification literature [Niu+16; CMR20; Li+22; Yun+24]. For such variables, the assumption of unimodality seems reasonable, since in regression it is common to model uncertainty with a unimodal distribution, such as a Gaussian [Dua+20; MPU21]. However, Anderson also identifies a second type, termed "assessed" ordered categorical variables, where an assessor provides a judgment of the grade of the ordered variable [SH22]. An example of this type is Likert scale surveys, which are commonly used in the social sciences [Lik32], or pain ratings in medicine. For this second type, which corresponds to the goodwill assessment use case, the inductive bias of unimodality appears rather arbitrary. Anderson further notes that for assessed ordered variables, it is not even a priori clear whether or not the ordering is relevant. Additionally, he notes that errors for assessed variables are likely to be greater, which aligns with the observed unconstrained predictive probability distributions in goodwill claim assessment (Figure 3.7). Similarily, in set-based uncertainty quantification (Section 2.3.1.2), there is a strong tendency to enforce contiguous sets for ordinal outcomes, either through the predictor itself [DMK23] or during the construction of the final prediction sets [LAP22; XGW23]. However, as mentioned, given the potential polarization in rating data, this also represents a strong inductive bias and may result in a loss of information.

As outlined in Section 3.2, rating data originating from human decision-makers inherently involves a degree of stochasticity, commonly referred to as aleatoric uncertainty in machine learning [HW21], as decision-makers may not consistently agree in their judgments. Understanding the degree of aleatoric uncertainty is crucial for improving decision-making processes. Since this type of uncertainty is irreducible, it cannot be circumvented directly. However, it can provide valuable insights. For example, in the context of goodwill claim assessment, aleatoric uncertainty could be utilized to de-bias training data by identifying decision alternatives that contribute to high aleatoric uncertainty, determining which alternatives are more appropriate at a given point in time, and subsequently refining the training data based on this analysis. Nonetheless, the simplest immediate action for handling high aleatoric uncertainty is to have the model abstain from making a prediction and instead forward the query to a human expert. The expert can then make a consolidated decision by considering all likely decision alternatives, a strategy that has been shown to enhance human decision-making [Cre+24]. Likewise, detecting epistemic uncertainty is crucial for reliable decision-making. Since the model lacks sufficient knowledge about a query with high epistemic uncertainty, the only viable strategy is to forward the query to a human expert to obtain a label. This label can then be incorporated into the model training process to improve its performance on similar cases in the future.

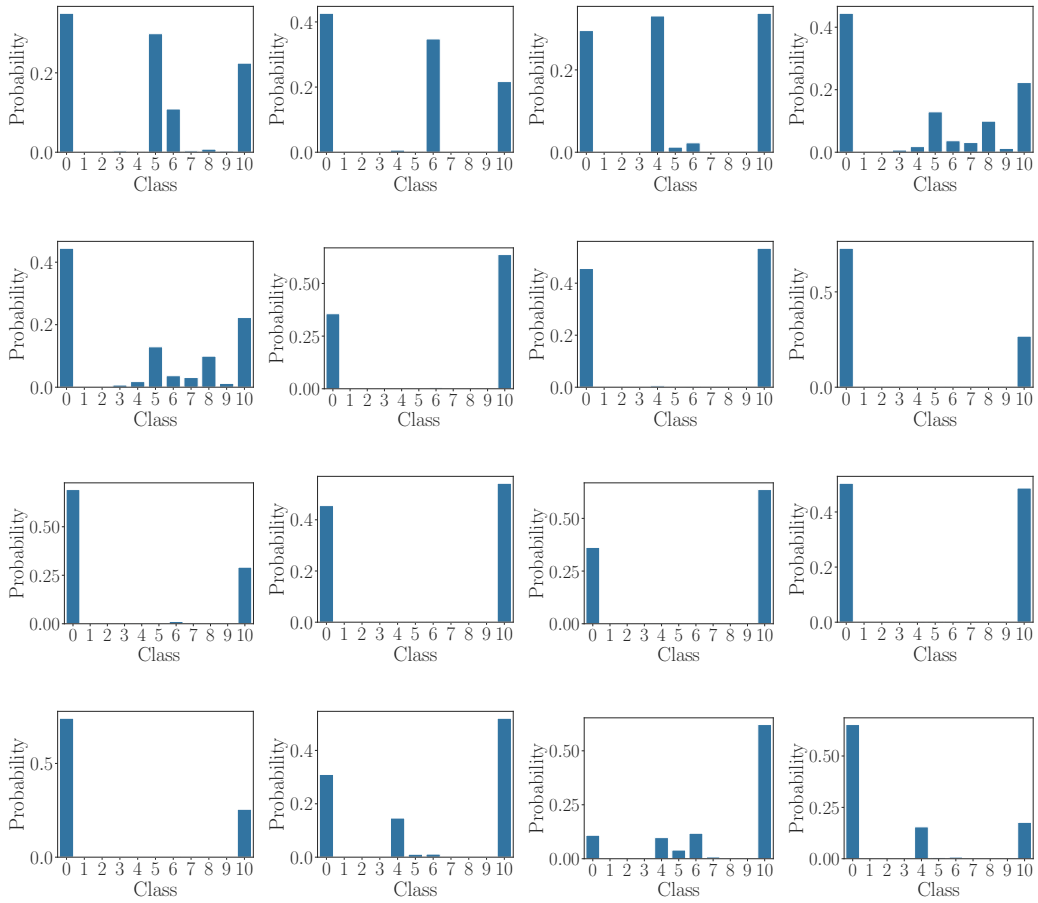


Fig. 3.7: Different multimodal predictive probability distributions (including extreme bimodal ones) are obtained from a CatBoost classifier [Pro+18] trained with the CE loss on a goodwill claim assessment dataset. Contributions $\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$ are ordinally encoded as $\mathcal{Y} = \{0, 1, 2, \ldots, 10\}$.

To distinguish between aleatoric and epistemic uncertainty, entropy and variance-based decompositions based on Bayesian inference (Section 2.3.2) are widely used and flexible methods for nominal classification and regression [Dep+18]. However, these methods have been criticized for not fulfilling several theoretical axioms [Wim+23]. These approaches have also been successfully applied to tree-based models [SH20; MPU21], which are particularly relevant for the goodwill claim assessment use case, as they continue to represent the state of the art for tabular data [SA22]. Furthermore, the Bayesian approach is conceptually appealing in the context of expert ratings, as it can be seen as an attempt to recover the original expert judgments or at least subsets from the aggregated single probabilistic decision. This is achieved through the instantiation of an ensemble of models based on different subsamples from the overall observed decisions (Figure 3.8). This recovery allows for capturing the agreement of a diverse set of expert groups, which is required to quantify epistemic uncertainty, rather than only aleatoric uncertainty based on a single probabilistic prediction. This approach contrasts with the common goal in decision science, which is to aggregate expert opinions [McA+21]. Instead, it embraces diverse opinions in the form of a larger hypothesis space, leading to a second-order probability distribution that enables the quantification of epistemic uncertainty. However, there is currently no specific instantiation of the Bayesian approach tailored for discrete ordered outcomes, such as expert rating data. This represents a significant gap that will be addressed within this thesis.

Selective Classification To leverage predictive uncertainty, an important downstream task is selective classification [GE17], also referred to as classification with abstention or rejection option [Hen+24]. In general, an uncertainty-based *rejector*, denoted as r, is a binary function that evaluates a query $\boldsymbol{x}$ based on the predictor h and its corresponding predictive uncertainty u. The rejector determines whether to reject the query using a predefined threshold τ:

$$r(\boldsymbol{x} \mid h)_\tau := \begin{cases} 1 & \text{if } u(\boldsymbol{x} \mid h) < \tau, \\ 0 & \text{otherwise.} \end{cases}$$

The model's output is then defined as:

$$(h, r)(\boldsymbol{x}) := \begin{cases} h(\boldsymbol{x}) & \text{if } r(\boldsymbol{x} \mid h) = 1, \\ \emptyset & \text{if } r(\boldsymbol{x} \mid h) = 0. \end{cases}$$

In cases where the rejector rejects the query (i.e., $r(\boldsymbol{x} \mid h) = 0$), the model abstains and returns an empty response ($\emptyset$) instead of providing a prediction.

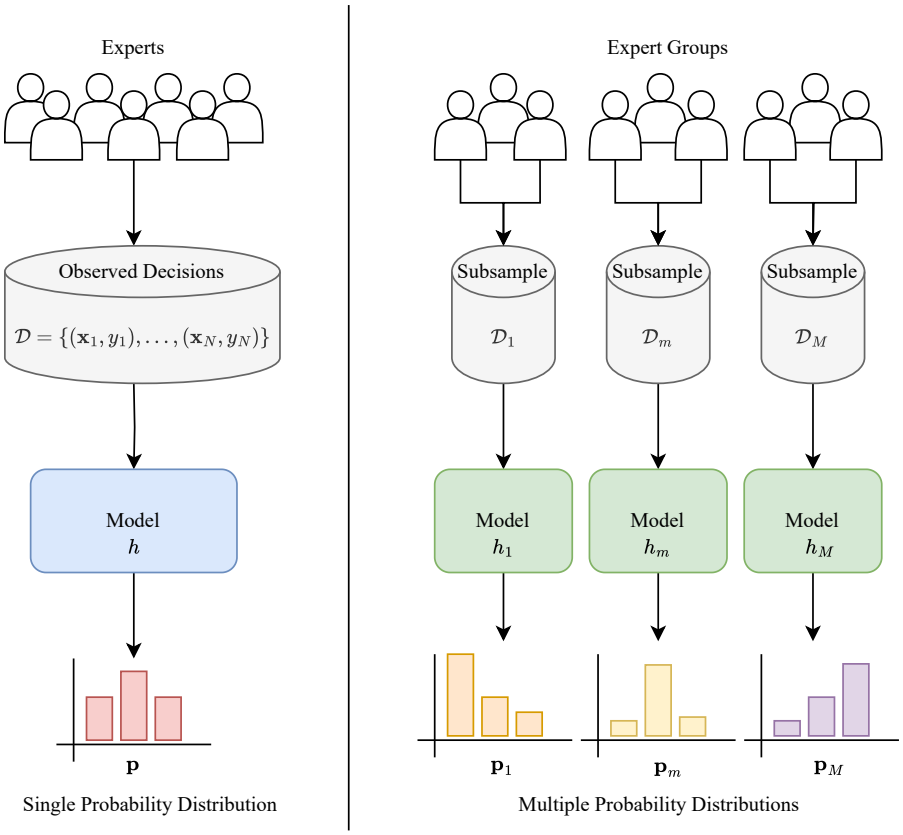


Fig. 3.8: Comparison of a single model trained on observed decisions with a Bayesian ensemble approach, where different models are trained on different subsamples of the data, allowing for the (partial) recovery of diverse expert opinions through multiple predictive probability distributions. This ensemble approach enables the estimation of both aleatoric and epistemic uncertainty.

In general, selective classification has two primary objectives: first, to achieve high accuracy on the queries it chooses to process, defined as

$$\mathcal{A} = \frac{TA}{TA + FA}, \tag{3.5}$$

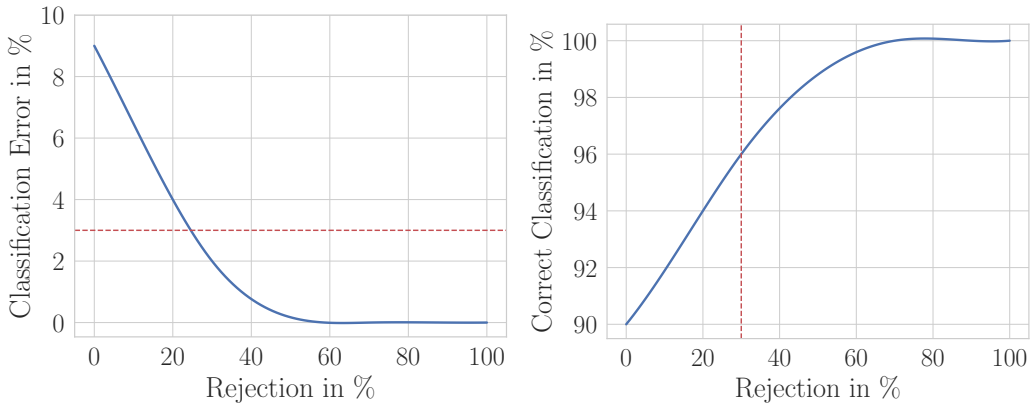
where TA stands for *True Accept* and FA for *False Accept*; and second, to maximize coverage by making predictions on as many queries as possible [Hen+24], defined as

$$\phi = \frac{TA + FA}{TA + FA + FR + TR}, \tag{3.6}$$

where TR and FR represent the *True Reject* and *False Reject* cases, respectively. Furthermore, the rejection rate is defined as $1 - \phi$.

However, these two objectives are inherently in conflict. This trade-off between risk and coverage is commonly referred to as the *Risk-Coverage* (RC) trade-off [EW10]. The uncertainty threshold τ serves as a hyperparameter that can be adjusted to balance the model's risk and coverage. A lower threshold increases the model's coverage by rejecting fewer queries, whereas a higher threshold reduces the model's risk by rejecting more queries. The optimal threshold is use-case dependent and can be fine-tuned by domain experts in a HITL manner [Hup22; Wyn+19]. A common method to evaluate the performance of selective classifiers is to plot the *rejection curve* [NZH10], which illustrates the trade-off between risk and coverage as a function of the uncertainty threshold τ. The rejection curve is typically plotted in a two-dimensional space, with the rejection rate (inverse of coverage) on the x-axis and the risk (e.g., misclassification rate) on the y-axis. Depending on the chosen measure of prediction quality or risk, the curve should either increase or decrease monotonically (Figure 3.9).

A third objective for selective classification arises in the context of observational human decision data: obtaining accurate labels for cases that exhibit high uncertainty. In this framework, queries deemed too uncertain (i.e., $r(\boldsymbol{x} \mid h) = 0$), resulting in an empty response ($\emptyset$), can be delegated to human experts. For instances of epistemic uncertainty, experts provide labels to enhance the model's knowledge by addressing gaps in its understanding. In contrast, for cases of aleatoric uncertainty, experts focus on mitigating bias or consolidating decisions by accounting for the inherent stochasticity in the data. This approach aligns with the principles of HITL systems [Mos+23], which emphasize the critical role of human oversight in automated decision-making processes [and23]. By leveraging human expertise, the model can learn from expert-provided labels and improve its knowledge iteratively over time. This aligns with the concept of *continual learning* [Wan+24], where an ML model is continuously updated with new data while simultaneously retaining



(a) A monotonically decreasing rejection curve based on the misclassification rate and a classification error threshold.

(b) A monotonically increasing rejection curve based on accuracy and a coverage threshold.

Fig. 3.9: Exemplary monotonically decreasing (3.9a) and increasing (3.9b) rejection curves based on the misclassification rate and accuracy, respectively, from which a decision maker can choose the corresponding threshold τ for the rejector, depending on the desired risk or coverage to attain (dashed red lines).

its previously learned knowledge. Proper labels can be obtained through an active learning-like approach, where an expert oracle, typically a human expert, is queried for correct decisions [NSH22; Mos+23]. Ideally, the human expert oracle possesses several key attributes that enhance the effectiveness of oversight and improve upon existing biased decisions. These attributes include suitable epistemic access to relevant aspects of the situation, self-control, and appropriate intentions for their role [Ste+24b]. As mentioned, obtaining correct labels is beneficial for cases with high aleatoric uncertainty, where de-biasing the data is necessary, as well as for cases with high epistemic uncertainty, where labeled data may be unavailable. Although queries with high uncertainty are not intended to be answered directly by the predictor, providing a subset of likely alternatives through a *Decision Support System* (DSS) can assist the expert oracle in reaching a more informed final decision while reducing cognitive load [SR24; Cre+24]. See Figure 3.10 for an overview of the process.

Notably, there is a tension between the *Bias Mitigation* and *Uncertainty Representation and Quantification* components. The more effective the *Bias Mitigation* component is, the higher the initial coverage of the model will be, and the less uncertainty will be present in the data. This implies that the model will be able to make more confident predictions, potentially reducing the need for uncertainty quantification and selective classification. Conversely, if the *Bias Mitigation* component is not effective, the model may struggle to make accurate predictions, resulting in higher uncertainty and an

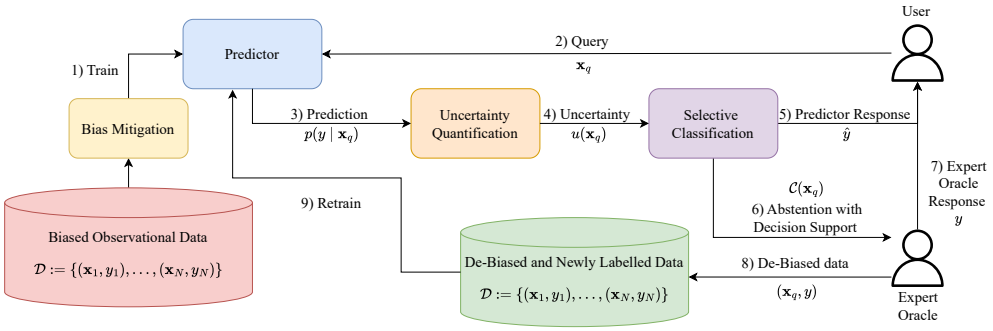


Fig. 3.10: Active learning-like process to either directly answer or to de-bias and label incoming queries based on uncertainty quantification and selective classification.

increased need for uncertainty quantification and selective classification, as well as de-biasing through human experts.

In the context of goodwill claim assessment, a selective classification approach as outlined could offer several advantages. Given that goodwill claim assessment is a high-stakes process, abstaining from making predictions on uncertain cases can help reduce the overall risk of incorrect decisions (3.5). Moreover, delegating uncertain cases to human experts for consolidation and accurate labeling allows the system to benefit from this data for future model training, thereby improving its performance over time. However, it is crucial to ensure that the labeling process is effective and does not perpetuate existing biases [Ste+24b]. Even if the degree of automation is initially low (3.6), it can be gradually increased as data quality improves, reducing reliance on human experts over time. Additionally, automating relatively straightforward cases can free up human experts to focus on analyzing complex and potentially controversial examples, thereby enhancing overall decision quality. This approach could also help balance the workload of human experts. Ultimately, the effectiveness of a selective classification approach depends significantly on the quality of the rejector, which in turn relies on the accuracy of the uncertainty representation and quantification.

Machine Learning Model Naturally, the machine learning model is at the center of the proposed framework and plays a vital role in it. Given the use case, the focus is on the tabular data modality, where *Gradient Boosted Trees* (GBTs) [Fri01], with popular implementations such as XGBoost [CG16], LightGBM [Ke+17], and CatBoost [Pro+18], remain the state-of-the-art model class [YK24; UL24; SA22; McE+23; GOV22]. Nonetheless, advances are being made in neural network-based models for tabular data, particularly through the use of pre-trained transformer-

based foundation models [Vas+17] and *in-context learning* [Hol+23; KGV24; Hol+25; Qu+25; Heg+23; WH25].

The specific choice of predictor in the proposed framework is strongly influenced by the *Bias Mitigation* and *Uncertainty Representation and Quantification* components, as well as their respective implementations. For instance, if a data-centric approach is employed and the human decision data is de-biased and cleaned prior to model training, by leveraging domain expertise, exploratory data analysis, or automated methods [NJC21], any off-the-shelf learner from the regime of ordinal classification methods might be a viable solution (Section 2.2.4). However, if de-biasing is intended to occur during model training by modeling human decision data in a weaker manner, such as through soft labels [DM19], sets [HC15], or credal sets [LH21a], more sophisticated loss functions are required, capable of handling imprecise data. If the focus is on uncertainty quantification, proper scoring rules [GR07; Eps69] as loss functions, potentially with added calibration [Men+23], become the go-to solution, as they are theoretically designed to provide unbiased, truthful predictive probabilities in expectation over a population [Men+23] and, consequently, truthful uncertainty representations.

However, what does "truthful predictive probabilities" mean in the context of rating data? In ordinal classification, truthful predictive probabilities are typically associated with unimodal probability distributions, as incentivized by the RPS [Eps69; Mur70; Gal23], which is considered a proper scoring rule for ordinal outcomes. However, given the differentiation between "grouped continuous" and "assessed" ordered categorical variables [And84], the assumption of unimodal distributions may not hold in the context of "assessed" ordered variables and goodwill claim assessment in particular, as discussed above. In models for rating data, the inductive bias of unimodality, which is prevalent in ordinal classification, appears arbitrary and reduces variance in a way that does not necessarily reflect the true uncertainty in the data.

Furthermore, an important aspect of selecting the type of predictor is the inherent trade-off between high accuracy and reduced error distances. As is commonly acknowledged, predictors that strongly focus on reducing error distances are likely to sacrifice exact hit-rate [Kas+24; Kra+01; BST09]. Thus, tuning this trade-off for a particular use case is another challenge that must be addressed. In the case of expert ratings, exact hit-rate appears to be similarly important to reducing error distances. All in all, in prescriptive analytics, particularly in a behavioral cloning setting, the flexibility of the learner, e.g., its ability to accommodate different loss functions or modeling capabilities, may be more important than achieving the highest predictive power.

Explainability Due to the non-interpretability of high-performing machine learning models, such as tree ensembles [Bre01; Fri02; CG16; Ke+17; Pro+18] or deep neural networks [LBH15; GBC16], there appears to be reluctance in using machine learning for high-stakes prescriptive analytics use cases. Instead, methods deemed interpretable, such as manually crafted decision rules, are often favored [Lep+20]. Naturally, if a predictor functions as a black box that completely obscures its reasoning, it will not foster trust among stakeholders of the ML-based system. XAI (Section 2.4) may serve as a crucial building block for addressing the aforementioned reluctance to adopt ML in prescriptive analytics.

When working with observational ratings, as in the considered use case, XAI is invaluable on multiple levels. First, it provides a global view of the model's behavior, which is crucial for understanding the overall decision-making process and ensuring alignment with expert expectations. For example, in the context of goodwill claim assessment, global feature importance can help ensure that the model is not relying on "spurious correlations", such as the amount of goodwill claimed, which could lead to biased predictions. By analyzing global feature importance, stakeholders can identify potential biases in the model by comparing the global feature importances to their own mental models and taking corrective actions if necessary. Furthermore, local explanations enable debugging the model on a case-by-case basis. Stakeholders can examine the specific reasons behind individual decisions and validate the model's predictions against their expectations at a more granular level. Explanations also play a crucial role in detecting and mitigating biases in the data [VFK21; Nto+20] by allowing the discovery of biased decision patterns. Additionally, local explanations make machine decisions auditable, which may be a legal or compliance requirement, such as the EU GDPR's "right to explanation" [GF17; Kam21]. Interestingly, human experts often face limitations in time and cognitive capacity, which can result in less thorough justifications for their decisions. In this context, there is significant potential to enhance the overall transparency of the decision-making process through the use of XAI technology. By providing clear and understandable explanations for algorithmic decisions, XAI can help bridge the gap in transparency that may exist in human decision-making. Making the uncertainty related to a query transparent may even further increase transparency [Bha+21].

In the context of goodwill claim assessments, the use of both global and local explanations appears essential to safeguard transparent and unbiased decision making. Regarding explanation types, SHAP [LL17] is particularly appealing as it provides both local and global explanations. Additionally, rule-based approaches are attractive due to their natural fit with existing rule-based systems and their widespread use in prescriptive analytics. Counterfactual explanations are another

interesting option, as they have been successfully applied in similar contexts, such as credit applications [McG+18]. Counterfactuals can be especially valuable to customers, as they allow for *what-if* analyses, enabling users to explore how changes to their attributes might affect the outcome. However, this capability may also lead to *strategic classification* [Har+16], where individuals manipulate their attributes to achieve a more favorable classification outcome.

Ordinal Target Awareness A cross-cutting concern that affects all components of the presented framework, to a certain extent, is ordinal target-awareness. Ignoring the ordinal structure of the target can result in suboptimal performance across various aspects, including data de-biasing, the machine learning model, and uncertainty quantification. Acknowledging the ordinal structure of the target and carefully considering the potential trade-off between hit-rate and error-distance is crucial. The specific challenges associated with this concern have already been discussed for each component and will not be reiterated here.

Human-in-the-Loop (HITL) Incorporating human expertise and domain knowledge is another cross-cutting concern, which is essential to fully exploit the potential of a behavioral cloning-based prescriptive ML-based system [Wu+22; Mos+23; Che+23]. It is in particular crucial to align a model with the values and goals of non-technical experts or, in the case of the goodwill use case, with organizational strategies. Chen et al. [Che+23] propose a taxonomy along two axes to translate non-technical expert feedback into machine learning model adjustments or updates: (1) levels of expert feedback, which can be either domain-level or observation-level feedback, and (2) types of model updates, which can involve changes to the dataset, the loss function, or the parameter space.

Domain-level feedback refers to high-level conceptual input provided by domain experts, such as the importance of specific features or the need to account for particular constraints. This type of feedback can guide the selection of data and features, the design of the model architecture, or the choice of the loss function. For example, in the context of goodwill claim assessment, domain-level feedback might suggest that certain features, such as the amount of goodwill claimed, should be treated differently due to their potential to introduce bias. Additionally, specific timeframes may need to be excluded from the data due to outdated decision strategies or special non-representative circumstances, such as a pandemic.

Observation-level feedback, in contrast, focuses on specific instances or predictions made by the model. This type of feedback can be used to refine the model's predictions or to identify and address biases in the data. For instance, in the goodwill

claim assessment use case, observation-level feedback might involve pinpointing cases where the model's predictions deviate from expert expectations or where certain features or decisions are under- or over-represented.

In both cases, XAI serves as an invaluable tool to facilitate the translation of expert feedback into actionable model adjustments and to validate these adjustments effectively. Concretely, expert feedback must be implemented in the *Bias Mitigation* component of the framework by leveraging the methods discussed in the corresponding section. Moreover, the *Selective Classification* component also relies on human input to define an appropriate threshold for the risk-coverage trade-off. This threshold is critical for determining the level of uncertainty at which the model should abstain from making predictions and defer to human experts. The selection of this threshold can have a substantial impact on the model's performance and its ability to effectively balance risk and accuracy.

Furthermore, there is a growing body of literature focused on the development and deployment of machine learning models in a highly automated manner, akin to the processes used in modern software system development [KL22; KKH23]. However, a prescriptive machine learning system, particularly one based on behavioral cloning, may necessitate a more cautious, stepwise development and deployment approach [Lav+22], along with continuous human oversight. This approach involves transitioning a model through several stages: development, shadow mode (where the model operates silently in the background, enabling comparisons between human and model decisions using XAI), decision support (where the model assists decision-makers by suggesting actions and providing explanations through XAI), and finally, automated decision-making (where the model autonomously makes decisions, remains auditable through XAI, and is constrained in its automation scope through selective classification) (Figure 3.11). Notably, all stages allow for reverting to the previous stage if decision contexts change. Each of these stages is guided by human experts who define bias mitigation strategies to align model decisions with organizational goals, evaluate and monitor decisions, and ultimately determine an appropriate risk-coverage trade-off. All in all, human oversight and involvement are key factors for a successful deployment of a behavioral cloning system, as the lack of factual outcome data requires consensus among different system stakeholder groups about the quality and usefulness of the system.

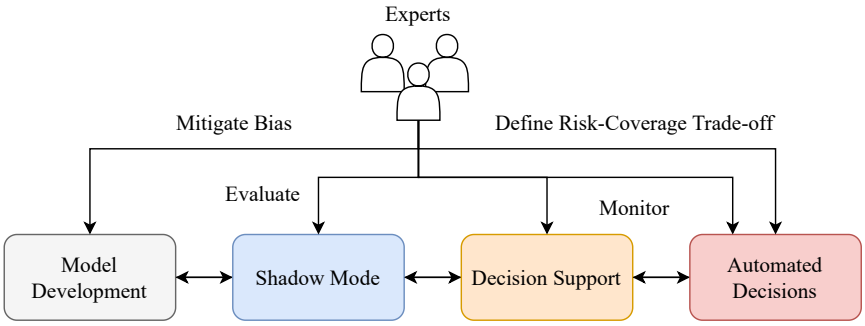


Fig. 3.11: Staggered rollout process of a prescriptive machine learning system, guided by human experts.

4 Contributions of this Thesis

In this chapter, the contributions of this thesis will be presented, focusing on how they address the different challenges outlined in the previous chapter and implement the conceptual framework introduced herein. Refer to Table 4.1 for an overview.

This Ph.D. thesis consists of the following contributions:

1. **Stefan Haas** and Eyke Hüllermeier. “A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part VI*. vol. 13718. Lecture Notes in Computer Science. Springer, 2022, pp. 170–184

   Discussion of Contribution 1: In Section 4.1, the first publication introduces the problem of using machine learning for goodwill claim assessments and proposes a tailored model that is ordinal, hierarchical, and cost-sensitive. In the spirit of learning a practicable prescriptive decision model rather than merely replicating past decisions [Hül21], the proposed model enables the strategic incentivization of the learner in a specific direction (e.g., customer-friendly or manufacturer-friendly). This is achieved through the specification of different cost matrices, which influence instance weights accordingly. Furthermore, the model addresses data imbalance in goodwill decisions, which, as discussed, is likely introduced by a bias toward extreme decisions. This is accomplished by implementing a sequential hierarchy of two models. The first layer determines the qualitative main rank (“no”,“partial”, or “full” contribution) using an ordinal and cost-sensitive model based on binary reduction [LL06]. The subsequent regression model then predicts the exact partial contributions, if required. When examining the proposed conceptual framework (Figure 3.4), the approach primarily contributes to the bias mitigation component by addressing data imbalance through the hierarchical structure and to strategic prescriptive decisions through cost-sensitivity (e.g., “How to decide if costs or customer satisfaction are considered more important?”). Additionally, the hierarchical model, which integrates both ordinal and regression models, demonstrates a robust awareness of ordinal targets. Human control and oversight are facilitated through the flexible specification of different cost matrices, which can be adapted to current contexts and requirements.

2. **Stefan Haas** and Eyke Hüllermeier. “Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests”. In: *International Journal of Data Science and Analytics* (2024)

   Discussion of Contribution 2: In Section 4.2, the second publication builds upon the initial model from the first contribution by incorporating uncertainty quantification. By integrating inductive conformal prediction [Pap08] into the hierarchical model, it introduces uncertainty awareness, producing either a contiguous set of predictions (qualitative main rank layer) or an interval (partial contribution regression layer) rather than a single output. To further enhance the model’s reliability and precision, a selective classification [GE17] technique is employed, allowing the model to process only those instances deemed to possess sufficient certainty. The balance between automation and risk is carefully negotiated through multi-objective optimization. This is achieved by adjusting significance levels and selective classification thresholds, which serve as adjustable parameters to fine-tune the risk-coverage trade-off [EW10]. The trade-off is visualized through rejection curves [NZH10], enabling human experts to select the most suitable balance between risk and coverage. This approach provides decision-makers with a more practical tool compared to marginal coverage guarantees. From a prescriptive analytics perspective, the approach also delivers an actionable outcome: either processing the query automatically when confidence is sufficiently high or delegating it to a human expert when it is not. The contribution primarily addresses the uncertainty quantification component of the conceptual framework (Figure 3.4), as it provides a method for quantifying uncertainty. By employing ordinal and regression-based predictors, the approach also ensures contiguous sets or intervals, which is in line with current conformal prediction approaches for ordinal classification [LAP22; DMK23].

3. **Stefan Haas**, Konstantin Hegestweiler, Michael Rapp, Maximilian Muschalik, and Eyke Hüllermeier. “Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments”. In: *Frontiers in Artificial Intelligence - AI in Business* 7 (2024)

   Discussion of Contribution 3: In Section 4.3, the third publication emphasizes the importance of XAI technology for ML-based prescriptive analytics to identify biases, enhance transparency, and validate machine decisions. Specifically, it proposes a process model to identify post-hoc explanation methods for an existing black-box model that are deemed useful by system stakeholders from the set of available explanation types (e.g., feature importance, counterfactuals) and validates this approach using the use case of automotive goodwill claim assessments. The process model is designed to bridge the gap between the variety of available

explanation types and the identification of those most useful for a specific use case. So far, the selection and evaluation of explanation methods have been largely exploratory, e.g., [Mal+21; ZRH22]. However, the proposed process model provides practitioners with structured yet flexible guidance throughout the process. In the case of the considered use case, the process model led to greater acceptance and trust in the prescriptive ML-based system, suggesting the transferability of the model to other use cases as well.

4. **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*. vol. 14174. Lecture Notes in Computer Science. Springer, 2023, pp. 3–18

   Discussion of Contribution 4: In Section 4.4, the fourth contribution introduces a novel class-wise unimodal label smoothing technique, which is inspired by traditional label smoothing and grounded in the Geometric distribution. This method adjusts the label information by redistributing a portion of the probability mass from the observed label to the neighboring classes in a unimodal manner. This approach is intended to handle label information in the goodwill claim assessment use case in a presumably more realistic and less stringent way. Moreover, two heuristics are introduced to mitigate the effects of concept drift over time and the rounding bias inherent from human expert judgments. Additionally, akin to the cost-sensitive approach employed in the first contribution, this smoothing method can be utilized to guide the predictor towards a desired outcome. By selectively redistributing the probability mass to one side, the model can be controlled to favor either customer or manufacturer interests, depending on the context. The proposed method improves on unimodal soft label approaches for ordinal classification in deep learning proposed so far, and detailed in Table 2.3 of Section 2.2.4.4, as it enables class-wise as well as asymmetric smoothing. Only the later proposed generalized triangular distribution-based soft labels provide similar flexibility in terms of shape control of the soft label distribution, class-wise, and asymmetric smoothing at the cost of requiring complex calculations to discretize the continuous triangular distributions [Var+23a]. The contribution further demonstrates the applicability of unimodal soft labeling to *Gradient Boosted Trees* (GBTs), which is highly relevant for practitioners. Notably, the method is not restricted to GBTs; it has also been implemented and successfully evaluated within the *dlordinal* library on image datasets. This library unifies numerous recent deep ordinal classification methodologies [Bér+25].[1] Natu-

[1] `https://github.com/ayrna/dlordinal`

rally, this contribution primarily focuses on the bias mitigation component of the conceptual framework (Figure 3.4) by proposing a method to address the bias introduced by human experts in the data.

5. **Stefan Haas** and Eyke Hüllermeier. "Uncertainty quantification in ordinal classification: A comparison of measures". In: *Int. J. Approx. Reason.* 186 (2025), p. 109479

   Discussion of Contribution 5: In Section 4.5, the fifth contribution explores uncertainty quantification for probabilistic ordinal classification. It enhances the uncertainty quantification component of the conceptual framework by explicitly addressing the ordinal nature of the data. Traditional uncertainty measures, such as Shannon entropy (2.21), confidence (2.22), or margin (2.23) are unsuitable for this task because they remain invariant under probability mass redistributions and, therefore, fail to capture the ordinal structure of the data. To the best of our knowledge, this is the first work to address uncertainty quantification specifically for the ordinal case, which has been largely overlooked compared to nominal classification and regression, despite its critical relevance in many high-stakes applications such as medicine and finance. This work proposes leveraging measures from the social sciences, particularly those designed for consensus assessment in Likert-scale surveys [AR22], as uncertainty measures in ordinal classification. These measures account for both the shape and the spread of the distribution, making them well-suited to the requirements of uncertainty quantification for ordinal data. Additionally, this paper introduces a novel approach that reduces multi-class ordinal uncertainty quantification to a series of sequentially ordered binary uncertainty quantification problems, wherein any uncertainty measure applicable to a Bernoulli distribution can be employed, e.g., entropy or variance. Furthermore, it proposes several axioms that a proper ordinal uncertainty measure should satisfy and demonstrates that the proposed methods, unlike nominal measures, adhere to these axioms. The theoretical contributions are further validated through extensive experiments on commonly used ordinal tabular benchmark datasets, where the proposed methods outperform standard nominal uncertainty quantification techniques. Specifically, they excel in capturing the requirements for exact hit rate and reduced error spread, which are critical for ordinal classification tasks.

6. **Stefan Haas** and Eyke Hüllermeier. "Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction". In: *Machine Learning* (2026)

   Discussion of Contribution 6: In Section 4.6, the sixth contribution extends the proposed ordinal binary uncertainty quantification method to distinguish

between aleatoric and epistemic uncertainty. It builds upon Sale et al.'s framework [Sal+24], which adapts entropy- and variance-based measures of uncertainty [Dep+18] to the binary case. The proposed ordinal approach, using entropy and variance as binary base measures, is extensively evaluated on commonly used ordinal tabular benchmark datasets. It is compared with label-wise and standard approaches for disentangling aleatoric and epistemic uncertainty. The results demonstrate that the ordinal approach outperforms other methods in error detection, achieving a higher exact hit rate (misclassification rate) and better error spread (mean absolute error). Additionally, the paper reveals that the common inductive bias in ordinal classification, where predictive probabilities are constrained to be unimodal by loss functions like QWK or squared EMD, negatively affects uncertainty quantification. This compression of predictive distributions reduces the information they convey, thereby hindering the accurate quantification of uncertainty. In contrast, the CE loss produces unbiased, truthful predictive probabilities in expectation. Given that ordinal classification has primarily focused on predictions rather than uncertainty quantification, this finding is both novel and significant. It encourages a re-evaluation of current practices, especially for high-risk applications in medicine and finance, where accurate uncertainty quantification is crucial. This work highlights the importance of future research into alternative loss functions that preserve the informativeness of predictive distributions.

Tab. 4.1: Mapping of thesis contributions to components of the conceptual framework (Figure 3.4). BM (Bias Mitigation), UQ (Uncertainty Quantification), SC (Selective Classification), XAI (Explainable Artificial Intelligence), OTA (Ordinal Target Awareness), HITL (Human-in-the-Loop).

No.	Publication	BM	UQ	SC	XAI	OTA	HITL
1	A Prescriptive ML Approach [SH22]	✓	-	-	-	✓	✓
2	Conformalized prescriptive ML [SH24]	-	✓	✓	-	✓	✓
3	Stakeholder-centric explns. [Ste+24a]	-	-	-	✓	-	✓
4	Rectifying Bias in Ordinal Data [SH23]	✓	-	-	-	✓	✓
5	Uncertainty in Ordinal Classif. [SH25]	-	✓	✓	-	✓	-
6	AU and EU for Ordinal Classif. [SH26]	-	✓	-	-	✓	-

4.1 A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain

Contributing Article

Stefan Haas and Eyke Hüllermeier. "A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part VI*. vol. 13718. Lecture Notes in Computer Science. Springer, 2022, pp. 170–184

Author Contribution Statement

The initial draft of the paper was written by the author, employing standard supervised machine learning techniques to replicate the goodwill claim decisions made by experts. Prof. Dr. Hüllermeier contributed his expertise in ordinal, hierarchical, and cost-sensitive learning, which facilitated the development of a custom prescriptive model tailored to the specific challenges of automotive goodwill claim assessment, including strong data imbalance, an ordinal target structure, and varying costs associated with errors. The author subsequently implemented the model and conducted all experiments. The entire paper underwent repeated revisions by both Prof. Dr. Hüllermeier and the author.

A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain

Stefan Haas[1]( ) and Eyke Hüllermeier[2]( )

[1] BMW Group, Munich, Germany
stefan.sh.haas@bmwgroup.com
[2] Institute of Informatics, University of Munich (LMU), Munich, Germany
eyke@lmu.de

Abstract. Car manufacturers receive thousands of goodwill requests for vehicle defects per year. At BMW, these requests for repair-cost contributions are either assessed automatically by a set of fixed rules or manually by human experts. To decrease manual effort, which is still around 50%, we propose a machine learning approach with the goal to discover so far unknown assessment patterns in human decisions. Since the assessment contribution data is heavily imbalanced, we structure the learning task hierarchically: The first layer's task is to predict the main rank of the request (no contribution, partial contribution, or full contribution). Then, in the case where partial contribution is suggested, the second layer predicts the concrete percentage using a regression model. To optimize our model and tailor it to certain strategies (e.g., customer friendly or more cost oriented), we make use of a custom-defined cost matrix. We also outline how the model can be used in a scenario in which it prescribes appropriate monetary contributions for requested repair-costs. This can initially happen in the form of a decision support system (DSS) and, in the next step, through automated decision making (ADM), where a certain part of goodwill requests is processed automatically by the prescriptive model.

Keywords: Prescriptive machine learning · Decision support systems · Automated decision making · Cost-sensitive learning · Hierarchical learning

1 Introduction

Rule-based expert systems are used widely in many fields, for example in industry to assess financial credit risks or in medicine to detect diseases such as breast cancer or diabetes [1,8]. They arguably constitute the simplest form of artificial intelligence (AI), storing rules carefully assembled by domain-knowledge in the form of if-then-else statements. They do not require any data and are *naturally interpretable* [2]. This makes them a natural fit for automating decision processes that need to be auditable, 100% accurate, and which comprise a certain risk, either financially or for life and limb.

M.-R. Amini et al. (Eds.): ECML PKDD 2022, LNAI 13718, pp. 170–184, 2023.
https://doi.org/10.1007/978-3-031-26422-1_11

One such financial rule-based expert system is the central Goodwill system of BMW. In cases of vehicle defects, dealers carry out goodwill repair on behalf of customers and in turn get compensated by the original equipment manufacturer (OEM) for their spare parts and labor efforts. Whether or not customers are eligible for goodwill compensation is decided automatically on the basis of a fixed set of expert rules. This automatic rule based assessment is only done in countries where no legal restrictions against it apply. In case the goodwill request is rejected in the first place, the final decision is transferred to a so-called *assessor*, a human after-sales goodwill expert, who manually looks at the individual case and determines the monetary contribution of the OEM, if any. Although a decision matrix to support this manual process is in place in many sales markets, it is still often a commercial gut decision and not standardized across markets.

The need for human intervention is due to several problems of a rule-based approach, notably the difficulty to maintain a coherent set of deterministic rules capturing all eventualities of a complex commercial use case. Therefore, the data-driven design of decision models by means of machine learning (ML) appears to be an appealing alternative to increase the degree of automation. Over the years, a good amount of historic human decision data has been collected, which can be leveraged in this regard. The goal hereby is to deduce so far unknown assessment patterns from observed human decisions that might be too complex to be put into rules in the first place. Supervised machine learning models can be trained on the observed decision data and later used in the manual decision process to *prescribe* certain monetary contributions. This can either happen in the form of a *decision support system* (DSS) or, if trust in the models is high enough, through *automated decision making* (ADM), which helps decrease manual human assessment effort and save costs in the long run.

The goodwill use case qualifies as what has recently been coined *prescriptive* machine learning [7]. In contrast to the common setting of *predictive* machine learning, the goal is not to predict some underlying ground-truth, but rather to learn models that stipulate appropriate decisions or actions to be taken in order to achieve a certain goal (i.e., to answer the question "How to make something happen?" rather than "What will happen?"). In fact, in the case of goodwill, there is nothing like a "right" monetary contribution. Instead, a decision is more or less appropriate, fair for the customer and strategically opportune for the company. Such decisions are supposed to ensure customer satisfaction while remaining economically reasonable from a manufacturer's perspective. In addition to increasing the degree of automation, prescriptive models may also contribute to the standardization, consistency, and objectivity of the decision process.

The main contribution of this paper is a prescriptive ML approach to goodwill assessment, which is based on real human decision data. In the next section, we describe the goodwill assessment problem in more detail. Next, we outline how prescriptive ML could be incorporated into the existing process. Then, we propose an ML method for goodwill assessments, which is specifically tailored to the use case and properties of the data. Finally, we conclude with related work, identify challenges and outline directions of future work.

2 The Vehicle Goodwill Assessment Process

Assessing goodwill requests is an important topic for manufacturers. In case of BMW, dealers yearly submit thousands of goodwill requests for vehicles that must be assessed. The question whether goodwill is granted or not, and which amount, is far from trivial. It is an individual *commercial decision* that must balance customer satisfaction and financial impact. In this regard, it is important to distinguish between *warranty*, which is a legal obligation for manufacturers, and *goodwill*, which is a non-obligatory service manufacturers provide to customers outside the *warranty* time window (usually after 3–5 years). The goal of compensating customers for product failures outside the *warranty* time window is primarily to safeguard customer satisfaction and loyalty with the brand.

At the OEM, handling goodwill on system level is currently a hybrid approach based on automatic and human manual assessment. The UML Use-Case diagram in Fig. 1 depicts the process and its actors.

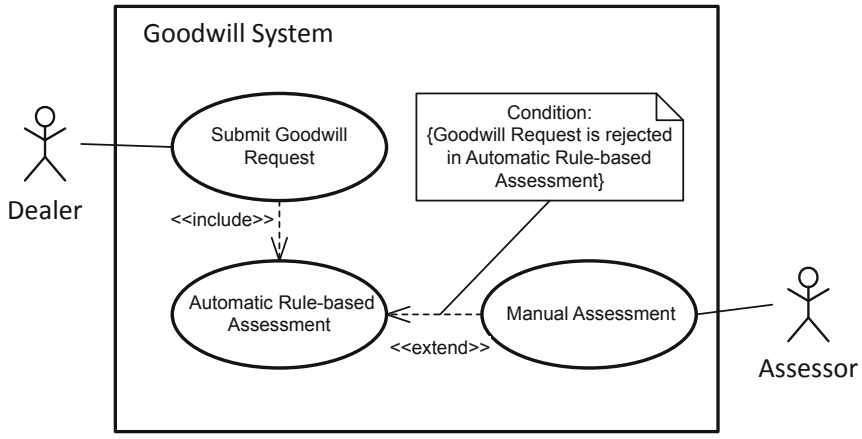


Fig. 1. UML use case diagram for the classic goodwill process.

The standard use case is as follows. Customers arrive at a dealership with a vehicle defect and request a repair from the dealer. Next, the dealer checks whether the manufacturer would grant goodwill for this particular defect by submitting a goodwill request on the behalf of the customer. The data the dealer has to enter ranges from certain vehicle information like vehicle mileage and age to estimated labor and parts costs for the repair itself. On system side, the request is first evaluated against a fixed set of rules (automatic rule-based assessment). If it goes through and goodwill is granted, the process is finished and the dealer will be compensated for the repair. If not, the goodwill request is further processed through a *manual assessment*. In this case, a human goodwill after-sales expert checks the request and makes the final decision. The manual assessment step only extends the automatic rule-based assessment in case of an automatic rejection in the first place but cannot be requested right from the beginning. In case of a manual assessment, the dealer also has the possibility

to send attachments (e.g., a video of rattling engine) and a free text comment along with the request.

In tangible terms, the result of the goodwill process is a percentage of the labor and parts cost contributions the dealer requests and the manufacturer is willing to pay. The set of possible contribution percentages ranges from 0 to 100% in steps of 10%: $C = \{0, 10, 20, \ldots, 100\}$. For instance, if the dealer has labor and parts costs of €1,149.82 and €903.30, respectively, and requests labor and parts cost contributions of 100%, the assessor decides which percentage of contribution is appropriate by taking all the provided information into account. He or she might first check the mileage and age of the vehicle, then the respective defect, whether the vehicle was regularly serviced, and so on. Based on these checks, he or she decides for a contribution, e.g., 50% for labor and 100% for parts. In our example, this would lead to a monetary compensation of the dealer of €574.91 for labor and €903.30 for parts.

To get an idea about the dimensions of automatic vs. manual goodwill assessments, Fig. 2 shows the overall proportion of automatic and manual goodwill assessments of some selected sales markets.

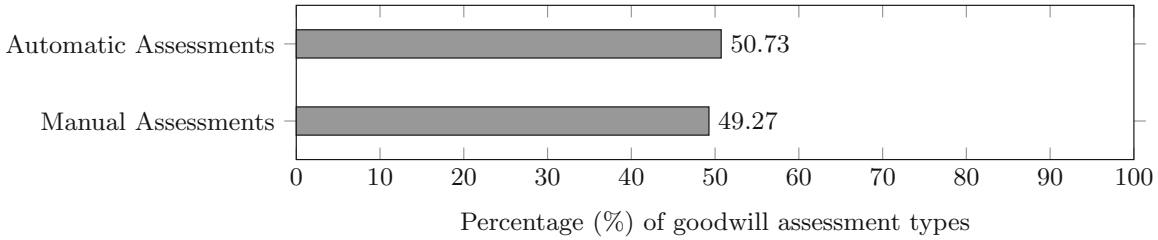


Fig. 2. Overall portions (%) of manual and automatic assessments.

Note that the period of data selection is veiled to allow no conclusions. The portion of goodwill requests that need to be assessed manually is almost as high (49.27%) as the portion of automatically processed goodwill requests (50.73%). In total numbers, 688,879 goodwill requests have been created so far, 349,488 of which were processed automatically by rules and 339,391 manually by a human expert.

Table 1 breaks down the goodwill numbers per selected National Sales Company (NSC). The NSC names have been anonymized here by letters (A to E), to prevent conclusions about goodwill strategies per country. The size of the sales market naturally influences the number of goodwill cases. From an assessment perspective it makes sense to look at the goodwill cases on a per sales market basis, since sales markets have their own goodwill strategies. Therefore, goodwill compensations is very market specific.

Table 1. Goodwill assessment numbers by National Sales Company (NSC).

NSC	Goodwill requests	Automatic	Manual	Degree of automation
A	35,624	20,998	14,626	58.94
B	76,461	48,666	27,795	63.65
C	84,030	47,278	36,752	56.26
D	437,656	200,831	236,825	45.89
E	55,108	31,715	23,393	57.55
$\sum$	688,879	349,488	339,391	⌀ 50.73 %

3 Prescriptive Machine Learning for Goodwill Assessment

In this section, we propose to extend the standard goodwill assessment process as outlined in the previous section, with prescriptive ML models. First, we describe how ML models could be integrated into the existing goodwill use case. Subsequently, we evaluate how well a complex human decision process such as goodwill assessment can be covered by supervised ML.

3.1 Enhancing the Goodwill Assessment Process

Figure 3 shows a goodwill use case extended by ML in comparison with the classic use case outlined in Fig. 1. The *prescriptive model assessment* can either be included in the *manual assessment* process or extend the *automatic rule-based assessment*.

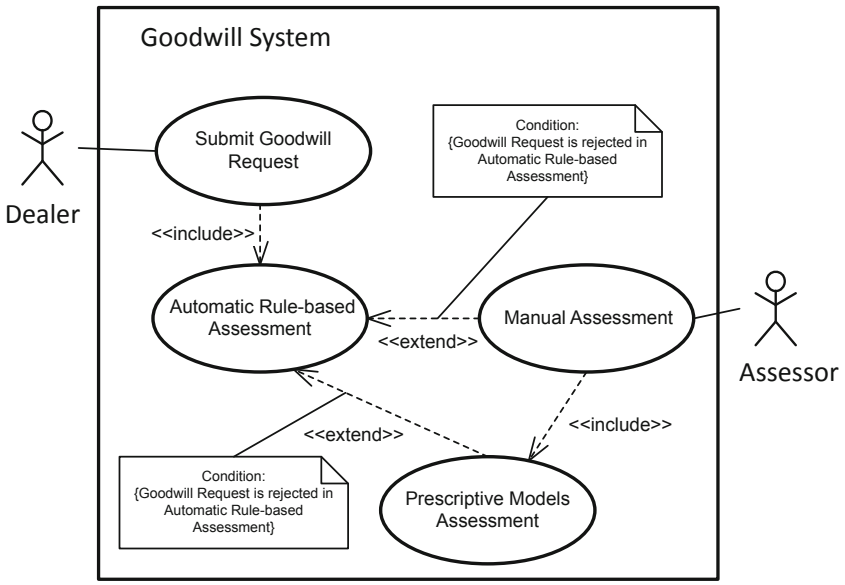


Fig. 3. UML use case diagram for the ML-enhanced goodwill process.

In the inclusion scenario, the prescriptive model supports the manual assessment through goodwill contribution suggestions that guide the assessor in his or her decision process. The prescriptive model serves as a *decision support system* (DSS) and only informs the assessor about the presumably most appropriate decision. Accepting the decision is not compulsory for the assessor, who still possesses the sovereignty over the goodwill decision. Nevertheless, the model suggestions could help to harmonize and standardize decisions from a business perspective. Including the prescriptive model assessment in the manual assessment might be a good starting point for making use of ML in the goodwill process, as the risk of wrong assessments is low and the final decision is still in the hands of an expert.

In the extension scenario, the model extends the automatic rule-based assessment and takes over cases not decidable by rules. The model assesses goodwill decisions automatically and supports the process through *automated decision making* (ADM). From a business perspective, this is the ultimate goal to aim for, as it will directly reduce process costs. However, this approach also comes with the greatest risk, as there is no human expert involved anymore who supervises the final decisions. Customer satisfaction and financial impact for the manufacturer are left to the machine. Leaving the final goodwill decision to a prescriptive model requires trust that can only be built through an evaluation by business experts over a long term period.

A combination of inclusion and extension is also conceivable. While ADM might be feasible in less complex cases, it might be advisable to just integrate the model as a DSS in more complex scenarios, leaving the final decision to a human expert. What exactly distinguishes less and more complex goodwill scenarios is still an open research question.

3.2 Prescriptive Machine Learning

The setting of prescriptive ML deviates from the standard setting of predictive ML in various ways [7]. This also includes the process of supervision. As already mentioned, in prescriptive ML, there is not necessarily something like a "ground-truth" or correct decision, and even if decisions might be compared in terms of quality or desirability of their implications, there is no guarantee that decisions made by human experts in the past were optimal. Therefore, taking them directly as targets for a supervised learning method might not be advisable [11]. In the case of goodwill, for example, a decision of 50% contribution appears to be somewhat overrepresented (cf. Fig. 4), letting one suspect that this is often taken as a default choice for a partial cost coverage, even if it might not necessarily be the most appropriate percentage. In the following, we will nevertheless assume that mimicking the expert is a reasonable strategy, at least as a first step toward a data-driven goodwill assessment, leaving more elaborate approaches for future work.

Under this premise, the problem is essentially reduced to a supervised learning task, with the observed human goodwill decisions

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

as training data. Instances are goodwill requests entered by the dealer and represented as a *feature vector* $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$. These instances are labeled by assessed contribution percentages, which serve as the target variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. The goal of the ML task is to learn a decision model $h^* \in \mathcal{H}$, where $\mathcal{H}$ is the class of candidate models (referred to as hypothesis space in the common setting of supervised learning). This model is a mapping $\mathcal{X} \to \mathcal{Y}$ supposed to approximate the training data and, more importantly, generalize well to new decision problems. Like in supervised learning, we model the performance of a model h in terms of a loss (error) function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, so that $l(y, \hat{y})$ denotes the penalty incurred by the learner for prescribing $\hat{y}$ when the expert decides y. The choice of a presumably optimal model h^* is commonly guided by the empirical risk

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

as an estimate of a model's performance. This measure is normally not minimized directly by the learner, however, because the empirical risk minimizer $h^* = \arg\min_{h \in \mathcal{H}} R(h)$ is knowingly prone to overfitting the training data, and hence to suboptimal generalization.

3.3 Human Goodwill Decision Data

Table 2 shows the features used for the ML task. In the first step, we will only look at the *hard facts*, such as vehicle mileage, vehicle age, the defect code, the costs, and the requested labor and part contributions. The raw data entered by the dealer will be enriched with further vehicle data that can be derived from the vehicle identification number (VIN), including the vehicle model type, the series, the motor series, the order country of the vehicle, the sales country of the vehicle, and whether the vehicle is a car or motorbike. The free-text dealer comment and attachments will be ignored for now, because they can be considered as "soft" facts. Besides, they are not immediately usable and require sophisticated post-processing techniques such as NLP. The rest of the data is a mixture of categorical and numerical data and qualifies as tabular data.

The features are pre-processed as follows: Numeric data is scaled using *min-max-scaling* (e.g., Parts, Labor and Total Costs), low cardinality categorical features are encoded using *one-hot-encoding* (e.g., Customer Type or Requested Labor and Parts Contributions), and high cardinality features are *hashed* (e.g., Defect Code or Vehicle Series).

Turning our attention to the target variable, Fig. 4 shows how the overall contributions are distributed over the possible percentages $\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$. Obviously, the data is heavily imbalanced, and contributions other than 0% and 100% are rarely used. Among the rare contributions, the 50% decision sticks out and appears a bit more frequently, whereas 90% is the least frequent contribution. As already said, this may reflect a common human pattern: If not being exactly sure what to grant, people tend to opt for a compromise in the middle. Another pattern one can observe is a kind of "generous rounding" to

Table 2. Features used for model training.

Attribute	Data type	Description
Vehicle Mileage	Numeric (continuous)	12,500
Vehicle Age	Numeric (continuous)	48
Enquiry Indicator	Categorical (ordinal)	Request after or before the repair
Warranty Stage	Categorical (nominal)	Standard or Extended Goodwill
Product Type	Categorical (nominal)	Car or Motorbike
Regular Service	Categorical (nominal)	Yes or No
Sales Country	Categorical (nominal)	NL
Order Country	Categorical (nominal)	BE
External Guarantee	Categorical (nominal)	Yes or No
Vehicle registered to customer	Categorical (nominal)	Yes or No
Vehicle Model Type	Categorical (nominal)	FG81
Vehicle Series	Categorical (nominal)	G21
Motor Series	Categorical (nominal)	N57T
Mobility provided	Categorical (nominal)	Yes or No
Defect Code	Categorical (nominal)	1178031500
Defect Code (Main and sub group only)	Categorical (nominal)	1178
Shared last expenses	Categorical (nominal)	Yes or No
Customer Type	Categorical (nominal)	Regular, Transit or International
Requested Labor Contribution (per cent)	Categorical (nominal)	60%
Requested Parts Contribution (per cent)	Categorical (nominal)	60%
Dealer Labor Contribution (per cent)	Categorical (nominal)	40%
Dealer Parts Contribution (per cent)	Categorical (nominal)	40%
Parts Costs	Numeric (continuous)	€903.30
Labor Costs	Numeric (continuous)	€1,149.82
Requested Open Time Units	Numeric (discrete)	5
Dealer Open Time Units	Numeric (discrete)	2
Additional service costs, e.g., replacement car	Numeric (continuous)	€460.30
Total Costs	Numeric (continuous)	€3,682.89

"meaningful" contributions, namely, 0%, 30%, 50%, 70%, 100%. Other contributions, such as 10% and 90%, are even more rare, probably because these are considered somewhat pedantic. In any case, the rare contributions are likely to carry important information, as they reflect subtle human instinct, and they are key to safeguard customer satisfaction. There is also an apparent tendency to contribute rather than not contribute from manufacturer's perspective, as the 100% bar is noticeably higher than the 0% bar. This is the case for labor as well as parts. However, for parts the tendency is stronger than for labor.

3.4 Hierarchical Cost-Sensitive Learning

From the description of the task and the data, it becomes clear that goodwill assessment comes with a number of important challenges from a machine learning perspective. First, looking at the scale of the target variable (contribution in percentage), the problem is somehow in-between ordinal classification and

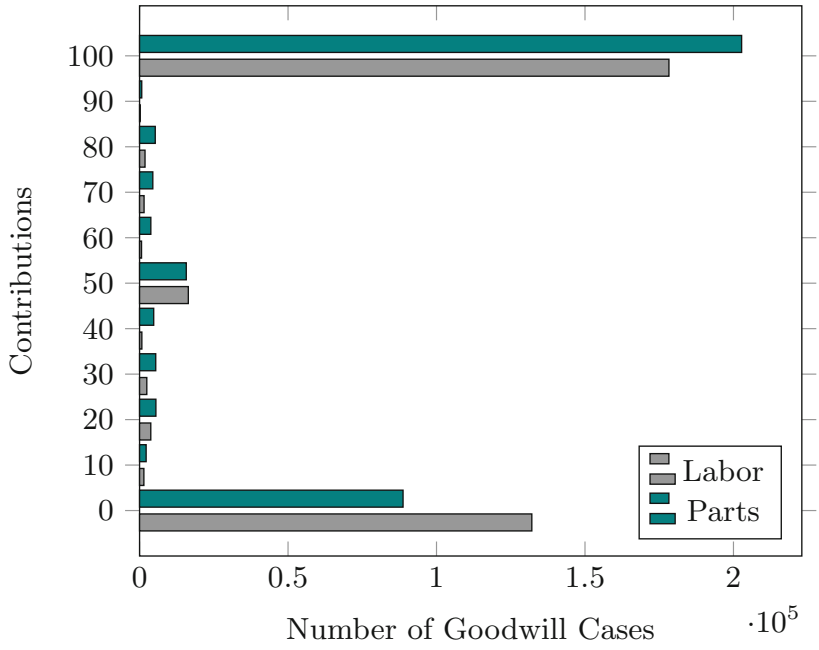


Fig. 4. Distribution of goodwill contributions for Labor and Parts at BMW.

regression: In principle, the target is numerical, but not all numbers between 0 and 100 are deemed valid prescriptions. Therefore, one may also think of tackling the task as a problem of ordinal classification with 11 class labels sorted in increasing order from lowest (0%) to highest (100%).

Related to the interpretation of the scale is the question of how a suitable loss function should look like. Obviously, a standard measure such as misclassification rate (0/1 loss) is inappropriate, even if the task is treated as a classification problem, because the loss function should take the linear structure of the contribution scale into account. Squared or absolute error as commonly used in regression do not appear to be perfect choices either, as one may argue that there is not only a quantitative but also a *qualitative* difference between the 0% decision, the 100% decision, and the decision of a partial contribution. This suggests a cost-sensitive approach, in which a cost (loss) function $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is explicitly defined in "tabular" form. As an additional advantage, this allows for incentivising the learner in a strategic way, e.g., to constructing more customer-friendly or more cost-oriented decision models.

Another challenge is the class imbalance. Imbalanced data makes learning more difficult, and many algorithms have a tendency to compromise the accuracy of small classes in favor of bigger classes [12]. This would be especially problematic in the case of goodwill assessment, enforcing extreme decisions at the cost of partial contributions. Common approaches to deal with imbalanced data include up-sampling of the minority classes or down-sampling of the predominant classes in order to balance the data [13]. Similar effects can be achieved by adding weights to the training examples, making the underrepresented examples more important and the overrepresented less.

To tackle both problems, cost-sensitivity and imbalance, we propose a hierarchical approach with a qualitative (categorical) first layer and a quantitative second layer. In the first layer, we solve an ordinal 3-class classification (or ranking) problem, distinguishing between classes NO (no contribution, rank 1), PARTIAL (partial contribution, rank 2), and FULL (full contribution, rank 3). Obviously, this problem is more balanced, because all contributions between 10% and 90% are collected in a single class.

In the case where an instance is assigned to PARTIAL in the first layer, it is forwarded to the second layer, where the concrete percentage of contribution is determined. Thus, while an instance $\boldsymbol{x}$ is mapped to a rank $r(\boldsymbol{x}) \in \{1, 2, 3\}$ in the first layer, $\boldsymbol{x}$ is mapped to any of the numbers $\{10, 20, \ldots, 90\}$ in the second layer. The latter task can be formalized as a (constrained) regression problem.

The first problem, where an example $(\boldsymbol{x}, y)$ consists of an input vector $\boldsymbol{x} \in \mathcal{X}$ and an ordinal label $y \in \mathcal{Y} = \{1, 2, ..., K\}$ (in our case $\{\text{NO}, \text{PARTIAL}, \text{FULL}\}$, i.e., $K = 3$), provides us with the opportunity to use the cost-sensitive ranking framework presented in [9]. This framework allows one to specify a *cost matrix* in a flexible way, which is especially convenient in our case. In fact, by utilizing a custom defined $K \times K$ cost matrix $\mathcal{C}$, we can configure the mislabeling cost according to our strategy, e.g., rather customer-friendly or more cost-oriented from manufacturer's perspective. The cost of predicting an example $(\boldsymbol{x}, y)$ as rank k is given by the entry $\mathcal{C}_{y,k}$ in the cost matrix. Table 3 shows two distinct strategies for goodwill assessments. The cost matrix on the left side shows a customer-friendly strategy, where the learner is strongly penalized when prescribing NO instead of FULL ($\mathcal{C}_{3,1} = 30$). On the right side, the cost matrix implements a more cost-orientated approach, where the learner is penalized the most for the decision FULL instead of NO ($\mathcal{C}_{1,3} = 30$). Note that the result of the regression model for the PARTIAL values ($k = 2$) will be mapped back to the interval $\mathcal{C}_{2,2} = [0, 5]$ to also integrate the regression into the overall cost-sensitive ranking framework. By the width of the interval, we can configure how much importance we give to the exact prediction of the values of the regression layer. Figure 5 visualizes the structure of the proposed hierarchical approach.

Table 3. Different assessment strategies specified by different cost functions: customer-oriented with higher penalization of contributions that are loo low (left) vs. manufacturer-oriented with higher penalization of contributions that are too high (right).

		Prescribed		
		NO	PARTIAL	FULL
Actual	NO	0	5	10
	PARTIAL	10	[0,5]	5
	FULL	30	10	0

		Prescribed		
		NO	PARTIAL	FULL
Actual	NO	0	10	30
	PARTIAL	5	[0,5]	10
	FULL	10	5	0

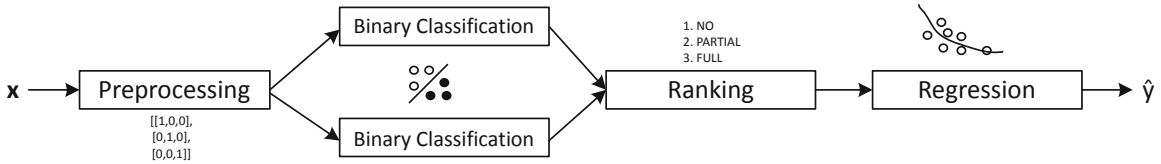


Fig. 5. Overview of the hierarchical cost-sensitive approach.

The approach [9] to ordinal classification is based on a reduction to weighted binary classification. More specifically, a binary classifier

$$f : \mathcal{X} \times \{1, \ldots, K-1\} \to \{0, 1\}$$

is trained that accepts *extended* instances $(\boldsymbol{x}, k)$ as input. As output, the classifier is supposed to produce 1 (answer "yes") if the true rank of $\boldsymbol{x}$ exceeds k and 0 (answer "no") otherwise. The actual rank of a query instance can then be determined by applying the following ranking rule:

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

To train the classifier, the original data is extended as follows: Each original example $(\boldsymbol{x}, y)$ is turned into extended examples $(\boldsymbol{x}^k, y^k)$ with weights $w_{y,k}$, where[1]

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

The weights $w_{y,k}$ control the importance of an example during the training phase of the binary classifier. The higher the cost difference between two adjacent ranks, the larger the weights and therefore the importance of a particular example.

Incorporating domain knowledge, we propose the following small modification of the ranking rule (2): As the proposed contribution essentially never exceeds the contribution q requested for $\boldsymbol{x}$, we set

$$r(\boldsymbol{x}) = \min \big\{ 1 + f(\boldsymbol{x}, 1) + f(\boldsymbol{x}, 2), q \big\}. \tag{3}$$

For the second layer of our model, any regression method can in principle be used. For the exact inference of the partial contribution values, we round and constrain the regression model's output to the set of possible contributions $\{10, \ldots, 90\}$. Also, like for the prescription of ranks, we make sure that the prescription does not exceed the requested contribution q:

$$\hat{y} = \min \left\{ \lfloor \frac{f(\boldsymbol{x})}{10} \rceil \cdot 10, q \right\} \tag{4}$$

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

4 Evaluation and Results

In this section, we evaluate our hierarchical cost-sensitive approach on BMW's goodwill data sets. For training the classifier f (and ranker r) in the first layer, a learning algorithm is needed that is able to handle weighted examples. In our experimental study, we used extreme gradient boosting (XGBoost) [3], a versatile method that proved to work very well on tabular data and also outperforms deep neural networks in this context [10]. Another advantage is that XGBoost can be used for both classification and regression, hence we could use it for training the first as well as the second layer of our model.

Tables 4 and 5 show the results of a ten-fold cross validation in terms of the mean and standard deviation of various performance metrics. The first metric of interest is the cost of the model's prescriptions according to the underlying cost function — here, we present results for the cost matrix (a) in Table 3 (those for matrix (b) look very similar). The middle part of the matrix, i.e., the cost for assessments involving a partial contribution, is filled with the absolute error of the regression model scaled to the specified interval (in this case $[0, 5]$). As the cost values are measured on an abstract scale without interpretable dimension, we also report the mean accuracy (ACC) for the ranking part and the mean absolute error (MAE) for the regression model (on a scale from 10 to 90), thereby making the results more tangible. Overall, our model shows a quite satisfactory performance.

Table 4. Evaluation metric results obtained for Labor.

	Ranking		Regression		Costs	
NSC	ACC	SD	MAE	SD	C	SD
A	0.887	0.032	0.942	0.24	1.133	0.303
B	0.904	0.014	5.094	0.524	1.018	0.221
C	0.926	0.028	4.519	0.454	0.725	0.271
D	0.857	0.009	1.306	0.19	1.321	0.09
E	0.881	0.047	7.161	1.755	1.064	0.398
Mean	**0.891**	0.026	**3.8044**	0.6326	**1.0522**	0.2566
Median	**0.887**	0.028	**4.519**	0.454	**1.064**	0.271

As already explained, the cost function can be used to tailor a decision model to certain strategies, e.g., making it more customer-friendly or more manufacturer-friendly (cost-oriented). To evaluate this feature, we looked at the confusion matrices obtained for the cost functions in Table 3. As can be seen in Table 6, the confusion matrix for the customer-friendly cost matrix is indeed more geared to the right, showing a tendency toward higher ranks and consequently higher contributions. In contrast, the matrix for the cost-oriented strategy is more geared towards the left side, with lower ranks and thus less contributions.

Table 5. Evaluation metric results obtained for Parts.

	Ranking		Regression		Costs	
NSC	ACC	SD	MAE	SD	C	SD
A	0.889	0.035	1.265	0.249	1.059	0.452
B	0.869	0.016	5.691	0.485	1.215	0.158
C	0.949	0.023	6.522	0.711	0.552	0.183
D	0.872	0.011	4.625	0.313	1.154	0.078
E	0.887	0.055	7.041	1.732	1.001	0.51
Mean	**0.8932**	0.028	**5.0288**	0.698	**0.9962**	0.2762
Median	**0.887**	0.023	**5.691**	0.485	**1.059**	0.183

Table 6. Different parts ranking confusion matrix depending on the assessment strategy (for NSC A): customer-oriented (left) vs. manufacturer-oriented (right).

		Prescribed		
		NO	PARTIAL	FULL
Actual	NO	494	47	45
	PARTIAL	0	286	34
	FULL	2	13	541

		Prescribed		
		NO	PARTIAL	FULL
Actual	NO	526	40	20
	PARTIAL	6	295	19
	FULL	11	34	511

5 Conclusion and Future Work

In this paper, we described the existing rule-based and manual goodwill assessment process at BMW and how it can be extended through prescriptive machine learning models. This can either happen in the form of a decision support system, automated decision making, or a combination of both. Furthermore, we proposed a hierarchical, cost-sensitive approach for learning prescriptive models from human goodwill decisions, which accounts for the specific structure of the decision space, counteracts class imbalance, and allows for tailoring strategies to different value systems and market situations (e.g., customer friendly vs. cost oriented).

Motivated by our encouraging results, we plan to address the following challenges in future work.

– *Trust and Explanation*: We noticed that business experts do not immediately trust a prescriptive ML solution. Therefore, involving business experts in the development and evaluation process is important, not only to improve the ML

solution itself, but also to foster trust in it. Explainability will play a key role in this regard, making machine learning more transparent and accessible to all stakeholders involved [5]. In fact, decisions need to be explained, and different parties may have different needs for explanation. For a dealer, feedback about the most important attribute that led to the rejection of the request might be enough, whereas an auditor needs to understand the whole reasoning process in detail.

- *Uncertainty*: Although the decision models we trained perform very well, showing the high potential of automated decision making, not all decisions appear to be perfect all the time. Therefore, it would be desirable to increase the uncertainty-awareness of decision models, so that final decisions could be transferred to the human expert in cases of high uncertainty [6].
- *Weak supervision*: As already mentioned, human goodwill decisions might be biased in one way or the other and should not necessarily be taken as a gold standard. Additionally, the data may contain concept drift due to strategy changes in the assessment process over time. Therefore, past decisions should be considered and modeled as *weak* information about the target rather than an incontestable ground truth, suggesting the use of methods for weakly supervised learning [14] in prescriptive modeling.
- *Fairness:* Another important question concerns the notion of fairness in the goodwill decision process. There might be different strategies toward fairness, depending on the sales market. For instance, some markets might want to treat all customers equally, independently of the money they spent for a vehicle, whereas others might want to prefer customers with higher priced vehicles in the goodwill process. It needs to be investigated whether or not models can be tailored to such strategies automatically, or if a manual intervention is required [4].

References

1. Abu-Naser, S.S., Bastami, B.G.: A proposed rule based system for breasts cancer diagnosis. World Wide J. Multidisc. Res. Dev. **2**(5), 27–33 (2016)
2. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. **70**, 245–317 (2021)
3. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
4. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), pp. 329–338. Association for Computing Machinery, New York, NY, USA (2019)
5. Hong, S.R., Hullman, J., Bertini, E.: Human factors in model interpretability: industry practices, challenges, and needs. Proc. ACM Hum. Comput. Interact. (CSCW1). **4**, 1–26 (2020)
6. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**(3), 457–506 (2021). https://doi.org/10.1007/s10994-021-05946-3

7. Hüllermeier, E.: Prescriptive machine learning for automated decision making: Challenges and opportunities. CoRR abs/2112.08268 (2021). https://arxiv.org/abs/2112.08268
8. Karthikeyan, R., Geetha, P., Ramaraj, E.: Rule based system for better prediction of diabetes. In: 3rd International Conference on Computing and Communications Technologies (ICCCT), pp. 195–203 (2019)
9. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19. MIT Press (2006)
10. Shwartz-Ziv, R., Armon, A.: Tabular data: deep learning is not all you need. Inf. Fusion **81**, 84–90 (2022)
11. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: learning from logged bandit feedback. In: Proceedings of ICML, International Conference on Machine Learning, pp. 814–823 (2015)
12. Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of ICML 24th International Conference on Machine Learning, pp. 935–942. NY, USA, New York (2007)
13. Zhang, N.N., Ye, S.Z., Chien, T.Y.: Imbalanced data classification based on hybrid methods. In: Proceedings of ICBDR 2nd International Conference on Big Data Research, pp. 16–20. Association for Computing Machinery, New York, NY, USA (2018)
14. Zhou, Z.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**, 44–53 (2018)

4.2 Conformalized Prescriptive Machine Learning for Uncertainty-aware Automated Decision Making: The Case of Goodwill Requests

Contributing Article

Stefan Haas and Eyke Hüllermeier. "Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests". In: *International Journal of Data Science and Analytics* (2024)

Author Contribution Statement

The author conceived the idea of using inductive conformal prediction to make the initial model, introduced in the previous publication, uncertainty-aware. The specific approach for integrating conformal prediction with the hierarchical model, as well as the selective classification strategy, was developed and implemented by the author. Additionally, the author conducted all experiments. The entire paper underwent repeated revisions by both Prof. Dr. Hüllermeier and the author.

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

REGULAR PAPER

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

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

Received: 30 January 2023 / Accepted: 21 May 2024

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

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

1 Introduction

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

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

Stefan Haas
stefan.sh.haas@bmwgroup.com

Eyke Hüllermeier
eyke@lmu.de

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

[2] BMW Group, Munich, Germany

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

Published online: 14 June 2024

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

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

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

2 Machine learning for automated decision making

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

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

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

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

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

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

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

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

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

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

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

3 Uncertainty in automated decision making

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

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

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

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

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

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

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

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

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

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

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

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

4 Reliable decision making using conformal prediction

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

4.1 Conformal prediction for uncertainty quantification

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

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

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

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

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

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

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

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

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

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

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

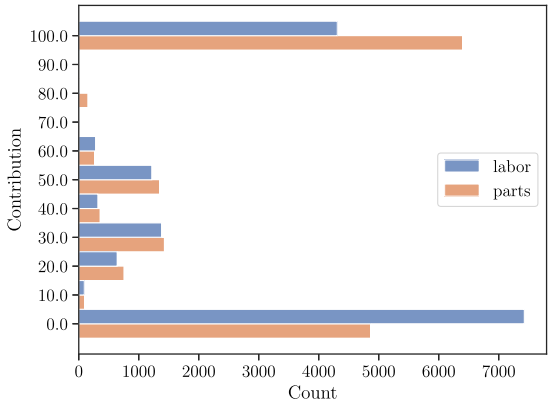
Interestingly, the prediction set can also be empty ($\Gamma^\epsilon(\boldsymbol{x}) = \emptyset$). This happens in cases where a query $\boldsymbol{x}$ cannot be combined into a sufficiently conforming tuple $(\boldsymbol{x}, y)$ with any of the candidates y, e.g., because $\boldsymbol{x}$ itself is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

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

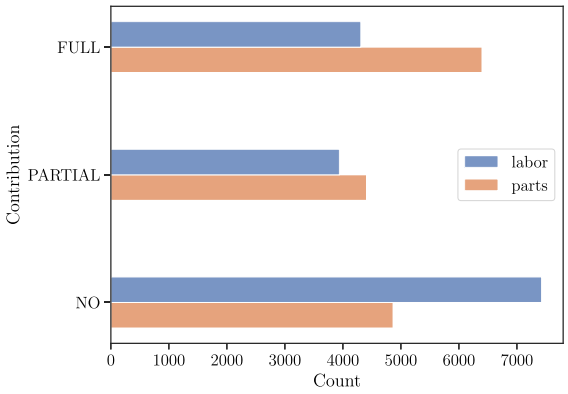
4.2 The hierarchical assessment model

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

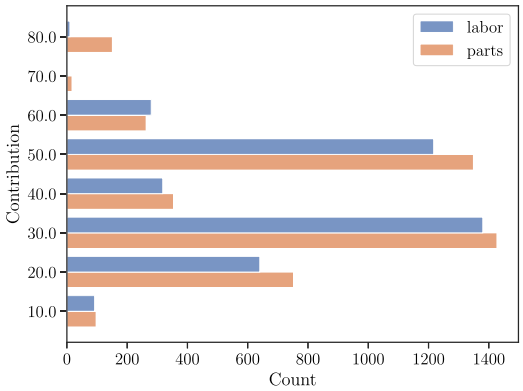
This hierarchical approach to goodwill assessment was chosen because the data is heavily imbalanced, with many 0 and 100% contributions on the one side and fewer, more widely distributed partial contributions on the other side [13]



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



(b) Collapsed main ranks distribution for ranking layer.



(c) Partial contributions distribution for regression layer.

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

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

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

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

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

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

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

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

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

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

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

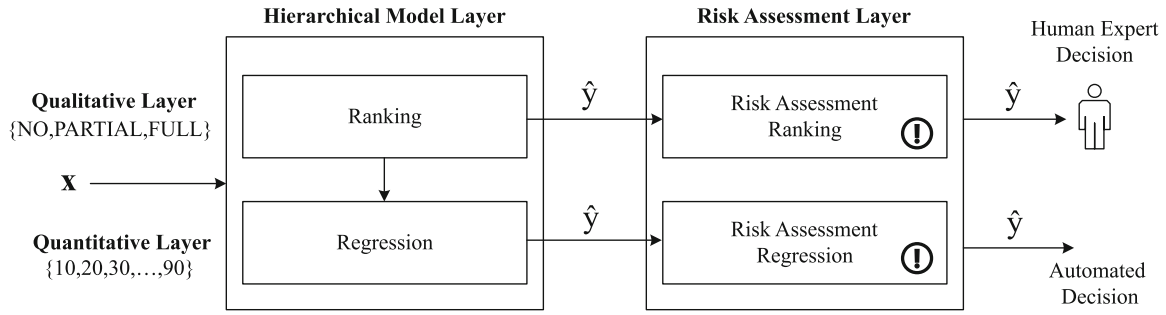


Fig. 2 Overview of uncertainty-aware goodwill assessments with reject option

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

4.3 Conformalizing the hierarchical model

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

4.3.1 Conformalizing the ranking layer

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

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

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

The prediction set for the ranking layer is given by

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

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

4.3.2 Conformalizing the regression layer

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

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

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

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

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

4.4 Risk quantification using conformal prediction

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

4.4.1 Quantifying risk

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

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

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

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

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

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

4.5 Selective uncertainty-aware automated decision making

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

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

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

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

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

4.6 The risk vs. degree of automation trade-off

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

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

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

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

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

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

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

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

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

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

where

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

Moreover, the DoA is defined as follows:

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

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

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

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

Table 1 Characteristics of the goodwill data set

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

5 Evaluation

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

5.1 The goodwill data set

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

5.2 Implementation

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

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

Table 2 Cost matrix for the ranking layer

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

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

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

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

5.3 Evaluation of conformal prediction

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

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

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

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

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

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

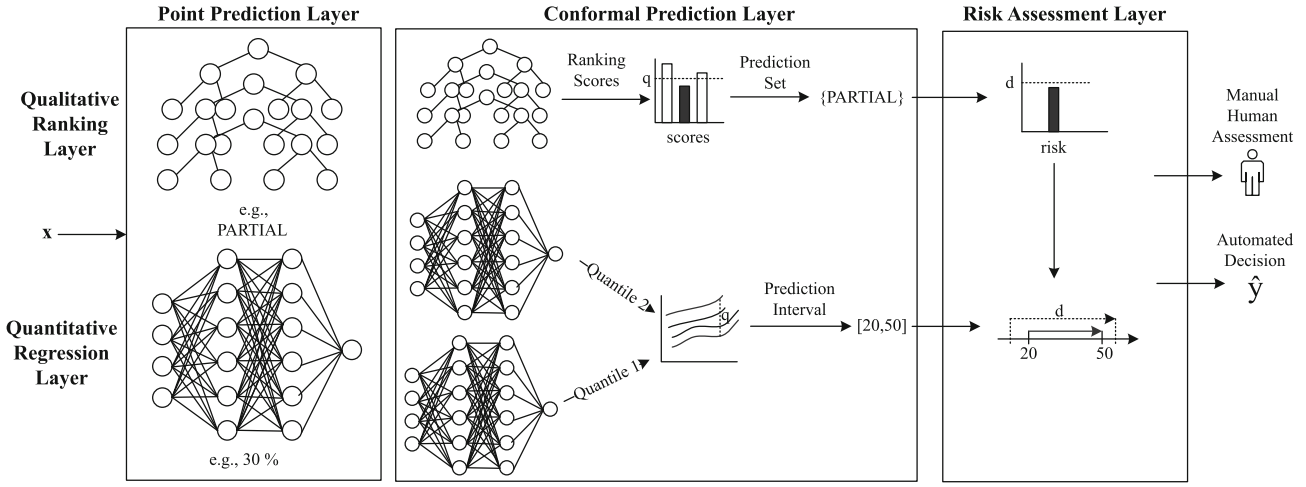
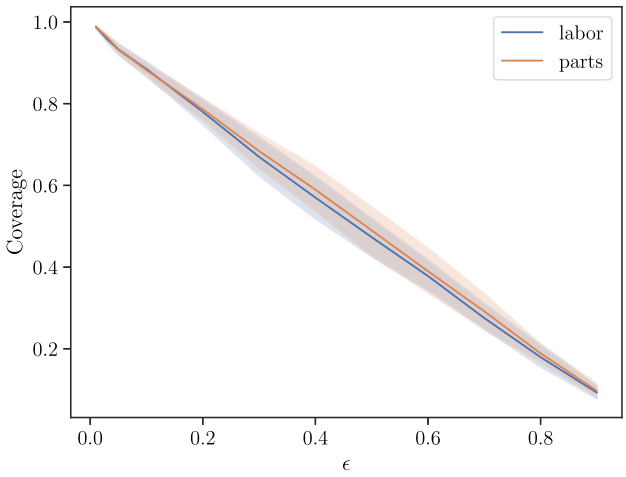
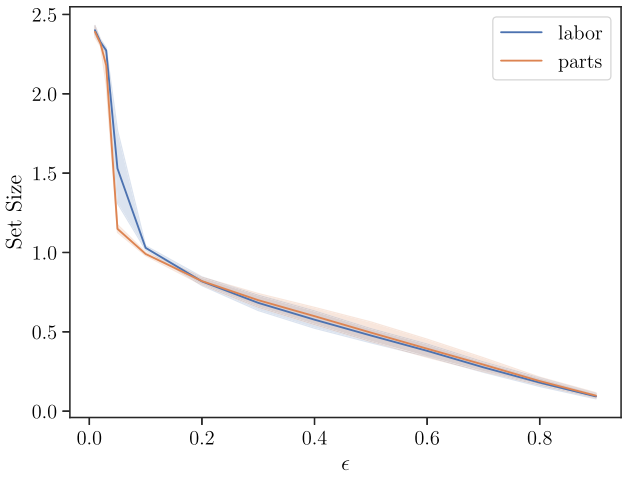


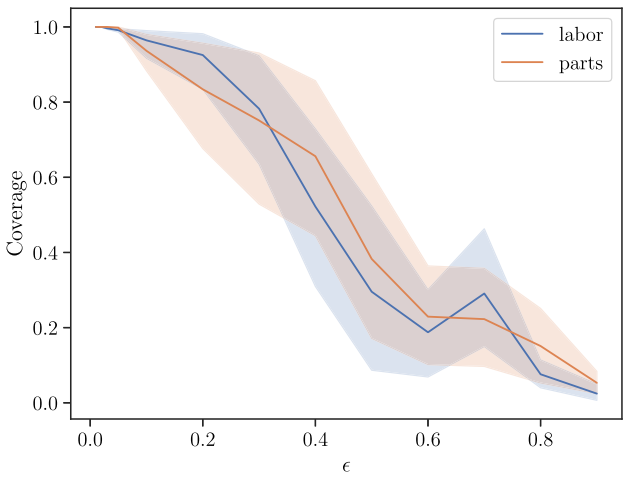
Fig. 3 Overview of the inference architecture



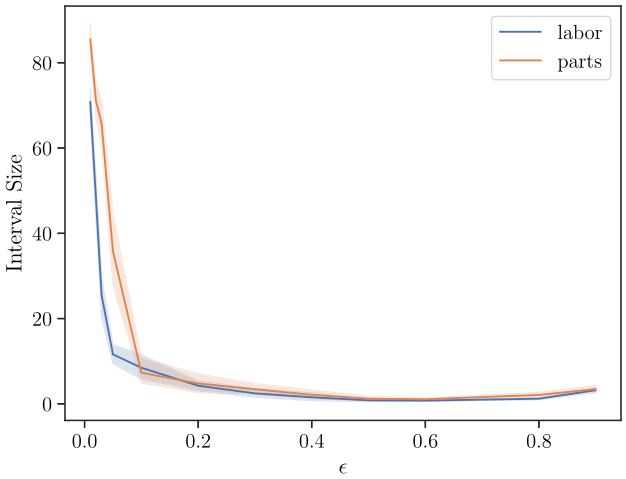
(a) Coverage of conformal predictor for ranking.



(b) Set sizes of conformal predictor for ranking.



(c) Coverage of conformal predictor for regression.



(d) Interval sizes of conformal predictor for regression.

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

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

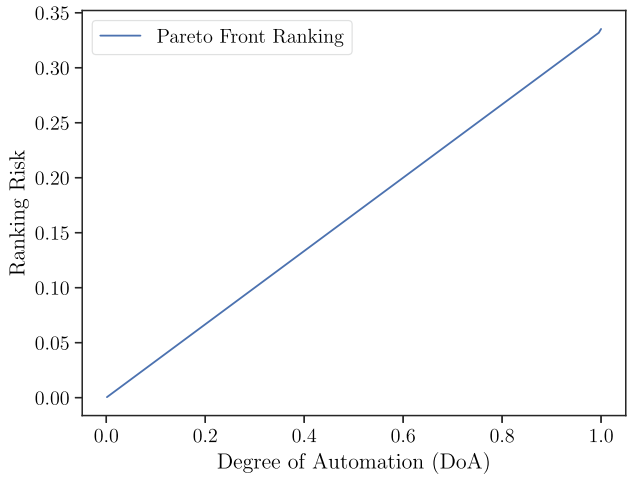
5.4 Evaluation of selective uncertainty-aware Pareto optimization

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

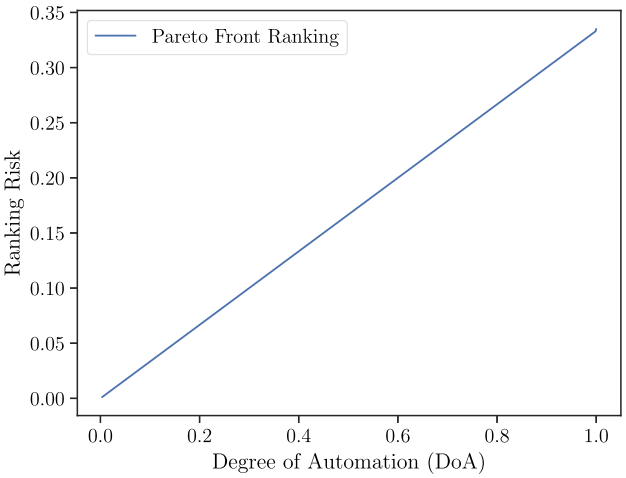
In each random search trial, we train the hierarchical model using the training data set (13,164 examples), then calibrate our conformal predictors with the calibration data set (693 examples) and evaluate our model's conformal and point predictions using the test set (1540 examples). Next, we determine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

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

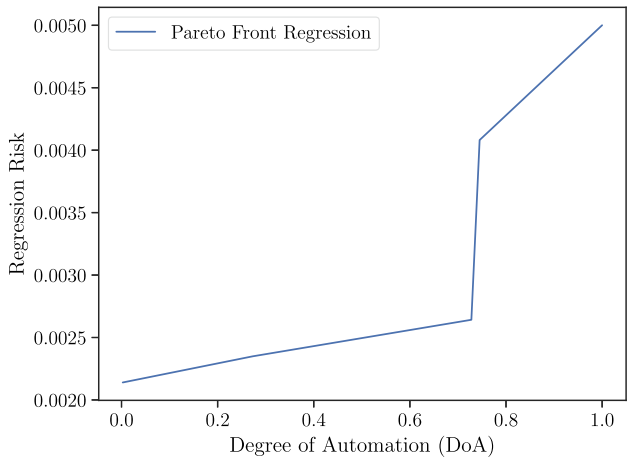
Design space - Random search	
ϵ_{rank}	$\{\epsilon_{\text{rank}} \in \mathbb{R} \,\vert\, 0 \leq \epsilon_{\text{rank}} \leq 1\}$
δ_{rank}	$\{\delta_{\text{rank}} \in \mathbb{R} \,\vert\, 0 \leq \delta_{\text{rank}} \leq 1\}$
ϵ_{reg}	$\{\epsilon_{\text{reg}} \in \mathbb{R} \,\vert\, 0 \leq \epsilon_{\text{reg}}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



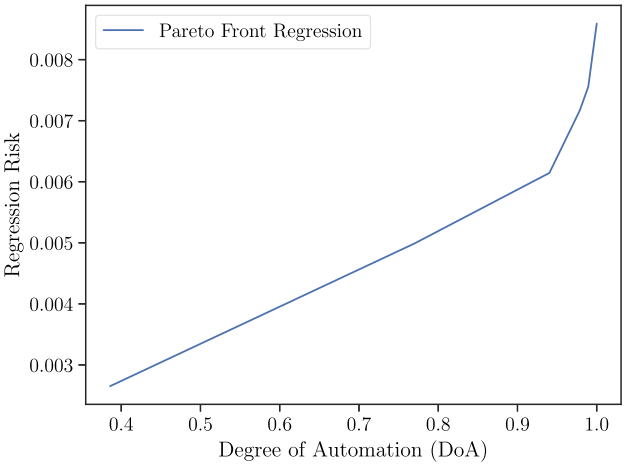
(a) Risk vs. DoA for the Labor Ranking Layer.



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



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

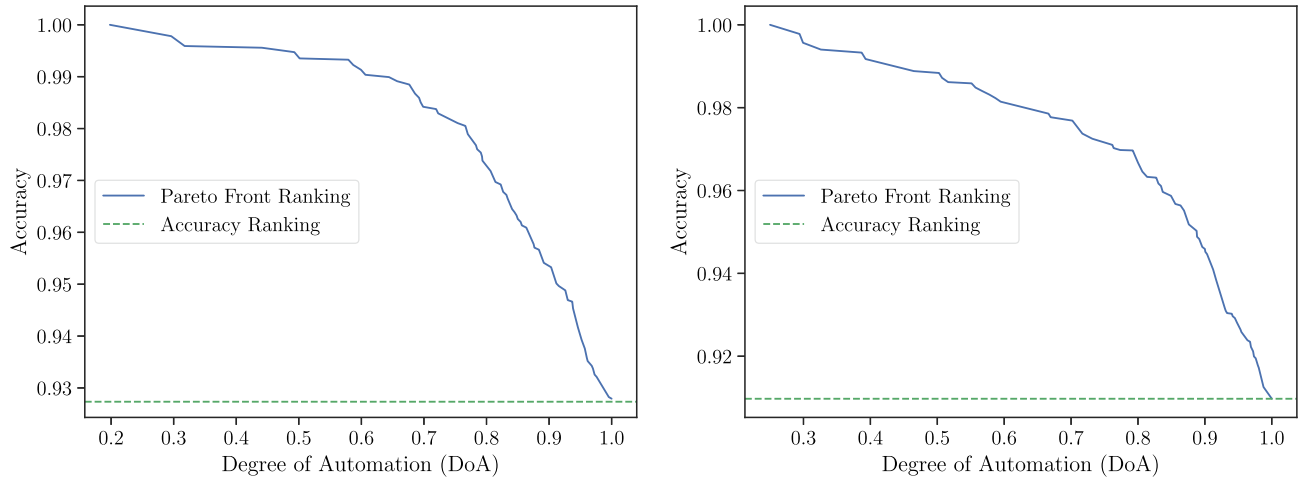


(d) Risk vs. DoA for the Parts Regression Layer.

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

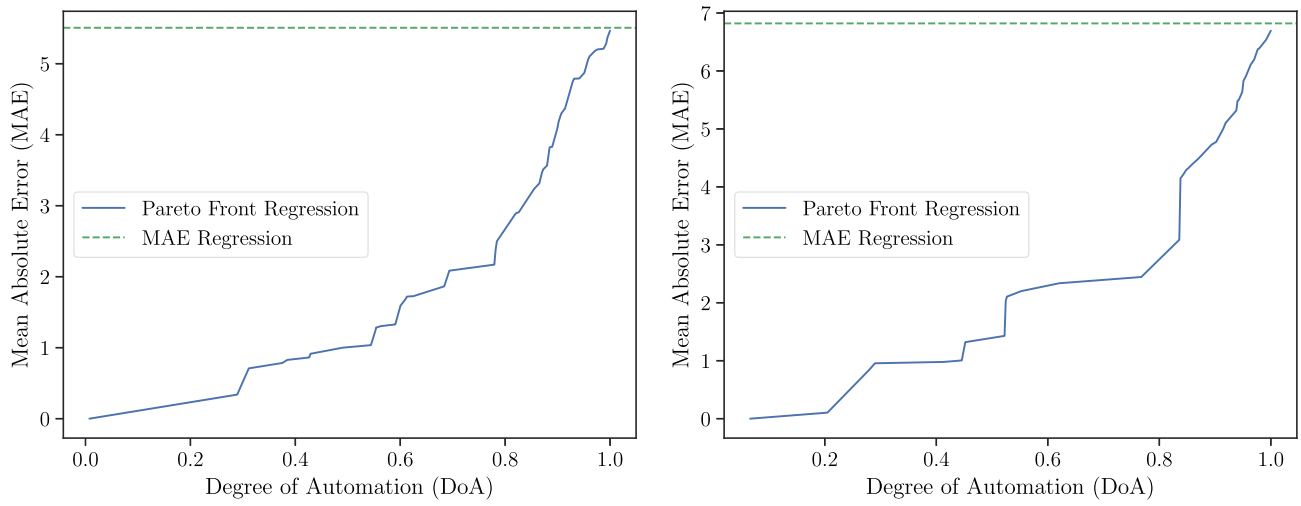
optimization problem. We hereby first look at the risk vs. degree of automation trade-off for ranking and regression in Fig. 5 also known as *risk coverage trade-off*. The degree of automation that is achievable in the ranking layer hereby linearly increases with increasing risk. Requests whose risk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

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



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

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

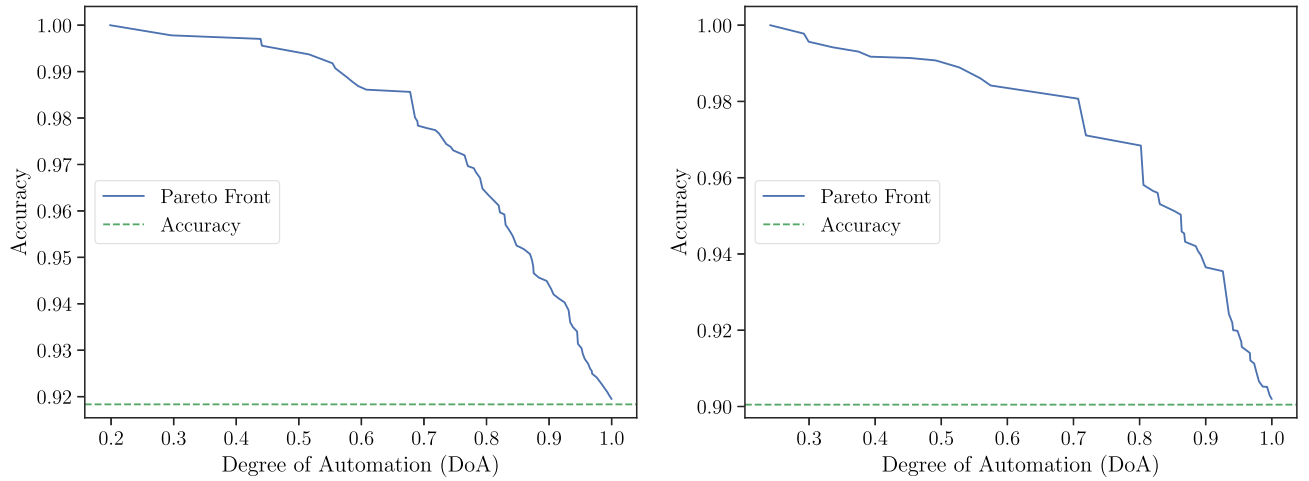


(a) MAE vs. DoA for Labor. **(b)** MAE vs. DoA for Parts.

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

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

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



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

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

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

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

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

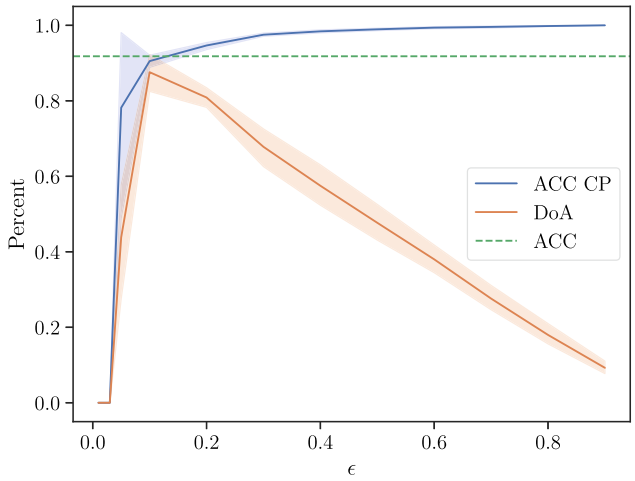
ϵ_{rank}	δ_{rank}	ϵ_{reg}	δ_{reg}	Accuracy (ACC)	Degree of automation (DOA)
0.085349	0.207636	0.728141	0.957674	0.901948	1.000000
0.214033	0.204153	0.338893	0.211275	0.958098	0.805844
0.196397	0.985753	0.783504	0.078822	0.980716	0.707143
0.313255	0.082391	0.619651	0.057391	0.990753	0.491558
0.764676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

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

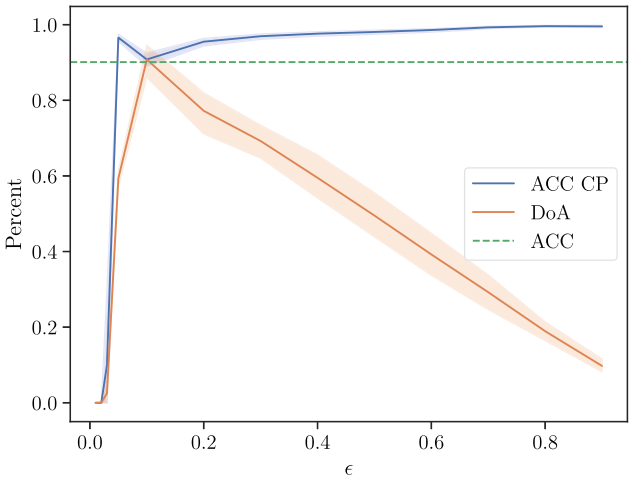
5.5 The effect of the significance level ϵ

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

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

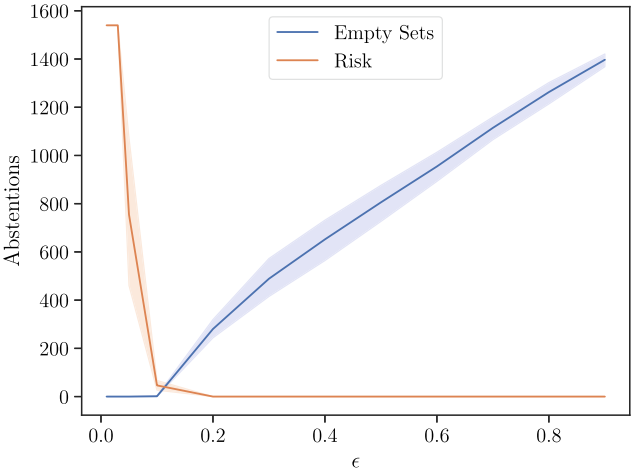


(a) Accuracy and Degree of Automation for labor.

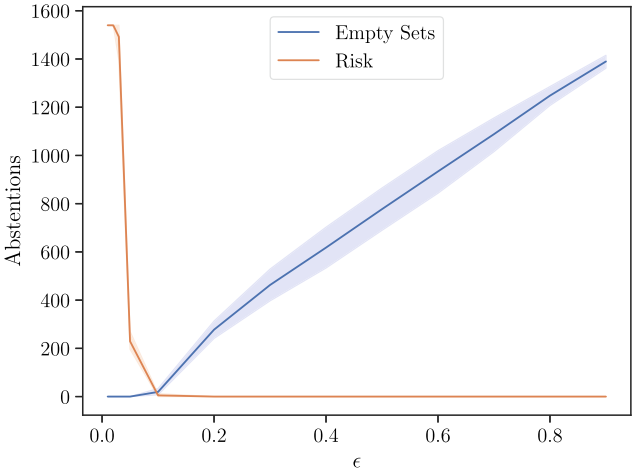


(b) Accuracy and Degree of Automation for Parts.

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



(a) Abstention types for labor.



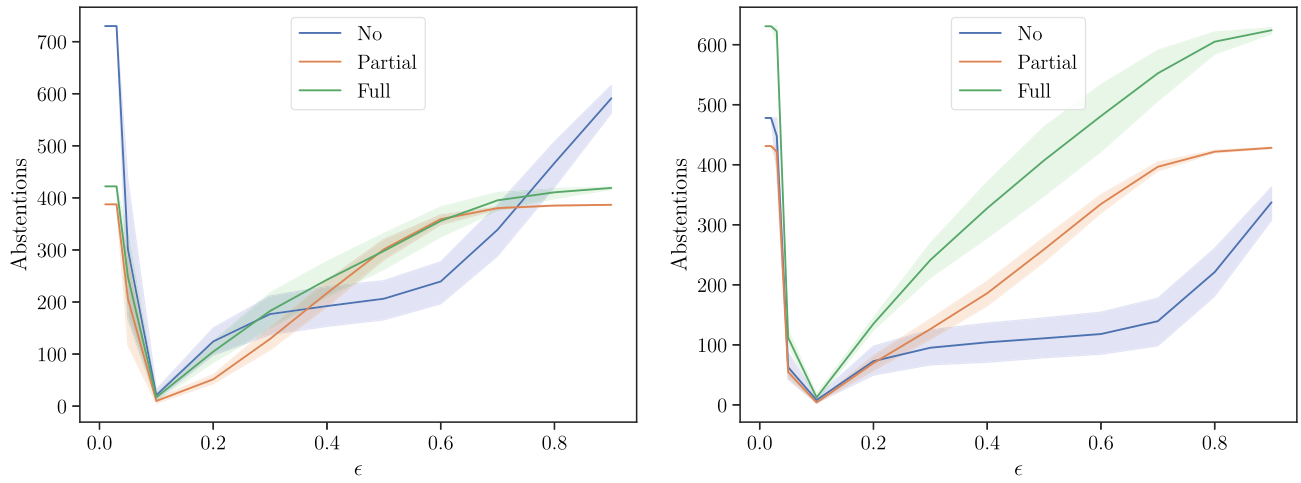
(b) Abstention types for parts.

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

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

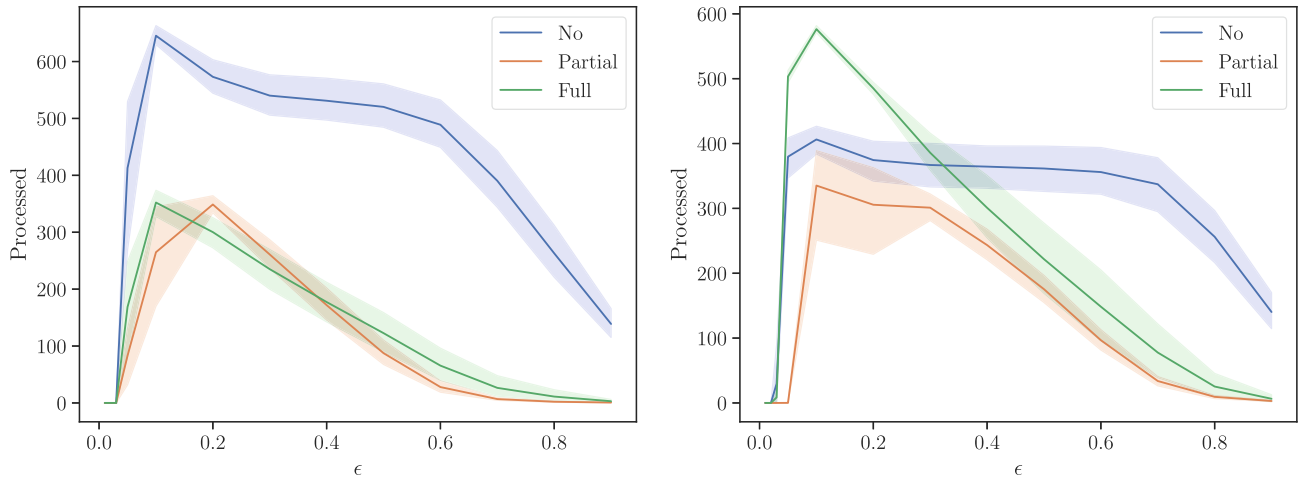
Figure 10 displays the corresponding reasons for abstentions depending on the significance level ϵ. For larger ϵ values, abstentions are exclusively caused by empty sets. In that case, few predicted cases fall below the required quantile threshold $\hat{q}$. For instance, if $\epsilon = 0.9$ only 10%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

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



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

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



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

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

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

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

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

6 Conclusion and future work

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

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

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

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

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

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

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

Funding Open Access funding enabled and organized by Projekt DEAL.

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

Declarations

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

References

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

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

4.3 Stakeholder-centric explanations for black-box decisions: An XAI process model and its application to automotive goodwill assessments

Contributing Article

Stefan Haas, Konstantin Hegestweiler, Michael Rapp, Maximilian Muschalik, and Eyke Hüllermeier. "Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments". In: *Frontiers in Artificial Intelligence - AI in Business* 7 (2024)

Author Contribution Statement

The concept of an XAI process model, designed to guide practitioners in selecting an appropriate XAI method, originated from the Master's thesis of Konstantin Hegestweiler, which was supervised by the author alongside Dr. Michael Rapp and Maximilian Muschalik. The initial manuscript, written by the author, was an extended and enhanced version of this Master's thesis. Additionally, the author conducted further experiments for the technical evaluation. The paper was subsequently revised by Dr. Michael Rapp, Maximilian Muschalik, Prof. Dr. Hüllermeier, and the author.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 24 October 2024
DOI 10.3389/frai.2024.1471208

OPEN ACCESS

EDITED BY
Asif Gill,
University of Technology Sydney, Australia

REVIEWED BY
Babajide J. Osatuyi,
Penn State Erie, The Behrend College, United States
Aleksandr Raikov,
National Supercomputer Center, China

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

RECEIVED 26 July 2024
ACCEPTED 30 September 2024
PUBLISHED 24 October 2024

CITATION
Haas S, Hegestweiler K, Rapp M, Muschalik M and Hüllermeier E (2024) Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments. *Front. Artif. Intell.* 7:1471208. doi: 10.3389/frai.2024.1471208

Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments

Stefan Haas[1,2]*, Konstantin Hegestweiler[1,2], Michael Rapp[1], Maximilian Muschalik[1,3] and Eyke Hüllermeier[1,3]

[1]Institute of Informatics, LMU Munich, Munich, Germany, [2]BMW Group, Munich, Germany, [3]Munich Center for Machine Learning, Munich, Germany

Machine learning has made tremendous progress in predictive performance in recent years. Despite these advances, employing machine learning models in high-stake domains remains challenging due to the opaqueness of many high-performance models. If their behavior cannot be analyzed, this likely decreases the trust in such models and hinders the acceptance of human decision-makers. Motivated by these challenges, we propose a process model for developing and evaluating explainable decision support systems that are tailored to the needs of different stakeholders. To demonstrate its usefulness, we apply the process model to a real-world application in an enterprise context. The goal is to increase the acceptance of an existing black-box model developed at a car manufacturer for supporting manual goodwill assessments. Following the proposed process, we conduct two quantitative surveys targeted at the application's stakeholders. Our study reveals that textual explanations based on local feature importance best fit the needs of the stakeholders in the considered use case. Specifically, our results show that all stakeholders, including business specialists, goodwill assessors, and technical IT experts, agree that such explanations significantly increase their trust in the decision support system. Furthermore, our technical evaluation confirms the faithfulness and stability of the selected explanation method. These practical findings demonstrate the potential of our process model to facilitate the successful deployment of machine learning models in enterprise settings. The results emphasize the importance of developing explanations that are tailored to the specific needs and expectations of diverse stakeholders.

KEYWORDS

eXplainable AI (XAI), prescriptive machine learning, decision support systems (DSS), SHapley Additive exPlanations (SHAP), goodwill assessment

1 Introduction

With the growing access to large amounts of data and the widespread availability of computational resources, the idea of using *machine learning* (ML) methods to guide human experts toward more rational, objective, and accurate decisions, rather than relying solely on their experience and intuition, becomes increasingly prevalent in many application domains. However, in high-stake domains, where decisions can come with severe consequences, there is often a reluctance to use ML methods. For example, this includes applications in healthcare, where decisions may significantly impact human lives,

and use cases in finance or industry that come with the risk of economic loss (Burkart and Huber, 2021; Adadi and Berrada, 2018). Concerns about adopting ML-driven technology are often attributed to the black-box characteristics of high-performance models, such as ensembles of decision trees or neural networks, which cannot easily be inspected, verified, and rectified by humans. Motivated by safety-critical applications, where the ability to understand a model's behavior is crucial for its successful adoption and acceptance by humans, there is a growing demand for *explainable artificial intelligence* (XAI). Besides the development of novel and inherently interpretable supervised ML methods (e.g., Rudin, 2019; Lou et al., 2012, 2013; Ustun and Rudin, 2016), this direction of research has led to various algorithmic solutions aimed at increasing the transparency of existing black-box approaches through *post-hoc* explanations (e.g., Ribeiro et al., 2016, 2018; Lundberg and Lee, 2017; Guidotti et al., 2018a; Plumb et al., 2018; Ming et al., 2018), which explain the inner workings and decision-making process of a trained machine learning model, after the model has already been developed and deployed. One can further distinguish between model-specific or model-agnostic methods, where an explanation method is limited to a specific model class or is model independent, respectively (Burkart and Huber, 2021). In the following, we focus on the latter, as model-agnostic, *post-hoc* approaches allow us to improve on existing models, which are proven to provide robust and accurate predictions.

Our research is driven by a real-world application in the automotive domain, where an ML-based system should support the assessment of goodwill requests. The goodwill process enables car dealers to request monetary compensation for reparations from the manufacturer on behalf of their customers. It qualifies as a high-risk business use case, as bad decisions either negatively affect customer satisfaction or harm the manufacturer's financial interests. Since the manual assessment of goodwill requests is tedious and time-consuming as automotive manufacturers receive up to several tens of thousands of goodwill requests per year, ML provides a tempting opportunity to reduce manual efforts and save costs. Moreover, due to the availability of tens of thousands or even hundreds of thousands of past goodwill requests and their respective outcome, supervised machine learning techniques can be used and have been shown to succeed in closely capturing expert decisions (Haas and Hüllermeier, 2023). However, despite these promising results, the opaqueness of existing models , due to their complex non-interpretable hierarchical structure and usage of gradient boosting, prevents their employment in practice. It is considered a significant limitation by stakeholders, who naturally want to limit the risk of unexpected behavior and therefore demand auditability of the models.

The explanatory needs of different stakeholders are typically context-dependent and may vary between different interest groups. For this reason, a single explanation method cannot always be expected to satisfy the requirements of different stakeholders across a wide variety of applications. As a result, the task of developing interpretable supervised ML systems can only partially be solved from an algorithmic perspective. Instead, it must be considered with high priority during a system's design, development, and evaluation phases. To our knowledge, no complete framework for developing XAI solutions deliberately tailored to different interest groups has yet been proposed in the literature. Instead, as elaborated in Section 2 below, existing publications tend to focus on specific aspects of the topic, such as algorithms, technical evaluation methods, visualization approaches, or user studies. As an important step toward closing the gap between these different research directions, we investigate an end-to-end approach considering all necessary steps for developing an XAI system, starting with stakeholder identification and requirements engineering over implementation to evaluation and user feedback. In summary, the contributions of our work are the following:

- In Section 3, we first discuss the real-world problem of automated goodwill assessment that further motivates the need for explainable ML systems in high-stake domains.
- In Section 4, we propose a streamlined and holistic process model for developing *post-hoc* explainable decision support systems based on findings from interdisciplinary literature and practical considerations.
- In Section 5, we demonstrate how the proposed process model can be applied to the previously introduced real-world scenario and validate its usefulness to meet the explanatory needs of different stakeholders.

By following a stakeholder-centric approach to XAI, we aim to overcome the reluctance to use ML-based solutions in an exemplary business context and hypothesize that our results can be transferred to similar domains. Concretely, we want to validate whether following this process model helps us to overcome the skepticism of ML usage in our exemplary high-stake business process. In detail, we would like to know whether increased stakeholder-centered transparency through XAI methods actually eases the introduction of ML into this high-stake process.

2 Related work

As our goal is to propose a process model deeply rooted in the XAI literature, this section provides a broad overview of existing work on the topic. Developing and evaluating transparent ML systems is an interdisciplinary effort, ranging from machine learning over human-computer interaction and visual analytics to the social sciences. Consequently, several comprehensive surveys exist that aim to consolidate this vast field of research (e.g., Burkart and Huber, 2021; Dwivedi et al., 2023; Minh et al., 2022; Adadi and Berrada, 2018; Guidotti et al., 2018b; Ali et al., 2023; Longo et al., 2024). However, these surveys are far from being an actionable guidance for practitioners in terms of how to approach the topic of XAI in concrete (high-stake) domain implementations.

Nevertheless, many existing publications focus on specific aspects of XAI instead. On the one hand, this includes work on technical aspects of the topic, such as the algorithmic details of different evaluation methods (e.g., Mc Grath et al., 2018; Molnar et al., 2020) and approaches for evaluating them quantitatively (e.g., Lopes et al., 2022; Bodria et al., 2021; Doshi-Velez and Kim, 2017). On the other hand, because XAI's primary goal is to satisfy the explanatory needs of human users and overcome their

skepticism about ML-based technology, research efforts have also been devoted to relevant aspects of human-computer interaction. Among others, contributions in this particular direction include studies on how knowledge about ML models should be presented to users visually (e.g., Hudon et al., 2021). In addition, the challenges of gathering feedback from users and measuring their satisfaction in ML systems are also frequently addressed in user studies (e.g., Kenny et al., 2021). A survey-based methodology for guiding the human evaluation of explanations with the goal to simplify human assessments of explanations is presented by Confalonieri and Alonso-Moral (2024). However, again, this study only focuses on the human evaluation part, neglecting all other parts of XAI system development.

The focus on stakeholder perspective and needs is, amongst others, emphasized by Langer et al. (2021). To our knowledge, Vermeire et al. (2021) are the only ones that address the problem of bridging the gap between stakeholder needs and explanation methods from a practicable and actionable angle. Concretely, they propose explanation ID cards and questionnaires to map explanaibilty methods to user needs. However, their methodology does not cover further technical or user-centered assessments of the matched explanation methods, which may be required in high-stakes settings to ensure reliable and useful explanations. Furthermore, an empirical validation of their proposed method is still missing. In line with our work, XAI tools and processes found in the literature are mapped to common steps in software engineering in Clement et al. (2023). Though the different software engineering phases also appear reasonable in an XAI context, starting out from requirements analysis over design implementation and evaluation over to deployment, the phases rather serve as a structure for the survey than an actionable methodology for practitioners developing XAI systems. Similarly, in Amershi et al. (2019), a general software engineering approach for developing ML systems is derived from practical experience. However, it does not cover any aspects of transparency. A unified framework for designing and evaluating XAI systems, based on a categorization of design goals and corresponding evaluation measures according to different target groups, is presented in Mohseni et al. (2021). However, the framework lacks guidance in terms of concrete XAI method selection. Moreover, it is worth mentioning that the European Commission provides a loose set of requirements for trustworthy AI systems (Floridi, 2019). In addition to the valuable insights provided by the publications mentioned above, we also rely on the taxonomies outlined in Burkart and Huber (2021), Adadi and Berrada (2018), Guidotti et al. (2018b), Arrieta et al. (2020), Meske et al. (2022), and Markus et al. (2021).

Similar to our work, research on XAI is often motivated by specific applications and use cases. Case studies have been conducted in many domains, including the insurance industry (van Zetten et al., 2022), finance (Purificato et al., 2023; Zhu et al., 2023), the public sector (Maltbie et al., 2021), auditing (Zhang et al., 2022), and healthcare (Gerlings et al., 2022). Usually, these studies can be grouped into either purely technically focused studies, without end-user or domain expert involvement, (e.g., Zhu et al., 2023; Orji and Ukwandu, 2024) or studies where feedback regarding the explanations and their comprehensibility is also collected from domain experts or end-users (e.g., van Zetten et al., 2022; Maltbie et al., 2021). The study presented by Baum et al. (2023) stands out as it follows the conceptual model presented by Langer et al. (2021), which considers explanation approaches and information as a means to satisfy different stakeholder desiderata (e.g., interests, expectations, needs, etc.) in particular contexts. Baum et al. adapt this conceptual model in a more practical way by starting with the different stakeholders, which they consider the main context of the explanation, and their particular needs. Based on this, explanation information and concrete XAI methods can then be derived. However, the study lacks empirical validation.

Moreover, beyond these logical paradigms, there are cognitive semantic interpretations that address non-formalisable (black box) aspects of AI. XAI can be conceptualized as a hybrid space where human and machine cognition interact distinctly. For instance, Miller (2019) discusses the importance of cognitive approaches in XAI, highlighting how cognitive semantics can make AI systems more understandable and trustworthy. Several researchers have proposed unique convergent methodologies from a wide array of disciplines (e.g., cognitive modeling, neural-symbolic integration) to ensure XAI's purposefulness and sustainability.

Due to most of the presented works only focusing on specific aspects of XAI and the lack of a coherent methodological framework for XAI system development, which was, for instance, amongst others acknowledged by Bhatt et al. (2020), Langer et al. (2021), and Vermeire et al. (2021), we see an urgent need for a holistic XAI system development process model providing guidance to deploy XAI systems in practice. Even more, as Bhatt et al. (2020) notice that the majority of XAI deployments are not for end users affected by the model but rather for machine learning engineers, who use explainability to debug the model itself, which shows a severe gap between explainability in practice and the goal of transparency for all involved stakeholders.

3 Application domain

As mentioned earlier, the process model proposed in this work is motivated by a real-world application in the automotive domain, where an ML system should support human decision-makers. In the following, we outline the requirements of said application and motivate the need for explainable machine learning models in the respective domain.

3.1 Warranty and goodwill in the automotive industry

Warranty and goodwill are essential aspects of after-sales management in the automotive industry. Vehicles are often costly, so customers have high expectations regarding the reliability of these products. Even if significant efforts are put into quality control, due to the vast number of vehicles sold by *original equipment manufacturers* (OEMs), many warranty claims and goodwill requests must unavoidably be dealt with each year.

Warranty—in contrast to goodwill—is a legal obligation of the OEM. If a customer notices a defect within a legally defined period of time, the manufacturer must rectify the problem at his own

expense. If no adequate solution can be provided, the customer may even have the option to withdraw from the purchase contract. However, it should be noted that the exact legal provisions for warranty may vary from country to country.

Goodwill describes an OEM's willingness to offer repairs, replacements, or financial compensations in the event of defects beyond the scope of warranty. There are no legal obligations here, i.e., an OEM can freely choose a strategy according to which goodwill requests should be handled. However, many manufacturing companies consider goodwill a vital tool to increase customer loyalty. From an OEM's point of view, compensations paid in response to goodwill requests can be understood as marketing investments that may positively affect the loyalty of its existing customers.

Since the duties that come with warranty are clear and legally binding, it is relatively straightforward to process warranty claims automatically, e.g., via rule-based systems. Only in difficult cases, or if the warranty process should be audited, it might be necessary for human experts to check individual claims manually. When we refer to manual expert activity within this study, we refer to *expert judgement*, in which single automotive after-sales experts or assessors leverage the accumulated knowledge, skills, and intuition they have developed over time to make a decision. This is in contrast to *networked expertise*, where the skills and knowledge of multiple experts are combined, or a more guided approach like *quality function deployment (QFD)*. Unfortunately, in the case of goodwill assessments, it is much more challenging to achieve a high degree of automation. For example, even though the car manufacturer employs a rule-based system to deal with goodwill requests in an automated manner, a large fraction of the received requests require a manual examination by human experts. Among 688, 879 goodwill requests considered in Haas and Hüllermeier (2023), only 349, 488 (50.73%) could be processed automatically, whereas 339, 391 (49.27%) demanded a manual assessment. Consequently, there is great potential to increase the degree of automation in the goodwill assessment process through machine learning techniques.

3.2 The use of machine learning in the goodwill assessment process

Supporting human assessors responsible for goodwill decisions through machine learning techniques is appealing from an OEM's perspective, as it can potentially reduce labor costs and foster a standardized goodwill strategy. Unlike decisions made by humans, which are often based on personal experience and intuition rather than being purely rational, assessments provided by ML models are deterministic. This helps to prevent cases where similar goodwill requests result in vastly different responses, which may damage the OEM's reputation. Figure 1 illustrates two different approaches considered in Haas and Hüllermeier (2023) for integrating ML models into the goodwill assessment process. The models can either be used for *automated decision-making* (ADM), where goodwill requests are processed automatically without human intervention, or as a *decision support system* (DSS), which merely provides recommendations to human experts and keeps them in control of the final decision. Whereas, ADM has a greater potential for cost savings, it does also come with a higher risk of incorrect decisions than the DSS approach, since no human supervision takes place.

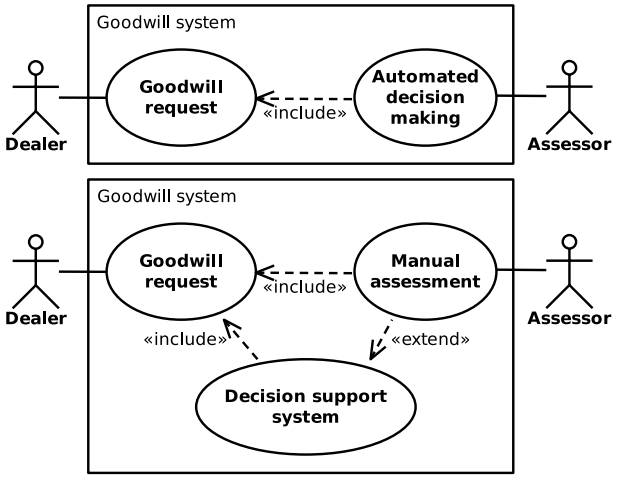


FIGURE 1
Integration of machine learning models into a goodwill assessment system using automated decision-making **(Top)** or a decision support system **(Bottom)**.

Regardless of whether an ADM or a DSS approach is pursued, we consider the problem of providing automated goodwill decisions as a *prescriptive machine learning* problem, a term that has recently been coined by Hüllermeier (2021). It emphasizes differences between the tasks of predicting an outcome and prescribing some sort of action or decision in a certain situation. The former is commonly considered in the standard setting of supervised learning, which assumes a kind of objective ground truth (used as a reference to assess the prediction). In the prescriptive setting, on the other side, there is normally nothing like a "true" or "correct" decision or action—in general, not even the optimality of a prescription can be verified retrospectively, because consequences can only be observed for the one decision made, but not for those other actions that have not been taken.

This lack of ground truth is inherent to goodwill decisions, too, as it cannot be guaranteed that decisions made by experts in the past have always beneficial regarding the OEM's business strategy in the long run. Nevertheless, mimicking the behavior of experts appears to be a natural strategy, as historical data $\mathcal{D} = \{(\vec{\mathbf{x}}_1, y_1), \ldots, (\vec{\mathbf{x}}_n, y_n)\}$, which incorporates information about goodwill requests $\in \mathcal{X}$ and corresponding decisions $y \in \mathcal{Y}$, can easily be used for supervised machine learning. On the one hand, a goodwill request is represented in terms of several *features*. They describe the properties of a vehicle, such as its age, mileage, or whether it was serviced regularly. In addition, they may provide information about a defect that was encountered, including the type of malfunction and the expected repair costs. On the other hand, the possible outcomes of a goodwill assessment depend on the OEM's business strategy. For example, BMW requires assessors to decide for a percentage between 0%, in which case the manufacturer does not offer any compensation, and 100%, which means that the manufacturer fully bears the repair costs. To support the work of

the assessors at BMW, Haas and Hüllermeier (2023) propose an ordinal classification method that models the outcome of goodwill decisions in terms of the compensation (multiples of 10%) as target variable $y \in \{0\%, 10\%, \ldots, 100\%\}$.

3.3 The need for explaining automated goodwill decisions

The previously mentioned ordinal classification method, developed at BMW and discussed in detail in Haas and Hüllermeier (2023), can be considered a *black-box model*. Even though it is able to achieve high accuracy compared to the historical decisions of human assessors, the model's opaqueness poses several challenges for its successful adoption in a business context. Due to its complexity originating from the usage of gradient boosted trees in combination with a hierarchical cost-sensitive framework (Haas and Hüllermeier, 2023), the model can neither be analyzed by human experts as a whole, nor does it provide any information about why certain decisions have been made. This leads to several issues regarding the acceptance and trustworthiness of the automated goodwill system. First, the lack of transparency impedes the ability of domain experts to audit the model and ensure that it adheres to the OEM's goodwill strategy. Second, because no reasons are given for a particular decision, it is hard to reason about cases where the system and human assessors disagree. This makes it difficult to provide valuable feedback that may help to improve the model and hinders the discovery of inconsistencies or biases in human decision-making.

Nevertheless, modern black-box models are valued for achieving state-of-the-art performance. Moreover, there is no legal obligation in goodwill for complete transparency of the assessment process. In settings like these, a solution that overcomes the aforementioned shortcomings while retaining the existing model is desirable. This motivates the use of *post-hoc explanation methods* that can provide insights into an existing black-box model. In particular, model-agnostic explanation approaches are appealing in this regard. They are intended to work with any ML model, regardless of the technical principles it relies on. Figure 2 provides a high-level overview of the interaction between a black-box model and an associated *post-hoc* explainer that aims to clarify the model's behavior. Section 4.2 discusses the characteristics and goals of commonly used explanation methods in more detail.

Due to the unavoidable risk of incorrect decisions in an ML-driven assessment process, in the following, we focus on using machine learning models in the context of decision support systems rather than for automated decision-making. Integrating explanation methods into a DSS, which by design requires human practitioners to closely interact with the automation system, facilitates its employment in high-stake domains and opens the door to the *human-in-command* (HIC) approach (Floridi, 2019) outlined in Figure 3. In this approach, a goodwill assessor can consult an explainable decision support system to safeguard his or her decisions. The assessor and the ML model decide independently on a given goodwill request. If the recommendation provided by the latter differs from the manual assessment, the assessor must be able to obtain a human-understandable explanation for the model's outcome to decide whether it is appropriate to revise the own decision.

4 A process model for developing *post-hoc* explanation systems

As argued in Section 1, there is an urgent need for increased transparency and trust in black-box machine learning models to be used in high-stake domains. Among others, transparency and trust are two of the main goals of XAI (see, e.g., Burkart and Huber, 2021; Arrieta et al., 2020; Lipton, 2018; Fiok et al., 2022). However, selecting the best-suited XAI tools for a specific use case from the vast amount of available methods can be challenging. Usually, not all available solutions can satisfy the explanatory needs of stakeholders equally. Hence, a deliberate selection of suitable tools and a careful evaluation of feedback received from stakeholders is crucial to meet the expectations in an XAI system. For this reason, we propose a process model for developing an explainable decision support system (eDSS) using a *design-science-research* approach (Simon, 1988). An overview of the iterative procedure, including the individual phases it consists of, is shown in Figure 4.

The focus of the process model is to identify and validate suitable *post-hoc* XAI methods, which allow for turning an ML-based DSS into an eDSS. The process starts with an existing black-box model and the intended result is a *post-hoc* explanation system that is tailored to the problem domain and the explanatory needs of the system's stakeholders. The different phases of the proposed process model are, on the one hand, motivated by the XAI literature review presented in Section 2 and the herein identified gaps and requirements, but are also grounded in several complementary theoretical perspectives from the fields of stakeholder theory, human-computer interaction (HCI), and decision support systems (DSS), which further justify the phases themselves and their sequence. At the core of the process model is a strong emphasis on stakeholder engagement, which is informed by stakeholder theory (Freeman and McVea, 2005; Mitchell et al., 1997; Mahajan et al., 2023). Stakeholder theory posits that organizations should consider the needs and interests of all parties affected by their decisions and actions, not just their shareholders, which in turn leads to a broader perspective, long term sustainability, ethical considerations, shared value creation, and eventually a competitive advantage. In the context of XAI system development, this translates to actively involving diverse stakeholder groups, such as end-users, domain experts, policymakers, and management, throughout the design and evaluation process, which is also common sense in XAI research (Kim et al., 2024; Langer et al., 2021; Longo et al., 2024; Baum et al., 2023). For instance, Baum et al. (2023) consider the different stakeholders and their needs as the main context of XAI system development that needs to be elucidated first. The explanation design phase of the process model is informed by principles and theories from the field of human-computer interaction (HCI). Specifically, the model draws on research on cognitive fit (Vessey, 1991) and mental models (Johnson-Laird, 1983) to ensure that the explanations generated by the eDSS are aligned with the mental representations and information processing capabilities of the target end-users and stakeholders, and hence useful and actionable. Additionally, the stakeholder-centric

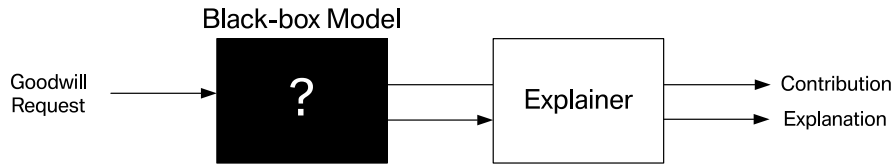


FIGURE 2
Model-agnostic *post-hoc* explainer for the black-box goodwill decision model.

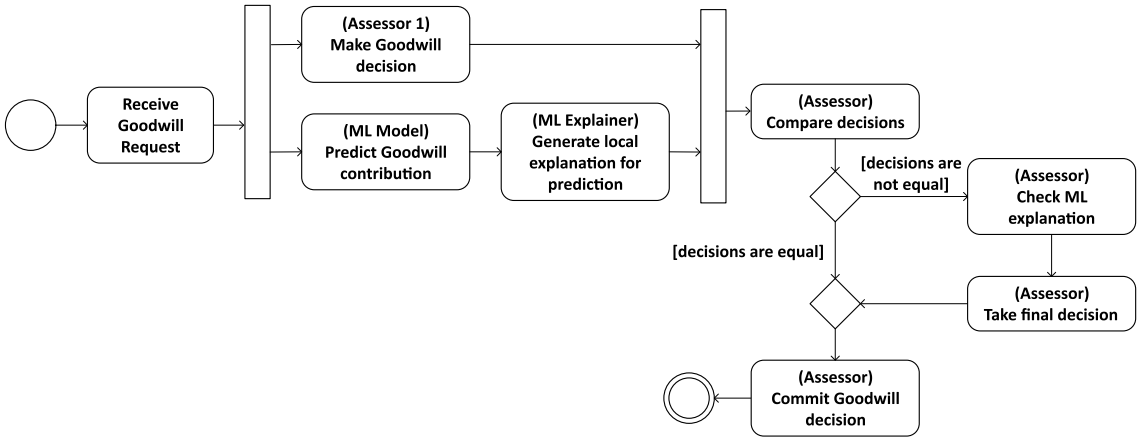


FIGURE 3
Human-in-command approach for an explainable decision support system.

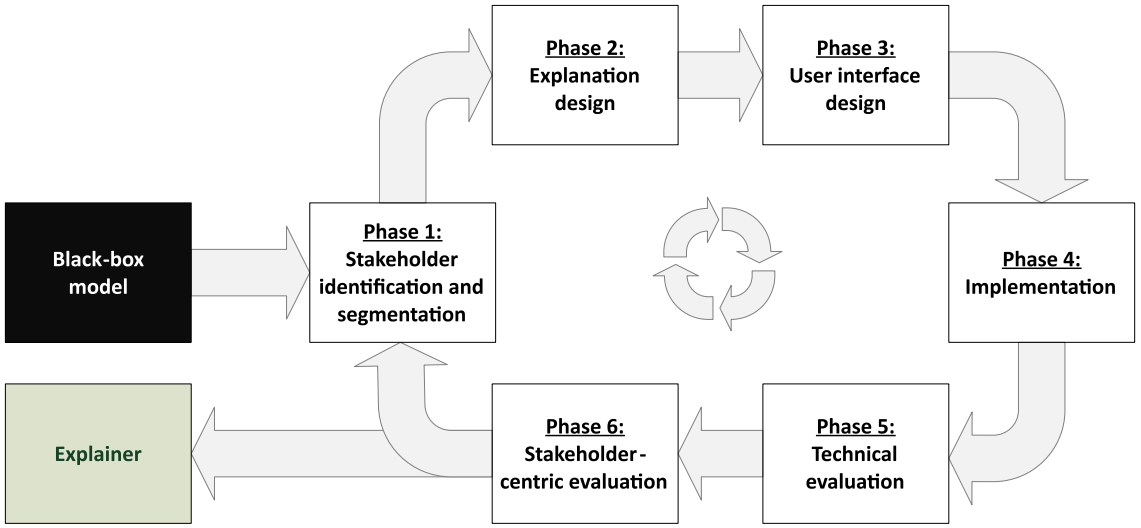


FIGURE 4
Iterative process model for developing an explainable decision support system, based on an existing black-box model.

evaluation phase is grounded in user-centered design approaches (Norman, 2002; Mao et al., 2005), which emphasize the importance of feedback from end-users to inform the design and refinement of interactive systems. By incorporating qualitative and quantitative assessments of stakeholder satisfaction and comprehension, the process model aims to develop explanations that are not only technically sound but also meaningful and useful to the intended users. There is also consensus in XAI research that a solid validation of an XAI system requires both a user-centered and a technical evaluation (Mohseni et al., 2021; Longo et al., 2024; Lopes et al., 2022). The overall structure of the process model, with its focus on developing an explainable decision support system,

is informed by classic theories and frameworks from the field of decision support systems (Keen, 1980; Sprague, 1980). DSS research has long emphasized the importance of user involvement, information presentation, and the integration of human judgment with analytical models to support complex decision-making (Shim et al., 2002; Power, 2002). By adapting these DSS principles to the context of XAI, the proposed process model ensures that the resulting eDSS not only provides accurate predictions but also supports stakeholders in understanding, trusting, and appropriately using the ML-based decision support system (Turban et al., 2010; Arnott and Pervan, 2005). This is also in line with Burkart and Huber (2021), who suggest to consider three aspects for building a useful explanation system: *Who* should be addressed by the explanations, *what* aspects of an ML system should be explained, and *how* should the explanation be presented. In the following subsections, we elaborate on the individual phases of our process model related to these fundamental questions.

4.1 Phase 1: Stakeholder identification and segmentation

Complex computer systems typically have several stakeholders that finance, design, build, use, or audit the system. Developing an eDSS should therefore start with identifying these interest groups, which may have varying expectations in the system and demand for different types of explanations (Gerlings et al., 2022; Kim et al., 2024). In the literature, the stakeholders of ML-based systems are usually separated into three main groups (see, e.g., Burkart and Huber, 2021; Mohseni et al., 2021; Arrieta et al., 2020; Meske et al., 2022), albeit named inconsistently. We rely on the terminology introduced by Hong et al. (2020):

- *Model consumers* or users are the persons affected by the decisions of an ML system. They can interact with the system passively or actively. In the former case, decisions are merely presented to the users, e.g., informing them about the approval or rejection of a loan. In the latter case, the predictions and explanations provided by the system should support human decision-makers, e.g., the person in charge of approving or rejecting a loan. In general, model consumers are not necessarily technical experts. And if they interact with a system passively, they can most likely not be considered domain experts.
- *Model builders* are responsible for developing and operating an ML model. They are proficient in ML but typically not domain experts.
- *Model breakers* are domain experts who have the necessary knowledge to verify that a model behaves correctly and meets the desired goals from a business perspective. However, they are usually not ML experts.

4.2 Phase 2: Explanation design

Once the interest groups of a system have been identified, the next step is to determine which aspects of an ML system need to be explained to each. Following Clement et al. (2023), we refer to this process as the "explanation design phase". Possible explanations can hereby differ in their scope and the technical principles they are based on Burkart and Huber (2021). As the usefulness of available explanation methods depends on the application context and the needs of the stakeholders, their individual goals and limitations must be considered for a well-informed choice. Figure 5 provides an overview of the technical differences between commonly used explanation methods discussed below. In the literature, different XAI methods are often characterized by the scope of the explanations they provide (see, e.g. Burkart and Huber, 2021; Adadi and Berrada, 2018; Molnar et al., 2020; Bodria et al., 2021):

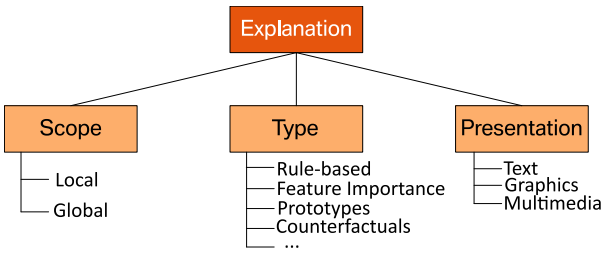


FIGURE 5
Taxonomy of commonly used explanation methods and presentation forms.

- *Global explanations* aim to provide a comprehensible representation of an entire ML model. Their goal is to make the overall behavior of a model transparent by capturing general patterns used by it.
- *Local explanations* focus on individual predictions provided by an ML system. They aim to disclose the reasons for why a particular decision has been made.

As previously mentioned, the preferred scope of explanations depends on the target audience and the application context. For example, product managers might be more interested in global explanations, as they allow them to verify a model's behavior by comparing the patterns it uses to their mental model. In contrast, human decision-makers might prefer local explanations, which can help them make specific decisions.

The most suitable explanation method also depends on the type of data used for training a model, such as tabular data, images or text (Bodria et al., 2021). As the application presented in Section 3 requires the handling of tabular data, we restrict ourselves to this particular scenario, where the following types of explanations are commonly used:

- *Rule-based* models and the conceptually related decision trees are often considered as inherently interpretable (Burkart and Huber, 2021). Hence, it is a natural choice to use rule-based representations for explaining black-box models (Guidotti et al., 2018a).

- *Feature importance* methods provide a ranking of the features found in the data, based on their contribution to a model's decisions (Ribeiro et al., 2016; Lundberg and Lee, 2017).
- *Prototypes* are the minimum subset of data samples that can be viewed as a condensed representation of a larger data distribution. Prototypes can either be obtained for general concepts found in the data or chosen based on their similarity to a particular example at hand (Bien and Tibshirani, 2011).
- *Counterfactuals* provide additional information about a model's predictions in the form of "what-if" scenarios. For example, they can expose the minimal changes of the input required to obtain a different outcome (Mc Grath et al., 2018; Molnar et al., 2020; Wachter et al., 2017). Unlike the other types of explanations listed above, counterfactuals cannot explain a model globally.

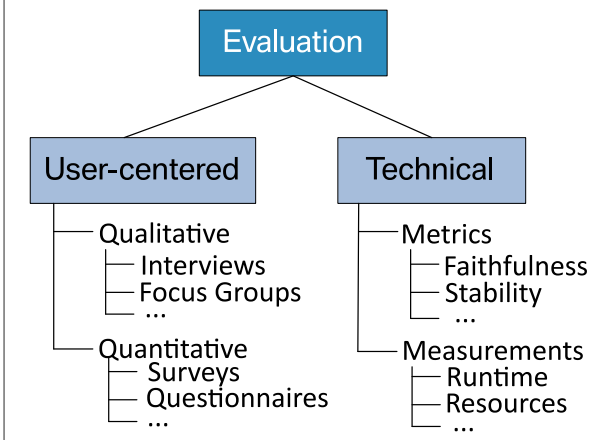


FIGURE 6
Taxonomy of technical and stakeholder-centric evaluation methods.

4.3 Phase 3: User interface design

Once the most suitable technical methods for explaining an ML model's behavior to stakeholders have been identified, an appropriate representation of the explanations must be found. Following Clement et al. (2023), we refer to this phase as the "user interface design". As the form in which explanations are presented to the target audience may significantly influence their intelligibility and usefulness, it is crucial to our process model. Burkart and Huber (2021) distinguish between the following types of representations:

- *Textual explanations* rely on natural language to inform the user, e.g., using complete sentences or bullet lists to justify why a particular decision was made. Textual descriptions can be intuitive because humans tend to explain their decisions verbally.
- *Graphical explanations* make use of visual illustrations, such as plots or diagrams. They may convey complex information in a condensed manner and are supported by many software libraries (e.g., Nori et al., 2019).
- *Multimedia explanations* may combine several types of representation forms, including text, graphics, audio, and video.

Again, the type of representation that best fits the stakeholders' explanatory needs is context-dependent. For example, human decision-makers who must present decisions to customers might prefer a textual description over a visual one. If the information provided by an XAI system is given in text form, they can more easily adopt the explanation and verbally communicate it to the customer. This might ease their work significantly compared to a graphical representation, where they must first extract the essential information and reformulate it in an appropriate verbal response.

4.4 Phase 4: Implementation

After one has decided on XAI methods and corresponding representation forms that are most promising to fulfill the demands in a particular use case, the technical groundwork must be laid for further testing the pursued solution. Generally, this requires implementing the selected explanation methods, integrating them with an existing ML model, and deploying the resulting software. As these steps highly depend on the infrastructure used in a particular application context, it is impossible to provide general advice on the implementation phase of our process model. So, instead, we continue with the technical and user-centric evaluation to be conducted afterward.

4.5 Phase 5: Technical evaluation

In the literature, there is a consensus that the evaluation of an XAI system should comprise a technical and a stakeholder-centric evaluation (Lopes et al., 2022; Mohseni et al., 2021). This obligation is also underpinned by several case studies that employ qualitative and quantitative methods to assess the correctness and suitability of explanations in a given setting (e.g., van Zetten et al., 2022; Maltbie et al., 2021). Moreover, Doshi-Velez and Kim (2017) provide a taxonomy for categorizing XAI evaluation methods. They distinguish between "functionally grounded" approaches based on formally defined metrics and "application-" or "human-grounded" techniques, where humans rate the quality of explanations. Similarly, Figure 6 provides an overview of commonly used evaluation techniques that we consider technical or stakeholder-centric. In the following, we first focus on the former before we continue with the latter in the subsequent section.

Technical evaluation methods aim to ensure the soundness of explanations. This is crucial because faulty behavior of an XAI system may fool an expert into making wrong decisions with severe consequences in high-stake domains. Bodria et al. (2021) highlight the following metrics for safeguarding the functional correctness of explanations:

- *Stability* validates how consistent the explanations provided by an XAI method are for similar examples.
- *Faithfulness* assesses how closely an explanation method can approximate the decisions of a black-box model.

Additional evaluation metrics for use in XAI are constantly proposed (see, e.g., Belaid et al., 2022 for a more extensive overview). For example, we also take runtime and usage of computational resources into account in Section 5.5.

4.6 Phase 6: Stakeholder-centric evaluation

A conceptionally sound and, according to technical criteria, properly working *post-hoc* explanation system might still not entirely fulfill the expectations and demands of individual stakeholders. For this reason, an essential building block of our process model is to evaluate an XAI system's usefulness with regard to the previously identified interest groups. As stressed by Lopes et al. (2022), this second evaluation phase aims to ensure the system's trustworthiness, measure the users' satisfaction, and verify the understandability and usability of the provided explanations. Because a purely technical approach cannot assess these qualitative goals, Doshi-Velez and Kim (2017) emphasize the need to gather feedback from humans working with the system in a real-world setting. When conducting such a user study, the technical background and (possibly lacking) domain knowledge of different interest groups must be considered to allow a realistic assessment of the explanations' comprehensibility. After all, if an explanation is not understandable from an end-user's perspective or is communicated inadequately, this may hamper the ML system's usefulness and trustworthiness.

One challenge of user-centric studies is to gather feedback from humans about their, most likely subjective, opinions regarding predefined goals in a structured and comparable way. Unfortunately, transcripts of personal interviews or reports written by participants (see, e.g., van Zetten et al., 2022; Maltbie et al., 2021; Cahour and Forzy, 2009) can be difficult to analyze. As an alternative, we advocate using Likert-scale questionnaires (see, e.g., van Zetten et al., 2022; Bussone et al., 2015), as discussed in Section 5.6.

5 Case study on automotive goodwill assessment

To demonstrate how the process model introduced in the previous section can be used in practice, we applied it to the application outlined in Section 3. Our goal was to extend an existing black-box model for goodwill assessment in the automotive domain with a *post-hoc* explanation system tailored to the needs of different stakeholders. Moreover, evaluating a conceptual method artifact and its effect on a real-world situation through a case study is a common evaluation method in design science research (Peffers et al., 2012).

5.1 Phase 1: Stakeholder identification and segmentation

According to the first step of our process model, we started by identifying the different stakeholders of the goodwill system. Based on our knowledge about the business use case at hand and discussions with representatives from potential interest groups in focus group meetings, we identified the following stakeholders:

- *IT specialists* employed by the OEM are responsible for developing, maintaining, and operating the goodwill system and its underlying ML model. They are technical experts but not domain experts.
- *Business specialists at the OEM* steer and control the company's global goodwill strategy from a business perspective and are responsible for all operational tasks. They are domain experts but not technical experts. Moreover, they collaborate closely with business specialists from *national sales companies* (NSCs), as described below.
- *Business specialists at NSCs* define guidelines for handling goodwill requests specific to a particular market and supervise the assessors operating in the respective area. They work closely with the parent organization's business specialists and, similar to the latter, are domain experts rather than technical experts.
- *Assessors* are domain experts who decide if the OEM should contribute to the costs of individual goodwill requests. Their decisions are based on the information available about a specific request and adhere to the guidelines established by business specialists. Moreover, assessors are active consumers of the ML system's recommendations.
- *Internal revisionists* audit the goodwill process. As goodwill does not come with legal obligations, they primarily ensure compliance with the OEM's strategic goals and guidelines.
- *Managers* responsible for quality control must ensure an efficient, fair, and transparent goodwill assessment process that benefits customer loyalty and, at the same time, keeps costs at an acceptable level.

Section 4.1 suggests assigning stakeholders to one of three groups: model consumers, model builders, and model breakers. The organizational structure outlined above matches this segmentation quite well. Assessors, who decide on goodwill requests and should actively be supported by the ML model, can be considered model consumers. IT specialists working on the ML system's technical aspects fulfill the roles of model builders. Finally, the responsibilities of business specialists at the OEM and NSCs are complementary. Like internal revisionists and managers, they are most interested in the ML system behaving consistently with their respective goals. Consequently, we consider them model breakers. Figure 7 illustrates the assignment of the goodwill system's stakeholders to distinct interest groups.

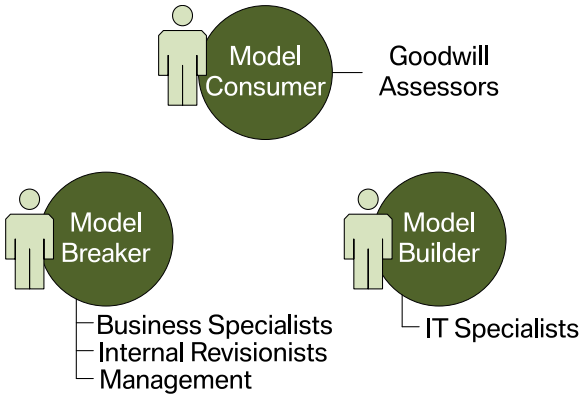


FIGURE 7
Segmentation of the goodwill system's stakeholders into interest groups.

5.2 Phase 2: Explanation design

After identifying and segmenting the goodwill system's stakeholders, our process model's next phase aims at identifying XAI methods that can satisfy their explanatory needs. When dealing with tabular data, we consider feature importance methods, prototypes, and rule-based explanations as technically suitable approaches. We conducted a five-point Likert-scale survey (Likert, 1932) to assess their usefulness regarding the stakeholders' expectations. In this survey, each explanation method was described on a non-technical level. In addition, we provided real-world examples of how the resulting explanations might be presented. Based on this information, we asked participants to what degree different explanations meet their requirements. For illustration, one of the questions included in the explanation design survey is shown in Figure 8.

To ensure the understandability of the web-based survey by non-technical users and due to the limited availability of all stakeholders, it was iteratively refined together with model consumer and breaker team leads in focus group sessions before it was sent to the final pool of stakeholders. The survey was answered by 36 persons working on goodwill assessment in a single market where the decision support system was planned to be deployed. Among the participants were 16 model consumers, eight model breakers, and 12 model builders, representing the majority of the target audience in the considered market. Figure 9 shows how many participants from the different interest groups agreed with the usefulness of potential explanation methods according to a five-point Likert-scale.

We conducted a Shapiro-Wilk test (Shapiro and Wilk, 1965) to check for an approximately normal distribution of answers per group. For none of the stakeholder groups and explanation methods, the p-values exceeded the significance level $\alpha = 0.05$. Consequently, the null hypothesis that the answers per group and method are normally distributed was rejected. Due to the non-normal data distribution, we conducted a non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952) to identify any statistically significant differences between the median answers of different stakeholder groups regarding the usefulness of individual XAI methods. The null hypothesis that the median is the same across all groups could not be rejected for counterfactuals and rule-based explanations (with $\alpha = 0.05$). However, it was rejected for prototypes and feature importance methods. To discover which groups of stakeholders assess the usefulness of these explanation methods differently than the others, we finally conducted a *post-hoc* Dunn (1964) test. It revealed that the answers of the model users regarding the usefulness of local feature importance methods differ from those of the other groups to a statistically relevant degree (with $\alpha = 0.05$). Table 1 summarizes the results of our analysis regarding the perceived helpfulness of explanation methods per stakeholder group. We conclude that all stakeholders of the goodwill system—especially model builders and breakers—consider local feature importance methods as the most promising XAI approach.

5.3 Phase 3: User interface design

According to the previously conducted design study, all stakeholders of the goodwill system expect that explanations based on feature importance can best satisfy their requirements and provide valuable insights into the system's behavior. Hence, we focused on this particular type of explanation during the user design phase that lays the conceptual groundwork for the remaining steps of our process model. In particular, it requires identifying the information the selected approach can provide from a technical standpoint and exploring possibilities to present it to the user.

To explain goodwill decisions by disclosing the impact of individual features, we planned to employ *Shapley additive explanations* (SHAP) (Lundberg and Lee, 2017). This method derives feature importance scores from so-called *Shapley values* originating from game theory (Shapley, 1953). Unlike related methods such as LIME (Ribeiro et al., 2018) or permutation feature importance (Breiman, 2001), it provides theoretical properties well-suited for explaining ML models (Covert et al., 2020). As necessary in our use case, SHAP and the closely-related Kernel SHAP approximation method are model-agnostic *post-hoc* approaches that can be used with any black-box decision model. Moreover, an open-source implementation of these methods, including support for different visualizations, is available.[1]

SHAP provides local explanations in the form of an additive feature attribution function (Lundberg and Lee, 2017; Molnar, 2022)

$$g(z') = \phi_0 + \sum_{j=1}^{d} \phi_j z'_j,$$

where g is the local linear surrogate explanation model and $z' \in \{0,1\}^M$ is a data point represented by M binary features also called *simplified features*. In the simplified features, a value of 1 means that the feature is present whereas a value of 0 indicates absence. The importance of the j-th feature is specified by the absolut value of the Shapley value $\phi_j \in \mathbb{R}$. Its sign indicates whether the feature has

1 https://github.com/slundberg/shap

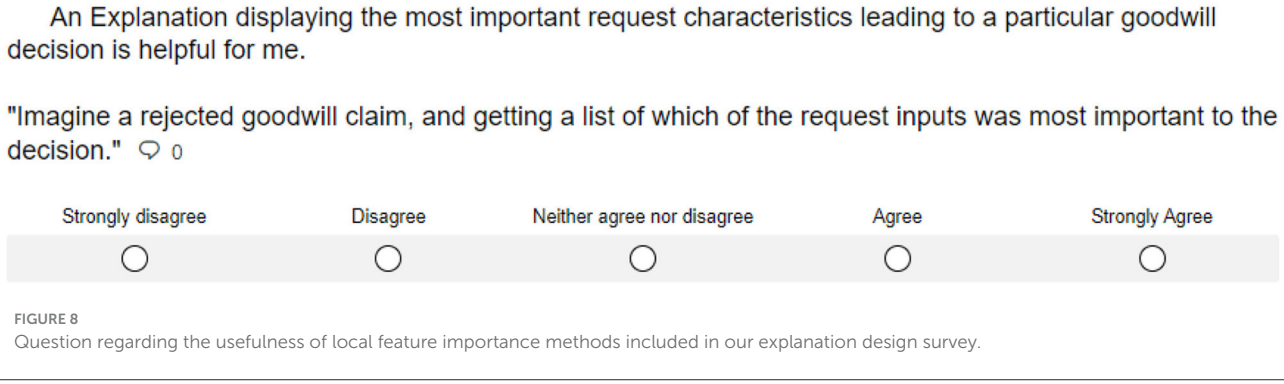


FIGURE 8
Question regarding the usefulness of local feature importance methods included in our explanation design survey.

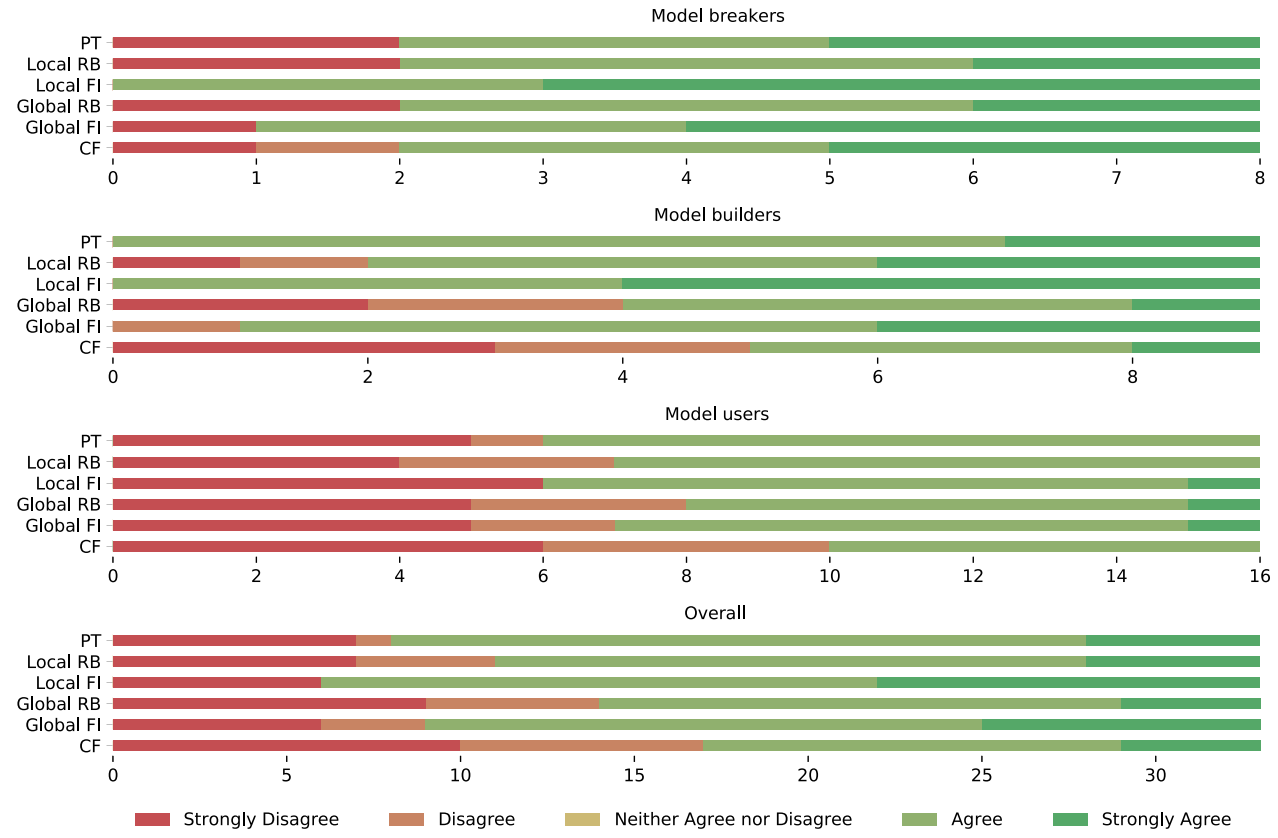


FIGURE 9
Frequency of agreement with the usefulness of prototypes (PT), rule-based explanations (RB), feature importance (FI), and counterfactuals (CF) by stakeholder group.

a positive or negative impact on the point prediction $\hat{y}$. This impact needs to be interpreted relative to a baseline $\mathbb{E}_x\left[\hat{f}(x)\right]$ that denotes the average of all model predictions.

In practice, the exact computation of Shapley values is often computationally infeasible, as 2^d feature subsets must be evaluated. To overcome this limitation, Kernel SHAP employs a sampling strategy for approximating Shapley values. For each data point x to be explained, the model is re-evaluated using a limited number of feature subsets (simplified features). Features that are missing from a subset (are set to 0) are withheld from the decision model. Unfortunately, individual feature values can only be removed from a data point if the model can handle missing values. Otherwise, they must be replaced by randomly sampled values to break the relationship between feature values and target variables (Covert et al., 2020). In case of tabular data, an absent feature equals replacement by a random feature value from the data. By adjusting the number of re-evaluations or samples, Kernel SHAP's computational demands and approximation quality can be traded off (see Section 5.5.3).

In the end, the linear explanation model g is trained by optimizing the following weighted sum of squared errors loss function L:

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} \left(\hat{f}(h_x(z')) - g(z')\right)^2 \pi_x(z')$$

The estimated weights of the linear model g are then the Shapley values $\phi_j \in \mathbb{R}$. $\hat{f}$ is the original model and h_x a helper function mapping simplified features to corresponding values from the actual instance x to be explained ($h_x : \{0,1\}^M \rightarrow \mathbb{R}^M$). π_x is the SHAP kernel providing a weight for each simplified feature vector. The basic idea is hereby to give small (few 1's) and large (many 1's) vectors the highest weights, as they provide the most information regarding the effect of individual features (isolated and total).

To obtain a global explanation for the model, the absolute Shapley values per j-th feature can simply be averaged over the data:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_j^{(i)}|$$

As outlined in Section 3, we utilize an ordinal classification method to decide on the percentage of goodwill costs to be taken by the OEM. In this context, features with negative Shapley values result in less compensation to be paid. In contrast, positive values correlate with a higher contribution. During the user interface design, we considered the following textual and graphical representations (see Figures 10, 11 for examples) to disclose the positive and negative factors that lead to a particular goodwill decision:

- We refer to a simple enumeration of the most influential features according to their Shapley values as the *text baseline*. It is restricted to features with positive (negative) values greater (smaller) than the quantile $q = 0.85$ ($q = 0.15$). The features are grouped by the sign of their Shapley values and sorted by their size in decreasing order.
- *Decision-logic-enhanced text* compares features supporting the financial claims that come with a goodwill request to those speaking against them or favoring a lower financial contribution. As before, only the most influential features favoring or contradicting a request are given in sorted order.
- *Force plots* visualize the contribution of individual features to a prediction based on their Shapley values. For this purpose, the positive or negative impact of each feature is shown relatively to the final prediction and the baseline value on a one-dimensional scale.
- Like the textual representations above, *text-enriched decision plots* provide a description of features sorted by their importance, albeit independently of whether they influence a prediction positively or negatively. However, similar to force plots, the contribution of each feature to the final prediction is shown graphically and put in relation to the baseline value.

TABLE 1 Median agreement with the usefulness of XAI methods per stakeholder group.

Explanation method	Stakeholder group	Useful?
Local feature importance	Model breaker/builder	Strongly agree
	Model user	Agree
Global feature importance	All	Agree
Prototypes	All	Agree
Local rule-based	All	Agree
Global rule-based	All	Agree
Counterfactuals	All	Agree

5.4 Phase 4: Implementation

Once the requirements in the explanation system have been identified, and one has settled for a technical approach that meets these demands, it must be implemented and integrated into the existing ecosystem. Figure 12 outlines the software architecture of the goodwill system. Dealers submit goodwill requests on behalf of their customers via the *dealer frontend*. As described in Section 3, requests are handled by a *rule-based assessment* if possible. Otherwise, a *manual assessment* must be performed. It starts with the invocation of the *ML prediction service* that recommends the compensation to be paid by the OEM for a particular goodwill request. In addition, the prediction service asynchronously triggers the *ML explanation service* by dispatching an explanation request to a FIFO queue monitored by the latter. Separating prediction and explanation into distinct micro-services is favorable as the execution of Kernel SHAP can be computationally costly and time-consuming. With micro-services, the underlying hardware can be scaled independently. Moreover, there is no need to provide explanations immediately after a new goodwill request arrives since it typically takes time until a human assessor can inspect them. Shapley values computed by the explanation service are stored in a central database. They are accessible through a web application called the *explanation dashboard*. Offering a standalone application for accessing explanations enables one to adjust to different stakeholder groups more flexibly. For example, assessors are most interested in explanations for pending goodwill requests. In contrast, other stakeholders like auditors or business experts might want to inspect goodwill decisions made in the past.

5.5 Phase 5: Technical evaluation

As the next step of our process model, a technical evaluation of the previously implemented explanation system should be conducted to ensure that it generates sound explanations. Such an evaluation is crucial as faulty explanations may trick human decision-makers into making wrong decisions. As part of our case study, we verify if the explanations based on Kernel SHAP fulfill two well-established evaluation metrics, namely *stability*, and

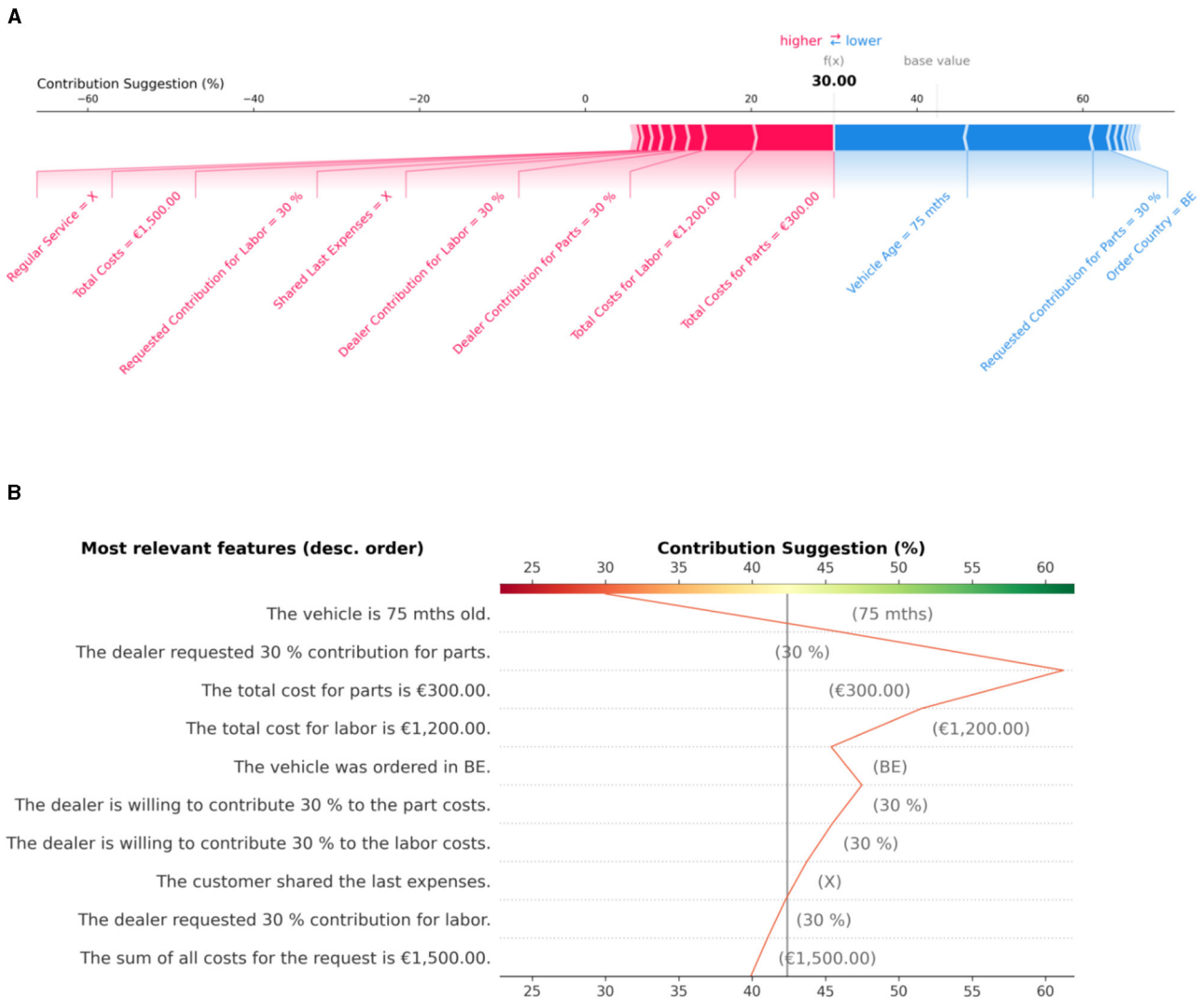


FIGURE 10
Graphical explanations of a goodwill decision based on feature importance. **(A)** The force plot displays feature values in favor of a higher contribution as red arrows, with their length indicating the magnitude of the contribution. Conversely, blue arrows represent feature values that contribute to a lower output value. Importantly, the importance of each feature value is always measured in relation to the base value in SHAP. **(B)** The text-enriched decision plot lists feature value descriptions in descending order of importance on the *y*-axis and the model's output on the *x*-axis. It is centered on the *x*-axis at the base value and shows through the oscillating line how the model's prediction changes through the contribution of the different feature values from bottom to top.

faithfulness (Bodria et al., 2021; Belaid et al., 2022; Alvarez-Melis and Jaakkola, 2018; Rong et al., 2022). *Fidelity* (Bodria et al., 2021), another common evaluation metric, which measures how well an interpretable surrogate model reflects the predictions of the original black-box model, is given by the Shapley value's *efficiency* property $\sum_{j=1}^{M} \phi_j = h(\vec{\mathbf{x}}) - \mathbb{E}\left[h(\vec{\mathbf{x}})\right]$ (Lundberg and Lee, 2017), which states that the feature contributions must add up to the difference of the prediction for $\vec{\mathbf{x}}$ and the average or base value ($\mathbb{E}\left[h(\vec{\mathbf{x}})\right]$). Hence, there is no need to assess this experimentally. In addition, to ensure that the implementation adheres to operational constraints, we measure the *computation time* and *memory consumption* needed to generate explanations. The literature lists many more metrics like *completeness*, *actionability*, *compactness*, *interpretability*, and *plausibility*, among others (Markus et al., 2021; Zhou et al., 2021). However, quantifying them can be challenging without incorporating user feedback, as they often involve subjective judgments and context-specific considerations that are not easily captured through technical means alone. That's why we focus on the established technical key metrics *stability* and *faithfulness* for Kernel SHAP here.

5.5.1 Stability

The stability of an explanation in the context of machine learning models is a crucial concept that refers to how sensitive the explanation is to small changes in the model's input. Explanation stability is an important consideration because it helps assess the reliability and robustness of the explanations provided by a machine learning model. If the explanations are highly sensitive to minor input perturbations, it can raise

concerns about the trustworthiness and consistency of the model's decision-making process.

Stability can be assessed in terms of the *Lipschitz constant*

$$L_x = \max_{x' \in \mathcal{N}_x} \frac{||e_x - e_{x'}||}{||x - x'||}.$$

The test instance for which an explanation should be provided is denoted by x, whereas e_x is the corresponding explanation in the form of Shapley values. We normalize both of these vectors by the sum of their elements. Moreover, $\mathcal{N}_x$ denotes a neighborhood consisting of instances x' similar to x (Bodria et al., 2021; Alvarez-Melis and Jaakkola, 2018).

Based on domain knowledge, we explore the neighborhood x' of a test instance x by applying random changes to some of its numerical features. This procedure is carried out for *mileage* (±100) with an interquartile range (IQR) of 72, 017.25, *vehicle age in month* (±1) with an IQR of 26.0, *labor costs* (±10) with an IQR of 415.0, *parts costs* (±10) with an IQR of 1, 150.0, and *open time costs* (±1) with an IQR of 34.96. For these relatively small changes we do not necessarily expect any changes in the model's predictions or the corresponding explanations.

Table 2 shows the results of our stability evaluation. Large values indicate great instability, meaning that for similar inputs quite different explanations are generated. In addition to the stability, its mean, and its standard deviation, we report the fraction of test instances for which predictions have changed compared to its neighbors. Finally, the table also includes the fraction of instances for which the top-2, -3, and -5 most important features according to Shapley values have changed due to the perturbations in some numerical features of neighboring instances. We observe that explanations of goodwill contributions to labor costs are far more unstable than those related to part costs according to the Lipschitz constant. For both of these explainers, the top-2 and top-3 most important features remain unaffected for the vast majority of test instances. However, for about 50% of the instances the top-5 ranks change, which indicates the limitations of Kernel SHAP's stability. Nevertheless, we consider this explanation method to be stable enough for our use case, because of the small number of changes in predictions and top-2 feature importance rankings.

A

\+ The total **cost for parts is €300.00.**
\+ The total **cost for labor is €1,200.00.**
\+ The **dealer is willing to contribute 30%** to the part costs.
\+ The **dealer is willing to contribute 30%** to the labor costs.

\- The vehicle is **75 mths old.**
\- The vehicle was **ordered in BE.**
\- An **external gurantee** for the vehicle **does not exist.**
\- There is **no information about BEST Type.**

→ Recommendation: 30% Contribution for parts

B

The dealer **requested 30.0 % contribution** for parts.
This is indicated by:

1. The vehicle is **75 mths old.**
2. The vehicle was **ordered in BE.**
3. An **external gurantee** for the vehicle **does not exist.**
4. There is **no information about BEST Type.**

We propose a **contribution of 30 % for parts**, but
also ask to review the indicators for a higher contribution:

1. The total **cost for parts is €300.00.**
2. The total **cost for labor is €1,200.00.**
3. The **dealer is willing to contribute 30%** to the part costs.
4. The **dealer is willing to contribute 30%** to the labor costs.

FIGURE 11
Textual explanations of a goodwill decision based on feature importance. **(A)** The text baseline approach displays the feature values contributing the most positively (+) as well as negatively (−) grouped and with descending importance, as well as the final recommendation by the model. **(B)** The decision-logic-enhanced text also groups the feature values with regards to their positive or negative contribution, but also puts the model's prediction into relation to what the dealer requested from the manufacturer on behalf of the end customer.

5.5.2 Faithfulness

The faithfulness of an explanation assesses how well the explanation approximates the true behavior of the underlying black-box machine learning model (Alvarez-Melis and Jaakkola, 2018). It measures how well the explanation captures the actual decision-making process of the model, rather than just providing a simplified or approximate representation. When dealing with explanations based on feature importance, their faithfulness can be evaluated by using so-called *deletion curves* (Petsiuk et al., 2018). According to this method, feature values are removed from test instances successively, depending on the importance of the corresponding features. The values of the most important features are removed first and after each deletion the model's prediction error is measured. The intuition behind this procedure is the following: If a particular feature is considered highly important by a feature importance method, its removal should lead to a drastic increase in prediction error. In contrast, the prediction error should only slightly deteriorate if one of the least important features is removed. When removing multiple features with decreasing importance, this should cause the prediction error to increase monotonically. Unfortunately, the used black-box models cannot handle tabular data from which individual features have been removed. To overcome this limitation, we sample from the marginal feature distribution to simulate the removal of features as suggested by Covert et al. (2021).

Figure 13 illustrates the faithfulness of the feature importance rankings that explain goodwill contributions to labor and part costs, respectively. In both cases, we observe that the removal of the most important feature already results in a significant change of the deletion curve. Moreover, the removal of additional features results in a monotonically increasing deletion curve until a plateau is finally reached. This testifies the faithfulness of the explanations provided by Kernel SHAP. For the labor costs, the average prediction error increases faster. However, in the limit, the prediction error is not affected as much as for the part costs.

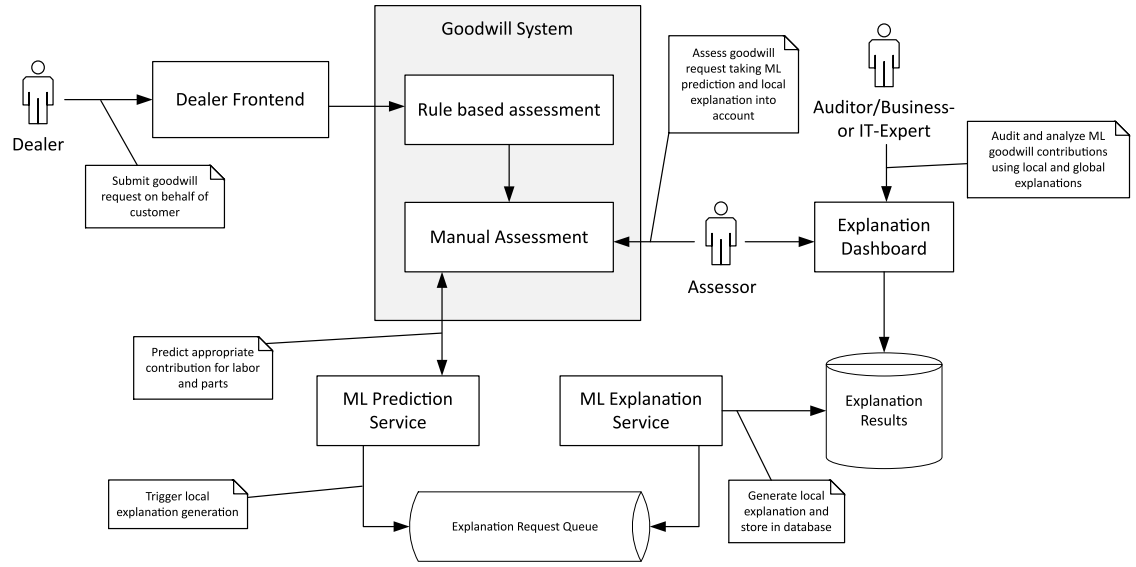


FIGURE 12
Software architecture of the decision support system for goodwill assessment.

TABLE 2 Stability of the Kernel SHAP explainer over a subset of 100 test samples.

Explainer	Stability	Prediction changes	Top-2 FI changes	Top-3 FI changes	Top-5 FI changes
Labor	1,026.4 ±1,507.4	0.01	0.04	0.12	0.47
Parts	544.0 ±939.6	0.00	0.00	0.18	0.53

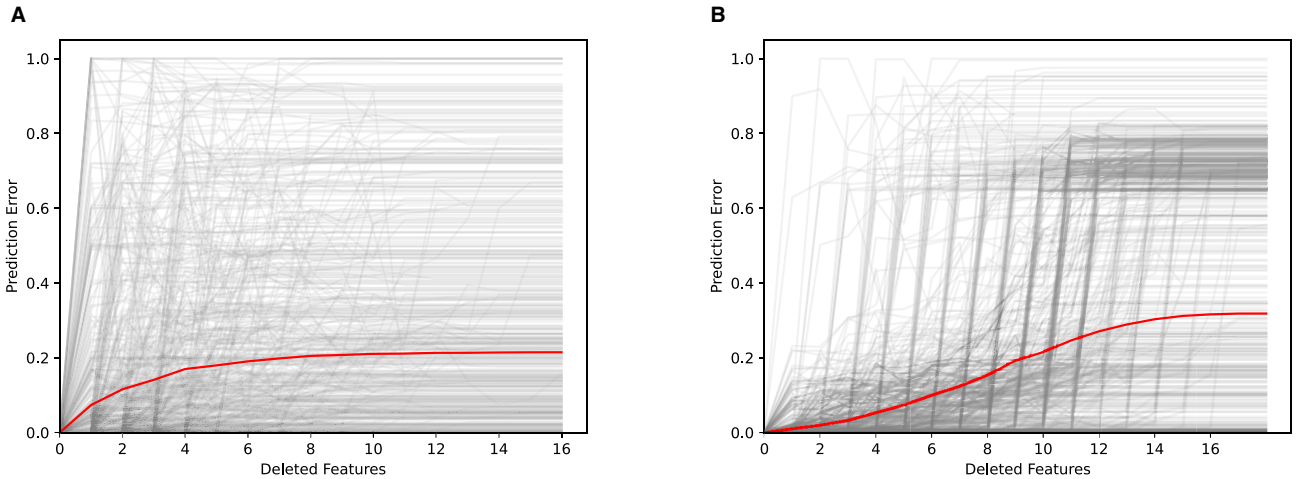


FIGURE 13
Deletion curves obtained for labor and part costs based on 578 randomly selected test instances. Features are removed by randomly sampling 10, 000 replacement values. The prediction error is computed as the relative error compared to the maximal possible error, which may vary depending on the prediction and therefore is normalized. **(A)** Labor. **(B)** Parts.

5.5.3 Runtime and memory consumption

The runtime and memory consumption of Kernel SHAP, apart from the underlying data and number of features, mainly depend on the size of the dataset and the number of times the model is re-evaluated, respectively, simplified features are sampled (`nsamples` parameter in the Kernel SHAP implementation) when explaining a prediction. In our use case, we have to deal with 26 features in total. As a result, the memory consumption

TABLE 3 Maximum runtime and resource consumption of Kernel SHAP for 100 samples.

Explainer	Max. runtime	Max. memory	Max. CPU
Labor	24.92 s	9.99 GB	1,659 mc
Parts	24.15 s	9.02 GB	1,713 mc

of Kernel SHAP is the most limiting factor. We therefore enforced a memory limit at around 10 GB to keep the memory consumption at an acceptable level. As a result, the implementation was deployable on a high density cluster environment without the need to provide dedicated machines with larger main memory.

Table 3 shows the runtime and memory consumption of Kernel SHAP when generating explanations of the contribution to labor and part costs, respectively. The algorithm was provided with a dataset consisting of 100 instances. It was configured to perform 3,000 re-evaluations or samples per explanation. In our use case, an average runtime of 25 s is acceptable, because explanations are provided to human assessors asynchronously instead of in real-time. The CPU utilization of ∼1.7 millicores is moderate. The test was carried out on a machine with 8 vCPUs and 28 GB main memory.

5.6 Phase 6: Stakeholder-centric evaluation

To evaluate the suitability of the considered explanation designs and the overall satisfaction with the explainable decision support system, we conducted a second web-based survey. Like the previous survey, it was iteratively refined together with non-technical stakeholders in focus group sessions before it was sent out to all stakeholders to ensure that the survey was also understandable for non-technical users and that the explanations' design was as clear as possible, e.g., with descriptive labels and meaningful exemplary cases. It addressed the same stakeholders as the first survey. In total, 23 stakeholders participated (11 model consumers, six model builders, six model breakers). Again, we relied on a Likert-scale questionnaire. The first part of the survey focused on the considered representations of explanations (cf. Figures 10, 11), whereas the second part aimed at evaluating the decision support system as a whole.

5.6.1 Preferences regarding the different explanation designs

The survey asked all stakeholders to pick their favorite representation of explanations among the four considered variants. Figure 14 illustrates how many stakeholders preferred each of the available options. To identify any statistically significant deviations from a uniform distribution ($H_0 : \tau = 0.25$), a right-sided binomial test was conducted for each option vs. the other options using a significance level of $\alpha = 0.05$. In addition, the same test was applied to the overall preferences of all stakeholders. When focusing on model users, the p-values obtained for the decision-logic-enhanced text visualization were smaller than α, which leads to a rejection of the null hypothesis and indicates a statistically significant preference for this representation form. The same result was obtained when considering the overall preferences of all stakeholders. Furthermore, the Wald confidence intervals were (30.71%, 69.29%) for all stakeholders and (39.22%, 89.67%) when focusing on the model users. Because even the lower bound of these confidence intervals is greater than $\tau = 0.25$, we consider the preference for the decision-logic-enhanced text design to be very strong. We also evaluated the comprehensibility and actionability of this preferred option using a Kruskal-Wallis test (with $\alpha = 0.05$). According to the reuslts, all stakeholder groups agree that this particular form of explanations is understandable, easy to comprehend, and helps making decisions.

5.6.2 Acceptance of the explainable decision support system

Besides the evaluation of different representation forms, we were also eager to testify if explanations based on feature importance are suited to increase the stakeholders' trust in the decision support system and if they believe that the system will have a positive impact on their task performance. Table 4 shows the questions included in our survey regarding these goals. The frequency distribution of the answers received for these questions are depicted in Figure 15. It should be noted that the null hypothesis of the non-parametric Kruskal-Wallis test, which states that the median is the same across all stakeholder groups, holds for all questions in Table 4, i.e., all stakeholders agree that the provided explanations increased their trust in the decision support system from which they believe that it will positively impact their task performance.

6 Discussion and conclusion

This paper presented a process model rooted in the XAI literature. It covers all the necessary steps for developing a *post-hoc* explanation system that enhances the transparency and trustworthiness of an existing black-box decision system. To demonstrate the usefulness of the proposed methodology, we applied it to a real-world problem in the automotive domain, which encompasses several characteristics like multiple stakeholder groups and a need for increased automation in conjunction with transparency, which are certainly present in other domains as well. Concretely, this study aimed to increase the trust and acceptance of stakeholders in an ML-based goodwill system. By following the process model, we were able to identify an XAI method, together with a suitable representation of the explanations it provides, that meets the requirements of different stakeholder groups. According to a final survey, all stakeholders agree that the selected and implemented XAI approach increases their trust in the decision system and can be expected to improve the performance of employees working with the system. From a design science research perspective, we believe that through our successful case study we have demonstrated our process model's *ease of use*, *efficiency*, *generality* and *operationality*, which are common evaluation

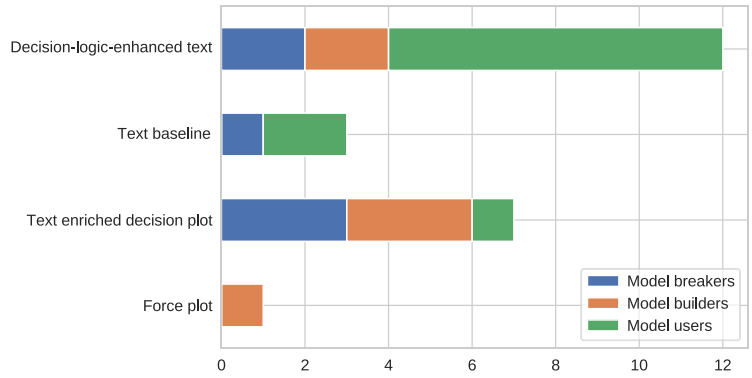


FIGURE 14
Number of stakeholders preferring the considered representation forms.

TABLE 4 Questions regarding the trust in the eDSS and its impact on task performance, as well as the median answers among all stakeholder groups.

Statement	Answer
The explanation increased my trust in the decision support system.	Agree
I would follow the contribution suggestion for the cases because of the explanation.	Agree
I could finish my task faster with the help of this explanation.	Agree

criteria for method type artifacts (Sonnenberg and Vom Brocke, 2012). We further believe that our proposed process model can be transferred to other domains facing similar challenges, as presented in this study, such as multiple stakeholder groups and a tailored model requiring model-agnostic, *post-hoc* explanation methods for different stakeholder groups. In the following, we elaborate on some findings and limitations we identified during our study.

6.1 The importance of stakeholder involvement

The results of both surveys that we conducted in the course of our study emphasize the importance of stakeholder involvement in the XAI development process. Initially, we did neither anticipate the potential of XAI methods based on feature importance to meet their expectations nor their preference for text-based explanations.

Regarding the considered XAI methods, we expected that stakeholders favor rule-based explanations because a rule-based decision system is already used in the domain. Most probably, their choice for feature importance methods can be explained by the bad experiences with the decade-old and hence overly complex rule system, which might not be considered interpretable anymore. Moreover, although we expected counterfactual explanations to be less valuable for assessors working at the OEM, we saw them as an attractive solution for car dealers and their customers. After all, learning how changes in goodwill requests would affect the outcome of the goodwill process would allow them to maximize the compensation paid by the manufacturer. Finally, we expected that model breakers, i.e., managers, business specialists, and revisionists, would be more interested in a global perspective on the decision-making process than in analyzing individual goodwill requests. However, there appears to be a general preference across all stakeholders to inspect specific cases and draw conclusions from them instead of being provided with global explanations.

Another interesting outcome of our case study was the stakeholders' preference for text-based explanations over graphical representations, although the former are restricted to rankings of features and cannot convey information about their absolute importance. Nevertheless, many users, particularly model consumers, i.e., assessors responsible for goodwill decisions, preferred to be provided with textual information. These results may indicate that text-based feedback is perceived as natural by users without a technical background and can be understood more easily, even without previous training.

6.2 Effects on the acceptance of machine learning

The feedback we obtained from different interest groups via the previously discussed surveys indicates that their trust in the decision support system has increased. Compared to the initial reluctance of stakeholders to rely on a black-box model, the employment of XAI positively impacted the acceptance of ML-based technology. On the one hand, we attribute this newfound openness to the increase in transparency achieved through XAI. On the other hand, we believe that the involvement of stakeholders in the design and development process positively influenced their attitude toward the system.

Furthermore, we noticed that the possibility to analyze recommendations made by the ML model fosters discussions about the model's fairness and possible biases in human goodwill

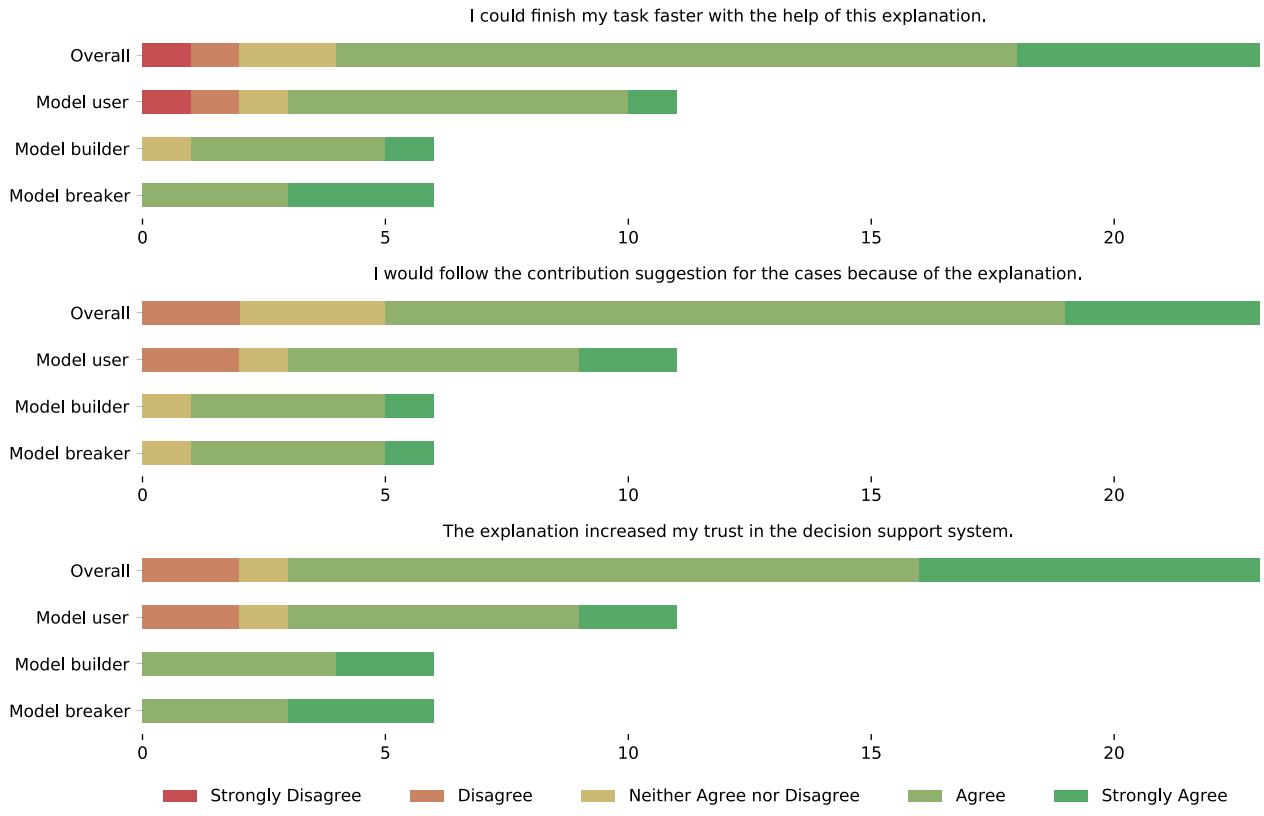


FIGURE 15
Frequency of agreement with statements regarding the trust in the eDSS and its impact on task performance per stakeholder group.

decisions. This suggests that XAI technologies can help to encourage fairness and increase awareness of unwanted biases in decision processes. However, increased trust in automated decision-making may also lead to over-reliance on the system, which is not desired in a high-stake business context built around the human-in-command principle. Instead, the goal should be an interplay between critically thinking human experts and the decision support system. As a countermeasure, the assessment process could be monitored to detect trends toward unilateral decisions that indicate algorithm aversion (Dietvorst et al., 2015) or automation bias (Lee and See, 2004).

6.3 Limitations and future work

Since the choice of suitable XAI approaches is very domain-specific, the process model proposed in this paper can only provide rough guidance. Consequently, it needs to be tailored to the specific use case, e.g., by considering appropriate explanation methods and presentation forms. Providing more guidance and even tool support to practitioners with regards to suitable explanation methods and designs depending on the domain, e.g., healthcare, finance, or the public sector, could be an interesting future avenue of research. As we have seen with the preference for textual explanation representation within this study, suitable methods and designs can be very domain-specific and contrary to common assumptions.

Moreover, the current process model only focuses on identifying, implementing, and evaluating *post-hoc* explanation methods that help to gain insights into an existing black-box model. In addition, future work may also deal with use cases where the goals of XAI should be considered from the start of the development process. In such cases, inherently interpretable white-box models can also play an important role and must therefore be taken into account.

The results of the first survey regarding the different explanation methodologies may also indicate that many stakeholders may not have fully understood the differences between the various explanation methods. This is evidenced by the agreement that all explanations are useful, but little difference in preferences among the methods. The purely textual web-based survey format could have been a limiting factor in this case. The second survey, which incorporated both textual and visual representations of the explanation methods, led to more nuanced results. This suggests that presenting explanations in a more tangible way, with more concrete domain-specific examples that stakeholders can relate to, appears beneficial.

In general, gathering feedback from human stakeholders remains a cumbersome and challenging task due to their

limited availability and ML/XAI expertise, which may also explain the primary usage of XAI by developers (Bhatt et al., 2020). Hence, there is a severe risk of biased feedback results originating from poorly designed XAI surveys, leading to misguided XAI systems. Pre-validating designs and surveys in focus groups, as done in our study, may be a way to prevent larger misconceptions and misunderstandings among stakeholders. However, automating, validating, and easing the collection of user feedback may be an important avenue for future research (Confalonieri and Alonso-Moral, 2024), as collecting stakeholder feedback is of utmost importance when developing XAI systems. Guidance in terms of XAI survey creation, visualization, and validation could reduce the risk of misconceptions and misguided XAI systems.

In terms of stakeholder segmentation, as discussed in Section 5.1, a more structured and fine-grained approach may also be beneficial, particularly to further split the model breaker stakeholders into more distinct interest groups. Model breakers usually encompass several interest groups, each of which may have distinct explanation needs, whereas the builder and user groups appear more homogeneous. Due to time and resource constraints, user segmentation was not carried out to the full extent in this study.

In terms of computational efficiency, the utilization of Kernel SHAP was not an issue in this study, where explanations could be generated in an asynchronous way. However, for applications that require real-time explanations, the usage of Kernel SHAP could be problematic due to the high memory usage and runtime as demonstrated in Section 5.5.3. Here, more efficient SHAP estimators may be required.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: private company owned data. Requests to access these datasets should be directed to stefan.sh.haas@bmwgroup.com.

Author contributions

SH: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. KH: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MR: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. MM: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. EH: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study received funding from BMW. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

SH and KH were employed at BMW AG.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Ali, S., Abuhmed, T., El-Sappagh, S. H. A., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99:101805. doi: 10.1016/j.inffus.2023.101805

Alvarez-Melis, D., and Jaakkola, T. S. (2018). "Towards robust interpretability with self-explaining neural networks," in *Proc. International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates), 7786–7795.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., et al. (2019). "Software engineering for machine learning: a case study," in *Proc. IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice* (New York City, NY: IEEE), 291–300.

Arnott, D., and Pervan, G. (2005). A critical analysis of decision support systems research. *J. Inf. Technol.* 20, 67–87. doi: 10.1057/palgrave.jit.2000035

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fus.* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Baum, D., Baum, K., Gros, T. P., and Wolf, V. (2023). "XAI requirements in smart production processes: a case study," in *Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part I, volume 1901 of Communications in Computer and Information Science*, ed. L. Longo (Cham: Springer), 3–24.

Belaid, M. K., Hüllermeier, E., Rabus, M., and Krestel, R. (2022). Compare-xAI: toward unifying functional testing methods for *post-hoc* XAI algorithms into an interactive and multi-dimensional benchmark. *arXiv* [preprint]. doi: 10.1007/978-3-031-44067-0_5

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., et al. (2020). "Explainable machine learning in deployment," in *FAT* '20: Conference on Fairness, Accountability, and Transparency*, eds. M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna (Barcelona: ACM), 648–657.

Bien, J., and Tibshirani, R. (2011). Prototype selection for interpretable classification. *Ann. Appl. Stat.* 5, 2403–2424. doi: 10.1214/11-AOAS495

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., and Rinzivillo, S. (2021). Benchmarking and survey of explanation methods for black box models. *arXiv* [preprint] arXiv:2102.13076. doi: 10.48550/arXiv.2102.13076

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Burkart, N., and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res*. 70, 245–317. doi: 10.1613/jair.1.12228

Bussone, A., Stumpf, S., and O'Sullivan, D. (2015). "The role of explanations on trust and reliance in clinical decision support systems," in *Proc. International Conference on Healthcare Informatics* (New York City, NY: IEEE), 160–169.

Cahour, B., and Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Saf. Sci*. 47, 1260–1270. doi: 10.1016/j.ssci.2009.03.015

Clement, T., Kemmerzell, N., Abdelaal, M., and Amberg, M. (2023). XAIR: a systematic metareview of explainable AI (XAI) aligned to the software development process. *Mach. Learn. Knowl. Extract*. 5, 78–108. doi: 10.3390/make5010006

Confalonieri, R., and Alonso-Moral, J. M. (2024). An operational framework for guiding human evaluation in explainable and trustworthy artificial intelligence. *IEEE Intell. Syst*. 39, 18–28. doi: 10.1109/MIS.2023.3334639

Covert, I. C., Lundberg, S. M., and Lee, S.-I. (2020). "Understanding global feature contributions with additive importance measures," in *Proc. International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates), 17212–17223.

Covert, I. C., Lundberg, S. M., and Lee, S.-I. (2021). Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res*. 22, 9477–9566.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol*. 144:114. doi: 10.1037/xge0000033

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv* [preprint]. doi: 10.48550/arXiv.1702.08608

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics* 6, 241–252. doi: 10.1080/00401706.1964.10490181

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., et al. (2023). Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comp. Surv*. 55, 1–33. doi: 10.1145/3561048

Fiok, K., Farahani, F. V., Karwowski, W., and Ahram, T. (2022). Explainable artificial intelligence for education and training. *J. Defense Model. Simul*. 19, 133–144. doi: 10.1177/15485129211028651

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nat. Mach. Intell*. 1, 261–262. doi: 10.1038/s42256-019-0055-y

Freeman, R. E., and McVea, J. (2005). *A Stakeholder Approach to Strategic Management. The Blackwell Handbook of Strategic Management* (Wiley), 183–201.

Gerlings, J., Jensen, M. S., and Shollo, A. (2022). "Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare," in *Handbook of Artificial Intelligence in Healthcare, Vol. 2*, eds. C.-P. Lim, Y.-W. Chen, A. Vaidya, C. Mahorkar, and L. C. Jain (Springer Nature), 169–198.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018a). Local rule-based explanations of black box decision systems. *arXiv* [preprint]. doi: 10.48550/arXiv.1805.10820

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018b). A survey of methods for explaining black box models. *ACM Comp. Surv*. 51, 1–42. doi: 10.1145/3236009

Haas, S., and Hüllermeier, E. (2023). "A prescriptive machine learning approach for assessing goodwill in the automotive domain," in *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* (Cham), 170–184.

Hong, S. R., Hullman, J., and Bertini, E. (2020). "Human factors in model interpretability: industry practices, challenges, and needs," in *Proc. ACM Human-Computer-Interaction* (New York, NY: Association for Computing Machinery), 1–26.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., and Sénécal, S. (2021). "Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence," in *Proc. Information Systems and Neuroscience*, 237–246.

Hüllermeier, E. (2021). Prescriptive machine learning for automated decision making: challenges and opportunities. *arXiv* [preprint]. doi: 10.48550/arXiv.2112.08268

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, inferEnce, and Consciousness*. Cambridge, MA: Harvard University Press.

Keen, P. G. (1980). Decision support systems: a research perspective. *Decis. Support Syst*. 11, 23–27. doi: 10.1016/B978-0-08-027321-1.50007-9

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. *Artif. Intell*. 294:103459. doi: 10.1016/j.artint.2021.103459

Kim, M., Kim, S., Kim, J., Song, T., and Kim, Y. (2024). Do stakeholder needs differ? - Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. *Int. J. Hum. Comput. Stud*. 181:103160. doi: 10.1016/j.ijhcs.2023.103160

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc*. 47, 583–621. doi: 10.1080/01621459.1952.10483441

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell*. 296:103473. doi: 10.1016/j.artint.2021.103473

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Fact*. 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol*. 22:55.

Lipton, Z. C. (2018). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., et al. (2024). Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* 106:102301. doi: 10.1016/j.inffus.2024.102301

Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). XAI systems evaluation: a review of human and computer-centred methods. *Appl. Sci*. 12, 9423. doi: 10.3390/app12199423

Lou, Y., Caruana, R., and Gehrke, J. (2012). "Intelligible models for classification and regression," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computing Machinery), 150–158.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). "Accurate intelligible models with pairwise interactions," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computing Machinery), 623–631.

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Proc. International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates), 4768–4777.

Mahajan, R., Lim, W. M., Sareen, M., Kumar, S., and Panwar, R. (2023). Stakeholder theory. *J. Bus. Res*. 166:114104. doi: 10.1016/j.jbusres.2023.114104

Maltbie, N., Niu, N., Van Doren, M., and Johnson, R. (2021). "XAI tools in the public sector: a case study on predicting combined sewer overflows," in *Proc. ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (New York, NY: Association for Computing Machinery), 1032–1044.

Mao, J.-Y., Vredenburg, K., Smith, P. W., and Carey, T. (2005). The state of user-centered design practice. *Commun. ACM* 48, 105–109. doi: 10.1145/1047671.1047677

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inform*. 113:103655. doi: 10.1016/j.jbi.2020.103655

Mc Grath, R., Costabello, L., Le Van, C., Sweeney, P., Kamiab, F., Shen, Z., et al. (2018). "Interpretable credit application predictions with counterfactual explanations," in *Proc. Neural Information Processing Systems-Workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy* (Red Hook, NY: Curran Associates).

Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Inf. Syst. Manag*. 39, 53–63. doi: 10.1080/10580530.2020.1849465

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell*. 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Ming, Y., Qu, H., and Bertini, E. (2018). RuleMatrix: visualizing and understanding classifiers with rules. *IEEE Trans. Vis. Comput. Graph*. 25, 342–352. doi: 10.1109/TVCG.2018.2864812

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artif. Intell. Rev*. 55, 3503–3568. doi: 10.1007/s10462-021-10088-y

Mitchell, R. K., Agle, B. R., and Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. *Acad. Manag. Rev*. 22, 853–886. doi: 10.2307/259247

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transact. Interact. Intell. Syst*. 11, 1–45. doi: 10.1145/3387166

Molnar, C. (2022). *Interpretable Machine Learning. 2nd Edn*. Available at: https://christophm.github.io/interpretable-ml-book/

Molnar, C., Casalicchio, G., and Bischl, B. (2020). "Interpretable machine learning-a brief history, state-of-the-art and challenges," in *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* (Cham), 417–431.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). InterpretML: a unified framework for machine learning interpretability. *arXiv* [preprint]. doi: 10.48550/arXiv.1909.09223

Norman, D. A. (2002). *The Design Of Everyday Things*. New York, NY: Basic Books.

Orji, U., and Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Mach. Learn. Appl.* 15:100516. doi: 10.1016/j.mlwa.2023.100516

Peffers, K., Rothenberger, M., Tuunanen, T., and Vaezi, R. (2012). "Design science research evaluation," in *Proc. of the International Conference on Design Science Research in Information Systems and Technology, DESRIST* (Berlin, Heidelberg: Springer), 398–410.

Petsiuk, V., Das, A., and Saenko, K. (2018). "RISE: randomized input sampling for explanation of black-box models," in *British Machine Vision Conference*, 151–163.

Plumb, G., Molitor, D., and Talwalkar, A. (2018). "Model agnostic supervised local explanations," in *Proc. International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates), 2520–2529.

Power, D. (2002). *Decision Support Systems: Concepts and Resources for Managers*. Westport; Connecticut; London: QUORUM BOOKS.

Purificato, E., Lorenzo, F., Fallucchi, F., and De Luca, E. W. (2023). The use of responsible artificial intelligence techniques in the context of loan approval processes. *Int. J. Hum. Comput. Interact.* 39, 1543–1562. doi: 10.1080/10447318.2022.2081284

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why should I trust you?": explaining the predictions of any classifier," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY: Association for Computing Machinery), 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). "Anchors: high-precision model-agnostic explanations," in *Proc. AAAI Conference on Artificial Intelligence*, 1527–1535.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., et al. (2022). Towards human-centered explainable AI: User studies for model explanations. *arXiv* [preprint]. doi: 10.48550/arXiv.2210.11584

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611. doi: 10.1093/biomet/52.3-4.591

Shapley, L. S. (1953). "A value for n-person games," in *Contributions to the Theory of Games, Vol. 28*, eds. H. Kuhn, and A. Tucker (Princeton, NJ: Princeton University Press), 307–317.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., and Carlsson, C. (2002). Past, present, and future of decision support technology. *Decis. Support Syst.* 33, 111–126. doi: 10.1016/S0167-9236(01)00139-7

Simon, H. A. (1988). The science of design: creating the artificial. *Design Issues* 4, 67–82. doi: 10.2307/1511391

Sonnenberg, C., and Vom Brocke, J. (2012). "Evaluation patterns for design science research artefacts," in *Proc. European Design Science Symposium, EDSS* (Springer), 71–83.

Sprague Jr, R. H. (1980). A framework for the development of decision support systems. *MIS Q.* 4, 1–26. doi: 10.2307/248957

Turban, E., Sharda, R., and Delen, D. (2010). *Decision Support and Business Intelligence Systems, 9th Edn.* Prentice Hall Press.

Ustun, B., and Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.* 102, 349–391. doi: 10.1007/s10994-015-5528-6

van Zetten, W., Ramackers, G., and Hoos, H. (2022). Increasing trust and fairness in machine learning applications within the mortgage industry. *Mach. Learn. Appl.* 10:100406. doi: 10.1016/j.mlwa.2022.100406

Vermeire, T., Laugel, T., Renard, X., Martens, D., and Detyniecki, M. (2021). "How to choose an explainability method? Towards a methodical implementation of XAI in practice," in *Workshop Proc. European Conference on Machine Learning and Knowledge Discovery in Databases* (Cham).

Vessey, I. (1991). Cognitive fit: a theory-based analysis of the graphs versus tables literature. *Decis. Sci.* 22, 219–240. doi: 10.1111/j.1540-5915.1991.tb00344.x

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. J. Law Technol.* 31:841. doi: 10.2139/ssrn.3063289

Zhang, C. A., Cho, S., and Vasarhelyi, M. (2022). Explainable artificial intelligence (XAI) in auditing. *Int. J. Account. Inf. Syst.* 46:100572. doi: 10.1016/j.accinf.2022.100572

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: a survey on methods and metrics. *Electronics* 10:593. doi: 10.3390/electronics10050593

Zhu, X., Chu, Q., Song, X., Hu, P., and Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Sci. Manag.* 6, 123–133. doi: 10.1016/j.dsm.2023.04.003

4.4 Rectifying Bias in Ordinal Observational Data Using Unimodal Label Smoothing

Contributing Article

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*. vol. 14174. Lecture Notes in Computer Science. Springer, 2023, pp. 3–18

Author Contribution Statement

The suggestion to treat historical expert ratings in a weaker manner was proposed by Prof. Dr. Hüllermeier. The specific class-wise unimodal label smoothing approach, based on the Geometric distribution, was developed by the author. Additionally, the author designed the smoothing heuristics presented in the paper to address imbalances and concept drift biases introduced by human raters. Furthermore, the label smoothing approach, along with all experiments and heuristics, was implemented by the author. The entire paper underwent multiple rounds of revision in collaboration between Prof. Dr. Hüllermeier and the author.

Rectifying Bias in Ordinal Observational Data Using Unimodal Label Smoothing

Stefan Haas[1( )] and Eyke Hüllermeier[2,3]

[1] BMW Group, Munich, Germany
stefan.sh.haas@bmwgroup.com
[2] Institute of Informatics, LMU Munich, Munich, Germany
[3] Munich Center for Machine Learning, Munich, Germany

Abstract. This paper proposes a novel approach for modeling observational data in the form of expert ratings, which are commonly given on an ordered (numerical or ordinal) scale. In practice, such ratings are often biased, due to the expert's preferences, psychological effects, etc. Our approach aims to rectify these biases, thereby preventing machine learning methods from transferring them to models trained on the data. To this end, we make use of so-called label smoothing, which allows for redistributing probability mass from the originally observed rating to other ratings, which are considered as possible corrections. This enables the incorporation of domain knowledge into the standard cross-entropy loss and leads to flexibly configurable models. Concretely, our method is realized for ordinal ratings and allows for arbitrary unimodal smoothings using a binary smoothing relation. Additionally, the paper suggests two practically motivated smoothing heuristics to address common biases in observational data, a time-based smoothing to handle concept drift and a class-wise smoothing based on class priors to mitigate data imbalance. The effectiveness of the proposed methods is demonstrated on four real-world goodwill assessment data sets of a car manufacturer with the aim of automating goodwill decisions. Overall, this paper presents a promising approach for modeling ordinal observational data that can improve decision-making processes and reduce reliance on human expertise.

Keywords: Prescriptive machine learning · Ordinal classification · Ordinal regression · Label smoothing · Observational data · Unimodal distribution

1 Introduction

Our starting point is rating data, where cases $\boldsymbol{x}$ are associated with a score or rating y, typically taken from an ordinal scale. In credit scoring, for example, a customer's credit worthiness could be rated on the scale $\mathcal{Y} = \{\text{poor, fair, good, very good, excellent}\}$; similar examples can be found in finance [10,16] or medicine [6,18]. Our real-world example, to which we will return later on in the experimental part, is the assessment of goodwill requests by a car manufacturer, where a human goodwill after-sales expert decides about the percentage of the

G. De Francisci Morales et al. (Eds.): ECML PKDD 2023, LNAI 14174, pp. 3–18, 2023.
https://doi.org/10.1007/978-3-031-43427-3_1

labor and parts cost contributions the manufacturer is willing to pay. In our case, the decision is a contribution between 0 and 100%, in steps of 10%, i.e., $\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$ — note that this scale is somewhat in-between cardinal and ordinal, and could in principle be treated either way.

From a machine learning (ML) perspective, rating data has (at least) two interesting properties. First, ML models learned on such data are *prescriptive* rather than predictive in nature [11]. In particular, given a case $\boldsymbol{x}$, there is arguably nothing like a *ground-truth* rating y. At best, a rating could be seen as fair from the point of view of a customer, or opportune from the point of view of a manufacturer. For machine learning, the problem is thus to learn a prescriptive model that stipulates "appropriate" ratings or actions to be taken to achieve a certain goal, rather than a predictive model targeting any ground-truth.

Second, rating data is often biased in various ways. This is especially true for observational data where labels or ratings are coming from human experts and may be geared towards the expert's preferences and views. For example, the distribution of ratings in our goodwill use case (cf. Fig. 1) clearly shows a kind of "rounding effect": Experts prefer ratings of 0%, 50%, and 100%; ratings in-between (20% or 30%, 70% or 80%) are still used but much less, while values close to these preferred ones, such as 10% or 90%, are almost never observed — presumably, these "odd" ratings are rounded to the closest "even" ratings. Consequently, such data should not necessarily be taken as a gold standard. On the contrary, it might be sub-optimal and may not necessarily suggest the best course of action to be taken in a given context.

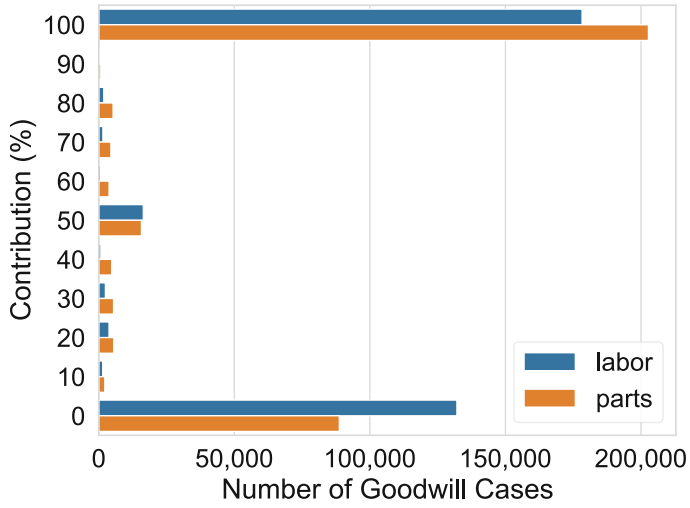


Fig. 1. Distribution of goodwill contributions for labor and parts at the car manufacturer.

To tackle this problem, our idea is to "weaken" the rating data through reallocation, turning a deterministic observation y into a (probability) distribution on $\mathcal{Y}$; this idea is inspired by a technique known as *label smoothing* [19]. For example, an observed rating of 50% could be replaced by a distribution assigning probabilities of 0.05, 0.2, 0.5, 0.2, 0.0.5, respectively, to 30%, 40%, 50%, 60%,

and 70%, suggesting that the actually most appropriate rating is not necessarily 50%, but maybe another value close by. Learning from such data can be seen as a specific form of *weakly supervised learning* [22].

More concretely, we propose a novel label smoothing approach based on the geometric distribution, which, compared to previous methods (cf. Sect. 2), enables more transparent and flexible re-distribution of probability mass. The approach is specifically tailored to probabilistic prescriptive ordinal classification, where a high degree of model configurability is required to correct bias in observational data, surpassing regularization aspects of previous methods by far. Our contributions can be summarized as follows:

- **Novel unimodal smoothing method:** In Sect. 3, we introduce our new unimodal label smoothing method. We first outline the basic smoothing approach and then extend it to a smoothing-relation based approach. This allows for flexible class-wise re-distribution of probability mass to inject domain knowledge into the standard cross-entropy loss.
- **Practically motivated heuristics:** Additionally, we present two heuristic smoothing functions to deal with common issues in observational data, namely concept drift and data imbalance (cf. Sect. 3.4).
- **Application to a real world automated decision making (ADM) use-case:** In Sect. 4, we apply and evaluate our proposed methods on the aforementioned use-case. To this end, we leverage real-world observational goodwill assessment data sets of a car manufacturer.

2 Related Work

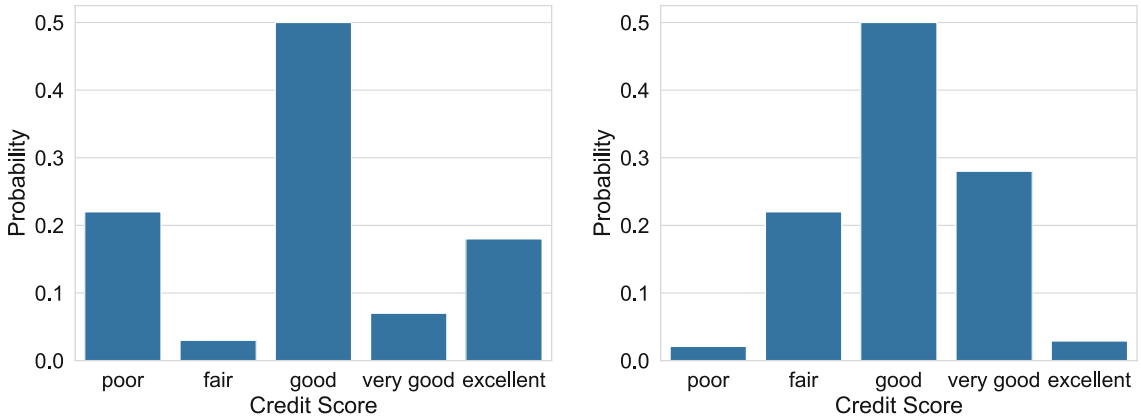
So-called *label smoothing* is a popular method in machine learning, specifically in deep learning [15,19], which is meant to reduce overconfidence in one-hot encoded (0/1) deterministic labels, thereby serving as a kind of regularizer and preventing the learner from over-fitting the (possibly noisy) training data. Label smoothing removes a certain amount of probability mass from the observed label and spreads it uniformly across the classes. That is, an observation $(\boldsymbol{x}_i, y_i)$ is turned into a training example $(\boldsymbol{x}_i, p_i^{LS})$, where p_i^{LS} is a probability distribution on $\mathcal{Y}$:

$$p_i^{LS}(k) = (1-\alpha)y_{i,k} + \alpha\frac{1}{K}\,,$$

with $K = |\mathcal{Y}|$ the number of classes, $y_{i,k} = 1$ for the observed class and $= 0$ otherwise, and $\alpha \in (0,1)$ a smoothing parameter. *Label relaxation* is a generalization of label smoothing, in which the single smoothed distribution is replaced by a larger set of candidate distributions [12]. While a uniform distribution of probability mass is a meaningful strategy for standard (nominal) classification, where classes have no specific order, this is arguably less true for *ordinal classification*, also called *ordinal regression* in statistics [9,17], where classes have a

linear order: $y_1 \prec y_2 \prec \cdots \prec y_K$. In this setting, one may rather expect a *unimodal* distribution of the classes, where the observed label is the single mode of the distribution, and classes closer to the mode are considered more likely than classes farther away. In ordinal classification, unimodality is not only a natural property for smoothing, but of course also for prediction [1,2,4,5]; see Fig. 2 for an illustration.

Liu et al. [13] propose to use the Binomial and Poisson distribution to redistribute the probability mass of one-hot encoded (0/1) labels in a unimodal fashion. However, the authors admit that both distributions are problematic: In the case of Poisson, it is not easy to flexibly adjust the shape, and for the Binomial distribution, it is difficult to flexibly adjust the position of the peak and the variance. Therefore, they propose another smoothing function $e^{\frac{-|k-j|}{\tau}}$ based on the exponential function, followed by a softmax normalization to turn the result into a discrete probability distribution on $\mathcal{Y}$. Here, $\tau > 0$ is a smoothing factor that determines the "peakedness" of the function, j the index of the observed class in the one-hot encoded label y_i (where the value is 1) and k the k-th class. However, how much probability mass is assigned to the mode and the rest of the classes is not transparent and might require significant experimentation effort. Vargas et al. propose unimodal smoothing methods based on the continuous Beta and Triangular distributions [20,21] where parameters need to be pre-calculated upfront depending on the current class and the overall number of classes. The Binomial and Poisson distribution have previously also been used to constrain the output of neural networks to unimodailty, where their usage appears more natural than for label smoothing. For instance, Beckham and Pal [2] use the Binomial and Poisson distributions as the penultimate layer in a deep neural network to constrain the output to unimodality before sending it through



(a) Multimodal distribution of credit scoring probabilities.

(b) Unimodal distribution of credit scoring probabilities.

Fig. 2. Exemplary multimodal (left) and unimodal (right) output distributions of credit scoring probabilities. The multimodal distribution on the left appears unnatural since the data underlies a natural order. One would rather expect a monotonic decrease of probability from the *mode* of the distribution, like it is shown on the right.

a final softmax layer. A quite similar approach was previously proposed by da Costa et al. [4,5].

3 Unimodal Label Smoothing Based on the Geometric Distribution

In the following, we introduce our novel unimodal label smoothing approach based on the geometric distribution. We begin with a motivation, explaining why smoothing degenerate one-point distributions is meaningful, especially in the setting of prescriptive ML primarily dealing with observational data.

3.1 Motivation

As already mentioned previously, our focus is on prescriptive probabilistic ordinal classification, where past observations are given in the form of data

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

with $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^m$ a *feature vector* characterizing a case, and $y_i \in \mathcal{Y}$ the corresponding *label* or observed rating. The set of class labels has a natural linear order: $y_1 \prec y_2 \prec \cdots \prec y_K$. In standard (probabilistic) supervised learning, the goal is then to learn a probabilistic predictor $\hat{p} : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ that performs well in terms of a loss (error) function $l : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \to \mathbb{R}_+$, and training such a predictor is guided by (perhaps regularized variants of) the empirical risk

$$R(\hat{p}) := \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{p}(\boldsymbol{x}_i))$$

as an estimate of the true generalization performance. The de-facto standard loss function for nominal probabilistic multi-class classification is the cross-entropy loss

$$H(y_i, \hat{p}_i) = -\sum_{k=1}^{K} p(y_i = k|\boldsymbol{x}_i) \log(\hat{p}(y_i = k|\boldsymbol{x}_i))\,,$$

where class labels are one-hot encoded as degenerate one-point distributions $p_i \in \mathbb{P}(\mathcal{Y})$ with $p_i(y_i|\boldsymbol{x}_i) = 1$ and $p_i(y|\boldsymbol{x}_i) = 0$ for $y \neq y_i$.

Since all classes apart from the ground-truth or observed label are set to zero, the cross-entropy loss then boils down to log-loss

$$H(y_i, \hat{p}_i) = -\log(\hat{p}(y_i|\boldsymbol{x}_i)).$$

Obviously, this only makes sense if the labels can be considered incontestable ground truth. Since this is not warranted to that extend in ordinal observational data, replacing this degenerate one point distributions with more realistic

smoothed unimodal *surrogate* distributions p^S is required to prevent the before shown degeneration of the cross-entropy loss.

$$H(p_i^S, \hat{p}_i) = -\sum_{k=1}^{K} p^S(y_i = k|\boldsymbol{x}_i) \log(\hat{p}(y_i = k|\boldsymbol{x}_i))$$

Furthermore, surrogate distributions may even serve to correct wrong inflationary decisions or inject domain knowledge into the learning process, which is a requirement in prescriptive ML scenarios and at the heart of this paper.

3.2 Basic Unimodal Label Smoothing

The geometric distribution models the probability that the k-th trial is the first success for a given success probability θ and trials $k \in \{1, 2, 3, \ldots\}$.

$$p(k) = (1 - \theta)^{(k-1)}\theta$$

Due to its monotonically decreasing curve, it's well suited to model an unimodal probability distribution. The shape of the distribution hereby heavily depends on the "success" probability θ. We may think of the original label of a training instance as the success probability θ and the future mode of our new unimodal distribution. The more probability mass we want to allocate to the original label of our training instance, the more peaked or degenerate the distribution will look like. In a standard scenario with one-hot encoded labels, the complete probability mass of 1 is initially assigned to the ground truth or observed label. To take away probability mass from the label, we introduce a smoothing factor $\alpha \in (0, 1)$. The probability assigned to the mode of the probability distribution is then defined as $(1 - \alpha)$ (cf. Eq. 1). The probability of the rest of the classes is modeled as a two-sided geometric distribution decreasing monotonically from the mode. Below is the raw, non-normalized version of our unimodal smoothing approach based on the geometric distribution with j as the index of the observed class in the one-hot encoded label y_i (where the value is 1):

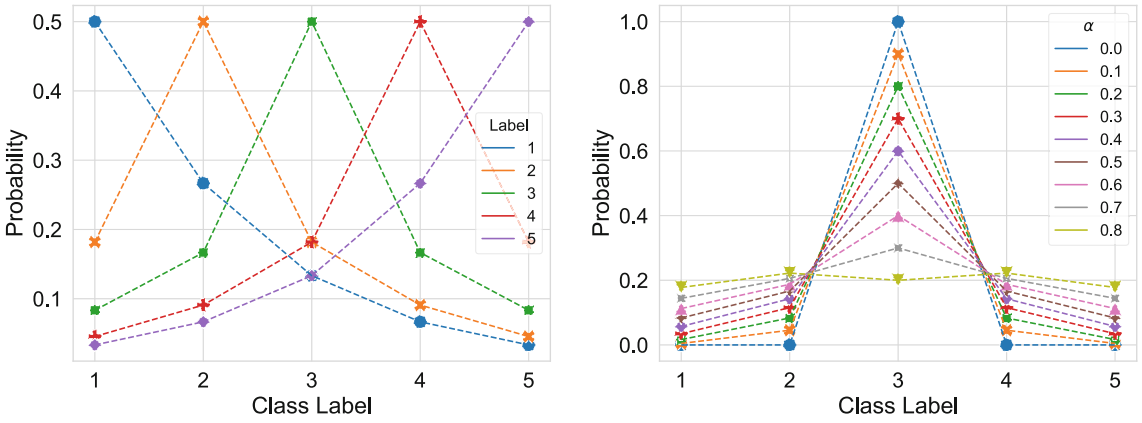
$$p_i^G(k) = \alpha^{|j-k|}(1 - \alpha) \tag{1}$$

Since the geometric distribution has infinite support we need to truncate and normalize it so that the probabilities sum to 1. We do this by introducing a normalizing constant G_i:

$$G_i = p_i^G(k \neq j) = \sum_{k \neq j} \alpha^{|j-k|}(1 - \alpha).$$

The normalized version of our smoothing approach with $\sum_{k=1}^{K} p_i^G(k) = 1$ then looks as follows:

$$p_i^G(k) = \begin{cases} 1 - \alpha & \text{if } k = j \\ 1/G_i \ \alpha^{|j-k|+1}(1 - \alpha) & \text{if } k \neq j \end{cases}.$$



(a) Smoothing of probability mass for different labels $y \in \{1, 2, 3, 4, 5\}$ and a fixed smoothing factor of $\alpha = 0.5$.

(b) Smoothing of probability mass for $y = 3$ and varying smoothing factors $\alpha \in \{0, 0.1, 0.2, 0.3, \ldots, 0.8\}$.

Fig. 3. The above figures illustrate our proposed unimodal smoothing approach based on the geometric distribution.

Note that we do not normalize the mode of the distribution $(1 - \alpha)$ since we want to keep transparent how much probability mass is allocated to the ground truth or observed label. Figures 3a and 3b illustrate how our unimodal smoothing approach based on the geometric distribution looks like for different classes and different smoothing factors respectively.

3.3 Class-Wise Unimodal Label Smoothing Using a Smoothing Relation

The basic smoothing approach presented in the previous subsection does not distinguish between classes and all classes are smoothed the same. Furthermore, it assumes a rather symmetric smoothing where probability mass is distributed to the left and right side of the mode (if possible). To achieve a higher degree of configurability in terms of smoothing, we introduce a so called *smoothing relation* (cf. table 1) that allows to define how strong the label or mode is smoothed per observed class index j (α_j), as well as the fraction of outstanding probability mass that is supposed to be distributed to the left $(F_{l,j})$ and right $(F_{r,j})$ of the mode, with $F_{l,j} + F_{r,j} = 1$. An extended smoothing function allowing class-wise smoothing based on a smoothing relation (cf. Table 1) is displayed below:

$$p_i^G(k) = \begin{cases} 1 - \alpha_j & \text{if } k = j \\ 1/G_i \; F_{l,j} \; \alpha_j^{(j-k)+1}(1 - \alpha_j) & \text{if } k < j \\ 1/G_i \; F_{r,j} \; \alpha_j^{(k-j)+1}(1 - \alpha_j) & \text{if } k > j \end{cases}, \quad \text{with } F_{l,j} + F_{r,j} = 1$$

Table 1. Two exemplary smoothing-relations to configure unimodal re-distribution of probability mass.

j	1	2	3	4	5
α	0	0.2	0.3	0.4	0.5
F_l	0	1	1	1	1
F_r	0	0	0	0	0

(a) Cautious smoothing relation.

j	1	2	3	4	5
α	0.5	0.4	0.3	0.2	0
F_l	0	0	0	0	0
F_r	1	1	1	1	0

(b) Generous smoothing relation.

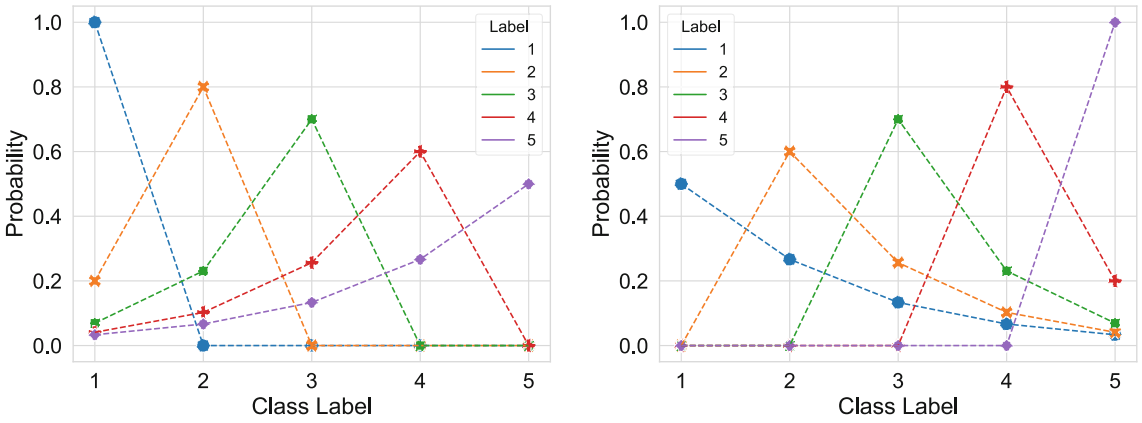
The adapted normalization constant G_j then looks as follows:

$$G_i = \begin{cases} \sum_{k>j} \alpha_j^{(k-j)}(1-\alpha_j) & \text{if } F_l, j = 0 \\ \sum_{k<j} \alpha_j^{(j-k)}(1-\alpha_j) & \text{if } F_r, j = 0 \\ \sum_{k\neq j} \alpha_j^{|j-k|}(1-\alpha_j) & \text{otherwise} \end{cases}.$$

In this case, one can particularly define how much of the outstanding probability mass is assigned left or right of the mode. This, in the extreme case, even enables unimodal one-sided label smoothing by distributing probability mass only to one side. This extreme scenario is shown in Table 1, where in the left smoothing-relation smoothing is only performed to the left side of the mode, with increasing α and in the right smoothing-relation only to the right, with decreasing α. Smoothing only to the left side of the mode indicates a more cautious smoothing, for instance, in our credit scoring example, probability mass is then re-distributed from higher ratings to lower ratings. The other way round, smoothing to the right indicates a more generous approach, where probability mass is re-distributed from lower ratings to higher ratings. Hence, through using this approach, probability mass can be flexibly re-distributed to correct any biases in the underlying observational data, e.g., too cautious or generous credit rating assessments in the past. Figure 4 shows the smoothing curves for the cautious (Fig. 4a) respectively generous smoothing (Fig. 4b).

3.4 Unimodal Smoothing Heuristics for Prescriptive Machine Learning

The basic smoothing approach outlined in Subsect. 3.2 smooths the distribution for every class the same, which may be a too simplified assumption. In contrast, the smoothing relation approach introduced in Subsect. 3.3 provides more flexibility, but on the other side also requires detailed knowledge about present biases and the domain. Hence, in the following we want to look at two generally applicable smoothing heuristics to deal with two common issues in observational data: data imbalance and concept drift.



(a) Left-tailed ("cautious") class-wise smoothing.

(b) Right-tailed ("generous") class-wise smoothing.

Fig. 4. Class-wise left- or right-tailed smoothing using the smoothing relations in Table 1.

Unimodal Smoothing Based on Class Priors. Observational data is often strongly imbalanced (cf. Fig. 1) [10]. From a prescriptive ML point of view, "correcting" over-proportionally used labels or ratings stronger through smoothing them more than infrequently used ones might be a reasonable correction. In observational data, the more inflationary a rating is used, the less meaningful it appears. The other way round, one may assume that rare ratings are selected more carefully and more thought might have been put into their selection. Moreover, this may also counteract class imbalance as the probability mass assigned to inflationary used ratings is reduced compared to more seldomly used ratings.

Hence, a simple smoothing heuristic may be to vary the smoothing factor depending on the class prior.

$$s_i(\alpha) = \alpha \cdot \frac{p(y_i)}{\max\limits_{y \in \{y_1, y_2, \ldots, y_K\}} p(y)} \in [0, \alpha]$$

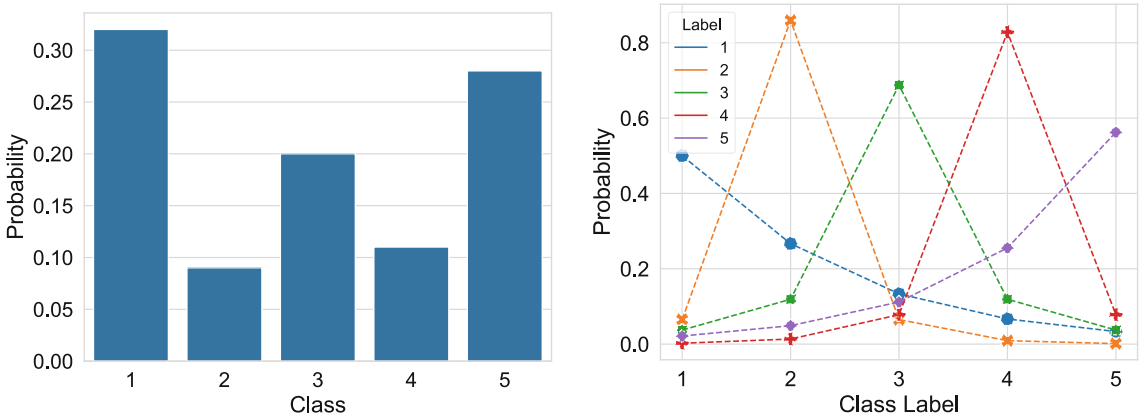
The single smoothing factor α is hereby replaced by a smoothing function $s_i(\alpha)$ depending on α and the class prior $p(y_i)$ normalized by the max class prior. The equation below shows the adapted unimodal smoothing approach dependent on prior class probabilities.

$$p_i^G(k) = \begin{cases} 1 - s_i(\alpha) & \text{if } k = j \\ 1/G_i \; s_i(\alpha)^{|j-k|+1}(1 - s_i(\alpha)) & \text{if } k \neq j \end{cases}$$

The normalizing constant G_i also needs to be updated accordingly:

$$G_i = p_i^G(k \neq j) = \sum_{k \neq j} s_i(\alpha)^{|j-k|}(1 - s_i(\alpha))$$

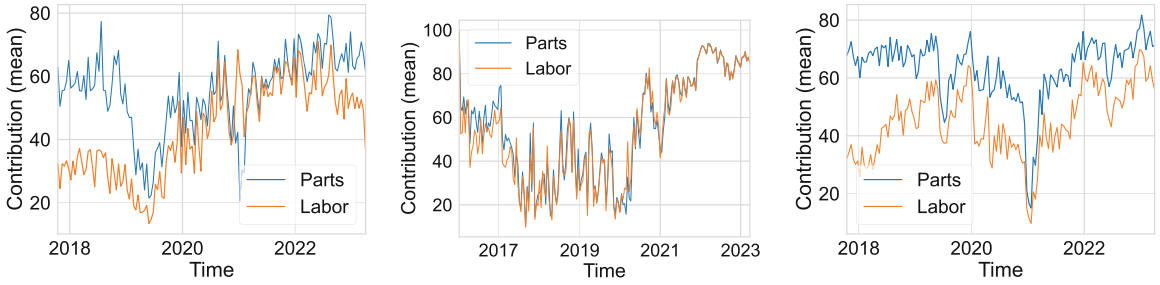
Figure 5 illustrates the class-wise smoothing approach on exemplary imbalanced class prior probabilities. As one can see, class 1 has the highest prior



(a) Exemplary imbalanced prior probabilities for a five classes problem.

(b) Class-wise smoothing of probability mass according to the priors on the left for different labels $y \in \{1, 2, 3, 4, 5\}$ and a fixed smoothing factor of $\alpha = 0.5$.

Fig. 5. Class-wise unimodal smoothing of probability mass depending on class priors.



(a) Market data set A. (b) Market data set B. (c) Market data set C.

Fig. 6. Goodwill decision mean values for parts and labor contributions over time, entailing concept drift and shift.

probability and is smoothed the most. Whereas class 2 has the lowest prior probability and is smoothed the least.

Time Based Unimodal Smoothing. Another very typical bias in observational data is concept drift or shift, where the target variable which a model tries to predict changes its statistical properties over time [14]. Typically, ratings conducted by human experts like credit rating assessments or candidate rating in human resources will not remain static over time. Strategies will change dynamically depending on market situations. This is also visible in our goodwill assessment data sets, where mean contribution ratings for labor and parts repair costs change dynamically over time for some markets (cf. Fig. 6).

Hence, we propose another simple linear smoothing function that will smooth older instances stronger than more recent ones, whereas $t_i \in T$ are time stamps

Table 2. Goodwill assessment data set sizes. All data sets have 26 features (18 categorical and 8 numeric) and a single label with 11 classes ($\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$).

Market	A	B	C	D
# Instances	17,652	27,390	43,286	13,832

accompanying each human expert rating:

$$s_i(\alpha) = \alpha \cdot \frac{\max\limits_{t \in T} t - t_i}{\max\limits_{t \in T} t - \min\limits_{t \in T} t} \in [0, \alpha]$$

4 Evaluation

In the following, we want to evaluate our proposed smoothing approaches on four ordinal real world goodwill assessment data sets of a car manufacturer (cf. Table 2), with the goal to predict appropriate monetary contributions for parts and labor repair costs on an ordinal scale from 0 to 100% ($\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$). The different data sets are taken from different national sales markets and reflect the different goodwill assessment strategies of the national sales companies (NSC) of the car manufacturer. At the moment, goodwill requests are to a large extend assessed manually by human experts [10]. However, the long term aim of the car manufacturer is to increase automation and process goodwill requests through automated decision making (ADM) [10]. The attributes of the data instances entail information about the vehicle and the case, for instance, vehicle age, mileage, requested costs, defect code, whether the vehicle was regularly serviced, etc. [10]. The data is of mid-size tabular nature which makes us rely on Gradient Boosted Trees (GBT) for our evaluation implementation [8]. Concretely, we make use of eXtreme Gradient Boosting (XGBoost) [3]. However, our proposed smoothing approaches are not limited to GBTs and could, for instance, also be used in a deep learning context. Table 2 summarizes some characteristics of the goodwill data sets used for evaluation. In general, the data sets are in most cases heavily imbalanced with mostly 0, 50 and 100% ratings (cf. Fig. 1). As shown in Fig. 6, data sets A, B and C also contain some sort of drift in the target ratings.

4.1 Relation and Priors Based Smoothing Results

To evaluate the flexibility of the smoothing-relation based approach on goodwill assessment data, we make use of the smoothing-relations shown in Table 3, which are similar to the previous five classes example (cf. Table 1), but expanded to 11 classes or ratings for the case of goodwill assessment. On max, 50% of the probability mass is re-distributed to other classes ($\alpha = 0.5$).

Table 3. Cautious (top) and generous (bottom) smoothing-relations for goodwill assessment with 11 classes ($\mathcal{Y} = \{0, 10, 20, \ldots, 100\}$).

j	1	2	3	4	5	6	7	8	9	10	11
α	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
F_l	0	1	1	1	1	1	1	1	1	1	1
F_r	0	0	0	0	0	0	0	0	0	0	0

(a) Cautious smoothing-relation.

j	1	2	3	4	5	6	7	8	9	10	11
α	0.5	0.45	0.4	0.35	0.3	0.25	0.2	0.15	0.1	0.05	0
F_l	0	0	0	0	0	0	0	0	0	0	0
F_r	1	1	1	1	1	1	1	1	1	1	0

(b) Generous smoothing-relation.

Tables 4 and 5 show the results of a ten-fold cross validation evaluation of the cautious (+Cautious) respectively generous (+Generous) smoothing-relation in relation to a standard nominal classification (Base). Additionally, we also display the results of the smoothing heuristic based on the class priors (+Priors) with a max smoothing factor of $\alpha = 0.5$. In all cases, we display the mean as well as the standard deviation ($\pm$) of the ten folds. The evaluated standard metrics are accuracy (ACC), mean absolute error (MAE) and mean squared error (MSE). Since we are dealing with an ordinal classification problem that lies somewhere between classification and regression, classification as well as regression metrics are of interest [7]. The underpay and overpay metrics are domain specific metrics relevant from a goodwill assessment perspective, as they indicate how much money was payed less (underpay), respectively more (overpay), compared to the manual human assessments. One can clearly see that the cautious, respectively generous, smoothing-relations are reflected in the results. In case of the cautious strategy there is a strong tendency of underpayment, whereas in case of the generous strategy there is a strong tendency for overpayment. The class priors based smoothing approach trades-off accuracy for improved MAE and MSE metrics, which can be considered beneficial in ordinal classification.

4.2 Time Based Smoothing Results

In the time-based smoothing evaluation, we set the smoothing factor to $\alpha = 0.8$, which is a rather aggressive value that leads to almost uniform re-distribution of probability mass for the oldest training instances. The data is split into training and test data with a ratio of 90/10, whereas the test data entails the most recent 10% of the data. The small test data set size of 10% was chosen to use the same amount of data for testing as in the experiments performed above which used 10-fold cross validation and, even more important, to specifically focus on very

Table 4. Results of smoothing for labor contributions of different goodwill assessment data sets with a max smoothing factor of $\alpha = 0.5$.

Model	ACC	MAE	MSE	UNDERPAY	OVERPAY
Base	**0.906** ±0.008	6.76 ±0.61	586.85 ±57.78	−20,810.34 ±3,987.55	22,220.51 ±3,486.62
+Generous	0.902 ±0.007	7.06 ±0.56	615.11 ±53.74	**−11,788.37** ±2,630.4	34,593.04 ±4,824.63
+Cautious	0.904 ±0.007	6.94 ±0.52	606.49 ±46.65	−30,887.52 ±5,159.07	**14,395.95** ±1,757.34
+Priors	0.905 ±0.009	**6.74** ±0.74	**581.03** ±68.85	−17,901.34 ±6001.66	25,858.72 ±3,534.84
Base	**0.89** ±0.007	6.7 ±0.51	521.91 ±45.95	−200,513.25 ±22,181.56	555,628.97 ±110,741.03
+Generous	0.885 ±0.005	6.96 ±0.45	545.08 ±41.78	**−112,300.64** ±24,587.5	712,084.73 ±129,923.33
+Cautious	0.887 ±0.005	6.9 ±0.41	542.97 ±38.37	−365,626.9 ±47,568.54	**402,926.56** ±91,423.83
+Priors	0.886 ±0.003	**6.63** ±0.3	**505.23** ±28.57	−224,039.44 ±28,151.06	511,294.98 ±109,859.14
Base	**0.933** ±0.004	**4.53** ±0.33	**380.53** ±31.07	-34,406.94 ±5629.54	52,374.52 ±6,396.72
+Generous	0.929 ±0.003	4.81 ±0.31	409.64 ±29.87	**−19,201.82** ±3,462.9	72,995.39 ±6,103.36
+Cautious	0.93 ±0.003	4.87 ±0.3	416.79 ±30.74	−52,842.93 ±8,453.54	**38,187.68** ±5,016.32
+Priors	0.93 ±0.004	4.68 ±0.37	393.6 ±35.41	−27,316.0 ±3,750.45	59,746.98 ±4,594.85
Base	**0.862** ±0.007	7.93 ±0.59	580.46 ±56.44	−153,618.28 ±35,419.88	345,107.29 ±66857.93
+Generous	**0.862** ±0.01	7.88 ±0.62	575.39 ±53.25	**−62,970.68** ±24,501.17	415,654.6 ±83243.69
+Cautious	0.859 ±0.008	7.93 ±0.67	578.67 ±63.08	−222,597.46 ±29955.5	**270,679.0** ±53,456.44
+Priors	**0.862** ±0.008	**7.7** ±0.61	**554.95** ±55.18	−105,230.72 ±30,140.38	371,125.59 ±58,349.45

Table 5. Results of smoothing for parts contributions of different goodwill assessment data sets with a max smoothing factor of $\alpha = 0.5$.

Model	ACC	MAE	MSE	UNDERPAY	OVERPAY
Base	**0.896** ±0.009	6.98 ±0.82	579.34 ±78.5	−31,744.94 ±9306.37	84,160.92 ±21,575.87
+Generous	0.892 ±0.008	7.16 ±0.7	594.27 ±66.21	**−15,378.27** ±5,296.54	102,037.97 ±19,866.94
+Cautious	0.895 ±0.006	7.13 ±0.63	598.47 ±61.55	−45,414.86 ±19,132.45	**72,023.1** ±14,395.4
+Priors	0.895 ±0.008	**6.96** ±0.7	**575.55** ±66.55	−36,270.52 ±15,750.53	78,736.36 ±20,136.15
Base	**0.894** ±0.006	6.24 ±0.35	477.5 ±31.8	−430,146.61 ±108977.81	1,122,151.24 ±176,516.56
+Generous	0.891 ±0.005	6.38 ±0.33	491.79 ±31.37	**−217,544.71** ±74,535.81	1,367,306.62 ±187,239.83
+Cautious	**0.894** ±0.005	6.22 ±0.31	481.64 ±30.41	−640,450.7 ±177,181.13	**894,571.07** ±166,891.27
+Priors	0.892 ±0.005	**6.1** ±0.24	**456.87** ±21.91	−546,358.25 ±165,084.6	967,882.1 ±143,394.63
Base	**0.884** ±0.003	4.24 ±0.16	243.91 ±12.32	−67,451.36 ±8,535.11	219,066.1 ±34,154.76
+Generous	0.882 ±0.006	4.28 ±0.21	247.98 ±13.92	**−38,309.93** ±6,165.02	245,536.55 ±37,405.61
+Cautious	0.883 ±0.005	**4.19** ±0.2	241.42 ±13.77	−83,698.36 ±15,123.86	**197,573.78** ±39,719.46
+Priors	**0.884** ±0.004	**4.19** ±0.16	**239.42** ±11.72	−63,682.4 ±8,578.55	217,069.17 ±35,658.62
Base	**0.87** ±0.009	7.16 ±0.57	514.27 ±53.2	−264,672.6 ±110,169.93	765,250.06 ±163,638.98
+Generous	0.867 ±0.007	7.38 ±0.52	534.6 ±51.22	**-128,742.79** ±41,807.12	890,592.62 ±174,359.46
+Cautious	0.866 ±0.006	7.28 ±0.45	523.84 ±39.17	−54,350.41 ±109,682.42	**631,753.19** ±123,490.03
+Priors	0.869 ±0.006	**7.03** ±0.53	**498.47** ±51.21	−222,986.36 ±67,758.96	778,818.12 ±160,019.91

recent data. For this evaluation, we focus on the three data sets that visually entail some sort of drift in the target rating over time (cf. figure 6). Tables 6 and 7 summarize the obtained time based smoothed results (+Time) for labor and parts contributions respectively in comparison to a nominal classification baseline (Base). One can clearly see that the time based smoothing approach increases the predictive performance of the models on our data sets for the majority of our metrics.

Table 6. Results of time based smoothing (+Time) compared to standard nominal classification (Base) for labor contributions of different goodwill assessment data sets ($\alpha = 0.8$).

Model	ACC	MAE	MSE	UNDERPAY	OVERPAY
Base	0.888	8.95	832.46	−20,166.82	31,891.31
+Time	**0.89**	**8.93**	**831.73**	**−15,223.71**	**24,952.21**
Base	0.827	8.92	668.35	−145,360.16	844,359.37
+Time	**0.829**	**8.45**	**633.88**	**−137,898.47**	**799,428.2**
Base	0.951	2.97	240.73	−31,880.59	29,368.8
+Time	**0.952**	**2.87**	**232.61**	**−28,603.83**	**27,001.83**

Table 7. Results of time based smoothing (+Time) compared to standard nominal classification (Base) for parts contributions of different goodwill assessment data sets ($\alpha = 0.8$).

Model	ACC	MAE	MSE	UNDERPAY	OVERPAY
Base	**0.845**	10.02	839.32	−72,368.92	**117,635.19**
+Time	0.844	**9.98**	**829.63**	**−58,450.89**	154,781.0
Base	**0.834**	8.73	655.28	−312,151.14	1,260,827.89
+Time	0.828	**8.57**	**643.12**	**−242,015.57**	**1,121,527.11**
Base	0.919	2.75	**146.03**	−85,673.42	**110,494.84**
+Time	**0.924**	**2.68**	147.0	**−75,981.27**	122,821.94

5 Conclusion

In this paper, we presented a novel unimodal label smoothing approach with the aim to rectify bias in ordinal observational data. We have demonstrated the effectiveness of the approach for the use case of automotive goodwill assessment. Through the usage of different smoothing-relations we can flexibly configure our models to be more cautious, respectively generous, with regards to goodwill assessments which is clearly indicated in strong underpayment, respectively strong overpayment, in comparison to a nominal classification baseline. The class priors based smoothing heuristic corrects inflationary used ratings through smoothing them stronger than less frequently used ratings which manifests in reduced MAE and MSE metrics compared to the baseline. Time based smoothing helps to reduce concept drift bias and outperforms standard nominal classification on the majority of our evaluated metrics. Overall we can say that, our proposed methods are effective and flexible tools to correct biased expert ratings and reduce reliance on human expertise.

References

1. Albuquerque, T., Cruz, R., Cardoso, J.S.: Quasi-unimodal distributions for ordinal classification. Mathematics **10**(6), 980 (2022)
2. Beckham, C., Pal, C.J.: Unimodal probability distributions for deep ordinal classification. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 411–419. PMLR (2017)
3. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016. pp. 785–794. ACM (2016)
4. da Costa, J.F.P., Alonso, H., Cardoso, J.S.: The unimodal model for the classification of ordinal data. Neural Netw. **21**(1), 78–91 (2008)
5. da Costa, J.P., Cardoso, J.S.: Classification of ordinal data using neural networks. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 690–697. Springer, Heidelberg (2005). https://doi.org/10.1007/11564096_70
6. Durán-Rosal, A.M., et al.: Ordinal classification of the affectation level of 3D-images in Parkinson diseases. Sci. Rep. **11**(1), 1–13 (2021)
7. Gaudette, L., Japkowicz, N.: Evaluation methods for ordinal classification. In: Gao, Y., Japkowicz, N. (eds.) AI 2009. LNCS (LNAI), vol. 5549, pp. 207–210. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01818-3_25
8. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data? arXiv preprint arXiv:2207.08815 (2022)
9. Gutiérrez, P.A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., Hervas-Martinez, C.: Ordinal regression methods: survey and experimental study. IEEE Trans. Knowl. Data Eng. **28**(1), 127–146 (2015)
10. Haas, S., Hüllermeier, E.: A prescriptive machine learning approach for assessing goodwill in the automotive domain. In: Amini, M., Canu, S., Fischer, A., Guns, T., Novak, P.K., Tsoumakas, G. (eds.) Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 13718, pp. 170–184. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-26422-1_11
11. Hüllermeier, E.: Prescriptive machine learning for automated decision making: challenges and opportunities. arXiv preprint arXiv:2112.08268 (2021)
12. Lienen, J., Hüllermeier, E.: From label smoothing to label relaxation. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2–9, 2021, pp. 8583–8591. AAAI Press (2021)
13. Liu, X., et al.: Unimodal regularized neuron stick-breaking for ordinal classification. Neurocomputing **388**, 34–44 (2020)
14. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G.: Learning under concept drift: a review. IEEE Trans. Knowl. Data Eng. **31**(12), 2346–2363 (2019)
15. Lukasik, M., Bhojanapalli, S., Menon, A.K., Kumar, S.: Does label smoothing mitigate label noise? In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 6448–6458. PMLR (2020)

16. Manthoulis, G., Doumpos, M., Zopounidis, C., Galariotis, E.: An ordinal classification framework for bank failure prediction: methodology and empirical evidence for US banks. Eur. J. Oper. Res. **282**(2), 786–801 (2020)
17. McCullagh, P.: Regression models for ordinal data. J. Royal Stat. Soc.: Ser. B (Methodological) **42**(2), 109–127 (1980)
18. Pérez-Ortiz, M., Cruz-Ramírez, M., Ayllón-Terán, M.D., Heaton, N., Ciria, R., Hervás-Martínez, C.: An organ allocation system for liver transplantation based on ordinal regression. Appl. Soft Comput. **14**, 88–98 (2014)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 2818–2826. IEEE Computer Society (2016)
20. Vargas, V.M., Gutiérrez, P.A., Barbero-Gómez, J., Hervás-Martínez, C.: Soft labelling based on triangular distributions for ordinal classification. Information Fusion (2023)
21. Vargas, V.M., Gutiérrez, P.A., Hervás-Martínez, C.: Unimodal regularisation based on beta distribution for deep ordinal regression. Pattern Recogn. **122**, 108310 (2022)
22. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Sci. Rev. **5**(1), 44–53 (2018)

4.5 Uncertainty Quantification in Ordinal Classification: A Comparison of Measures

Contributing Article

Stefan Haas and Eyke Hüllermeier. "Uncertainty quantification in ordinal classification: A comparison of measures". In: *Int. J. Approx. Reason.* 186 (2025), p. 109479

Author Contribution Statement

The author developed the foundational idea of the desired properties for a good uncertainty measure in ordinal probabilistic classification and proposed the use of so-called consensus measures, which are widely used for assessing Likert-scale surveys in the social sciences. Additionally, Prof. Dr. Hüllermeier suggested a binary decomposition method that enables common measures, such as entropy, to be adapted into an ordinal uncertainty measure. The author carried out the proofs demonstrating that the ordinal binary decomposition method possesses the same properties as consensus measures and can be regarded as a generic form of certain consensus measures. The experimental comparisons of the measures were implemented and evaluated by the author. The entire paper underwent multiple rounds of revision by both Prof. Dr. Hüllermeier and the author.

International Journal of Approximate Reasoning 186 (2025) 109479

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Uncertainty quantification in ordinal classification: A comparison of measures

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

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

ARTICLE INFO

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

ABSTRACT

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

1. Introduction

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

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

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

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

Available online 28 May 2025

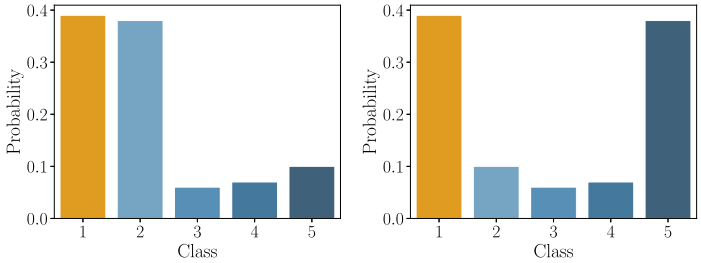


Fig. 1. Two very different distributions sharing the same Shannon entropy $H = 1.32$. In contrast, variance detects the higher dispersion on the right ($V = 3.25$) compared to the left ($V = 1.62$).

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

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

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

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

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

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

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

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

Our contributions can be summarized as follows:

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

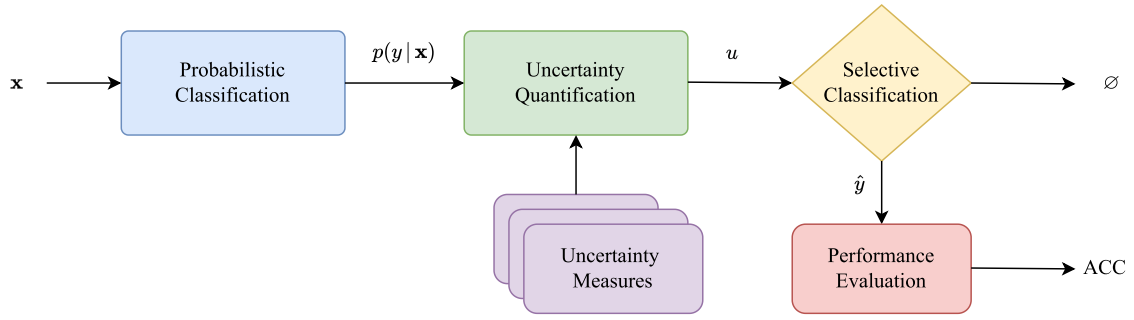


Fig. 2. Different uncertainty measures are evaluated for their ability to quantify uncertainty of predictions $p(y \mid \boldsymbol{x})$ over a set of (test) instances $\boldsymbol{x}$. The performance of these measures is assessed in a downstream selective classification task, where the learner abstains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

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

2. Learning probabilistic predictors

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. Uncertainty quantification for probabilistic predictors

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Measuring consensus, polarization and agreement in ordinal data

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

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

4.1. Properties of ordinal measures

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

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

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

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

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

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

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

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

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

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

4.2. Ordinal measures

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

4.2.1. The measure by Leik

We begin with Leik's measure of ordinal consensus [16], which computes the dispersion D as a measure of ordinal consensus for a probability (relative frequency) distribution $\boldsymbol{p}$ with K categories using the cumulative distribution $F_k(\boldsymbol{p}) = \sum_{1 \le t \le k} p_t$:

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

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

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

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

Formally, the following proposition can be shown.

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

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

4.2.2. The measure by Blair and Lacy

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

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

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

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

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

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

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

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

4.2.3. The measure by Tastle and Wierman

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

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

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

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

4.2.4. The measure by Van der Eijk

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

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

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

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

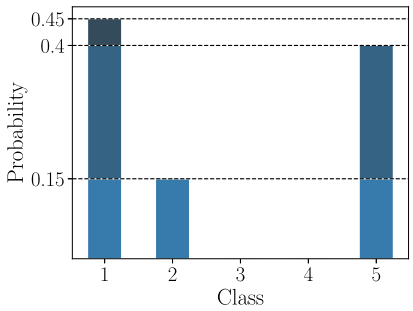


Fig. 3. Illustration of how Van der Eijk's measure of agreement reduces a probability distribution $\boldsymbol{p} = (0.45, 0.15, 0.0, 0.0, 0.4)$ horizontally into different layers based on the difference between the k-th and $(k-1)$-th smallest probabilities $(p_{(k)} - p_{(k-1)})$. The overall level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

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

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

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

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

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

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

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

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

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

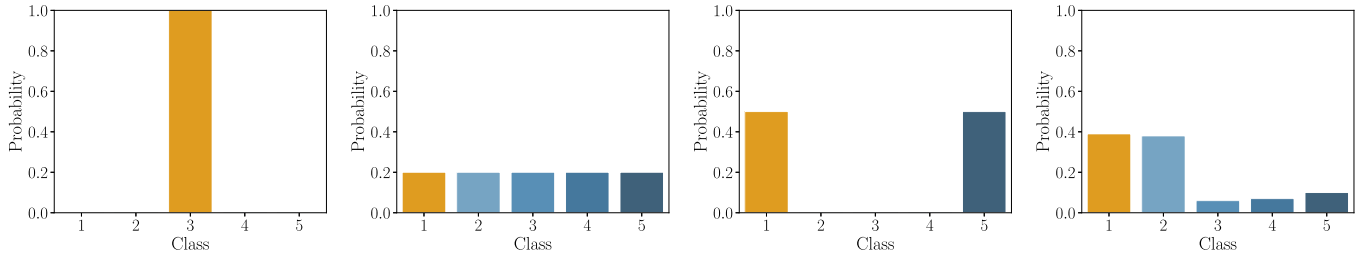
Formally, we can also show that the measure satisfies the axioms presented in Section 4.1.

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

4.3. The measure by Pavlopoulos and Likas

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

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

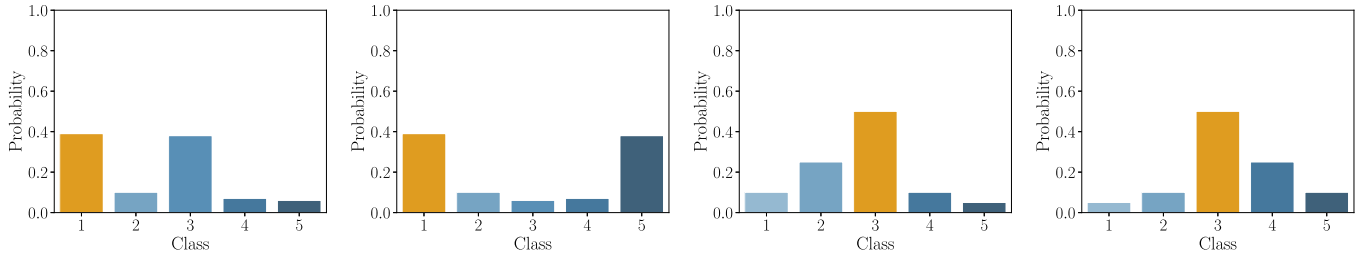


(a) $C_1 = 0.0$, $C_2 = 0.0$, Cns $= 0.0$, $C_A = 0.0$, DFU $= 0.0$

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

4.4. Ordinal uncertainty quantification using consensus measures

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

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

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

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

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

4.5. Variance

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

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

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

5. Binary decomposition for uncertainty quantification in ordinal classification

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

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

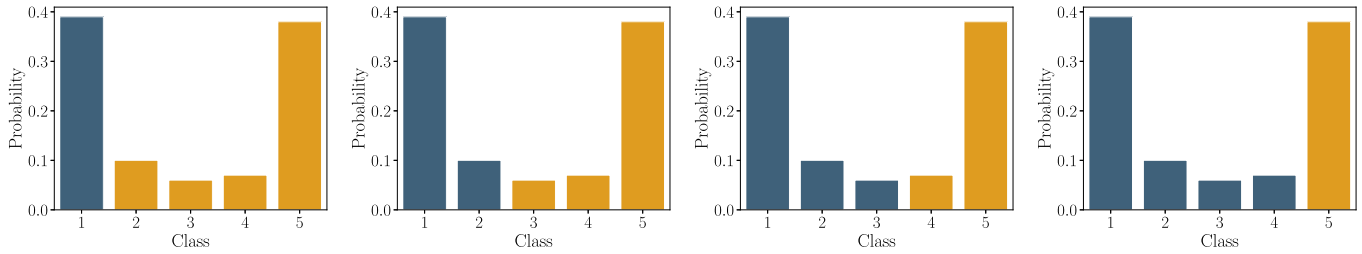
The principle of binary reduction can also be applied to uncertainty quantification [26]. In the ordinal case, it suggests a measure of the form

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6. Experiments with ordinal benchmark datasets

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

6.1. Choice of base learner and datasets

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

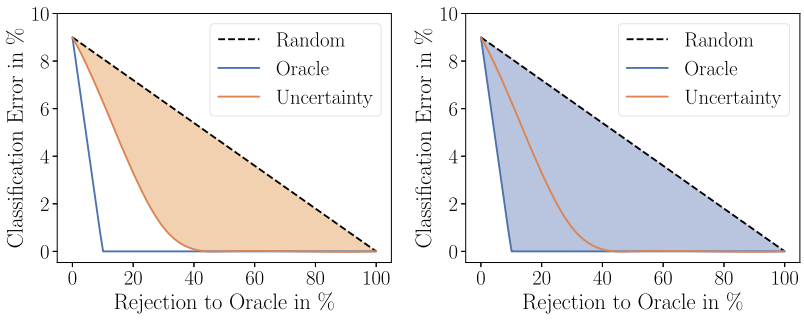
$$l_{CE}(\boldsymbol{y}, \boldsymbol{p}) = -\sum_{k=1}^{K} y_k \log(p_k), \tag{13}$$

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

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

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

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



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

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

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

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

6.2. Experimental setup

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

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

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

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

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

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

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

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

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

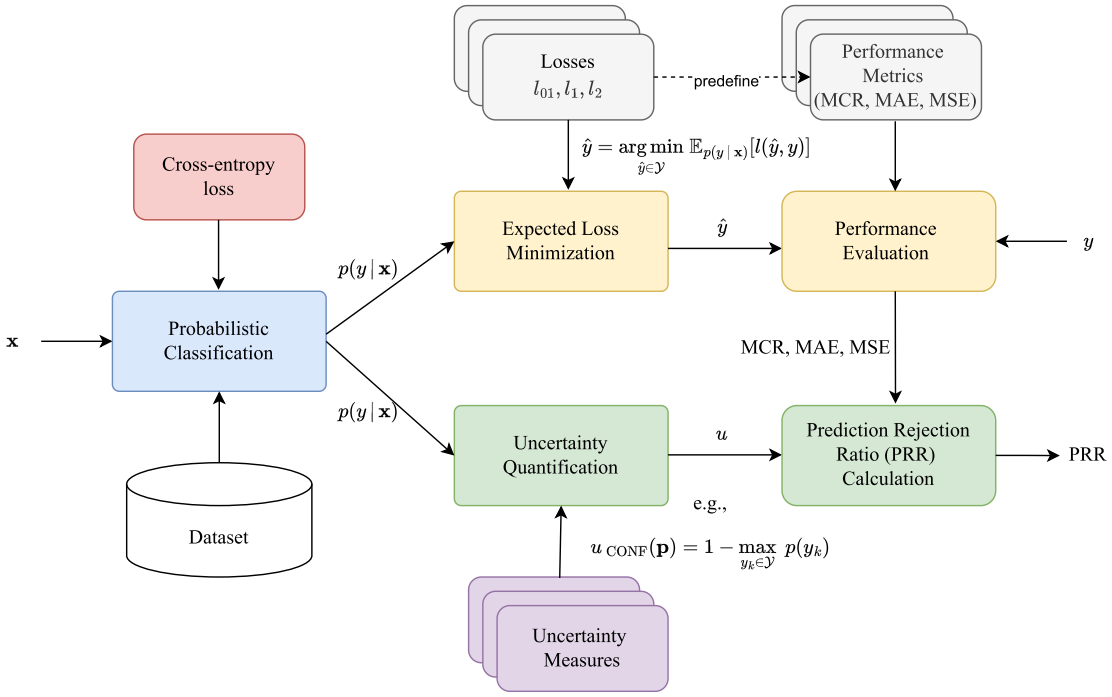


Fig. 7. Overview of the experimental approach: Final predictions $\hat{y}$ are derived using various Bayes estimators, while the predictive uncertainty u is quantified using different uncertainty measures. All these measures utilize the unbiased and realistic predictive probability distribution $\boldsymbol{p}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

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

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

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

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

6.3. Results and analysis

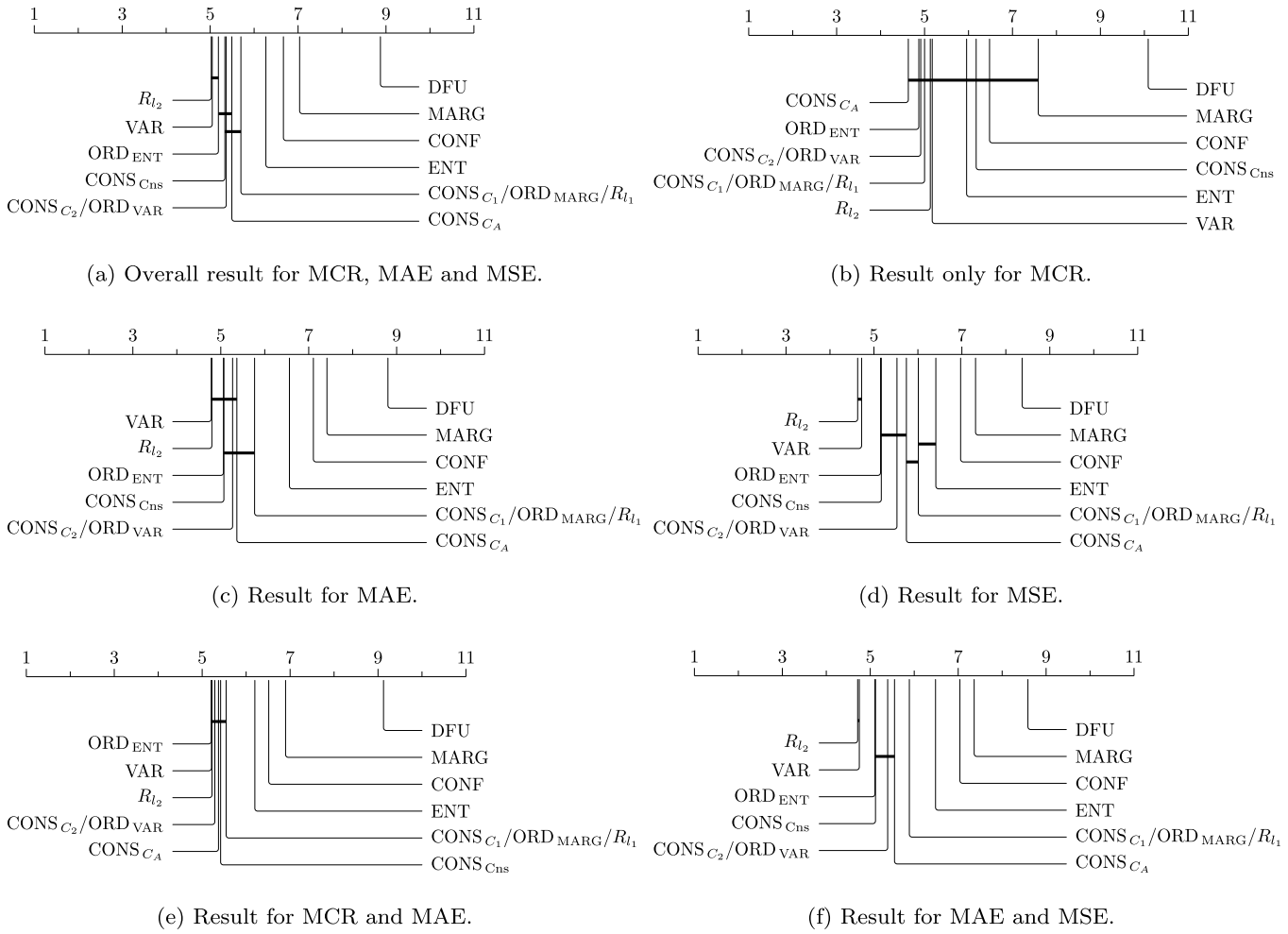
Table 3 displays the overall PRR results of a 10-fold cross validation on the selected ordinal benchmark datasets. In total, we evaluate fourteen uncertainty measures: CONF, MARG, ENT, VAR, CONS_{Cns} [18], CONS_{C_1} [16], CONS_{C_2} [17], ORD_{ENT}, ORD_{MARG}, ORD_{VAR}, R_{l_1}, R_{l_2}, CONS_{C_A}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

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

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

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

Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Gini}$	$CONS_{C_1}$	$CONS_{C_2}$	$CONS_{C_4}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Triazines	MCR	0.1368 ± 0.3073	0.1235 ± 0.2903	0.1788 ± 0.2762	0.1889 ± 0.2491	0.178 ± 0.26	0.1863 ± 0.2821	0.1666 ± 0.2632	0.1933 ± 0.2847	0.0669 ± 0.1725	0.1824 ± 0.271	0.1863 ± 0.2821	0.1666 ± 0.2632	0.1863 ± 0.2821	**0.2052** ± 0.2561
	MAE	0.1582 ± 0.3453	0.1411 ± 0.3361	0.238 ± 0.3072	0.3176 ± 0.2388	0.2896 ± 0.2725	0.2559 ± 0.3217	0.2531 ± 0.2891	0.2582 ± 0.3164	0.0427 ± 0.3001	0.276 ± 0.2974	0.2559 ± 0.3217	0.2531 ± 0.2891	0.2559 ± 0.3217	**0.3307** ± 0.2532
	MSE	0.1146 ± 0.3485	0.1287 ± 0.3558	0.206 ± 0.3239	0.3454 ± 0.2633	0.3133 ± 0.2826	0.2325 ± 0.3468	0.2483 ± 0.3128	0.229 ± 0.3475	-0.0017 ± 0.4853	0.2845 ± 0.327	0.2325 ± 0.3468	0.2483 ± 0.3128	0.2325 ± 0.3468	**0.3632** ± 0.2467
Machine CPU	MCR	0.7118 ± 0.1656	0.6856 ± 0.1807	0.7361 ± 0.1422	0.7976 ± 0.1446	0.7692 ± 0.1482	0.7626 ± 0.1573	0.775 ± 0.156	0.7814 ± 0.1505	0.5421 ± 0.3997	0.7846 ± 0.1356	0.7626 ± 0.1573	0.775 ± 0.156	0.7626 ± 0.1573	**0.7998** ± 0.1371
	MAE	0.6349 ± 0.1503	0.5975 ± 0.1694	0.6685 ± 0.1402	**0.7762** ± 0.1369	0.7184 ± 0.1313	0.7105 ± 0.1255	0.7298 ± 0.1258	0.7429 ± 0.1263	0.5784 ± 0.4175	0.7516 ± 0.1143	0.7105 ± 0.1255	0.7298 ± 0.1258	0.7105 ± 0.1255	0.7746 ± 0.1249
	MSE	0.5662 ± 0.1707	0.518 ± 0.1867	0.6018 ± 0.1658	**0.7541** ± 0.1427	0.6661 ± 0.1432	0.6561 ± 0.1402	0.6817 ± 0.1382	0.7021 ± 0.1357	0.5902 ± 0.4384	0.7184 ± 0.1289	0.6561 ± 0.1402	0.6817 ± 0.1382	0.6561 ± 0.1402	0.7478 ± 0.1287
Auto MPG	MCR	0.345 ± 0.1364	0.3317 ± 0.14	0.3658 ± 0.123	**0.4126** ± 0.0988	0.386 ± 0.1138	0.3829 ± 0.1137	0.3931 ± 0.1053	0.3909 ± 0.1048	0.1828 ± 0.1159	0.4037 ± 0.0973	0.3829 ± 0.1137	0.3931 ± 0.1053	0.3829 ± 0.1137	0.4029 ± 0.1083
	MAE	0.3485 ± 0.116	0.3307 ± 0.1206	0.3617 ± 0.1116	**0.4582** ± 0.1076	0.4402 ± 0.1063	0.4264 ± 0.0963	0.4389 ± 0.107	0.4353 ± 0.0951	0.2575 ± 0.2245	0.4469 ± 0.0982	0.4264 ± 0.0963	0.4389 ± 0.107	0.4264 ± 0.0963	0.4544 ± 0.1128
	MSE	0.3474 ± 0.2539	0.3301 ± 0.2549	0.3438 ± 0.2316	**0.4973** ± 0.1856	0.4795 ± 0.2055	0.4486 ± 0.2024	0.4683 ± 0.2057	0.4637 ± 0.1961	0.3548 ± 0.3224	0.4799 ± 0.1895	0.4486 ± 0.2024	0.4683 ± 0.2057	0.4486 ± 0.2024	0.4889 ± 0.1944
Pyrimidines	MCR	0.1434 ± 0.5549	0.1395 ± 0.596	**0.1839** ± 0.5618	-0.1734 ± 0.3926	-0.0529 ± 0.455	0.0512 ± 0.4001	-0.0086 ± 0.3905	-0.0123 ± 0.4874	0.0558 ± 0.3954	-0.0278 ± 0.3872	0.0512 ± 0.4001	-0.0086 ± 0.3905	0.0512 ± 0.4001	-0.1734 ± 0.3926
	MAE	0.106 ± 0.3315	0.0434 ± 0.3076	0.3371 ± 0.2252	0.2449 ± 0.3379	0.2608 ± 0.3647	0.3479 ± 0.2953	0.3486 ± 0.2926	**0.3663** ± 0.2219	0.2396 ± 0.3515	0.2957 ± 0.3093	0.3479 ± 0.2953	0.3486 ± 0.2926	0.3479 ± 0.2953	0.2358 ± 0.3439
	MSE	0.1688 ± 0.5395	-0.0235 ± 0.4597	0.2901 ± 0.6089	0.5872 ± 0.3016	0.5745 ± 0.2872	0.5694 ± 0.2538	**0.5965** ± 0.2379	0.5212 ± 0.3031	0.3067 ± 0.3976	0.5785 ± 0.2933	0.5694 ± 0.2538	0.5965 ± 0.2379	0.5694 ± 0.2538	0.5945 ± 0.2933
Abalone	MCR	0.2629 ± 0.0303	0.2422 ± 0.0302	0.2783 ± 0.0392	0.2874 ± 0.035	0.2889 ± 0.0259	0.2872 ± 0.027	0.2854 ± 0.0334	0.282 ± 0.0298	0.0461 ± 0.0823	0.2857 ± 0.038	0.2872 ± 0.027	0.2854 ± 0.0334	0.2872 ± 0.027	**0.2922** ± 0.0284
	MAE	0.2447 ± 0.0466	0.2118 ± 0.0479	0.2925 ± 0.0474	**0.3215** ± 0.0458	0.3039 ± 0.0454	0.295 ± 0.0474	0.3065 ± 0.0455	0.2991 ± 0.0426	0.1025 ± 0.0952	0.3159 ± 0.0456	0.295 ± 0.0474	0.3065 ± 0.0455	0.295 ± 0.0474	0.318 ± 0.0473
	MSE	0.2132 ± 0.0949	0.1706 ± 0.0931	0.2833 ± 0.0882	**0.3221** ± 0.0835	0.2779 ± 0.0978	0.2642 ± 0.0972	0.2932 ± 0.0867	0.286 ± 0.0852	0.1217 ± 0.1079	0.3149 ± 0.0826	0.2642 ± 0.0972	0.2932 ± 0.0867	0.2642 ± 0.0972	0.3039 ± 0.0945
Boston Housing	MCR	0.3612 ± 0.2102	0.3652 ± 0.2086	0.364 ± 0.2154	**0.3705** ± 0.2144	0.3623 ± 0.2119	0.363 ± 0.2126	0.3653 ± 0.2124	0.367 ± 0.2143	-0.0357 ± 0.2781	0.3688 ± 0.2096	0.363 ± 0.2126	0.3653 ± 0.2124	0.363 ± 0.2126	0.3667 ± 0.212
	MAE	0.355 ± 0.2066	0.3586 ± 0.2058	0.359 ± 0.2116	**0.3769** ± 0.2074	0.3616 ± 0.207	0.3616 ± 0.2062	0.3641 ± 0.2058	0.3694 ± 0.2064	-0.0224 ± 0.2827	0.3713 ± 0.2028	0.3616 ± 0.2062	0.3641 ± 0.2058	0.3616 ± 0.2062	0.3718 ± 0.2041
	MSE	0.3393 ± 0.1991	0.3425 ± 0.1993	0.3446 ± 0.2026	**0.378** ± 0.1882	0.3529 ± 0.1897	0.3515 ± 0.1892	0.3542 ± 0.1881	0.3647 ± 0.1865	0.0081 ± 0.3165	0.3668 ± 0.186	0.3515 ± 0.1892	0.3542 ± 0.1881	0.3515 ± 0.1892	0.3714 ± 0.1842
Stocks Domain	MCR	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
	MAE	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
	MSE	0.682 ± 0.0819	**0.6839** ± 0.0811	0.6812 ± 0.0812	0.6777 ± 0.0767	0.6835 ± 0.0806	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6777 ± 0.0783	0.031 ± 0.201	0.6808 ± 0.0802	0.6805 ± 0.0817	0.6803 ± 0.0817	0.6805 ± 0.0817	0.6777 ± 0.0767
Wisconsin Breast Cancer	MCR	0.2093 ± 0.3534	0.139 ± 0.3264	**0.2763** ± 0.3239	0.1976 ± 0.3016	0.1699 ± 0.3346	0.1838 ± 0.3426	0.2121 ± 0.3204	0.1557 ± 0.3397	-0.0961 ± 0.3344	0.2493 ± 0.29	0.1838 ± 0.3426	0.2121 ± 0.3204	0.1838 ± 0.3426	0.2008 ± 0.324
	MAE	0.1418 ± 0.2491	0.0913 ± 0.2475	0.2018 ± 0.2611	0.2263 ± 0.2923	0.2296 ± 0.2845	0.2228 ± 0.2627	0.2328 ± 0.2609	0.1713 ± 0.2517	0.1391 ± 0.2426	0.2394 ± 0.2555	0.2228 ± 0.2627	0.2328 ± 0.2609	0.2228 ± 0.2627	**0.251** ± 0.2897
	MSE	0.1149 ± 0.2686	0.0691 ± 0.2501	0.1634 ± 0.2834	0.1357 ± 0.3245	0.1429 ± 0.3008	0.1429 ± 0.2893	0.1528 ± 0.2964	0.1027 ± 0.2623	0.1731 ± 0.2083	0.1465 ± 0.3104	0.1429 ± 0.2893	0.1528 ± 0.2964	0.1429 ± 0.2893	**0.1806** ± 0.3045
Obesity	MCR	0.8883 ± 0.0845	0.8866 ± 0.0855	0.8874 ± 0.0843	**0.894** ± 0.0775	0.888 ± 0.0867	0.8893 ± 0.0829	0.8896 ± 0.0823	0.8872 ± 0.0834	0.5076 ± 0.3054	0.8892 ± 0.0819	0.8893 ± 0.0829	0.8896 ± 0.0823	0.8893 ± 0.0829	0.8933 ± 0.0775
	MAE	0.8856 ± 0.0819	0.8841 ± 0.083	0.8848 ± 0.0819	**0.8913** ± 0.0752	0.8855 ± 0.0844	0.8867 ± 0.0805	0.8871 ± 0.08	0.8846 ± 0.081	0.5047 ± 0.3041	0.8867 ± 0.0796	0.8867 ± 0.0805	0.8871 ± 0.08	0.8867 ± 0.0805	0.8906 ± 0.0751
	MSE	0.8839 ± 0.0808	0.8826 ± 0.0819	0.8831 ± 0.0809	**0.8896** ± 0.0744	0.8838 ± 0.0834	0.8851 ± 0.0795	0.8855 ± 0.0791	0.8828 ± 0.08	0.5054 ± 0.3047	0.8852 ± 0.0787	0.8851 ± 0.0795	0.8855 ± 0.0791	0.8851 ± 0.0795	0.8889 ± 0.0743
CMC	MCR	0.3143 ± 0.0738	**0.3146** ± 0.0775	0.306 ± 0.0665	0.2399 ± 0.0435	0.2282 ± 0.0461	0.2851 ± 0.0585	0.2772 ± 0.0532	0.2678 ± 0.0486	0.009 ± 0.069	0.2745 ± 0.0529	0.2851 ± 0.0585	0.2772 ± 0.0532	0.2851 ± 0.0585	0.2226 ± 0.0441
	MAE	0.2113 ± 0.0546	0.2218 ± 0.063	0.1973 ± 0.0436	0.2889 ± 0.0717	0.2926 ± 0.0705	0.2754 ± 0.0636	0.278 ± 0.0658	0.2778 ± 0.0695	0.0222 ± 0.146	0.2774 ± 0.0659	0.2754 ± 0.0636	0.278 ± 0.0658	0.2754 ± 0.0636	**0.2964** ± 0.0666
	MSE	0.0308 ± 0.0515	0.0411 ± 0.0578	0.0303 ± 0.0416	0.1405 ± 0.0795	0.1576 ± 0.0736	0.0807 ± 0.0541	0.089 ± 0.0602	0.1018 ± 0.0801	-0.0426 ± 0.1895	0.0907 ± 0.0601	0.0807 ± 0.0541	0.089 ± 0.0602	0.0807 ± 0.0541	**0.1739** ± 0.0704
Grub Damage	MCR	0.2406 ± 0.239	0.2157 ± 0.2222	0.2767 ± 0.2586	0.286 ± 0.2756	0.2384 ± 0.317	0.2553 ± 0.2915	0.3038 ± 0.3327	0.2871 ± 0.3163	0.072 ± 0.1933	**0.3176** ± 0.3154	0.2553 ± 0.2915	0.3038 ± 0.3327	0.2553 ± 0.2915	0.2359 ± 0.274
	MAE	0.0922 ± 0.2525	0.0739 ± 0.2451	0.1287 ± 0.2708	**0.2431** ± 0.2965	0.2121 ± 0.3041	0.1577 ± 0.2975	0.2169 ± 0.3508	0.2045 ± 0.3586	0.1267 ± 0.2378	0.2254 ± 0.3252	0.1577 ± 0.2975	0.2169 ± 0.3508	0.1577 ± 0.2975	0.2417 ± 0.2513
	MSE	0.159 ± 0.3607	0.1237 ± 0.3671	0.1764 ± 0.3528	0.2793 ± 0.2708	0.2606 ± 0.2436	0.1871 ± 0.329	0.2361 ± 0.3313	0.2159 ± 0.3824	0.1544 ± 0.2695	0.2375 ± 0.3425	0.1871 ± 0.329	0.2361 ± 0.3313	0.1871 ± 0.329	**0.2986** ± 0.2125
New Thyroid	MCR	0.9822 ± 0.0288	0.9822 ± 0.0288	0.9822 ± 0.0288	**1.0** ± 0.0	1.0 ± 0.0	0.9875 ± 0.02	0.9875 ± 0.02	1.0 ± 0.0	0.5203 ± 0.5054	0.9875 ± 0.02	0.9875 ± 0.02	0.9875 ± 0.02	0.9875 ± 0.02	1.0 ± 0.0
	MAE	0.9742 ± 0.0462	0.9742 ± 0.0462	0.9742 ± 0.0462	**0.9969** ± 0.0076	0.9969 ± 0.0076	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9969 ± 0.0076	0.5421 ± 0.4852	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9804 ± 0.0326	0.9938 ± 0.0153
	MSE	0.969 ± 0.0582	0.969 ± 0.0582	0.969 ± 0.0582	**0.9949** ± 0.0125	0.9949 ± 0.0125	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9949 ± 0.0125	0.5561 ± 0.475	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9758 ± 0.0425	0.9898 ± 0.0251
Balance Scale	MCR	0.8648 ± 0.0996	0.8627 ± 0.0937	0.8551 ± 0.1164	0.8642 ± 0.0703	0.8531 ± 0.0686	0.8602 ± 0.0818	0.8679 ± 0.077	**0.8817** ± 0.0669	0.0997 ± 0.2413	0.8729 ± 0.0743	0.8602 ± 0.0818	0.8679 ± 0.077	0.8602 ± 0.0818	0.8509 ± 0.068
	MAE	0.8072 ± 0.0953	0.8051 ± 0.0924	0.7949 ± 0.1074	0.8327 ± 0.0746	0.8217 ± 0.0767	0.8141 ± 0.076	0.8247 ± 0.0745	**0.8483** ± 0.0756	0.132 ± 0.2571	0.8309 ± 0.0713	0.8141 ± 0.076	0.8247 ± 0.0745	0.8141 ± 0.076	0.8206 ± 0.0706
	MSE	0.8032 ± 0.1032	0.8016 ± 0.1039	0.801 ± 0.1033	0.8303 ± 0.0607	0.8264 ± 0.0678	0.8142 ± 0.0754	0.8221 ± 0.0667	**0.8392** ± 0.055	0.0736 ± 0.2428	0.8278 ± 0.0593	0.8142 ± 0.0754	0.8221 ± 0.0667	0.8142 ± 0.0754	0.8336 ± 0.0733
Automobile	MCR	0.6564 ± 0.3921	0.6546 ± 0.3793	0.671 ± 0.3805	0.6911 ± 0.3496	0.6885 ± 0.3621	0.6832 ± 0.3676	0.6885 ± 0.3674	0.6965 ± 0.3608	0.4635 ± 0.17	0.6904 ± 0.3522	0.6832 ± 0.3676	0.6885 ± 0.3674	0.6832 ± 0.3676	**0.7036** ± 0.3567
	MAE	0.6215 ± 0.3772	0.6182 ± 0.3645	0.6284 ± 0.3672	0.6653 ± 0.3294	0.6581 ± 0.3442	0.6468 ± 0.3538	0.6498 ± 0.3541	0.6634 ± 0.3455	0.4099 ± 0.302	0.654 ± 0.3408	0.6468 ± 0.3538	0.6498 ± 0.3541	0.6468 ± 0.3538	**0.6803** ± 0.3373
	MSE	0.6003 ± 0.3861	0.5927 ± 0.3738	0.5974 ± 0.3762	0.6436 ± 0.3331	0.6299 ± 0.3495	0.6185 ± 0.3629	0.6187 ± 0.3637	0.6366 ± 0.3519	0.3452 ± 0.4262	0.6231 ± 0.3512	0.6185 ± 0.3629	0.6187 ± 0.3637	0.6185 ± 0.3629	**0.6582** ± 0.3392
Eucalyptus	MCR	0.447 ± 0.0761	0.4476 ± 0.077	0.4428 ± 0.0876	0.4439 ± 0.0876	**0.4571** ± 0.0833	0.4556 ± 0.0773	0.4503 ± 0.0846	0.4466 ± 0.08	0.0345 ± 0.136	0.442 ± 0.0872	0.4556 ± 0.0773	0.4503 ± 0.0846	0.4556 ± 0.0773	0.4523 ± 0.0818
	MAE	0.428 ± 0.0517	0.4267 ± 0.0577	0.4298 ± 0.0573	0.4369 ± 0.0488	**0.4446** ± 0.0545	0.4412 ± 0.0495	0.4393 ± 0.053	0.4365 ± 0.0507	0.0231 ± 0.1391	0.4341 ± 0.0497	0.4412 ± 0.0495	0.4393 ± 0.053	0.4412 ± 0.0495	0.4429 ± 0.0489
	MSE	0.3957 ± 0.0815	0.3962 ± 0.0863	0.4022 ± 0.0912	0.4182 ± 0.0819	0.4203 ± 0.0844	0.4158 ± 0.0788	0.417 ± 0.087	0.4164 ± 0.0842	0.0129 ± 0.1227	0.4139 ± 0.0829	0.4158 ± 0.0788	0.417 ± 0.087	0.4158 ± 0.0788	**0.4212** ± 0.0816
TAE	MCR	0.1154 ± 0.2666	0.0897 ± 0.2809	0.1422 ± 0.2643	0.1044 ± 0.3305	0.0589 ± 0.3033	0.154 ± 0.3105	0.1526 ± 0.3076	0.1108 ± 0.3173	**0.1902** ± 0.1997	0.1566 ± 0.333	0.154 ± 0.3105	0.1526 ± 0.3076	0.154 ± 0.3105	0.0423 ± 0.2976
	MAE	0.1 ± 0.2511	0.0766 ± 0.2699	0.133 ± 0.187	0.1766 ± 0.2973	0.1485 ± 0.2889	**0.2121** ± 0.2494	0.1958 ± 0.2452	0.1821 ± 0.3006	0.1478 ± 0.2066	0.1973 ± 0.2525	0.2121 ± 0.2494	0.1958 ± 0.2452	0.2121 ± 0.2494	0.1328 ± 0.3007
	MSE	0.0208 ± 0.3281	-0.0067 ± 0.3443	0.008 ± 0.2383	0.2304 ± 0.3097	0.2082 ± 0.2995	0.1955 ± 0.3094	0.2084 ± 0.2665	0.1936 ± 0.3486	**0.3243** ± 0.1727	0.2187 ± 0.2662	0.1955 ± 0.3094	0.2084 ± 0.2665	0.1955 ± 0.3094	0.1849 ± 0.2956
Heart (CLE)	MCR	0.5456 ± 0.1668	0.5276 ± 0.1618	**0.5763** ± 0.1566	0.5709 ± 0.1325	0.5534 ± 0.1482	0.5541 ± 0.1492	0.5583 ± 0.1375	0.5747 ± 0.1458	0.1375 ± 0.2534	0.5744 ± 0.1347	0.5541 ± 0.1492	0.5583 ± 0.1375	0.5541 ± 0.1492	0.5703 ± 0.142
	MAE	0.5029 ± 0.1576	0.4786 ± 0.1532	0.5285 ± 0.1573	**0.5575** ± 0.105	0.5344 ± 0.1225	0.5187 ± 0.1361	0.5298 ± 0.1238	0.5391 ± 0.1208	0.0889 ± 0.2527	0.5458 ± 0.1153	0.5187 ± 0.1361	0.5298 ± 0.1238	0.5187 ± 0.1361	0.5536 ± 0.1185
	MSE	0.3833 ± 0.1841	0.3494 ± 0.1832	0.4062 ± 0.1854	**0.4899** ± 0.1379	0.4479 ± 0.1438	0.4021 ± 0.1471	0.429 ± 0.1592	0.4304 ± 0.1412	0.0288 ± 0.2581	0.453 ± 0.1497	0.4021 ± 0.1471	0.429 ± 0.1592	0.4021 ± 0.1471	0.483 ± 0.1496
SWD	MCR	0.1983 ± 0.0991	0.1859 ± 0.1042	0.1848 ± 0.086	0.1949 ± 0.0967	0.2006 ± 0.102	**0.2019** ± 0.1052	0.1948 ± 0.0963	0.2018 ± 0.0977	0.043 ± 0.1066	0.1938 ± 0.0905	0.2019 ± 0.1052	0.1948 ± 0.0963	0.2019 ± 0.1052	0.1988 ± 0.1088
	MAE	0.1143 ± 0.0971	0.1152 ± 0.0931	0.1172 ± 0.0867	0.1372 ± 0.1034	0.1405 ± 0.103	0.1252 ± 0.1113	0.1285 ± 0.1002	0.1275 ± 0.1068	0.052 ± 0.1113	0.1312 ± 0.0953	0.1252 ± 0.1113	0.1285 ± 0.1002	0.1252 ± 0.1113	**0.1423** ± 0.1089
	MSE	0.1017 ± 0.0935	0.101 ± 0.0816	0.1154 ± 0.0946	0.1331 ± 0.1141	0.1401 ± 0.0978	0.1134 ± 0.0979	0.1166 ± 0.1073	0.1071 ± 0.1135	0.0498 ± 0.1022	0.1239 ± 0.1023	0.1134 ± 0.0979	0.1166 ± 0.1073	0.1134 ± 0.0979	**0.1403** ± 0.1032
ERA	MCR	0.1708 ± 0.0604	**0.2059** ± 0.0941	0.1377 ± 0.0886	0.0468 ± 0.0662	0.0801 ± 0.0635	0.131 ± 0.0633	0.101 ± 0.066	0.1777 ± 0.0575	0.0901 ± 0.1104	0.085 ± 0.0646	0.131 ± 0.0633	0.101 ± 0.066	0.131 ± 0.0633	0.0598 ± 0.0802
	MAE	0.0353 ± 0.0965	0.0409 ± 0.1273	0.0069 ± 0.076	0.0138 ± 0.0923	0.0477 ± 0.0839	0.0597 ± 0.0668	0.0358 ± 0.0806	0.0439 ± 0.0688	**0.0756** ± 0.1181	0.0131 ± 0.0887	0.0597 ± 0.0668	0.0358 ± 0.0806	0.0597 ± 0.0668	0.0221 ± 0.0972
	MSE	0.0586 ± 0.0932	**0.0725** ± 0.1066	0.0088 ± 0.1241	-0.0023 ± 0.1514	0.0222 ± 0.1352	0.0283 ± 0.1189	0.0174 ± 0.141	0.0241 ± 0.1206	0.0421 ± 0.1256	0.0031 ± 0.1457	0.0283 ± 0.1189	0.0174 ± 0.141	0.0283 ± 0.1189	0.0016 ± 0.152
ESL	MCR	0.2079 ± 0.1653	0.2119 ± 0.1854	0.2164 ± 0.1122	0.2738 ± 0.1225	**0.2953** ± 0.1443	0.2487 ± 0.1418	0.2484 ± 0.1228	0.2275 ± 0.1257	0.0123 ± 0.1516	0.2467 ± 0.1141	0.2487 ± 0.1418	0.2484 ± 0.1228	0.2487 ± 0.1418	0.2817 ± 0.1347
	MAE	0.1738 ± 0.1687	0.1778 ± 0.1887	0.182 ± 0.1401	**0.2887** ± 0.1285	0.2864 ± 0.1415	0.2305 ± 0.1371	0.2343 ± 0.1158	0.2197 ± 0.1146	0.0469 ± 0.1569	0.2524 ± 0.1218	0.2305 ± 0.1371	0.2343 ± 0.1158	0.2305 ± 0.1371	0.287 ± 0.1346
	MSE	0.1048 ± 0.2408	0.1098 ± 0.2698	0.1089 ± 0.2014	**0.275** ± 0.1536	0.2374 ± 0.1863	0.1689 ± 0.1811	0.18 ± 0.1441	0.1718 ± 0.1428	0.1026 ± 0.1814	0.227 ± 0.1351	0.1689 ± 0.1811	0.18 ± 0.1441	0.1689 ± 0.1811	0.2592 ± 0.1613
LEV	MCR	0.1559 ± 0.1345	0.1398 ± 0.1402	0.1571 ± 0.1182	0.1539 ± 0.11	0.1582 ± 0.1268	0.1605 ± 0.13	**0.1652** ± 0.1202	0.1619 ± 0.1271	-0.0153 ± 0.1251	0.1493 ± 0.1169	0.1605 ± 0.13	0.1652 ± 0.1202	0.1605 ± 0.13	0.1542 ± 0.1235
	MAE	0.1302 ± 0.1181	0.1079 ± 0.121	0.1426 ± 0.1144	0.1397 ± 0.1001	0.1311 ± 0.1096	0.1342 ± 0.1134	**0.1483** ± 0.1077	0.1457 ± 0.1143	0.0062 ± 0.1436	0.137 ± 0.1091	0.1342 ± 0.1134	0.1483 ± 0.1077	0.1342 ± 0.1134	0.1315 ± 0.1047
	MSE	0.1375 ± 0.1169	0.1048 ± 0.1145	0.1607 ± 0.1268	0.1546 ± 0.1063	0.1297 ± 0.1088	0.1394 ± 0.1076	0.161 ± 0.1116	**0.1666** ± 0.1204	0.0264 ± 0.182	0.1572 ± 0.1206	0.1394 ± 0.1076	0.161 ± 0.1116	0.1394 ± 0.1076	0.1335 ± 0.101
Red Wine	MCR	0.4037 ± 0.0738	0.3982 ± 0.0732	0.4169 ± 0.0708	0.4228 ± 0.0698	0.4074 ± 0.0677	0.4109 ± 0.0691	0.4146 ± 0.0687	0.4155 ± 0.0694	-0.052 ± 0.0356	**0.4237** ± 0.0688	0.4109 ± 0.0691	0.4146 ± 0.0687	0.4109 ± 0.0691	0.4171 ± 0.0673
	MAE	0.4021 ± 0.0719	0.396 ± 0.0727	0.4217 ± 0.0682	0.4321 ± 0.0659	0.4133 ± 0.0645	0.4144 ± 0.0667	0.4205 ± 0.0651	0.419 ± 0.0672	-0.062 ± 0.0413	**0.4322** ± 0.0661	0.4144 ± 0.0667	0.4205 ± 0.0651	0.4144 ± 0.0667	0.4251 ± 0.0639
	MSE	0.39 ± 0.0898	0.3827 ± 0.0925	0.4162 ± 0.0818	**0.4296** ± 0.0694	0.4049 ± 0.0783	0.4028 ± 0.082	0.4126 ± 0.076	0.4078 ± 0.0812	-0.0647 ± 0.0781	0.4283 ± 0.0737	0.4028 ± 0.082	0.4126 ± 0.076	0.4028 ± 0.082	0.4203 ± 0.0729
White Wine	MCR	0.3642 ± 0.0537	0.3583 ± 0.0509	0.3545 ± 0.053	0.3388 ± 0.0519	0.354 ± 0.0485	**0.3673** ± 0.0508	0.3583 ± 0.0495	0.3672 ± 0.0527	0.0477 ± 0.0641	0.3472 ± 0.053	0.3673 ± 0.0508	0.3583 ± 0.0495	0.3673 ± 0.0508	0.3584 ± 0.0487
	MAE	0.3479 ± 0.058	0.338 ± 0.0542	0.3495 ± 0.0604	0.338 ± 0.0566	0.3465 ± 0.0535	0.3564 ± 0.0555	0.3532 ± 0.0572	**0.3587** ± 0.0575	0.058 ± 0.0602	0.3454 ± 0.0592	0.3564 ± 0.0555	0.3532 ± 0.0572	0.3564 ± 0.0555	0.3523 ± 0.0507
	MSE	0.339 ± 0.0698	0.3249 ± 0.0674	0.3493 ± 0.0716	0.3409 ± 0.0662	0.3423 ± 0.0628	0.3498 ± 0.0641	0.3516 ± 0.0694	**0.3554** ± 0.0687	0.0574 ± 0.0624	0.3475 ± 0.0691	0.3498 ± 0.0641	0.3516 ± 0.0694	0.3498 ± 0.0641	0.3506 ± 0.0589



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

(b) Result only for MCR.

(c) Result for MAE.

(d) Result for MSE.

(e) Result for MCR and MAE.

(f) Result for MAE and MSE.

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

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

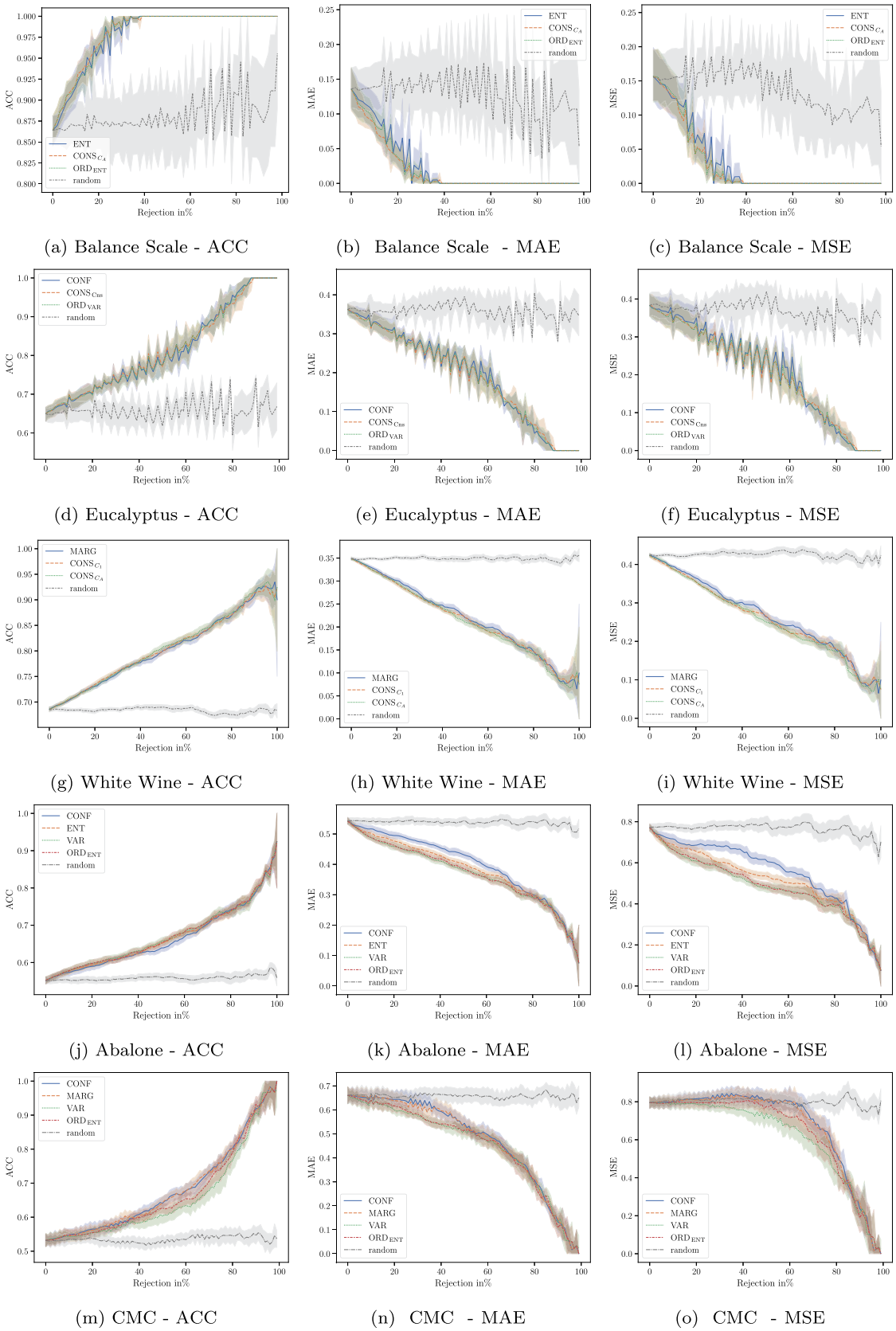
The DFU measure performs worst on all performance metrics and is often close or even worse than random rejection, which indicates that the probabilistic output of the predictor is mostly unimodal. Given unimodal probability distributions, DFU is not able to quantify any uncertainty at all, which might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

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

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

7. Case study: automotive goodwill claim assessment

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



(a) Balance Scale - ACC (b) Balance Scale - MAE (c) Balance Scale - MSE

(d) Eucalyptus - ACC (e) Eucalyptus - MAE (f) Eucalyptus - MSE

(g) White Wine - ACC (h) White Wine - MAE (i) White Wine - MSE

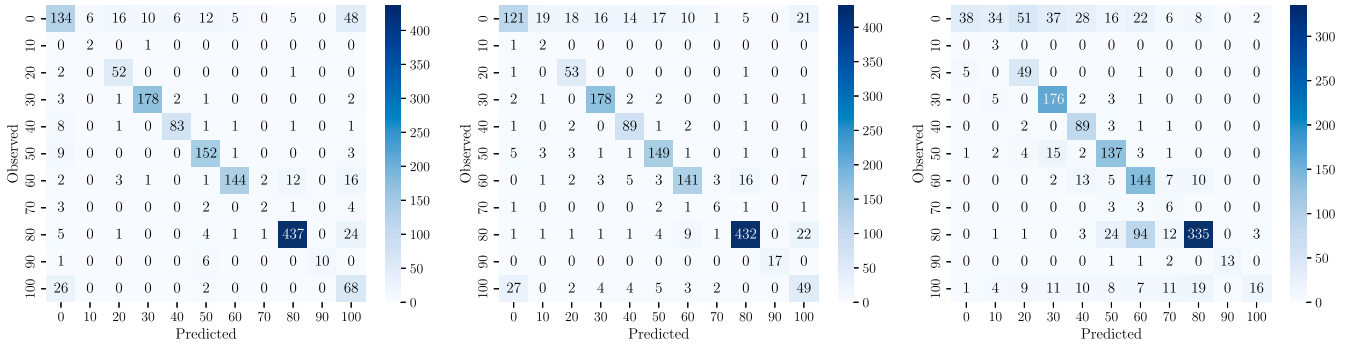
(j) Abalone - ACC (k) Abalone - MAE (l) Abalone - MSE

(m) CMC - ACC (n) CMC - MAE (o) CMC - MSE

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

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

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



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

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

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

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

7.1. Datasets

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

7.2. Experimental setup

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

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

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

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

7.3. Results and analysis

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

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

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

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

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

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

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

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

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

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

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

8. Conclusion and future work

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

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

Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Cns}$	$CONS_{C_1}$	$CONS_{C_2}$	$CONS_{C_4}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Market A (Parts)	ACC	0.7364	0.7245	**0.747**	0.6468	0.6541	0.6972	0.6943	0.6956	0.67	0.6916	0.6972	0.6943	0.6972	0.6484
	MAE	0.6416	0.6351	0.6433	0.685	0.6827	0.6731	0.6767	0.6707	0.6188	0.6821	0.6731	0.6767	0.6731	**0.6852**
	MSE	0.5102	0.504	0.5138	**0.6331**	0.6237	0.5799	0.5882	0.5778	0.5007	0.6	0.5799	0.5882	0.5799	0.6324
Market A (Labor)	ACC	0.8192	0.8097	**0.8284**	0.7238	0.7247	0.7807	0.7783	0.7785	0.7738	0.7743	0.7807	0.7783	0.7807	0.7253
	MAE	0.6886	0.6826	0.6924	0.7136	0.7071	0.7127	0.7169	0.709	0.6737	**0.72**	0.7127	0.7169	0.7127	0.7141
	MSE	0.5694	0.5643	0.5713	**0.6553**	0.6432	0.6212	0.6294	0.6169	0.5632	0.6372	0.6212	0.6294	0.6212	0.6552
Market B (Parts)	ACC	0.6434	0.6432	0.6398	0.6391	0.6563	**0.6567**	0.6562	0.6501	0.6475	0.6531	0.6567	0.6562	0.6567	0.64
	MAE	0.6163	0.6222	0.5991	0.713	**0.7193**	0.6899	0.6944	0.6903	0.6526	0.6988	0.6899	0.6944	0.6899	0.713
	MSE	0.5136	0.5246	0.4914	0.6858	**0.6877**	0.6268	0.636	0.6349	0.5718	0.6468	0.6268	0.636	0.6268	0.6852
Market B (Labor)	ACC	**0.7791**	0.7768	0.7775	0.7333	0.744	0.765	0.7617	0.759	0.7319	0.7566	0.765	0.7617	0.765	0.7348
	MAE	0.7762	0.7775	0.7684	0.833	**0.8336**	0.8248	0.8284	0.8256	0.7854	0.8311	0.8248	0.8284	0.8248	0.8332
	MSE	0.7326	0.7356	0.7241	**0.845**	0.8427	0.8125	0.8206	0.8161	0.7559	0.8281	0.8125	0.8206	0.8125	0.8449
Market C (Parts)	ACC	**0.6029**	0.6007	0.5893	0.4702	0.5127	0.5494	0.5347	0.5489	0.4028	0.5158	0.5494	0.5347	0.5494	0.4761
	MAE	0.5725	0.5562	0.5847	0.6038	0.6086	0.6074	0.6085	0.6083	0.5329	**0.6098**	0.6074	0.6085	0.6074	0.604
	MSE	0.4268	0.4059	0.453	**0.62**	0.5883	0.5418	0.5579	0.5399	0.5032	0.5796	0.5418	0.5579	0.5418	0.618
Market C (Labor)	ACC	0.7816	0.7796	**0.7818**	0.6888	0.7002	0.7348	0.7304	0.7284	0.716	0.7235	0.7348	0.7304	0.7348	0.6895
	MAE	0.8013	0.802	0.7933	0.7979	0.7989	0.8076	0.8071	0.8063	0.798	0.8059	**0.8076**	0.8071	0.8076	0.7978
	MSE	0.8225	0.8243	0.8138	**0.8589**	0.8568	0.8531	0.855	0.8531	0.8339	0.8568	0.8531	0.855	0.8531	0.8587
Market D (Parts)	ACC	**0.6803**	0.6734	0.6749	0.5005	0.5175	0.6057	0.5924	0.602	0.5424	0.5805	0.6057	0.5924	0.6057	0.5028
	MAE	0.5147	0.5206	0.5021	0.5206	0.5193	0.5589	0.5575	**0.5662**	0.5348	0.5555	0.5589	0.5575	0.5589	0.5212
	MSE	0.412	0.4202	0.4025	0.494	0.4814	0.4936	0.4994	**0.5079**	0.4907	0.5044	0.4936	0.4994	0.4936	0.494
Market D (Labor)	ACC	**0.754**	0.753	0.7511	0.6227	0.6409	0.6995	0.6919	0.6911	0.6588	0.6771	0.6995	0.6919	0.6995	0.6265
	MAE	0.7623	0.7587	0.763	0.7557	0.7553	**0.7752**	0.7749	0.7731	0.7586	0.7725	0.7752	0.7749	0.7752	0.7561
	MSE	0.7285	0.7229	0.7322	**0.7759**	0.7721	0.7689	0.772	0.766	0.7408	0.7754	0.7689	0.772	0.7689	0.7755
Market E (Parts)	ACC	0.6081	**0.6099**	0.5927	0.5794	0.5791	0.6015	0.5989	0.6	0.6055	0.5956	0.6015	0.5989	0.6015	0.5794
	MAE	0.6042	0.6067	0.5881	0.6056	0.6045	0.612	**0.6124**	0.6105	0.6041	0.612	0.612	0.6124	0.612	0.6056
	MSE	0.5163	0.5181	0.508	0.5398	0.5388	0.532	0.5349	0.5304	0.5169	0.537	0.532	0.5349	0.532	**0.5399**
Market E (Labor)	ACC	0.6188	**0.6223**	0.6014	0.5908	0.5929	0.6141	0.6102	0.6135	0.6217	0.6056	0.6141	0.6102	0.6141	0.5908
	MAE	0.6183	0.6224	0.6006	0.6206	0.6211	0.6268	0.6265	**0.6275**	0.6231	0.6255	0.6268	0.6265	0.6268	0.6207
	MSE	0.5731	0.5776	0.5618	0.5991	0.5986	0.5909	0.5936	0.5928	0.5798	0.5952	0.5909	0.5936	0.5909	**0.5992**
Market F (Parts)	ACC	0.7364	0.7245	**0.747**	0.6468	0.6541	0.6972	0.6943	0.6956	0.67	0.6916	0.6972	0.6943	0.6972	0.6484
	MAE	0.6416	0.6351	0.6433	0.685	0.6827	0.6731	0.6767	0.6707	0.6188	0.6821	0.6731	0.6767	0.6731	**0.6852**
	MSE	0.5102	0.504	0.5138	**0.6331**	0.6237	0.5799	0.5882	0.5778	0.5007	0.6	0.5799	0.5882	0.5799	0.6324
Market F (Labor)	ACC	0.8192	0.8097	**0.8284**	0.7238	0.7247	0.7807	0.7783	0.7785	0.7738	0.7743	0.7807	0.7783	0.7807	0.7253
	MAE	0.6886	0.6826	0.6924	0.7136	0.7071	0.7127	0.7169	0.709	0.6737	**0.72**	0.7127	0.7169	0.7127	0.7141
	MSE	0.5694	0.5643	0.5713	**0.6553**	0.6432	0.6212	0.6294	0.6169	0.5632	0.6372	0.6212	0.6294	0.6212	0.6552
Market G (Parts)	ACC	**0.7319**	0.7194	0.7286	0.6533	0.6633	0.7113	0.7031	0.7013	0.6483	0.697	0.7113	0.7031	0.7113	0.6536
	MAE	0.5769	0.5704	0.5783	0.6637	**0.6688**	0.6628	0.6661	0.6463	0.5446	0.6655	0.6628	0.6661	0.6628	0.6637
	MSE	0.4605	0.4546	0.4644	0.6388	0.6344	0.5869	0.6032	0.5758	0.451	0.6115	0.5869	0.6032	0.5869	**0.6389**
Market G (Labor)	ACC	0.6665	0.6642	**0.6691**	0.6379	0.637	0.6568	0.6535	0.6518	0.6595	0.6496	0.6568	0.6535	0.6568	0.6376
	MAE	0.6771	0.6766	0.6756	0.6745	0.6739	0.6789	**0.6789**	0.6772	0.6752	0.6781	0.6789	0.6789	0.6789	0.6741
	MSE	0.6092	0.6094	0.606	0.6235	0.6229	0.6175	0.6202	0.6178	0.6095	0.6216	0.6175	0.6202	0.6175	**0.6236**

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

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

Table 7
Ranks of measures for MCR on goodwill assessment.

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

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

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

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

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

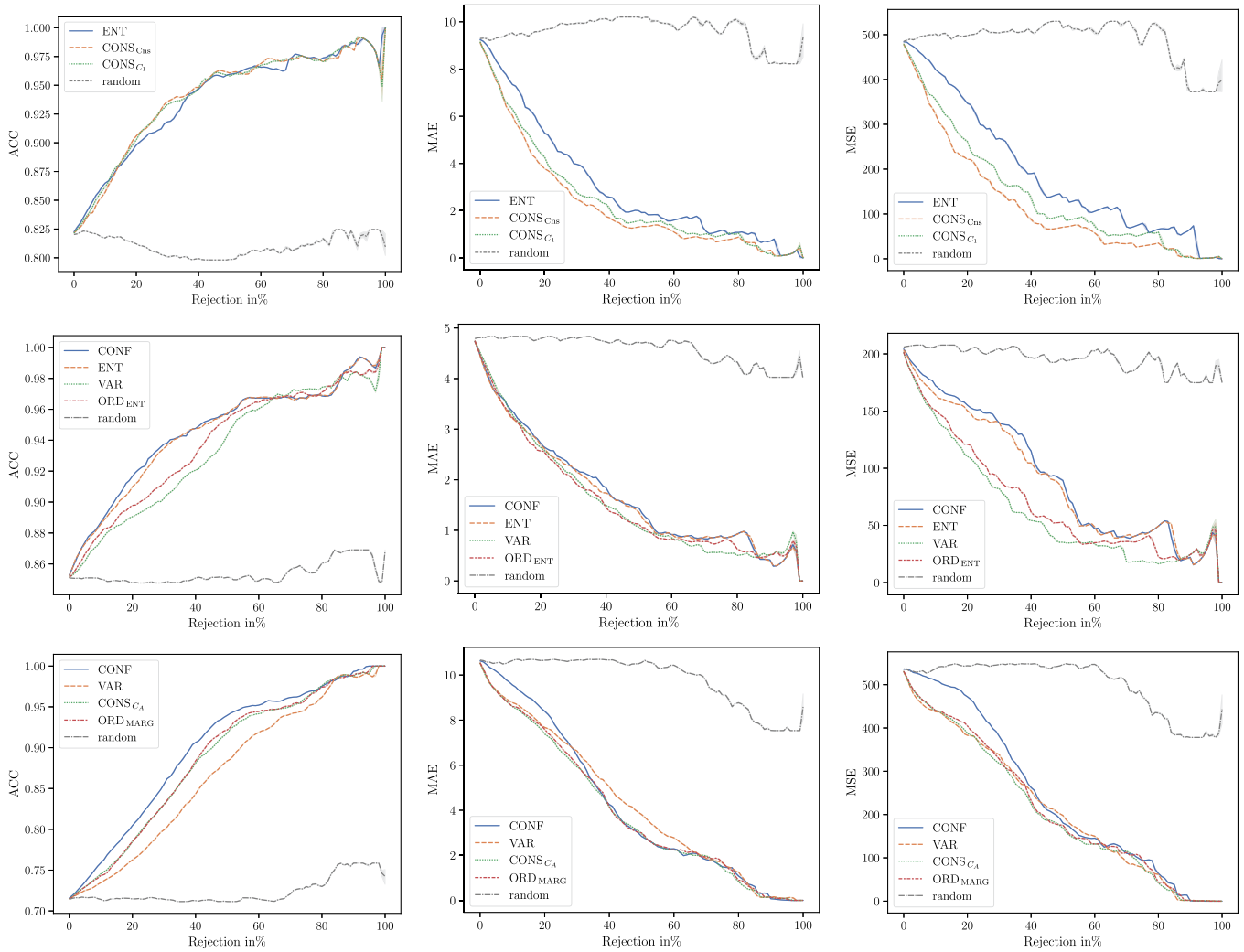


Fig. 11. Exemplary rejection curves for three of the goodwill claim assessment datasets displaying conventional uncertainty measures like entropy, margin and variance in comparison to consensus and ordinal binary decomposition-based measures.

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

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

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

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

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

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

CRediT authorship contribution statement

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

Declaration of competing interest

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

Appendix A. Proofs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A4: This trivially holds true.

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

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

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

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

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

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

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

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

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

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

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

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

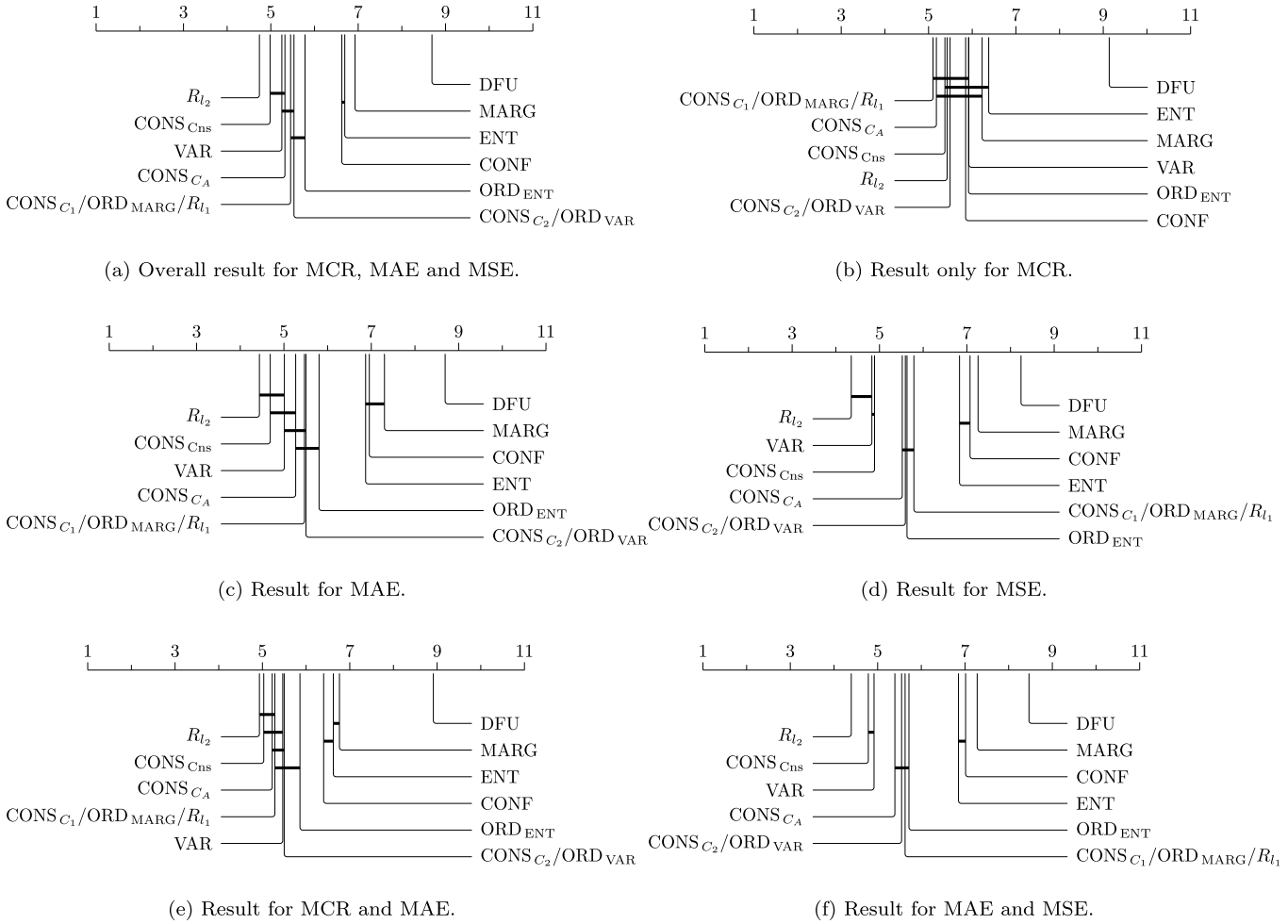
Appendix B. Prediction rejection ratios (PRRs) with multi-layer perceptron (MLP) as base learner

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

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

Table B.11
MLP parameters [61].

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



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

(b) Result only for MCR.

(c) Result for MAE.

(d) Result for MSE.

(e) Result for MCR and MAE.

(f) Result for MAE and MSE.

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

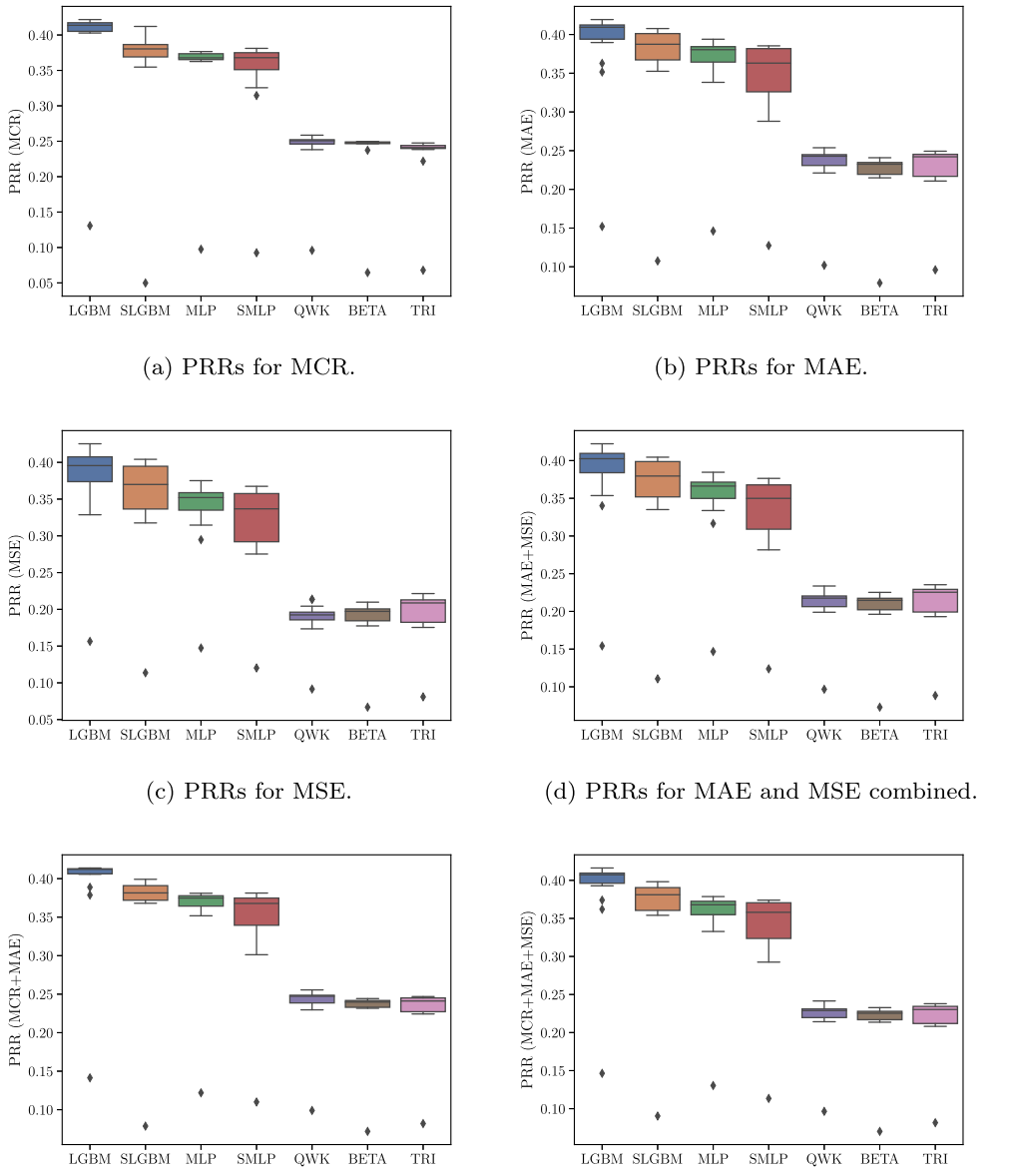
Appendix C. Comparison of prediction rejection ratios (PRRs) for different predictors

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

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

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

Dataset	Metric	CONF	MARG	ENT	VAR	CONS_{CH}	CONS_{C_1}	CONS_{C_2}	CONS_{C_3}	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Triazines	ACC	0.0721 ± 0.2426	0.041 ± 0.2415	0.0953 ± 0.2401	0.0868 ± 0.2936	0.0877 ± 0.2796	0.0854 ± 0.2672	**0.0959** ± 0.2816	0.0952 ± 0.2779	0.0736 ± 0.278	0.0938 ± 0.2822	0.0854 ± 0.2672	0.0959 ± 0.2816	0.0854 ± 0.2672	0.0884 ± 0.2803
	MAE	0.1833 ± 0.1919	0.1504 ± 0.1925	0.2145 ± 0.1663	0.2388 ± 0.1931	0.238 ± 0.1938	0.2184 ± 0.1993	0.2317 ± 0.2079	0.234 ± 0.2007	**0.2681** ± 0.2501	0.2317 ± 0.2066	0.2184 ± 0.1993	0.2317 ± 0.2079	0.2184 ± 0.1993	0.2429 ± 0.1916
	MSE	0.2114 ± 0.1935	0.1877 ± 0.1898	0.2484 ± 0.1808	0.2775 ± 0.2289	0.2757 ± 0.2146	0.2448 ± 0.205	0.2633 ± 0.2147	0.2625 ± 0.2114	**0.3055** ± 0.2256	0.2668 ± 0.218	0.2448 ± 0.205	0.2633 ± 0.2147	0.2448 ± 0.205	0.2761 ± 0.2301
Machine CPU	ACC	0.6171 ± 0.19	0.5714 ± 0.1991	0.6573 ± 0.185	**0.7228** ± 0.1246	0.6786 ± 0.1568	0.6829 ± 0.1624	0.6926 ± 0.1732	0.6958 ± 0.1677	0.6628 ± 0.1603	0.6905 ± 0.1614	0.6829 ± 0.1624	0.6926 ± 0.1732	0.6829 ± 0.1624	0.7079 ± 0.1307
	MAE	0.5758 ± 0.2401	0.5025 ± 0.2575	0.6554 ± 0.1853	**0.727** ± 0.1673	0.6945 ± 0.1841	0.6816 ± 0.1894	0.7023 ± 0.1854	0.6926 ± 0.1854	0.7058 ± 0.1062	0.7028 ± 0.1829	0.6816 ± 0.1894	0.7023 ± 0.1854	0.6816 ± 0.1894	0.722 ± 0.1745
	MSE	0.5313 ± 0.3124	0.4457 ± 0.3216	0.601 ± 0.203	0.7024 ± 0.1965	0.6861 ± 0.1958	0.6707 ± 0.2066	0.6818 ± 0.2018	0.6782 ± 0.1973	0.7041 ± 0.1034	0.6804 ± 0.2025	0.6707 ± 0.2066	0.6818 ± 0.2018	0.6707 ± 0.2066	**0.7051** ± 0.2007
Auto MPG	ACC	0.3533 ± 0.1097	0.3429 ± 0.0983	0.3501 ± 0.1542	0.379 ± 0.1612	0.3779 ± 0.1143	0.3847 ± 0.1068	0.3695 ± 0.1411	0.3759 ± 0.1327	0.1373 ± 0.2025	0.3665 ± 0.156	0.3847 ± 0.1068	0.3695 ± 0.1411	0.3847 ± 0.1068	**0.3874** ± 0.1283
	MAE	0.3419 ± 0.1405	0.3004 ± 0.129	0.3697 ± 0.1409	0.4164 ± 0.1584	0.413 ± 0.158	0.4086 ± 0.137	0.4081 ± 0.1456	0.4179 ± 0.1334	0.2236 ± 0.172	0.4028 ± 0.1464	0.4086 ± 0.137	0.4081 ± 0.1456	0.4086 ± 0.137	**0.4205** ± 0.1612
	MSE	0.3459 ± 0.1433	0.2847 ± 0.1365	0.3703 ± 0.1653	0.4122 ± 0.2452	0.4119 ± 0.2525	0.405 ± 0.2056	0.4079 ± 0.2318	**0.4221** ± 0.2119	0.2215 ± 0.2276	0.3971 ± 0.2248	0.405 ± 0.2056	0.4079 ± 0.2318	0.405 ± 0.2056	0.4123 ± 0.2541
Pyrimidines	ACC	0.0514 ± 0.3327	**0.0549** ± 0.2403	-0.0738 ± 0.341	-0.0941 ± 0.4573	-0.0652 ± 0.4214	-0.0254 ± 0.3109	-0.0883 ± 0.3331	-0.0018 ± 0.3369	-0.0649 ± 0.2524	-0.0741 ± 0.3557	-0.0254 ± 0.3109	-0.0883 ± 0.3331	-0.0254 ± 0.3109	-0.1133 ± 0.437
	MAE	0.2048 ± 0.3698	0.1852 ± 0.3425	0.1806 ± 0.4419	0.2266 ± 0.5205	0.2178 ± 0.4683	0.201 ± 0.4205	0.1724 ± 0.4429	**0.2422** ± 0.4212	0.2082 ± 0.2977	0.2157 ± 0.4665	0.201 ± 0.4205	0.1724 ± 0.4429	0.201 ± 0.4205	0.2063 ± 0.5018
	MSE	0.2272 ± 0.4368	0.2051 ± 0.4218	0.2463 ± 0.4471	0.3193 ± 0.4848	0.3118 ± 0.4018	0.2788 ± 0.358	0.2691 ± 0.3797	**0.32** ± 0.3752	0.2847 ± 0.3739	0.3093 ± 0.4146	0.2788 ± 0.358	0.2691 ± 0.3797	0.2788 ± 0.358	0.2961 ± 0.4655
Abalone	ACC	0.3465 ± 0.0652	0.3126 ± 0.0704	0.3285 ± 0.0615	0.3189 ± 0.054	0.3437 ± 0.0625	**0.3528** ± 0.0639	0.335 ± 0.0618	0.345 ± 0.0612	0.0225 ± 0.0667	0.3213 ± 0.0564	0.3528 ± 0.0639	0.335 ± 0.0618	0.3528 ± 0.0639	0.3387 ± 0.0547
	MAE	0.3674 ± 0.0522	0.2913 ± 0.0572	0.3936 ± 0.0461	0.3967 ± 0.0377	0.3998 ± 0.0441	0.398 ± 0.0448	0.4008 ± 0.0431	0.4034 ± 0.0425	0.0819 ± 0.0834	0.3963 ± 0.0401	0.398 ± 0.0448	0.4008 ± 0.0431	0.398 ± 0.0448	**0.406** ± 0.037
	MSE	0.4004 ± 0.072	0.2756 ± 0.06	0.4719 ± 0.0692	0.4913 ± 0.0619	0.4715 ± 0.0652	0.4561 ± 0.0669	0.4815 ± 0.0661	0.4772 ± 0.0688	0.1573 ± 0.0858	0.4871 ± 0.065	0.4561 ± 0.0669	0.4815 ± 0.0661	0.4561 ± 0.0669	**0.4916** ± 0.0581
Boston Housing	ACC	0.4113 ± 0.142	0.4189 ± 0.1396	0.4113 ± 0.1409	0.4367 ± 0.13	**0.4483** ± 0.1321	0.4315 ± 0.1295	0.431 ± 0.1275	0.4315 ± 0.1252	-0.1139 ± 0.2308	0.427 ± 0.1293	0.4315 ± 0.1295	0.431 ± 0.1275	0.4315 ± 0.1295	0.4461 ± 0.1289
	MAE	0.3897 ± 0.173	0.3949 ± 0.1696	0.3986 ± 0.1749	0.4435 ± 0.1656	0.4467 ± 0.1616	0.428 ± 0.1621	0.4296 ± 0.1612	0.4343 ± 0.1584	-0.0644 ± 0.2905	0.4282 ± 0.1668	0.428 ± 0.1621	0.4296 ± 0.1612	0.428 ± 0.1621	**0.4489** ± 0.1616
	MSE	0.3188 ± 0.2236	0.3241 ± 0.2253	0.3343 ± 0.2331	**0.408** ± 0.2107	0.3989 ± 0.2064	0.376 ± 0.2099	0.3823 ± 0.214	0.3917 ± 0.203	-0.0197 ± 0.3391	0.3847 ± 0.2187	0.376 ± 0.2099	0.3823 ± 0.214	0.376 ± 0.2099	0.4078 ± 0.2025
Stocks Domain	ACC	0.7104 ± 0.053	**0.712** ± 0.0521	0.7064 ± 0.0556	0.7045 ± 0.0532	0.7098 ± 0.0527	0.7101 ± 0.0529	0.7084 ± 0.0537	0.7088 ± 0.0532	0.0667 ± 0.1669	0.7052 ± 0.0544	0.7101 ± 0.0529	0.7084 ± 0.0537	0.7101 ± 0.0529	0.7068 ± 0.0526
	MAE	0.7053 ± 0.0621	**0.707** ± 0.0615	0.7013 ± 0.064	0.6994 ± 0.062	0.7048 ± 0.0618	0.705 ± 0.062	0.7034 ± 0.0627	0.7038 ± 0.0622	0.0772 ± 0.1834	0.7001 ± 0.0629	0.705 ± 0.062	0.7034 ± 0.0627	0.705 ± 0.062	0.7017 ± 0.0615
	MSE	0.7091 ± 0.0685	**0.7107** ± 0.0679	0.7051 ± 0.0704	0.7034 ± 0.0679	0.7086 ± 0.0682	0.7088 ± 0.0684	0.7072 ± 0.0691	0.7076 ± 0.0686	0.0855 ± 0.1749	0.704 ± 0.0692	0.7088 ± 0.0684	0.7072 ± 0.0691	0.7088 ± 0.0684	0.7056 ± 0.0679
Wisconsin Breast Cancer	ACC	0.2284 ± 0.2257	**0.2549** ± 0.2251	0.2476 ± 0.2515	0.254 ± 0.2291	0.2267 ± 0.2278	0.2302 ± 0.2322	0.2487 ± 0.2331	0.2511 ± 0.229	0.1021 ± 0.3648	0.2544 ± 0.2374	0.2302 ± 0.2322	0.2487 ± 0.2331	0.2302 ± 0.2322	0.2538 ± 0.2289
	MAE	0.1399 ± 0.2296	0.1505 ± 0.246	0.172 ± 0.2467	0.1946 ± 0.2395	0.1767 ± 0.2442	0.1674 ± 0.2337	0.1724 ± 0.2446	0.1721 ± 0.2422	0.0559 ± 0.179	0.1751 ± 0.2368	0.1674 ± 0.2337	0.1724 ± 0.2446	0.1674 ± 0.2337	**0.2016** ± 0.2371
	MSE	0.0228 ± 0.205	0.036 ± 0.2232	0.035 ± 0.2139	0.06 ± 0.1771	0.0353 ± 0.1877	0.0209 ± 0.1798	0.0328 ± 0.1921	0.0363 ± 0.1797	-0.0146 ± 0.1863	0.0339 ± 0.1872	0.0209 ± 0.1798	0.0328 ± 0.1921	0.0209 ± 0.1798	**0.0705** ± 0.1775
Obesity	ACC	0.6996 ± 0.1197	0.698 ± 0.1203	0.7004 ± 0.1201	0.7153 ± 0.1169	0.7098 ± 0.119	0.7074 ± 0.1187	0.7077 ± 0.1192	0.7086 ± 0.1189	0.3128 ± 0.0759	0.7091 ± 0.1189	0.7074 ± 0.1187	0.7077 ± 0.1192	0.7074 ± 0.1187	**0.7159** ± 0.1167
	MAE	0.6877 ± 0.1297	0.6861 ± 0.1301	0.6889 ± 0.1301	**0.709** ± 0.1246	0.7009 ± 0.1278	0.698 ± 0.1273	0.6986 ± 0.1281	0.7005 ± 0.1265	0.3406 ± 0.087	0.7003 ± 0.1279	0.698 ± 0.1273	0.6986 ± 0.1281	0.698 ± 0.1273	0.7089 ± 0.1237
	MSE	0.6509 ± 0.1803	0.6491 ± 0.1798	0.6524 ± 0.1807	**0.681** ± 0.1699	0.6683 ± 0.1758	0.6644 ± 0.1753	0.6655 ± 0.1763	0.6689 ± 0.1719	0.3695 ± 0.101	0.668 ± 0.1762	0.6644 ± 0.1753	0.6655 ± 0.1763	0.6644 ± 0.1753	0.6799 ± 0.1681
CMC	ACC	**0.3125** ± 0.0578	0.3075 ± 0.0594	0.3115 ± 0.0543	0.262 ± 0.0796	0.2523 ± 0.0808	0.2958 ± 0.0755	0.2962 ± 0.0787	0.2786 ± 0.0802	0.0065 ± 0.0625	0.2935 ± 0.0778	0.2958 ± 0.0755	0.2962 ± 0.0787	0.2958 ± 0.0755	0.247 ± 0.076
	MAE	0.1692 ± 0.0588	0.1823 ± 0.0587	0.1524 ± 0.0582	0.2856 ± 0.0732	**0.2938** ± 0.0728	0.2546 ± 0.0736	0.2608 ± 0.075	0.2647 ± 0.0743	0.0438 ± 0.1734	0.2614 ± 0.0727	0.2546 ± 0.0736	0.2607 ± 0.0749	0.2546 ± 0.0736	0.2921 ± 0.0774
	MSE	-0.0501 ± 0.0514	-0.0409 ± 0.0548	-0.0534 ± 0.0453	0.1019 ± 0.0932	**0.1219** ± 0.0901	0.0199 ± 0.0719	0.0328 ± 0.0772	0.0634 ± 0.0976	-0.0279 ± 0.1918	0.0378 ± 0.0784	0.0199 ± 0.0719	0.0328 ± 0.0772	0.0199 ± 0.0719	0.1215 ± 0.095
Grub Damage	ACC	0.1264 ± 0.2738	0.1306 ± 0.2719	0.1273 ± 0.2484	0.1327 ± 0.2264	0.1339 ± 0.2385	0.1372 ± 0.2516	0.133 ± 0.2534	0.1337 ± 0.2487	0.1091 ± 0.22	**0.1482** ± 0.2439	0.1372 ± 0.2516	0.133 ± 0.2534	0.1372 ± 0.2516	0.1351 ± 0.2253
	MAE	0.1719 ± 0.2461	0.1692 ± 0.2485	0.1861 ± 0.2196	0.2433 ± 0.1971	0.2411 ± 0.209	0.2157 ± 0.224	0.2259 ± 0.2219	0.2269 ± 0.2252	0.2375 ± 0.2462	**0.2466** ± 0.2182	0.2157 ± 0.224	0.2259 ± 0.2219	0.2157 ± 0.224	0.2389 ± 0.1984
	MSE	0.1743 ± 0.2098	0.1619 ± 0.2202	0.2294 ± 0.2574	0.2845 ± 0.2276	0.2798 ± 0.2236	0.2446 ± 0.2161	0.263 ± 0.2269	0.26 ± 0.2334	0.219 ± 0.2945	**0.2866** ± 0.2238	0.2446 ± 0.2161	0.263 ± 0.2269	0.2446 ± 0.2161	0.2736 ± 0.2305
New Thyroid	ACC	**0.9789** ± 0.0422	0.9789 ± 0.0422	0.9789 ± 0.0422	0.9342 ± 0.0607	0.9448 ± 0.0467	0.9789 ± 0.0422	0.9543 ± 0.0625	0.9543 ± 0.0625	0.099 ± 0.6567	0.9543 ± 0.0625	0.9789 ± 0.0422	0.9543 ± 0.0625	0.9789 ± 0.0422	0.9648 ± 0.0445
	MAE	0.9621 ± 0.0385	0.9621 ± 0.0385	0.9621 ± 0.0385	0.942 ± 0.0548	0.9558 ± 0.0393	**0.9793** ± 0.0415	0.9621 ± 0.054	0.9621 ± 0.054	0.2441 ± 0.6591	0.9621 ± 0.054	0.9793 ± 0.0415	0.9621 ± 0.054	0.9793 ± 0.0415	0.9759 ± 0.0301
	MSE	0.9877 ± 0.0245	0.9877 ± 0.0245	0.9877 ± 0.0245	0.9585 ± 0.0394	0.9585 ± 0.0394	**1.0** ± 0.0	0.9785 ± 0.0283	0.9785 ± 0.0283	0.2286 ± 0.6949	0.9785 ± 0.0283	1.0 ± 0.0	0.9785 ± 0.0283	1.0 ± 0.0	0.9785 ± 0.0283
Balance Scale	ACC	0.9794 ± 0.0227	0.9732 ± 0.0324	**0.9917** ± 0.0116	0.9917 ± 0.0116	0.9753 ± 0.0325	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9884 ± 0.0164	0.2731 ± 0.2291	0.9917 ± 0.0116	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9794 ± 0.0227	0.9794 ± 0.0227
	MAE	0.9794 ± 0.0227	0.9691 ± 0.0308	**0.9917** ± 0.0116	0.9917 ± 0.0116	0.9753 ± 0.0325	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9884 ± 0.0164	0.2783 ± 0.2339	0.9917 ± 0.0116	0.9794 ± 0.0227	0.9865 ± 0.0143	0.9794 ± 0.0227	0.9794 ± 0.0227
	MSE	0.97 ± 0.0301	0.9584 ± 0.0341	**0.9866** ± 0.0173	0.9849 ± 0.0204	0.9654 ± 0.0378	0.97 ± 0.0301	0.9786 ± 0.0233	0.98 ± 0.0311	0.2127 ± 0.2086	0.9866 ± 0.0173	0.97 ± 0.0301	0.9786 ± 0.0233	0.97 ± 0.0301	0.97 ± 0.0301
Automobile	ACC	0.6087 ± 0.169	0.6077 ± 0.1778	0.6159 ± 0.1666	0.6461 ± 0.1815	**0.6588** ± 0.1802	0.6278 ± 0.1757	0.6325 ± 0.1764	0.6399 ± 0.189	0.1865 ± 0.2471	0.6442 ± 0.1863	0.6278 ± 0.1757	0.6325 ± 0.1764	0.6278 ± 0.1757	0.6411 ± 0.1811
	MAE	0.5782 ± 0.1452	0.5744 ± 0.1714	0.593 ± 0.1223	**0.6597** ± 0.1187	0.6561 ± 0.1286	0.617 ± 0.1351	0.6179 ± 0.1331	0.6414 ± 0.1419	0.2662 ± 0.2664	0.6351 ± 0.1404	0.617 ± 0.1351	0.6179 ± 0.1331	0.617 ± 0.1351	0.6551 ± 0.1172
	MSE	0.4706 ± 0.2127	0.466 ± 0.2488	0.4894 ± 0.1886	**0.5927** ± 0.1659	0.5713 ± 0.1828	0.5192 ± 0.1923	0.518 ± 0.1916	0.5562 ± 0.199	0.3207 ± 0.3228	0.5419 ± 0.1982	0.5192 ± 0.1923	0.518 ± 0.1916	0.5192 ± 0.1923	0.5923 ± 0.164
Eucalyptus	ACC	0.3781 ± 0.1218	0.3764 ± 0.1245	0.3764 ± 0.1173	0.3867 ± 0.1089	**0.3923** ± 0.1165	0.3863 ± 0.1159	0.3868 ± 0.1139	0.3844 ± 0.1112	0.0567 ± 0.0865	0.3859 ± 0.1117	0.3863 ± 0.1159	0.3868 ± 0.1139	0.3863 ± 0.1159	0.3894 ± 0.1104
	MAE	0.3684 ± 0.0981	0.3666 ± 0.1014	0.3686 ± 0.0905	0.387 ± 0.0828	**0.391** ± 0.091	0.3835 ± 0.0887	0.3842 ± 0.0886	0.382 ± 0.0865	0.0736 ± 0.0971	0.3843 ± 0.0853	0.3835 ± 0.0887	0.3842 ± 0.0886	0.3835 ± 0.0887	0.39 ± 0.0846
	MSE	0.2739 ± 0.1054	0.2747 ± 0.1089	0.2716 ± 0.0992	0.2976 ± 0.1026	0.2986 ± 0.1075	0.2887 ± 0.1006	0.29 ± 0.1019	0.2887 ± 0.104	0.0735 ± 0.1641	0.2917 ± 0.1012	0.2887 ± 0.1006	0.29 ± 0.1019	0.2887 ± 0.1006	**0.3006** ± 0.1046
TAE	ACC	0.3197 ± 0.2681	0.3155 ± 0.27	**0.3699** ± 0.3052	0.2785 ± 0.2782	0.2337 ± 0.2417	0.3422 ± 0.2758	0.3452 ± 0.2962	0.3234 ± 0.3136	0.3133 ± 0.2544	0.3564 ± 0.3155	0.3422 ± 0.2758	0.3452 ± 0.2962	0.3422 ± 0.2758	0.2573 ± 0.2612
	MAE	0.2438 ± 0.2996	0.2337 ± 0.2839	0.2808 ± 0.354	0.2726 ± 0.3876	0.249 ± 0.3392	0.2676 ± 0.369	0.292 ± 0.3907	0.2932 ± 0.4059	**0.3828** ± 0.1444	0.3023 ± 0.4138	0.2676 ± 0.369	0.292 ± 0.3907	0.2676 ± 0.369	0.2512 ± 0.3604
	MSE	0.1901 ± 0.334	0.1741 ± 0.3109	0.2691 ± 0.3972	0.2444 ± 0.3838	0.2108 ± 0.3279	0.2213 ± 0.3685	0.2483 ± 0.3975	0.2431 ± 0.4024	**0.391** ± 0.2021	0.262 ± 0.4331	0.2213 ± 0.3685	0.2483 ± 0.3975	0.2213 ± 0.3685	0.2111 ± 0.3477
Heart (CLE)	ACC	0.5042 ± 0.1372	0.4765 ± 0.1319	0.5423 ± 0.1514	**0.5516** ± 0.1179	0.5317 ± 0.1152	0.5205 ± 0.1377	0.5324 ± 0.1272	0.5333 ± 0.1196	0.2361 ± 0.2612	0.5478 ± 0.1196	0.5205 ± 0.1377	0.5324 ± 0.1272	0.5205 ± 0.1377	0.546 ± 0.1148
	MAE	0.4347 ± 0.1905	0.4092 ± 0.1871	0.4843 ± 0.2058	**0.5233** ± 0.1548	0.5016 ± 0.1648	0.4802 ± 0.1861	0.4981 ± 0.174	0.4963 ± 0.164	0.2681 ± 0.2767	0.5125 ± 0.1676	0.4802 ± 0.1861	0.4981 ± 0.174	0.4802 ± 0.1861	0.5108 ± 0.1607
	MSE	0.3049 ± 0.2016	0.2876 ± 0.2024	0.3412 ± 0.2095	**0.4115** ± 0.1716	0.38 ± 0.1712	0.3476 ± 0.1871	0.3711 ± 0.1771	0.373 ± 0.1709	0.2217 ± 0.2833	0.3891 ± 0.1731	0.3476 ± 0.1871	0.3711 ± 0.1771	0.3476 ± 0.1871	0.3917 ± 0.1703
SWD	ACC	0.2174 ± 0.1341	0.2035 ± 0.1389	0.1873 ± 0.1239	0.2008 ± 0.1225	0.2119 ± 0.1324	0.2214 ± 0.1411	0.2015 ± 0.1292	0.2091 ± 0.1325	0.0415 ± 0.1136	0.195 ± 0.1249	0.2212 ± 0.141	0.2016 ± 0.1292	**0.2217** ± 0.1412	0.2164 ± 0.1402
	MAE	0.1613 ± 0.0897	0.1616 ± 0.1101	0.1315 ± 0.0816	0.1562 ± 0.0793	0.1705 ± 0.0774	0.1764 ± 0.0918	0.1526 ± 0.0846	0.1544 ± 0.0916	0.0227 ± 0.0968	0.1458 ± 0.082	0.1763 ± 0.0915	0.1525 ± 0.0848	0.1764 ± 0.0918	**0.1841** ± 0.0968
	MSE	0.138 ± 0.1001	0.1384 ± 0.1007	0.1135 ± 0.1014	0.1427 ± 0.0936	0.1584 ± 0.0762	0.153 ± 0.089	0.1307 ± 0.0999	0.1253 ± 0.1041	0.0126 ± 0.1016	0.1272 ± 0.0969	0.1528 ± 0.0887	0.1306 ± 0.1	0.1529 ± 0.0891	**0.1749** ± 0.0957
ERA	ACC	0.1484 ± 0.0781	**0.1821** ± 0.1054	0.1121 ± 0.0708	0.0047 ± 0.1062	0.0403 ± 0.1209	0.0952 ± 0.118	0.071 ± 0.1183	0.1488 ± 0.0731	0.0504 ± 0.0738	0.0481 ± 0.1096	0.0958 ± 0.1172	0.0735 ± 0.1179	0.098 ± 0.1156	0.012 ± 0.113
	MAE	0.0219 ± 0.067	-0.0038 ± 0.0893	-0.0113 ± 0.0854	-0.009 ± 0.1556	0.0116 ± 0.1469	0.0083 ± 0.1243	-0.0087 ± 0.1333	0.0104 ± 0.0947	**0.0581** ± 0.1277	-0.0255 ± 0.1349	0.0086 ± 0.1243	-0.008 ± 0.1332	0.0107 ± 0.1228	-0.0021 ± 0.1593
	MSE	0.0228 ± 0.105	-0.0043 ± 0.0952	-0.0061 ± 0.1183	0.0341 ± 0.1995	0.033 ± 0.1767	0.0109 ± 0.1596	0.0102 ± 0.1652	0.0079 ± 0.1307	**0.0503** ± 0.147	0.0059 ± 0.1697	0.0108 ± 0.1598	0.0112 ± 0.1659	0.0135 ± 0.1591	0.0382 ± 0.197
ESL	ACC	0.2265 ± 0.2031	**0.2292** ± 0.2049	0.1954 ± 0.201	0.1928 ± 0.1872	0.2216 ± 0.1887	0.225 ± 0.2076	0.2092 ± 0.2057	0.2139 ± 0.2056	-0.0048 ± 0.2012	0.1885 ± 0.1915	0.225 ± 0.2076	0.2092 ± 0.2057	0.225 ± 0.2076	0.2206 ± 0.1947
	MAE	0.2114 ± 0.1969	**0.2177** ± 0.2004	0.1775 ± 0.2006	0.1707 ± 0.1858	0.2013 ± 0.1796	0.2084 ± 0.2015	0.1904 ± 0.2018	0.1956 ± 0.2042	-0.0446 ± 0.2235	0.1648 ± 0.1918	0.2084 ± 0.2015	0.1904 ± 0.2018	0.2084 ± 0.2015	0.1981 ± 0.1837
	MSE	0.185 ± 0.2661	**0.1878** ± 0.2696	0.1476 ± 0.2701	0.1347 ± 0.2503	0.166 ± 0.2407	0.1767 ± 0.2708	0.1584 ± 0.2668	0.1643 ± 0.2749	-0.081 ± 0.2366	0.1261 ± 0.2597	0.1767 ± 0.2708	0.1584 ± 0.2668	0.1767 ± 0.2708	0.1582 ± 0.2513
LEV	ACC	0.1576 ± 0.0858	0.1371 ± 0.099	0.1584 ± 0.0962	0.1709 ± 0.1113	0.1746 ± 0.0791	0.1735 ± 0.0821	0.1724 ± 0.0888	0.1692 ± 0.0769	-0.0128 ± 0.118	0.1661 ± 0.1051	0.1735 ± 0.0821	0.1724 ± 0.0888	0.1734 ± 0.0822	**0.1812** ± 0.0897
	MAE	0.1523 ± 0.0756	0.1271 ± 0.0828	0.1669 ± 0.1055	**0.1795** ± 0.1115	0.1684 ± 0.0867	0.1716 ± 0.081	0.1785 ± 0.0988	0.1753 ± 0.0792	0.0202 ± 0.1325	0.1757 ± 0.1132	0.1716 ± 0.0808	0.1785 ± 0.0989	0.1716 ± 0.0807	0.1788 ± 0.0948
	MSE	0.1662 ± 0.0855	0.1304 ± 0.0807	0.1946 ± 0.132	**0.2056** ± 0.1313	0.1778 ± 0.1046	0.1914 ± 0.0939	0.2033 ± 0.1257	0.2047 ± 0.1066	0.0296 ± 0.1797	0.2051 ± 0.136	0.191 ± 0.0936	0.2033 ± 0.1259	0.191 ± 0.0936	0.1943 ± 0.105
Red Wine	ACC	0.2949 ± 0.0705	0.2941 ± 0.0744	0.2807 ± 0.0583	0.2831 ± 0.053	**0.3107** ± 0.0682	0.3036 ± 0.0657	0.296 ± 0.0606	0.2913 ± 0.0652	-0.025 ± 0.0925	0.2788 ± 0.0575	0.3036 ± 0.0657	0.296 ± 0.0606	0.3036 ± 0.0657	0.3038 ± 0.0582
	MAE	0.2456 ± 0.0787	0.2418 ± 0.0808	0.2458 ± 0.0718	0.2611 ± 0.0745	**0.2849** ± 0.0845	0.2673 ± 0.0815	0.2679 ± 0.0803	0.2535 ± 0.0818	-0.043 ± 0.0956	0.2527 ± 0.0751	0.2673 ± 0.0815	0.2679 ± 0.0803	0.2673 ± 0.0815	0.2797 ± 0.0782
	MSE	0.2058 ± 0.106	0.1998 ± 0.0991	0.2168 ± 0.1211	0.2462 ± 0.1319	**0.268** ± 0.1345	0.2398 ± 0.12	0.2464 ± 0.1341	0.222 ± 0.1211	-0.065 ± 0.0991	0.2315 ± 0.1322	0.2398 ± 0.12	0.2464 ± 0.1341	0.2398 ± 0.12	0.2666 ± 0.1306
White Wine	ACC	0.2471 ± 0.029	0.2377 ± 0.0318	0.2371 ± 0.0392	0.2402 ± 0.0404	0.2479 ± 0.0315	**0.2547** ± 0.0304	0.2457 ± 0.0374	0.2502 ± 0.0351	0.054 ± 0.0578	0.2396 ± 0.0409	0.2547 ± 0.0304	0.2457 ± 0.0374	0.2547 ± 0.0304	0.2489 ± 0.0354
	MAE	0.2233 ± 0.0486	0.2076 ± 0.0495	0.2295 ± 0.0552	0.2443 ± 0.0444	0.248 ± 0.0394	0.2478 ± 0.0427	0.2436 ± 0.0455	0.2451 ± 0.0449	0.0553 ± 0.0534	0.2406 ± 0.0472	0.2478 ± 0.0427	0.2436 ± 0.0455	0.2478 ± 0.0427	**0.2539** ± 0.0421
	MSE	0.1913 ± 0.0902	0.1695 ± 0.0873	0.2175 ± 0.0973	0.2519 ± 0.0725	0.2478 ± 0.0707	0.2355 ± 0.0775	0.2414 ± 0.0788	0.2389 ± 0.0811	0.0567 ± 0.0689	0.2443 ± 0.0782	0.2355 ± 0.0775	0.2414 ± 0.0788	0.2355 ± 0.0775	**0.2586** ± 0.0717



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

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

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

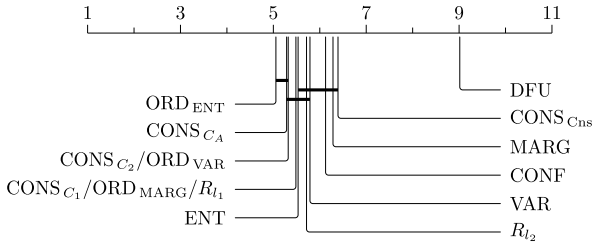
Fig. C.13. PRR values obtained over all tabular ordinal benchmark datasets and uncertainty measures grouped by underlying base learner.

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

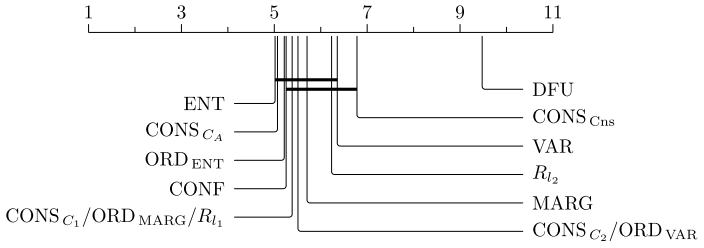
Table C.13 displays the average results of the different predictors over all datasets in terms of predictive performance (ACC, 1-OFF, MAE, MSE, and QWK) as well as calibration (negative log-likelihood (NLL), Brier Score (BS), and expected calibration error (ECE)). In summary, LGBM and SLGBM generally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

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

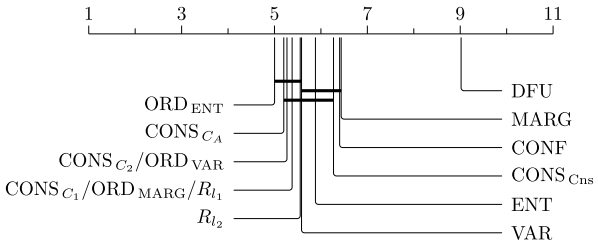
Predictor	ACC (↑)	1-OFF (↑)	MAE (↓)	MSE (↓)	QWK (↑)	NLL (↓)	BS (↓)	ECE (↓)
LGBM	**0.627** ±0.196	0.898 ±0.114	0.526 ±0.378	0.961 ±1.031	0.673 ±0.246	**1.145** ±0.591	0.520 ±0.244	0.071 ±0.047
SLGBM	0.625 ±0.198	**0.906** ±0.108	**0.506** ±0.352	**0.851** ±0.848	**0.689** ±0.233	1.693 ±1.145	**0.517**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



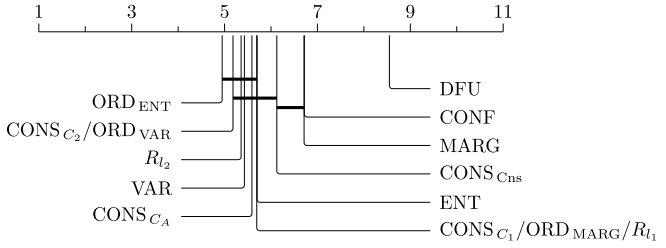
(a) Overall result for MCR, MAE and MSE.



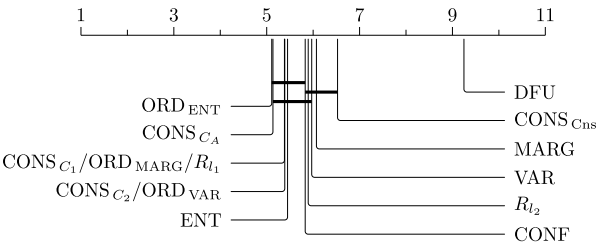
(b) Result only for MCR.



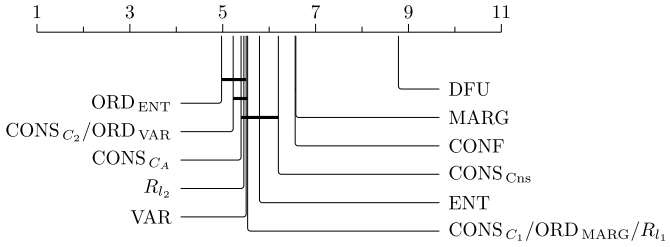
(c) Result for MAE.



(d) Result for MSE.



(e) Result for MCR and MAE.



(f) Result for MAE and MSE.

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

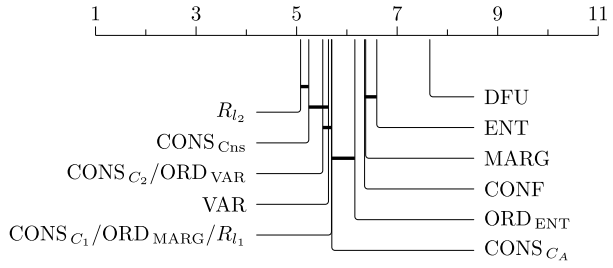
larger calibration issues in relation to cross-entropy loss, as indicated by higher NLL, BS, and ECE values. This appears to negatively impact uncertainty quantification in ordinal classification and in turn leads to smaller PRR values.

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

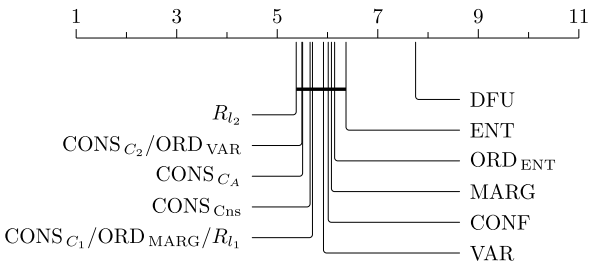
In this section, we present additional experimental results using *A Simple Approach to Ordinal Classification* with LightGBM as a binary base learner (SLGBM) [24] instead of LightGBM with CE loss (cf. Section 6). As shown in Appendix C, the simple approach to ordinal classification leads to increased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

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

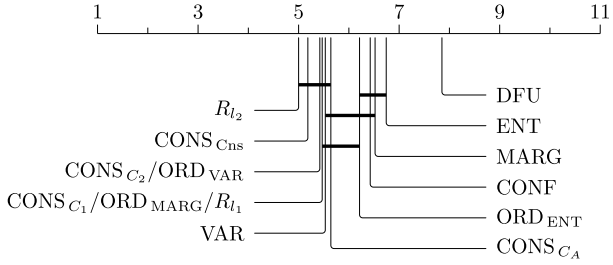
Dataset	Metric	CONF	MARG	ENT	VAR	$CONS_{Chi}$	$CONS_{C_1}$	$CONS_{C_2}$	$CONS_{C_3}$	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Triazines	ACC	0.3444 ± 0.331	0.3432 ± 0.2928	**0.4018** ± 0.2813	0.3489 ± 0.2471	0.3972 ± 0.2174	0.3413 ± 0.3261	0.3515 ± 0.2948	0.3794 ± 0.226	0.0071 ± 0.1513	0.3722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



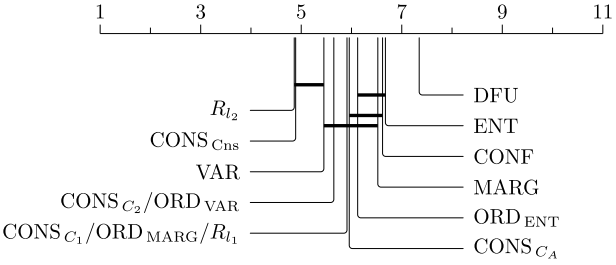
(a) Overall result for MCR, MAE and MSE.



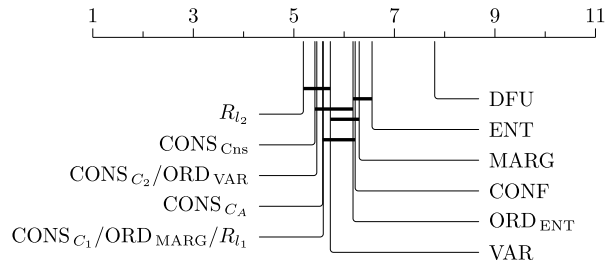
(b) Result only for MCR.



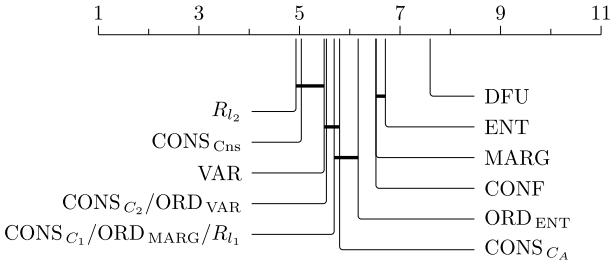
(c) Result for MAE.



(d) Result for MSE.



(e) Result for MCR and MAE.



(f) Result for MAE and MSE.

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

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

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

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

Data availability

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

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

Dataset	Metric	CONF	MARG	ENT	VAR	CONS_{Dis}	CONS_{C_1}	CONS_{C_2}	CONS_{C_3}	DFU	ORD_{ENT}	ORD_{MARG}	ORD_{VAR}	R_{J_1}	R_{J_2}
Triazines	ACC	0.2064 ± 0.3448	0.2067 ± 0.3594	0.2138 ± 0.2875	0.07 ± 0.4394	0.095 ± 0.4334	0.0979 ± 0.4061	0.1188 ± 0.4056	**0.214** ± 0.3077	-0.0423 ± 0.3118	0.1157 ± 0.3964	0.0979 ± 0.4061	0.1188 ± 0.4056	0.0979 ± 0.4061	0.0989 ± 0.4238
	MAE	0.2958 ± 0.3199	0.247 ± 0.3199	**0.3137** ± 0.2742	0.27 ± 0.2947	0.258 ± 0.3152	0.3132 ± 0.335	0.3068 ± 0.3174	0.2787 ± 0.2986	0.16 ± 0.1645	0.3061 ± 0.3072	0.3132 ± 0.335	0.3068 ± 0.3174	0.3132 ± 0.335	0.2178 ± 0.3358
	MSE	0.2251 ± 0.2632	0.1913 ± 0.2675	0.2279 ± 0.2376	0.2554 ± 0.3023	**0.2571** ± 0.2861	0.2416 ± 0.2851	0.2408 ± 0.2853	0.1877 ± 0.2737	0.0944 ± 0.3228	0.2417 ± 0.2842	0.2416 ± 0.2851	0.2408 ± 0.2853	0.2416 ± 0.2851	0.212 ± 0.2814
Machine CPU	ACC	0.3461 ± 0.3784	0.3499 ± 0.3806	0.3397 ± 0.369	0.3932 ± 0.3681	0.3779 ± 0.3895	0.3692 ± 0.3824	0.3731 ± 0.3783	0.3866 ± 0.3848	**0.4284** ± 0.4006	0.3874 ± 0.378	0.3692 ± 0.3824	0.3731 ± 0.3783	0.3692 ± 0.3824	0.3952 ± 0.3673
	MAE	0.2744 ± 0.4141	0.283 ± 0.4246	0.2693 ± 0.4063	0.4364 ± 0.399	0.3984 ± 0.4515	0.3683 ± 0.4256	0.3755 ± 0.4233	0.3793 ± 0.4227	**0.4712** ± 0.3978	0.3858 ± 0.4175	0.3683 ± 0.4256	0.3755 ± 0.4233	0.3683 ± 0.4256	0.4293 ± 0.4071
	MSE	0.1971 ± 0.5017	0.2076 ± 0.5189	0.2077 ± 0.4949	0.393 ± 0.5023	0.3325 ± 0.5526	0.2908 ± 0.5059	0.302 ± 0.5101	0.2965 ± 0.5098	**0.4114** ± 0.4518	0.3139 ± 0.495	0.2908 ± 0.5059	0.302 ± 0.5101	0.2908 ± 0.5059	0.3811 ± 0.5103
Auto MPG	ACC	0.2142 ± 0.0869	0.2102 ± 0.0985	0.2133 ± 0.0786	0.2709 ± 0.0984	0.2591 ± 0.0925	0.2445 ± 0.0863	0.2497 ± 0.0963	0.2552 ± 0.0945	**0.304** ± 0.1506	0.255 ± 0.0769	0.2445 ± 0.0863	0.2497 ± 0.0963	0.2445 ± 0.0863	0.2719 ± 0.1006
	MAE	0.2152 ± 0.0912	0.2136 ± 0.0889	0.2053 ± 0.1024	0.2672 ± 0.1198	0.267 ± 0.1135	0.2546 ± 0.1035	0.2533 ± 0.1133	0.2521 ± 0.1117	**0.2985** ± 0.1773	0.2516 ± 0.1034	0.2546 ± 0.1035	0.2533 ± 0.1133	0.2546 ± 0.1035	0.2667 ± 0.1248
	MSE	0.071 ± 0.1635	0.0788 ± 0.1682	0.0735 ± 0.1698	0.1084 ± 0.1659	0.1128 ± 0.1773	0.0978 ± 0.1618	0.0953 ± 0.1635	0.0865 ± 0.16	**0.1788** ± 0.2043	0.089 ± 0.1604	0.0978 ± 0.1618	0.0953 ± 0.1635	0.0978 ± 0.1618	0.1085 ± 0.1719
Pyrimidines	ACC	-0.0073 ± 0.4975	-0.1132 ± 0.6203	0.1058 ± 0.4352	0.0813 ± 0.5463	0.0681 ± 0.4439	0.0399 ± 0.48	0.0602 ± 0.5042	0.0288 ± 0.4958	**0.1838** ± 0.4793	0.096 ± 0.4534	0.0399 ± 0.48	0.0602 ± 0.5042	0.0399 ± 0.48	0.12 ± 0.5264
	MAE	0.0766 ± 0.3678	0.0113 ± 0.3587	0.1021 ± 0.3763	0.1421 ± 0.4204	0.0351 ± 0.4102	0.0005 ± 0.4567	0.0587 ± 0.4062	0.0816 ± 0.3881	0.0123 ± 0.3409	0.0446 ± 0.3888	0.0005 ± 0.4567	0.0587 ± 0.4062	0.0005 ± 0.4567	**0.1502** ± 0.3569
	MSE	-0.0537 ± 0.4374	0.0454 ± 0.4705	-0.0608 ± 0.4675	0.0444 ± 0.5467	-0.0369 ± 0.549	-0.0982 ± 0.5931	-0.0565 ± 0.5747	-0.079 ± 0.4904	-0.0349 ± 0.4706	-0.0667 ± 0.4583	-0.0982 ± 0.5931	-0.0565 ± 0.5747	-0.0982 ± 0.5931	**0.0655** ± 0.5233
Abalone	ACC	0.167 ± 0.0855	0.1668 ± 0.0843	0.1648 ± 0.0847	0.1774 ± 0.0953	0.1739 ± 0.0902	0.1736 ± 0.0905	0.1727 ± 0.0906	0.176 ± 0.094	0.0672 ± 0.1249	0.1733 ± 0.0929	0.1736 ± 0.0905	0.1727 ± 0.0906	0.1736 ± 0.0904	**0.1783** ± 0.0964
	MAE	0.1533 ± 0.1251	0.1531 ± 0.125	0.1515 ± 0.1227	0.1624 ± 0.1299	0.1611 ± 0.1291	0.1597 ± 0.1289	0.159 ± 0.1281	0.1605 ± 0.1296	0.0483 ± 0.1313	0.1591 ± 0.1286	0.1597 ± 0.1289	0.159 ± 0.128	0.1597 ± 0.1289	**0.1637** ± 0.1324
	MSE	0.1213 ± 0.1743	0.1202 ± 0.1755	0.123 ± 0.1706	0.1272 ± 0.1667	0.127 ± 0.1727	0.1248 ± 0.1721	0.1244 ± 0.1695	0.1241 ± 0.1686	0.0066 ± 0.1428	0.125 ± 0.1673	0.1248 ± 0.1721	0.1244 ± 0.1694	0.1248 ± 0.1721	**0.1285** ± 0.1716
Boston Housing	ACC	0.3587 ± 0.2232	0.3592 ± 0.2235	0.3508 ± 0.2276	0.3876 ± 0.2397	**0.4095** ± 0.2474	0.3961 ± 0.2371	0.3905 ± 0.2399	0.3925 ± 0.2366	-0.1767 ± 0.211	0.3817 ± 0.2364	0.3961 ± 0.2371	0.3905 ± 0.2399	0.3961 ± 0.2371	0.405 ± 0.2434
	MAE	0.3153 ± 0.1943	0.3218 ± 0.1958	0.3149 ± 0.2118	0.3544 ± 0.2373	**0.3782** ± 0.2384	0.3592 ± 0.2233	0.3583 ± 0.2345	0.3579 ± 0.2288	-0.1223 ± 0.1858	0.35 ± 0.2354	0.3592 ± 0.2233	0.3583 ± 0.2345	0.3592 ± 0.2233	0.378 ± 0.2371
	MSE	0.2328 ± 0.2327	0.2406 ± 0.2368	0.2342 ± 0.2424	0.2766 ± 0.2461	**0.2986** ± 0.2511	0.2668 ± 0.2415	0.274 ± 0.2557	0.2732 ± 0.2497	-0.0354 ± 0.2498	0.2703 ± 0.2497	0.2668 ± 0.2415	0.274 ± 0.2557	0.2668 ± 0.2415	0.295 ± 0.2492
Stocks Domain	ACC	-0.0551 ± 0.2565	-0.0574 ± 0.2644	-0.0631 ± 0.2404	-0.053 ± 0.2428	-0.047 ± 0.2598	-0.0496 ± 0.2543	-0.0558 ± 0.245	-0.0496 ± 0.2459	**0.1443** ± 0.4103	-0.0566 ± 0.2387	-0.0496 ± 0.2543	-0.0558 ± 0.245	-0.0496 ± 0.2543	-0.0433 ± 0.2529
	MAE	-0.0607 ± 0.2625	-0.063 ± 0.2705	-0.0687 ± 0.2464	-0.0586 ± 0.2489	-0.0526 ± 0.266	-0.0552 ± 0.2602	-0.0614 ± 0.251	-0.0552 ± 0.2517	**0.1491** ± 0.4081	-0.0622 ± 0.2449	-0.0552 ± 0.2602	-0.0614 ± 0.251	-0.0552 ± 0.2602	-0.0489 ± 0.2587
	MSE	-0.0684 ± 0.2664	-0.0711 ± 0.2735	-0.0727 ± 0.2564	-0.0626 ± 0.2584	-0.0587 ± 0.2723	-0.0625 ± 0.2644	-0.0668 ± 0.2585	-0.062 ± 0.2563	**0.1453** ± 0.3992	-0.0659 ± 0.2547	-0.0625 ± 0.2644	-0.0668 ± 0.2585	-0.0625 ± 0.2644	-0.0559 ± 0.2632
Wisconsin Breast Cancer	ACC	0.3744 ± 0.2153	0.3546 ± 0.1934	**0.4269** ± 0.2601	0.3238 ± 0.274	0.3032 ± 0.2687	0.376 ± 0.2436	0.3822 ± 0.2527	0.4052 ± 0.261	-0.1898 ± 0.4986	0.3649 ± 0.2516	0.376 ± 0.2436	0.3822 ± 0.2527	0.376 ± 0.2436	0.3069 ± 0.276
	MAE	0.191 ± 0.1801	**0.1966** ± 0.1695	0.1942 ± 0.1926	0.0741 ± 0.1887	0.075 ± 0.1805	0.1458 ± 0.1712	0.1345 ± 0.1853	0.1689 ± 0.1899	-0.1297 ± 0.1813	0.1142 ± 0.1763	0.1458 ± 0.1712	0.1345 ± 0.1853	0.1458 ± 0.1712	0.0764 ± 0.1738
	MSE	0.1998 ± 0.2633	**0.2296** ± 0.2588	0.1721 ± 0.2311	0.0225 ± 0.1439	0.0185 ± 0.1416	0.0824 ± 0.1765	0.0811 ± 0.1635	0.1384 ± 0.1951	-0.1149 ± 0.2102	0.0659 ± 0.1546	0.0824 ± 0.1765	0.0811 ± 0.1635	0.0824 ± 0.1765	0.044 ± 0.1345
Obesity	ACC	0.1729 ± 0.5138	0.1941 ± 0.5197	0.1362 ± 0.5001	0.2807 ± 0.565	0.2802 ± 0.5589	0.2408 ± 0.5429	0.2411 ± 0.5438	0.258 ± 0.5524	**0.39** ± 0.4871	0.2298 ± 0.5408	0.2408 ± 0.5429	0.2411 ± 0.5438	0.2408 ± 0.5429	0.2855 ± 0.5659
	MAE	0.1709 ± 0.5167	0.1921 ± 0.5222	0.1353 ± 0.5031	0.2771 ± 0.5663	0.2771 ± 0.5606	0.2385 ± 0.5453	0.2389 ± 0.5461	0.2556 ± 0.5541	**0.3865** ± 0.4887	0.2281 ± 0.5429	0.2385 ± 0.5453	0.2389 ± 0.5461	0.2385 ± 0.5453	0.282 ± 0.5673
	MSE	0.1159 ± 0.4862	0.1388 ± 0.4905	0.0826 ± 0.4744	0.2341 ± 0.5299	0.2315 ± 0.5256	0.1898 ± 0.5108	0.1908 ± 0.5113	0.2104 ± 0.518	**0.3379** ± 0.4419	0.1812 ± 0.5072	0.1898 ± 0.5108	0.1908 ± 0.5113	0.1898 ± 0.5108	0.2383 ± 0.5316
CMC	ACC	0.2667 ± 0.0778	0.263 ± 0.0759	**0.2707** ± 0.0855	0.2492 ± 0.0853	0.2431 ± 0.0843	0.2596 ± 0.0859	0.26 ± 0.0836	0.2579 ± 0.0829	0.1854 ± 0.0799	0.2597 ± 0.0848	0.2596 ± 0.0859	0.2601 ± 0.0836	0.2597 ± 0.0859	0.2455 ± 0.0873
	MAE	0.1852 ± 0.1029	0.1816 ± 0.1021	0.1908 ± 0.1074	0.2039 ± 0.1056	0.2028 ± 0.1047	0.1973 ± 0.1068	0.1995 ± 0.1044	0.197 ± 0.1035	0.0643 ± 0.0686	0.2003 ± 0.1049	0.1973 ± 0.1068	0.1996 ± 0.1044	0.1974 ± 0.1068	**0.2048** ± 0.1095
	MSE	0.0434 ± 0.0994	0.0431 ± 0.0993	0.0489 ± 0.0961	0.0898 ± 0.1005	**0.0915** ± 0.1005	0.0636 ± 0.099	0.0709 ± 0.0977	0.0711 ± 0.0983	-0.0183 ± 0.0753	0.0736 ± 0.0983	0.0636 ± 0.099	0.0709 ± 0.0977	0.0637 ± 0.099	0.0904 ± 0.1039
Grub Damage	ACC	0.2412 ± 0.2663	**0.2611** ± 0.2905	0.2268 ± 0.2241	0.1932 ± 0.2722	0.2001 ± 0.2989	0.2141 ± 0.2712	0.1897 ± 0.2733	0.1887 ± 0.276	0.22 ± 0.1992	0.1908 ± 0.2722	0.2141 ± 0.2712	0.1897 ± 0.2733	0.2141 ± 0.2712	0.2213 ± 0.2655
	MAE	0.2292 ± 0.2824	0.214 ± 0.2849	0.2246 ± 0.2639	**0.3042** ± 0.2489	0.2949 ± 0.2512	0.2411 ± 0.2645	0.2644 ± 0.2558	0.2401 ± 0.2607	0.1989 ± 0.2189	0.2656 ± 0.2472	0.2411 ± 0.2645	0.2644 ± 0.2558	0.2411 ± 0.2645	0.2855 ± 0.2482
	MSE	0.1439 ± 0.269	0.1316 ± 0.2577	0.1671 ± 0.279	**0.2846** ± 0.315	0.2774 ± 0.2835	0.1826 ± 0.2849	0.2294 ± 0.2958	0.1653 ± 0.2883	0.1511 ± 0.3198	0.2282 ± 0.3022	0.1826 ± 0.2849	0.2294 ± 0.2958	0.1826 ± 0.2849	0.2192 ± 0.2714
New Thyroid	ACC	0.9571 ± 0.0413	0.9646 ± 0.0301	0.9471 ± 0.0475	0.9451 ± 0.0782	0.925 ± 0.0743	0.9646 ± 0.0301	0.9545 ± 0.04	**0.9651** ± 0.0435	0.2908 ± 0.2841	0.9651 ± 0.0435	0.9646 ± 0.0301	0.9545 ± 0.04	0.9646 ± 0.0301	0.9625 ± 0.0582
	MAE	0.9535 ± 0.0336	0.949 ± 0.0347	0.9501 ± 0.0457	0.9646 ± 0.0512	0.9445 ± 0.053	0.9636 ± 0.0381	0.957 ± 0.0398	**0.9708** ± 0.0258	0.4562 ± 0.2503	0.9708 ± 0.0258	0.9636 ± 0.0381	0.957 ± 0.0398	0.9636 ± 0.0381	0.9622 ± 0.0487
	MSE	0.9375 ± 0.0642	0.9216 ± 0.0802	0.9405 ± 0.0867	0.9727 ± 0.041	0.956 ± 0.0465	0.9664 ± 0.0535	0.9605 ± 0.0513	**0.9742** ± 0.0271	0.5876 ± 0.3271	0.9742 ± 0.0271	0.9664 ± 0.0535	0.9605 ± 0.0513	0.9664 ± 0.0535	0.9742 ± 0.0271
Balance Scale	ACC	0.9392 ± 0.0679	0.9299 ± 0.0656	0.9387 ± 0.0487	0.9363 ± 0.0516	0.7979 ± 0.1368	0.9392 ± 0.0679	0.9411 ± 0.047	**0.9437** ± 0.0449	0.1831 ± 0.2245	0.9387 ± 0.0487	0.9392 ± 0.0679	0.9411 ± 0.047	0.9392 ± 0.0679	0.9369 ± 0.0666
	MAE	**0.9468** ± 0.0666	0.9386 ± 0.0656	0.934 ± 0.0472	0.932 ± 0.0496	0.815 ± 0.1357	0.9468 ± 0.0666	0.9403 ± 0.044	0.9425 ± 0.0421	0.0392 ± 0.4347	0.934 ± 0.0472	0.9468 ± 0.0666	0.9403 ± 0.044	0.9468 ± 0.0666	0.9448 ± 0.0657
	MSE	**0.9427** ± 0.0642	0.9345 ± 0.0626	0.934 ± 0.0472	0.932 ± 0.0496	0.8115 ± 0.1389	0.9427 ± 0.0642	0.9403 ± 0.044	0.9425 ± 0.0421	0.0444 ± 0.44	0.934 ± 0.0472	0.9427 ± 0.0642	0.9403 ± 0.044	0.9427 ± 0.0642	0.9407 ± 0.0631
Automobile	ACC	0.4039 ± 0.2083	0.3575 ± 0.208	0.3944 ± 0.2328	0.3871 ± 0.2428	0.4057 ± 0.2179	**0.4281** ± 0.2125	0.4231 ± 0.2194	0.4127 ± 0.2217	0.0335 ± 0.3054	0.3808 ± 0.2369	0.4281 ± 0.2125	0.4231 ± 0.2194	0.4281 ± 0.2125	0.4003 ± 0.2364
	MAE	0.3791 ± 0.2187	0.3395 ± 0.225	0.372 ± 0.2478	0.3745 ± 0.2551	0.3852 ± 0.2347	**0.4015** ± 0.2241	0.3989 ± 0.2347	0.394 ± 0.2299	0.097 ± 0.3128	0.3641 ± 0.2539	0.4015 ± 0.2241	0.3989 ± 0.2347	0.4015 ± 0.2241	0.387 ± 0.2443
	MSE	0.3813 ± 0.2548	0.3604 ± 0.2789	0.3551 ± 0.2827	0.3715 ± 0.298	0.3683 ± 0.2773	**0.3826** ± 0.2643	0.3766 ± 0.2904	0.3725 ± 0.2869	0.1954 ± 0.2853	0.354 ± 0.307	0.3826 ± 0.2643	0.3766 ± 0.2904	0.3826 ± 0.2643	0.3773 ± 0.2784
Eucalyptus	ACC	0.3706 ± 0.1	0.3641 ± 0.0943	0.3607 ± 0.0943	0.3735 ± 0.1111	0.3752 ± 0.1054	**0.3767** ± 0.1003	0.3694 ± 0.1033	0.3723 ± 0.1057	0.0896 ± 0.1217	0.3661 ± 0.1113	0.3767 ± 0.1003	0.3694 ± 0.1033	0.3767 ± 0.1003	0.3766 ± 0.1052
	MAE	0.3778 ± 0.0866	0.3629 ± 0.0738	0.3819 ± 0.0776	**0.4029** ± 0.0927	0.3974 ± 0.0946	0.3945 ± 0.0873	0.3924 ± 0.0872	0.3917 ± 0.095	0.1264 ± 0.1097	0.3924 ± 0.0949	0.3945 ± 0.0873	0.3924 ± 0.0872	0.3945 ± 0.0873	0.4002 ± 0.0925
	MSE	0.3194 ± 0.0993	0.3053 ± 0.0959	0.3345 ± 0.0896	**0.3644** ± 0.0946	0.3479 ± 0.1054	0.3384 ± 0.1037	0.3444 ± 0.0972	0.3395 ± 0.1054	0.1595 ± 0.0999	0.3507 ± 0.0928	0.3384 ± 0.1037	0.3444 ± 0.0972	0.3384 ± 0.1037	0.3534 ± 0.1009
TAE	ACC	**0.1403** ± 0.3363	0.1152 ± 0.3569	0.1364 ± 0.3115	0.0605 ± 0.3617	0.0442 ± 0.401	0.0644 ± 0.42	0.097 ± 0.3412	0.1167 ± 0.3487	-0.0477 ± 0.3118	0.097 ± 0.3412	0.0644 ± 0.42	0.097 ± 0.3412	0.0644 ± 0.42	0.0897 ± 0.3479
	MAE	0.0243 ± 0.2865	0.0335 ± 0.3057	0.0089 ± 0.2685	0.014 ± 0.2789	0.0152 ± 0.3019	0.0397 ± 0.3007	0.0267 ± 0.2648	0.0185 ± 0.3052	-0.0786 ± 0.3735	0.0303 ± 0.2685	0.0397 ± 0.3007	0.0267 ± 0.2648	0.0397 ± 0.3007	**0.0877** ± 0.2685
	MSE	-0.0714 ± 0.3229	-0.0577 ± 0.3411	-0.0884 ± 0.3012	-0.0719 ± 0.307	-0.043 ± 0.31	-0.0322 ± 0.3198	-0.0697 ± 0.2965	-0.068 ± 0.3337	-0.1015 ± 0.4367	-0.0661 ± 0.3009	-0.0322 ± 0.3198	-0.0697 ± 0.2965	-0.0322 ± 0.3198	**0.0429** ± 0.3208
Heart (CLE)	ACC	0.3971 ± 0.2242	0.3758 ± 0.2277	0.4059 ± 0.2198	0.3473 ± 0.2027	0.3149 ± 0.2126	0.3762 ± 0.2097	0.3763 ± 0.2036	**0.4244** ± 0.1774	0.2186 ± 0.1934	0.3847 ± 0.2019	0.3762 ± 0.2097	0.3763 ± 0.2036	0.3762 ± 0.2097	0.3389 ± 0.2074
	MAE	0.3405 ± 0.238	0.317 ± 0.2501	0.3509 ± 0.239	0.3213 ± 0.2522	0.2859 ± 0.2681	0.3332 ± 0.2451	0.3315 ± 0.2426	**0.3703** ± 0.2013	0.1925 ± 0.2498	0.3411 ± 0.2326	0.3332 ± 0.2451	0.3315 ± 0.2426	0.3332 ± 0.2451	0.3088 ± 0.2491
	MSE	0.3101 ± 0.2305	0.2789 ± 0.2503	0.326 ± 0.2351	0.3161 ± 0.2696	0.272 ± 0.2845	0.3063 ± 0.2504	0.3089 ± 0.2511	**0.3437** ± 0.2111	0.2026 ± 0.335	0.3205 ± 0.2384	0.3063 ± 0.2504	0.3089 ± 0.2511	0.3063 ± 0.2504	0.2966 ± 0.2636
SWD	ACC	0.1853 ± 0.0955	0.1928 ± 0.0975	0.1722 ± 0.0739	0.1864 ± 0.0816	**0.1954** ± 0.0952	0.1877 ± 0.0964	0.1825 ± 0.0829	0.1772 ± 0.0831	0.0163 ± 0.0923	0.174 ± 0.0749	0.1877 ± 0.0964	0.1825 ± 0.0829	0.1877 ± 0.0964	0.1934 ± 0.0952
	MAE	0.1879 ± 0.0969	0.1963 ± 0.0975	0.1741 ± 0.0787	0.1905 ± 0.0856	**0.2024** ± 0.1009	0.192 ± 0.0984	0.1858 ± 0.0854	0.1792 ± 0.0865	0.0065 ± 0.0943	0.1762 ± 0.0803	0.192 ± 0.0984	0.1858 ± 0.0854	0.192 ± 0.0984	0.1979 ± 0.0985
	MSE	0.1876 ± 0.1183	0.1969 ± 0.118	0.1732 ± 0.106	0.1928 ± 0.1096	**0.2075** ± 0.1266	0.194 ± 0.1201	0.1863 ± 0.1086	0.1785 ± 0.1101	-0.0019 ± 0.1167	0.1759 ± 0.1067	0.194 ± 0.1201	0.1863 ± 0.1086	0.194 ± 0.1201	0.2005 ± 0.1199
ERA	ACC	0.1439 ± 0.1203	0.1255 ± 0.1142	0.1652 ± 0.1106	**0.181** ± 0.111	0.1387 ± 0.1281	0.1545 ± 0.1149	0.1672 ± 0.1127	0.1735 ± 0.0989	0.1034 ± 0.1044	0.1795 ± 0.1103	0.1545 ± 0.1149	0.1672 ± 0.1127	0.1545 ± 0.1149	0.1654 ± 0.1269
	MAE	0.0426 ± 0.148	0.0423 ± 0.1443	0.0398 ± 0.1399	0.1478 ± 0.1265	0.1388 ± 0.1559	0.1036 ± 0.1505	0.1063 ± 0.1452	0.1105 ± 0.1401	**0.167** ± 0.0988	0.1046 ± 0.1339	0.1036 ± 0.1505	0.1063 ± 0.1452	0.1036 ± 0.1505	0.1438 ± 0.1379
	MSE	-0.039 ± 0.1963	-0.0261 ± 0.2045	-0.0584 ± 0.1814	0.0828 ± 0.14	0.0793 ± 0.1813	0.0264 ± 0.1908	0.0184 ± 0.1749	0.0263 ± 0.181	**0.1836** ± 0.1313	0.013 ± 0.1649	0.0264 ± 0.1908	0.0184 ± 0.1749	0.0264 ± 0.1908	0.0844 ± 0.1492
ESL	ACC	0.0922 ± 0.1431	0.0879 ± 0.1542	0.0755 ± 0.1455	0.097 ± 0.1717	0.0977 ± 0.1914	0.1017 ± 0.1664	0.1024 ± 0.1743	0.1084 ± 0.1705	-0.0491 ± 0.1696	0.0842 ± 0.1722	0.1017 ± 0.1664	0.102 ± 0.1746	0.1013 ± 0.1668	**0.1137** ± 0.1794
	MAE	0.0744 ± 0.178	0.0707 ± 0.187	0.0574 ± 0.183	0.0717 ± 0.2019	0.0802 ± 0.2176	0.0818 ± 0.1963	0.0806 ± 0.2044	0.0837 ± 0.1997	-0.0963 ± 0.1582	0.0615 ± 0.2035	0.0818 ± 0.1963	0.0803 ± 0.2047	0.0814 ± 0.1967	**0.0906** ± 0.2084
	MSE	0.0146 ± 0.2086	0.0123 ± 0.2136	-0.0044 ± 0.2241	0.0094 ± 0.232	0.0276 ± 0.2318	0.023 ± 0.2204	0.0221 ± 0.23	0.02 ± 0.2223	-0.1868 ± 0.1086	0.0027 ± 0.2366	0.023 ± 0.2204	0.0218 ± 0.2303	0.0227 ± 0.2208	**0.029** ± 0.2298
LEV	ACC	0.0772 ± 0.1122	0.0904 ± 0.1233	0.0619 ± 0.1076	0.0669 ± 0.0991	**0.0993** ± 0.1215	0.0809 ± 0.1056	0.077 ± 0.1038	0.06 ± 0.1008	-0.036 ± 0.0888	0.0591 ± 0.101	0.0809 ± 0.1056	0.077 ± 0.1038	0.0809 ± 0.1056	0.0771 ± 0.1016
	MAE	0.0489 ± 0.1018	0.062 ± 0.1128	0.0362 ± 0.0973	0.0457 ± 0.0837	**0.076** ± 0.1055	0.0557 ± 0.0914	0.052 ± 0.0907	0.0379 ± 0.0838	-0.0023 ± 0.0945	0.0358 ± 0.0883	0.0557 ± 0.0914	0.052 ± 0.0907	0.0557 ± 0.0914	0.0542 ± 0.0862
	MSE	0.0118 ± 0.1021	0.0242 ± 0.1114	0.0029 ± 0.0967	0.0199 ± 0.0879	**0.0469** ± 0.0996	0.0233 ± 0.0901	0.0203 ± 0.091	0.0096 ± 0.0818	0.0443 ± 0.1486	0.0067 ± 0.0911	0.0233 ± 0.0901	0.0203 ± 0.091	0.0233 ± 0.0901	0.0257 ± 0.0871
Red Wine	ACC	0.1695 ± 0.0622	**0.1698** ± 0.0614	0.1679 ± 0.0622	0.1643 ± 0.0597	0.1679 ± 0.0623	0.1673 ± 0.0614	0.1681 ± 0.0615	0.1661 ± 0.0622	-0.0272 ± 0.1224	0.1668 ± 0.0614	0.1679 ± 0.0615	0.1676 ± 0.0613	0.1675 ± 0.0613	0.1646 ± 0.06
	MAE	**0.1529** ± 0.0615	0.1527 ± 0.0607	0.1501 ± 0.0633	0.1459 ± 0.0593	0.1505 ± 0.0622	0.1507 ± 0.0605	0.1507 ± 0.0622	0.1495 ± 0.0609	-0.0433 ± 0.1172	0.1488 ± 0.062	0.1513 ± 0.0608	0.1502 ± 0.0617	0.1508 ± 0.0604	0.1477 ± 0.0582
	MSE	**0.1281** ± 0.0717	0.1271 ± 0.0701	0.124 ± 0.076	0.1192 ± 0.0708	0.1248 ± 0.0716	0.1259 ± 0.07	0.1251 ± 0.0738	0.1249 ± 0.072	-0.0696 ± 0.1164	0.1225 ± 0.0744	0.1265 ± 0.0706	0.1246 ± 0.073	0.126 ± 0.0698	0.1224 ± 0.0676
White Wine	ACC	0.0931 ± 0.0476	0.0922 ± 0.0474	0.0932 ± 0.0465	0.0927 ± 0.0455	0.0928 ± 0.047	0.0933 ± 0.0472	0.0936 ± 0.047	**0.0942** ± 0.0473	0.0219 ± 0.0575	0.0929 ± 0.0459	0.0934 ± 0.0473	0.0934 ± 0.0469	0.0933 ± 0.0472	0.0926 ± 0.0465
	MAE	0.0977 ± 0.0545	0.0965 ± 0.0543	0.0981 ± 0.0543	0.0993 ± 0.0545	0.0985 ± 0.0548	0.0988 ± 0.0547	0.0992 ± 0.0548	**0.1** ± 0.0552	0.027 ± 0.0587	0.0989 ± 0.0545	0.0988 ± 0.0548	0.099 ± 0.0549	0.0988 ± 0.0548	0.099 ± 0.0548
	MSE	0.1065 ± 0.0756	0.1049 ± 0.0753	0.1076 ± 0.0765	**0.1118** ± 0.0798	0.1092 ± 0.078	0.1092 ± 0.0775	0.1099 ± 0.078	0.111 ± 0.0788	0.0405 ± 0.0631	0.1104 ± 0.079	0.1092 ± 0.0774	0.1097 ± 0.0781	0.1092 ± 0.0774	0.1109 ± 0.0789

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.6 Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction

Contributing Article

Stefan Haas and Eyke Hüllermeier. “Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction”. In: *Machine Learning* (2026)

Author Contribution Statement

Inspired by prior work on label-wise aleatoric and epistemic uncertainty quantification [Sal+24], the author proposed leveraging the previously developed ordinal binary decomposition method to separate aleatoric and epistemic uncertainty in ordinal classification. The author designed the specific ordinal approaches, implemented them, conducted all experiments, and authored the manuscript. The entire paper underwent multiple rounds of revision in collaboration between Prof. Dr. Hüllermeier and the author.

Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction

Stefan Haas[1,2*] and Eyke Hüllermeier[1,3,4]

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

*Corresponding author(s). E-mail(s): stefan.haas@campus.lmu.de; stefan.sh.haas@bmwgroup.com;
Contributing authors: eyke@lmu.de;

Abstract

Ordinal classification problems, where labels exhibit a natural order, are prevalent in high-stakes fields such as medicine and finance. Accurate uncertainty quantification, including the decomposition into aleatoric (inherent variability) and epistemic (lack of knowledge) components, is crucial for reliable decision-making. However, existing research has primarily focused on nominal classification and regression. In this paper, we introduce a novel class of measures of aleatoric and epistemic uncertainty in ordinal classification, which is based on a suitable reduction to (entropy- and variance-based) measures for the binary case. These measures effectively capture the trade-off in ordinal classification between exact hit-rate and minimial error distances. We demonstrate the effectiveness of our approach on various tabular ordinal benchmark datasets using ensembles of gradient-boosted trees and multi-layer perceptrons for approximate Bayesian inference. Our method significantly outperforms standard and label-wise entropy and variance-based measures in error detection, as indicated by misclassification rates and mean absolute error. Additionally, the ordinal measures show competitive performance in out-of-distribution (OOD) detection. Our findings highlight the importance of considering the ordinal nature of classification problems when assessing uncertainty.

Keywords: Ordinal Classification, Ordinal Regression, Uncertainy Quantification, Aleatoric Uncertainty, Epistemic Uncertainty, Binary Reduction

1 Introduction

Supervised machine learning models are increasingly used for high-stakes decision-making in domains such as medicine and finance. Consider predicting treatment effects (Rafique, Islam, & Kazi, 2021) or automating loan approvals (Uddin et al., 2023). Likewise, quantifying the *predictive uncertainty* associated with a query $\boldsymbol{x}_q$ becomes more and more important for reliable and safe decision-making. Information about the predictive uncertainty could, for instance, be used in a downstream selective classification (Geifman & El-Yaniv, 2017; Hendrickx, Perini, der Plas, Meert, & Davis, 2024) approach in which only certain enough queries are processed automatically while uncertain ones are delegated to human experts. This, in turn, reduces the risk of wrong predictions and increases the overall accuracy of the predictor (Haas & Hüllermeier, 2025a).

A common distinction in uncertainty quantification is drawn between so-called *aleatoric* and *epistemic* uncertainty (Hüllermeier & Waegeman, 2021; Senge et al., 2014). The latter refers to the uncertainty that arises due to a lack of knowledge or information, e.g., previously unseen medical cases in clinical diagnosis or novel traffic scenarios in autonomous driving. It can be reduced with additional training data or by selecting a better predictor or model for the specific task. In contrast, aleatoric uncertainty is irreducible and arises from the inherent randomness in the data, e.g., disagreement among physicians interpreting the same clinical image or differing evaluations of a portfolio among financial analysts. Identifying and measuring these uncertainties allows for more nuanced detection of issues related to the learning and prediction process. For example, information about aleatoric uncertainty empowers decision-makers to assess whether a dependable decision can be reached at all. Likewise, understanding epistemic uncertainty provides valuable insights into whether a given predictor is sufficiently informed to make reliable decisions or requires additional training data or a different type of model or model class. Concrete applications where this distinction has proven beneficial include active learning (Nguyen, Shaker, & Hüllermeier, 2022), where instances with high epistemic uncertainty are preferentially selected to improve the model, and out-of-distribution (OOD) detection (Lu et al., 2025), where high epistemic uncertainty typically indicates samples lying outside the previously seen data distribution.

In general, the focus of uncertainty quantification in machine learning has primarily been on nominal classification and regression, with the Bayesian approach prevailing in the literature, where epistemic uncertainty is represented by a second-order probability distribution representing the posterior over the hypothesis space: a probability distribution of probability distributions (Hüllermeier & Waegeman, 2021). In practice, such second-order distributions are commonly approximated through Monte Carlo sampling using ensembles (Malinin, Prokhorenkova, & Ustimenko, 2021; Sale et al., 2024; Shaker & Hüllermeier, 2020; Wimmer, Sale, Hofman, Bischl, & Hüllermeier, 2023) or other variational techniques such as dropout (Gal & Ghahramani, 2016) or drop connect (Mobiny, Nguyen, Moulik, Garg, & Wu, 2021) in deep learning.

A principled approach to measuring and separating aleatoric and epistemic uncertainty on the basis of classical information-theoretic measures of (Shannon) entropy (Shannon, 1948) was proposed by Depeweg, Hernandez-Lobato, Doshi-Velez, and

Udluft (2018). The approach has widely been adopted for nominal classification (Löhr, Ingrisch, & Hüllermeier, 2024; Malinin et al., 2021; Mobiny et al., 2021; Saberi, Shaker, Duguay, Scott, & Hüllermeier, 2024; Shaker & Hüllermeier, 2020; Shaker & Hüllermeier, 2021). However, particularly in high-stakes use-cases, the class labels $y \in \mathcal{Y}$ often exhibit a natural order relation, with $y_1 \prec y_2 \prec \cdots \prec y_K$. Think of credit scoring with $\mathcal{Y} = \{\text{poor, fair, good, very good, excellent}\}$ or any other rating application, such as disease severity in medicine or employee performance evaluation in human resources. Since ordinal classification lies somewhat between classification and regression, commonly used entropy and variance-based approaches are also applicable in the ordinal case (Malinin et al., 2021). However, these approaches may not ideally reflect the inherent trade-off within ordinal classification between exact hit-rate and minimized error distances, as has been shown by Haas and Hüllermeier (2025b) for entropy. Entropy is not a good choice for uncertainty quantification in ordinal classification, as it does not take into account the dispersion of a probability distribution and is invariant to the redistribution of probability mass (cf. Figure 1 for an illustration). To the best of our knowledge, no efforts have been devoted to disentangling aleatoric and epistemic uncertainty in the context of ordinal classification yet, which is crucial for reliable decision-making in many high-stakes use cases.

In this paper, we make the following key contributions:

- We explore methods to disentangle aleatoric and epistemic uncertainty in ordinal classification based on commonly used entropy and variance-based decomposition approaches.
- We propose a novel binary decomposition method that builds upon commonly used entropy and total variance-based decompositions for quantifying aleatoric and epistemic uncertainty, taking the ordinal structure into account.
- We compare our novel method with standard entropy- and total variance-based approaches, as well as with label-wise entropy and variance-based binary decompositions, in an extensive study on twenty-three common tabular ordinal benchmark datasets using ensembles of gradient-boosted trees and multi-layer perceptrons for approximate Bayesian inference, specifically evaluating performance in error detection and out-of-distribution detection.
- Moreover, we demonstrate that the cross-entropy (CE) loss, as a proper scoring rule, is advantageous over ordinal losses for uncertainty quantification in ordinal classification in general. While ordinal losses often induce compressed unimodal predictive probability distributions, which can serve as a beneficial inductive bias for predictions, particularly for discretized continuous ordinal targets, they negatively impact uncertainty quantification.

2 Learning Probabilistic Predictors

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

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

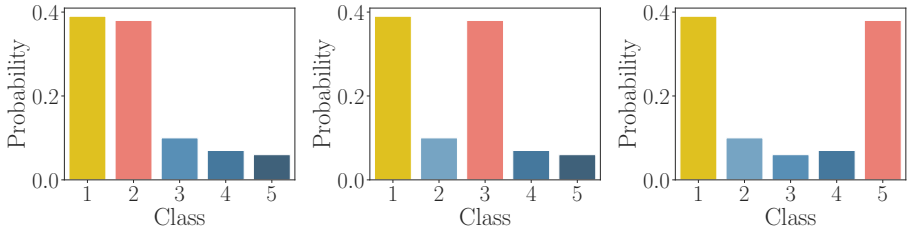


Fig. 1: Several probability distributions which lead to the same (Shannon) entropy ($\mathbb{H}$ = 1.9 with base 2). In ordinal classification, where the minimization of error distance is an important factor in addition to the exact hit-rate, this behavior has been shown to be problematic (Haas & Hüllermeier, 2025b). Obviously, the different distributions should be associated with different degrees of uncertainty, arguably increasing from left to right in this example.

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

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

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

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

where $l : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_+$ is a loss (error) function. Training probabilistic predictors is typically accomplished by minimizing the (perhaps regularized) empirical risk

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

as an estimate of the true risk (generalization performance). Then, the empirical risk minimizer

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

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

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

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

So-called (strictly) proper scoring rules (Gneiting & Raftery, 2007) are commonly used as loss functions l in (1). These have the nice theoretical property of being minimized (in expectation) by the true conditional probabilities, hence incentivizing the learner to produce well-calibrated probability estimates (2). An important example of such loss functions is the log-loss or cross-entropy loss (CE)

$$l_{CE}(\boldsymbol{p}, y) = -\sum_{k=1}^{K} [\![y = y_k]\!] \log(p_k)\,, \tag{3}$$

where $[\![\cdot]\!]$ denotes the indicator function.

In the context of ordinal classification, one might be tempted to prefer dedicated ordinal losses, such as the quadratic weighted kappa (QWK) (de La Torre, Puig, & Valls, 2018) or the squared Earth Mover's Distance (EMD) loss (Hou, Yu, & Samaras, 2016), which are designed to produce accurate predictions while accounting for the ordinal structure of $\mathcal{Y}$. One should note, however, that accurate prediction is not the same as faithful uncertainty representation, and indeed, from a probability estimation point of view, ordinal losses of that kind provide the wrong incentive. Imagine, for example, that the probability $p(\cdot \mid \boldsymbol{x}_q)$ is uniform over the three classes y_1, y_2, y_3. Then, CE is minimized (in expectation) by predicting exactly this distribution, whereas the L_1-loss, which takes the order of the classes into account, is minimized by the distribution that assigns probability 1 to y_2 (and hence suggests complete certainty).

This problem is confirmed by de La Torre et al. (2018) and Liu et al. (2020), who demonstrate that ordinal losses inherently bias the predictive probability distributions towards unimodality by penalizing more distant errors at the loss level. While this approach can be effective for loss minimization, it does not necessarily promote truthful uncertainty representation and, depending on the application and type of data, may introduce an undesirable inductive bias. The literature on ordinal classification places a strong emphasis on what Anderson (1984) refers to as "grouped continuous" ordered categorical variables, where an inherently continuous variable, such as age, is discretized into groups (Cao, Mirjalili, & Raschka, 2020; Q. Li et al., 2022; Niu, Zhou, Wang, Gao, & Hua, 2016; Yun et al., 2024). For such kind of variables, the assumption of unimodality may still appear to be reasonable. However, Anderson also identifies a second type, termed "assessed" ordered categorical variables, where an *assessor* provides a *judgment* (Haas & Hüllermeier, 2022). For this second type, the inductive bias of unimodality appears rather arbitrary. Anderson further notes that errors for assessed variables are likely to be greater. We refer to Appendix D for experiments demonstrating the superiority of proper scores (in particular the CE loss) for uncertainty quantification in ordinal classification compared to dedicated ordinal losses.

3 Aleatoric and Epistemic Uncertainty

A probabilistic predictor's uncertainty is purely aleatoric, as it fully commits to a single hypothesis h, which in turn fully commits to a single probability distribution

(2) when making a prediction. Hence, it does not represent any epistemic uncertainty about the hypothesis itself, nor about the probability (2). A popular framework that caters for the representation of epistemic uncertainty on top of aleatoric uncertainty is Bayesian inference. Here, instead of committing to a single hypothesis, a prior $p(h)$ is placed over the candidates $h \in \mathcal{H}$ (Gal & Ghahramani, 2016; Kendall & Gal, 2017). Learning essentially consists of updating the prior distribution $p(h)$ to the posterior distribution $p(h \mid \mathcal{D})$ in light of the training data $\mathcal{D}$:

$$p(h \mid \mathcal{D}) = \frac{p(h) \cdot p(\mathcal{D} \mid h)}{p(\mathcal{D})} \propto p(h) \cdot p(\mathcal{D} \mid h) \,,$$

where $p(\mathcal{D} \mid h)$ is the likelihood of the hypothesis h (i.e., the probability of the data given h) and $p(\mathcal{D})$ is the marginal probability of the data, which serves as a normalization factor. Intuitively, $p(h \mid \mathcal{D})$ captures the state of knowledge of the learner and hence its epistemic uncertainty. The more concentrated (or "peaked") the probability mass in a small region of $\mathcal{H}$, the less uncertain the learner is. Since every $h \in \mathcal{H}$ produces a probabilistic prediction, the belief about the outcome y is represented by a second-order probability: a probability distribution of probability distributions (Shaker & Hüllermeier, 2021) (cf. Figure 2 for an illustration).

More concretely, the predictive posterior distribution specifies the posterior probability of each outcome $y \in \mathcal{Y}$. It is defined in terms of the *expected* probability $p(y \mid \boldsymbol{x}, h)$, where the expectation is taken with respect to the posterior distribution $p(h \mid \mathcal{D})$:

$$p(y \mid \boldsymbol{x}) = \mathbb{E}_{p(h \mid \mathcal{D})}[\, p(y \mid \boldsymbol{x}, h) \,] = \int_{\mathcal{H}} p(y \mid \boldsymbol{x}, h) \, d\, p(h \mid \mathcal{D}) \tag{4}$$

In practice, as computing the exact expectation is intractable, (4) is commonly approximated using Monte Carlo integration techniques. In the simplest case, a finite ensemble $H = \{h_1, \ldots, h_M\}$ consisting of a set of M models is trained, and integration is replaced by the arithmetic average (Malinin et al., 2021; Shaker & Hüllermeier, 2020; Shaker & Hüllermeier, 2021):

$$p(y \mid \boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} p(y \mid \boldsymbol{x}, h_m) \tag{5}$$

3.1 Entropy

A principled and popular approach to measuring and separating aleatoric and epistemic uncertainty in (Bayesian) machine learning, though criticized for not satisfying certain theoretical axioms (Wimmer et al., 2023), is based on the classical information-theoretic measure of (Shannon) entropy (Depeweg et al., 2018). Total uncertainty (TU) is hereby measured in terms of the entropy of the posterior predictive distribution:

$$\mathrm{TU}(\boldsymbol{x}) = \mathbb{H}\big[p(y \mid \boldsymbol{x})\big] = \mathbb{H}\Big[\mathbb{E}_{p(h \mid \mathcal{D})}[\, p(y \mid \boldsymbol{x}, h) \,]\Big] \,, \tag{6}$$

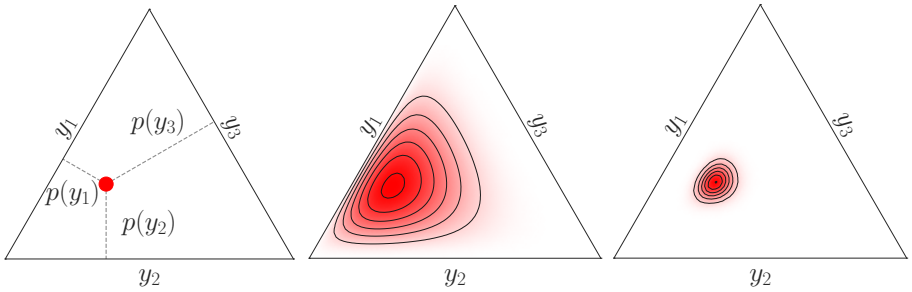


Fig. 2: Uncertainty awareness, illustrated on the probability simplex for $\mathcal{Y} = \{y_1, y_2, y_3\}$. From left to right: Aleatoric uncertainty without any epistemic uncertainty awareness, Bayesian representation of epistemic uncertainty in the form of a probability distribution over probability distributions, a more concentrated or "peaked" second-order distribution compared to the previous one.

where the Shannon entropy of a (discrete) probability distribution $\boldsymbol{p} = (p_1, \ldots, p_K)$ is given by

$$\mathbb{H}(\boldsymbol{p}) = -\sum_{k=1}^{K} p_k \cdot \log(p_k)\,.$$

By fixing a hypothesis $h \in \mathcal{H}$, the epistemic uncertainty is essentially removed, and only aleatoric uncertainty remains. Therefore, a natural measure of aleatoric uncertainty (AU) is the conditional entropy (i.e., the expected entropy of $p(y \mid \boldsymbol{x}, h)$, with the expectation taken with regard to the posterior $p(h \mid \mathcal{D})$):

$$\mathrm{AU}(\boldsymbol{x}) = \mathbb{E}_{p(h \mid \mathcal{D})}\mathbb{H}\big[\, p(y \mid \boldsymbol{x}, h) \,\big] = \int_{\mathcal{H}} p(h \mid \mathcal{D})\, \mathbb{H}\big[\, p(y \mid \boldsymbol{x}, h) \,\big]\; d\, h \tag{7}$$

Eventually, the epistemic uncertainty (EU) is measured in terms of the *mutual information* between hypotheses h and outcomes y, which is obtained as the difference between total and aleatoric uncertainty:

$$\underbrace{\mathbb{I}\big[y, h \mid \boldsymbol{x}, \mathcal{D}\big]}_{\mathrm{EU}(\boldsymbol{x})} = \underbrace{\mathbb{H}\big[p(y \mid \boldsymbol{x})\big]}_{\mathrm{TU}(\boldsymbol{x})} - \underbrace{\mathbb{E}_{p(h \mid \mathcal{D})}\mathbb{H}\big[\, p(y \mid \boldsymbol{x}, h) \,\big]}_{\mathrm{AU}(\boldsymbol{x})} \tag{8}$$

In practice, approximations of (6) and (7), and hence of (8), are again obtained by replacing integration over $\mathcal{H}$ with averaging over a finite ensemble:

$$\mathrm{TU}(\boldsymbol{x}) \approx \mathbb{H}\left[\frac{1}{M}\sum_{m=1}^{M} p(y \mid \boldsymbol{x}, h_m)\right] \tag{9}$$

$$\mathrm{AU}(\boldsymbol{x}) \approx \frac{1}{M}\sum_{m=1}^{M} \mathbb{H}\big[\, p(y \mid \boldsymbol{x}, h_m) \,\big] \tag{10}$$

Since labels in ordinal classification are of categorical nature, the entropy-based approach is also applicable to probabilistic ordinal classification. Note, however, that it does not take the ordinal structure into account and is invariant to permutations of the probability degrees (Haas & Hüllermeier, 2025b; Hüllermeier & Waegeman, 2021).

3.2 Variance

Another principled approach to separating aleatoric and epistemic uncertainty, also originally proposed by Depeweg et al. (2018), is based on the *law of total variance*. This measure is conceptually similar to the entropy measure defined in (8), making use of the variance $\mathbb{V}$ as the base measure. The total uncertainty can hereby be decomposed into its aleatoric and epistemic parts as follows:

$$\underbrace{\mathbb{V}_{p(y \mid \boldsymbol{x}, \mathcal{D})}[y \mid \boldsymbol{x}]}_{\mathrm{TU}(\boldsymbol{x})} = \underbrace{\mathbb{V}_{p(h \mid \mathcal{D})}\left[\mathbb{E}_{p(y \mid \boldsymbol{x}, h)}[y \mid \boldsymbol{x}]\right]}_{\mathrm{EU}(\boldsymbol{x})} + \underbrace{\mathbb{E}_{p(h \mid \mathcal{D})}\left[\mathbb{V}_{p(y \mid \boldsymbol{x}, h)}[y \mid \boldsymbol{x}]\right]}_{\mathrm{AU}(\boldsymbol{x})} \tag{11}$$

However, the above decomposition, is primarily applicable to numerical targets y, for which variance is well-defined, but not for categorical or ordinal targets. Therefore, this alternative measure has so far been primarily used for quantifying uncertainty in regression tasks, where the target variable is continuous or integer-valued (Malinin et al., 2021).

In practice, ordinal targets $y_1, \ldots, y_K$ are often encoded in terms of numbers $1, \ldots, K$, and thereby embedded in the metric space $(\mathbb{R}, |\cdot|)$. Obviously, this embedding is debatable, as it postulates equal distances between all neighbored ordinal categories — an assumption that, even if acceptable as an approximation in some cases, is disputable in general. Nevertheless, if one is willing to accept this assumption, the measures in (11) can be computed. Their ensemble-based approximations are given as follows:

$$\mathrm{AU}(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} p(k \mid \boldsymbol{x}, h_m) \cdot (k - \mu_m)^2 \,, \tag{12}$$

$$\mathrm{EU}(\boldsymbol{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \left[\mu - \mu_m\right]^2 , \tag{13}$$

$$\mathrm{TU}(\boldsymbol{x}) \approx \sum_{k=1}^{K} p(k \mid \boldsymbol{x}) \cdot (k - \mu)^2 \,, \tag{14}$$

with

$$\mu = \frac{1}{M} \sum_{m=1}^{M} \mu_m \,, \quad \mu_m = \sum_{k=1}^{K} p(k \mid \boldsymbol{x}, h_m) \cdot k \quad (m = 1, \ldots, M) \,.$$

As a dispersion measure, variance takes the distance or dispersion of probability mass into account. Thus, compared to entropy, it has the advantage of not being invariant to permutation of probability degrees. However, as already mentioned, the assumption of equal distances, a prerequisite for applying variance-based uncertainty

quantification in ordinal classification, remains debatable. In particular, for subjective or qualitative scales, such as perceived risk in risk assessment or reported pain levels in medicine, differences between categories are not necessarily uniform. For example, going from "no pain" to "mild pain" may not feel the same as going from "severe pain" to "unbearable pain." Therefore, the assumption of equal spacing appears unrealistic and overly simplistic. It is important to be aware of this limitation when applying variance-based uncertainty quantification in ordinal classification.

4 Uncertainty Measures through Binary Reduction

So far, we considered probabilistic multinomial classifiers $h : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ with $\mathcal{Y} = \{y_1, \ldots, y_K\}$. Such classifiers produce predictions

$$p\left(\cdot \mid \boldsymbol{x}, h\right) = \left(p(y_1 \mid \boldsymbol{x}, h), \ldots, p(y_K \mid \boldsymbol{x}, h)\right),$$

where $p(y_k \mid \boldsymbol{x}, h)$ denotes the probability of the class y_k predicted by h for the query instance $\boldsymbol{x}$. For any classifier h, let $h^{(k)} : \mathcal{X} \to \mathbb{P}(\{0, 1\})$ denote the (derived) binary classifier that predicts the Bernoulli distribution

$$p\left(\cdot \mid \boldsymbol{x}, h^{(k)}\right) = \left(p\big(0 \mid \boldsymbol{x}, h^{(k)}\big), p\big(1 \mid \boldsymbol{x}, h^{(k)}\big)\right),$$

with $p(1 \mid \boldsymbol{x}, h^{(k)}) = p(y_k \mid \boldsymbol{x}, h)$ and $p(0 \mid \boldsymbol{x}, h^{(k)}) = \sum_{1 \leq i \neq k \leq K} p(y_i \mid \boldsymbol{x}, h)$. In other words, $h^{(k)}$ predicts whether class y_k will occur as an outcome or not; it treats this class as positive and all other classes as negative, and adopts the probabilities for these two cases from the probabilities predicted by h.

Obviously, all uncertainty measures introduced in the previous section can also be computed for the binary case ($K = 2$). If U is any such measure, then we denote by $U^{(k)}$ the measure that is obtained by replacing the multinomial distributions $p(\cdot \mid \boldsymbol{x}, h)$ with the binary distributions $p(\cdot \mid \boldsymbol{x}, h^{(k)})$. In particular,

$$\begin{aligned}
\mathrm{TU}_{\mathbb{U}}^{(k)}(\boldsymbol{x}) &= \mathbb{U}\Big[\mathbb{E}_{p(h \mid \mathcal{D})}\left[p(\cdot \mid \boldsymbol{x}, h^{(k)})\right]\Big] &&\approx \mathbb{U}\left[\frac{1}{M}\sum_{m=1}^{M} p(\cdot \mid \boldsymbol{x}, h_m^{(k)})\right], \\
\mathrm{AU}_{\mathbb{U}}^{(k)}(\boldsymbol{x}) &= \mathbb{E}_{p(h \mid \mathcal{D})}\mathbb{U}\big[p(\cdot \mid \boldsymbol{x}, h^{(k)})\big] &&\approx \frac{1}{M}\sum_{m=1}^{M} \mathbb{U}\left[p(\cdot \mid \boldsymbol{x}, h_m^{(k)})\right],
\end{aligned}$$

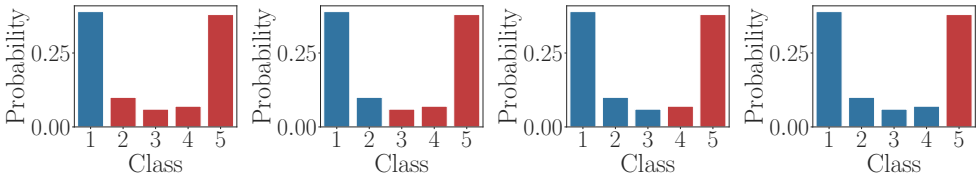
where $\mathbb{U} = \mathbb{H}$ (entropy-based) or $\mathbb{U} = \mathbb{V}$ (variance-based).

Given label-wise measures of that kind, Sale et al. (2024) propose to define overall measures of uncertainty for the original multinomial case as follows:

$$U_{\mathbb{U}}^{cat}(\boldsymbol{x}) = \sum_{k=1}^{K} U_{\mathbb{U}}^{(k)}(\boldsymbol{x}) \tag{15}$$

This approach is inspired by binary decomposition techniques for reducing a multinomial classification problem to several binary problems (Allwein, Schapire, & Singer, 2001), in particular the one-versus-rest decomposition (Rifkin & Klautau, 2004). Here, the same idea of reduction is applied to uncertainty measures. Intuitively, the uncertainty in the answer to the question "Which class will occur?" is defined as an aggregation of the uncertainties in the questions "Will class y_k occur, yes or no?", $k = 1, \ldots, K$. One obvious advantage of this approach is that measures like variance can be used in a theoretically sound manner, because variance is well defined for the binary case — as opposed to the multinomial case. Another advantage is that uncertainty quantification becomes somewhat more nuanced, for example because the overall uncertainty can be attributed to specific class labels (Sale et al., 2024).

A one-versus-rest decomposition is still invariant toward the permutation of class labels and does not take the ordinal structure of the problem into account. In particular, with variance as a base measure, the overall measure of uncertainty does not capture any notion of dispersion or distance. Thus, one-versus-rest decomposition seems to be appropriate for the multinomial but not for the ordinal case. Indeed, the arguably most natural reduction in the case of ordinal classification is not achieved by means of a one-versus-rest decomposition, but rather through binary splits of the ordinal scale, separating a lower part $\{y_1, \ldots, y_k\}$ of the scale from an upper part $\{y_{k+1}, \ldots, y_K\}$ (Frank & Hall, 2001; L. Li & Lin, 2006). In the following, we will refer to this approach as the *order-consistent split* (OCS) reduction. Again, this reduction allows for solving the original classification task — the correct ordinal category can be recovered from consistent answers to $K - 1$ queries of the form "Does the class label exceed level k?", $k = 1, \ldots, K - 1$ (see Figure 3 for an illustration). Moreover, it takes the ordinal structure of the problem into account (Huhn & Hüllermeier, 2008) and, unlike variance (cf. Section 3.2), does not require a cardinal scale with equally spaced categories — a requirement that may pose a significant limitation not only in theory but also in practice.



(a) Class 1 vs. classes 2 to 5. (b) Classes 1 and 2 vs. classes 3 to 5. (c) Classes 1 to 3 vs. classes 4 and 5. (d) Classes 1 to 4 vs. class 5.

Fig. 3: Example of an OCS decomposition with five classes.

Correspondingly, for any classifier h, let $h^{(\leq k)} : \mathcal{X} \to \mathbb{P}(\{0, 1\})$ now denote the (derived) binary classifier that predicts

$$p\left(\cdot \mid \boldsymbol{x}, h^{(\leq k)}\right) = \left(p\big(0 \mid \boldsymbol{x}, h^{(\leq k)}\big), p\big(1 \mid \boldsymbol{x}, h^{(\leq k)}\big)\right)$$

with

$$p\left(0 \mid \boldsymbol{x}, h^{(\leq k)}\right) = \sum_{1 \leq i \leq k} p(y_i \mid \boldsymbol{x}, h)$$
$$p\left(1 \mid \boldsymbol{x}, h^{(\leq k)}\right) = \sum_{k+1 \leq i \leq K} p(y_i \mid \boldsymbol{x}, h)\,.$$

Thus, $h^{(\leq k)}$ treats the classes $y_1, \ldots, y_k$ as negative and the classes $y_{k+1}, \ldots, y_K$ as positive, and again adopts the probabilities for these two cases from the probabilities predicted by h. Similar to above, we obtain

$$\mathrm{TU}_{\mathbb{U}}^{(\leq k)}(\boldsymbol{x}) = \mathbb{U}\Big[\mathbb{E}_{p(h \mid \mathcal{D})}\left[p(\cdot \mid \boldsymbol{x}, h^{(\leq k)})\right]\Big] \approx \mathbb{U}\left[\frac{1}{M}\sum_{m=1}^{M} p(\cdot \mid \boldsymbol{x}, h_m^{(\leq k)})\right],$$
$$\mathrm{AU}_{\mathbb{U}}^{(\leq k)}(\boldsymbol{x}) = \mathbb{E}_{p(h \mid \mathcal{D})}\mathbb{U}\big[p(\cdot \mid \boldsymbol{x}, h^{(\leq k)})\big] \approx \frac{1}{M}\sum_{m=1}^{M} \mathbb{U}\left[p(\cdot \mid \boldsymbol{x}, h_m^{(\leq k)})\right],$$

where $\mathbb{U} = \mathbb{H}$ (entropy-based) or $\mathbb{U} = \mathbb{V}$ (variance-based). Moreover, the overall measures of uncertainty for the original ordinal problem is given by

$$U_{\mathbb{U}}^{ord}(\boldsymbol{x}) = \sum_{k=1}^{K-1} U_{\mathbb{U}}^{(\leq k)}(\boldsymbol{x})\,. \tag{16}$$

For the special case of total uncertainty on first-order (predictive) probabilities, the above construction has already been considered by Haas and Hüllermeier (2025b), who argue for its suitability in the context of ordinal classification. They show, for example, that both $\mathrm{TU}_{\mathbb{H}}^{ord}$ and $\mathrm{TU}_{\mathbb{V}}^{ord}$ are maximized by the bimodal distribution that assigns probability $1/2$ to the extreme classes 1 and K, respectively, and not by the uniform distribution.

An important question is whether the binary reduction increases the computational complexity of decomposing uncertainty into AU, EU, and TU at inference time compared to established measures. For conventional approaches (cf. Sections 3.1 and 3.2), the computational cost is $\mathcal{O}(3 \cdot K \cdot M \cdot n) = \mathcal{O}(K \cdot M \cdot n)$, where n is the number of data samples, K the number of classes, M the number of predictors in the ensemble, and the constant factor 3 accounts for the three uncertainty measures (AU, EU, TU). The factor K arises because common measures such as Shannon entropy or variance require iterating over all K class probabilities.[1] In contrast, for the proposed binary reduction technique, the computational cost is

$$\mathcal{O}(3 \cdot 2 \cdot (K-1) \cdot M \cdot n) = \mathcal{O}(K \cdot M \cdot n),$$

where each binary split involves only two meta-classes, and the procedure is applied across $K - 1$ splits corresponding to the order-consistent splits of the ordinal scale.

[1] Note, however, that K can effectively be treated as a constant (and hence removed as a factor), because the number of levels in common ordinal classification tasks is relatively small.

Consequently, the binary reduction approach retains the same asymptotic complexity as conventional measures and incurs no additional computational costs.

5 Experiments on Ordinal Benchmark Datasets

In the following sections, we evaluate the approaches described above for disentangling aleatoric and epistemic uncertainty using common tabular ordinal benchmark datasets. Our focus is on how effectively the different measures enhance predictive performance and decision-making, taking into account standard ordinal classification metrics. Specifically, we assess the measures' effectiveness in error detection through rejection-based experiments and evaluate their performance in out-of-distribution detection.[2]

5.1 Experimental Setup

To approximate Bayesian inference, we create ensembles consisting of 10 independent gradient boosted trees (GBT) (Friedman, 2001) (refer to Appendix E for additional experimental results using ensembles of 10 Multi-Layer Perceptrons (MLPs) instead). As demonstrated by previous studies, increasing ensemble size improves calibration and reduces epistemic uncertainty by decreasing the variance among ensemble members (Lakshminarayanan, Pritzel, & Blundell, 2017; Snoek et al., 2019). However, returns diminish beyond moderate ensemble sizes (typically 5–10 members), while computational costs continue to increase. For this reason, we adopt this commonly used ensemble size in our uncertainty quantification experiments (Malinin et al., 2021; Wimmer et al., 2023). In our study, we use GBTs rather than deep neural networks as they provide state-of-the-art performance on tabular datasets (Grinsztajn, Oyallon, & Varoquaux, 2022; Shwartz-Ziv & Armon, 2022), and tabular datasets are prevalent in high-stakes settings like finance (Chang, Chang, & Wu, 2018) or medicine (Yıldız & Kalayci, 2025). Hence, GBTs are also highly relevant for practitioners. Furthermore, unlike Random Forests, they also enable flexible usage of loss functions, including proper scoring rules such as cross-entropy loss (Gneiting & Raftery, 2007).

According to Malinin et al. (2021), we set the *subsample* rate to 0.5 to induce stochasticity in the sequential training process and eventually in the resulting trees. In the context of gradient boosting, subsampling refers to using a subset of the training data to train each individual tree in the ensemble. All other parameters are left with the default values. Concretely, we use LightGBM (Ke et al., 2017) as a fast and popular gradient boosting library for our ensemble implementations with the cross-entropy (CE) loss for multi-class classification (refer to Appendix C for additional experimental results using other popular GBT libraries). This approach enables us to obtain conditional probability distributions $p(y \mid \boldsymbol{x_q})$, which serve as the foundation for evaluating various uncertainty measures.

Table 1 summarizes key attributes of the twenty-three ordinal benchmark datasets used in our evaluation (Bischl et al., 2025; Kelly, Longjohn, & Nottingham, 2023; Vanschoren, Van Rijn, Bischl, & Torgo, 2014). These datasets are widely used in ordinal

[2]The source code for the experiments is made avaliable at https://github.com/stefanahaas41/ordinal-aleatoric-epistemic-uncertainty

classification research and exhibit variability in size, number of features, number of classes, and imbalance ratio (IR), thereby providing a solid basis for a comprehensive evaluation of the proposed uncertainty measures. IR quantifies the degree of class imbalance, typically defined as the ratio between the number of instances in the class with the largest number of instances and that in the class with the smallest number of instances (Zhu, Guo, & Xue, 2020).

Table 1: Twenty-three common ordinal benchmark datasets used for evaluating the different uncertainty measures, including their imbalance ratio (IR).

Dataset	# Instances	# Features	# Classes	Class distribution	IR
Grub Damage	155	8	4	(49,41,46,19)	2.58
Obesity	2,111	16	7	(272, 287, 290, 290, 351, 297, 324)	1.29
CMC	1,473	9	3	(629, 333, 511)	1.89
New Thyroid	215	5	3	(150, 35, 30)	5.00
Balance Scale	625	4	3	(288, 49, 288)	5.88
Automobile	205	25	7	(3, 22, 67, 54, 32, 27)	22.33
Eucalyptus	736	19	5	(180,107,130,214,105)	2.04
TAE	151	5	3	(49, 50, 52)	1.06
Heart (CLE)	303	13	5	(164, 55, 36, 35, 13)	12.62
SWD	1,000	10	4	(32,352,399,217)	12.47
ERA	1,000	4	9	(92,142,181,172,158,118,88,31,18)	10.06
ESL	488	4	9	(2,12,38,100,116,135,62,19,4)	67.50
LEV	1,000	4	5	(93,280,403,197,27)	14.93
Red Wine	1,599	11	6	(10, 53, 681, 638, 199, 18)	68.10
White Wine	4,898	11	7	(20, 163, 1457, 2198, 880, 175, 5)	439.60
Triazines	186	60	5	(7, 10, 26, 86, 57)	12.29
Machine CPU	209	6	10	(115, 37, 21, 6, 8, 5, 3, 4, 4, 6)	38.33
Auto MPG	392	7	10	(13,78,73,58,53,48,37,22,4,6)	19.50
Boston Housing	506	13	5	(77, 239, 123, 36, 31)	7.71
Pyrimidines	74	27	10	(2, 2, 14, 14, 13, 5, 10, 4, 3, 7)	7.00
Abalone	4,177	8	10	(17, 431, 1648, 1388, 432, 125, 100, 29, 4, 3)	549.33
Wisconsin Breast Cancer	194	32	5	(67, 41, 43, 24, 19)	3.53
Stocks Domain	950	9	5	(158, 227, 272, 207, 86)	3.16

In terms of evaluating the predictive performance in relation to the quantified uncertainties, we rely on the two most popular metrics in the realm of ordinal classification: Accuracy (ACC) (or its inverse misclassification rate (MCR) or mean zero-one error (MZE)) and mean absolute error (MAE) (Gaudette & Japkowicz, 2009; Gutiérrez, Pérez-Ortiz, Sánchez-Monedero, Fernández-Navarro, & Hervás-Martínez, 2016). Another very popular performance measure for ordinal classification is the quadratic weighted kappa (QWK) (Cohen, 1968; de La Torre et al., 2018). However, QWK poses some challenges in rejection-based evaluation in uncertainty quantification when the required confusion matrix becomes sparse, which is why we exclude it here. The mean squared error (MSE) is also commonly used when evaluating ordinal classification, emphasizing larger error distances (Baccianella, Esuli, & Sebastiani, 2009; Gaudette & Japkowicz, 2009). There are also some dedicated performance measures for imbalanced ordinal classification, such as average MAE (AMAE) (Baccianella et al., 2009) and maximum MAE (MMAE) (Cruz-Ramírez, Hervás-Martínez, Sánchez-Monedero, & Gutiérrez, 2014). However, since our focus is on general uncertainty quantification and not specifically on uncertainty quantification for imbalanced data, we do not consider them here. Consequently, we believe that ACC and MAE best capture the fundamental trade-off in ordinal classification between exact hit rate and error distance minimization.

In general, we use 10-fold cross-validation for all our experiments to ensure robust and fair comparison of all uncertainty measures. In terms of preprocessing the datasets for the experimental evaluation, all categorical features were one-hot encoded and the ordinal labels $y_1, \ldots, y_K$ were integer encoded from $1, \ldots, K$.

5.2 Accuracy-Rejection Curves

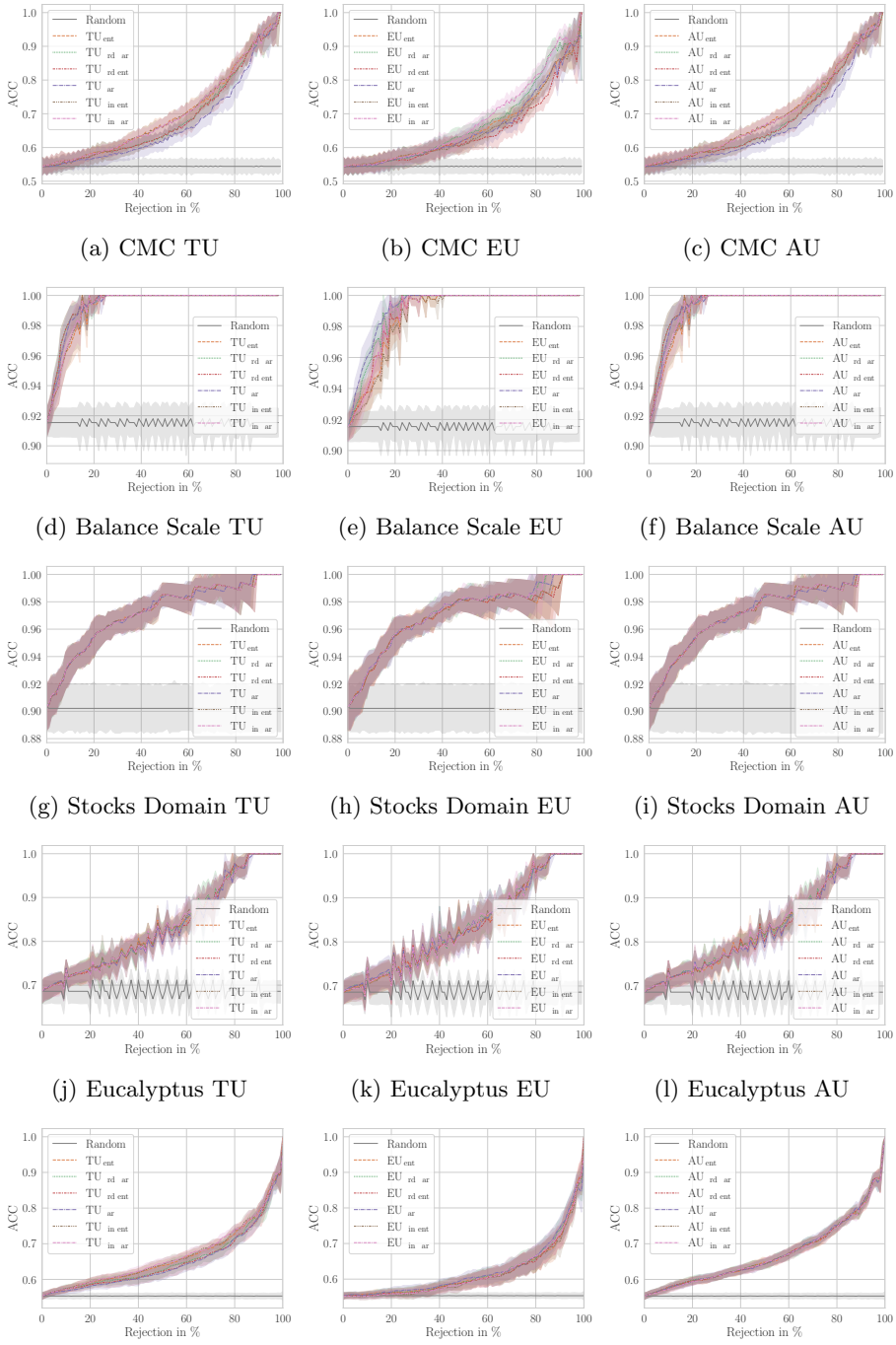
A common approach for evaluating the quality of uncertainty quantification methods are *accuracy-rejection curves*, which depict the accuracy of a predictor as a function of the percentage of rejections (Huhn & Hüllermeier, 2009; Nadeem, Zucker, & Hanczar, 2010). A predictor that is allowed to abstain from predicting a certain percentage p of queries will only predict the $(1-p)\%$ of queries that it feels most certain about. Ideally, the accuracy should increase (for performance measures that are supposed to be maximized like ACC) or the error metric should decrease (for performance measures that are supposed to be minimized like MAE) with increasing p, leading to a monotonically increasing or decreasing curve, unlike a flat random accuracy-rejection curve.

Figures 4 and 5 display accuracy-rejection curves for some selected ordinal benchmark datasets for ACC and MAE based on the different uncertainty types (AU, EU, and TU) and uncertainty measures: $\mathbb{H}$ (ent), $\mathbb{V}$ (var), $U_{\mathbb{H}}^{cat}$ (bin-ent), $U_{\mathbb{V}}^{cat}$ (bin-var), $U_{\mathbb{H}}^{ord}$ (ord-ent), and $U_{\mathbb{V}}^{ord}$ (ord-var). As one can clearly see, all measures are capable of quantifying the different uncertainty types properly, as all accuracy-rejection curves either increase or decrease monotonically for ACC and MAE, respectively. Hence, all the above-presented measures appear to be viable solutions for uncertainty quantification in ordinal classification. In spite of capturing different types of uncertainty, AU, EU, and TU lead to highly correlated curves, which is in line with observations in (Mucsányi, Kirchhof, & Oh, 2024). However, accuracy-rejection curves only provide a coarse visual way to assess the quality of uncertainty quantification methods and do not allow for rigorous statistical comparisons.

5.3 Prediction-Rejection-Ratios (PRRs)

To compare the different uncertainty quantification methods on a more fine-grained numerical level, *prediction-rejection ratios* (PRRs), as introduced by Malinin (2019), provide a good solution. Essentially, PRRs summarize rejection curves into a single numerical value. Specifically, the area between the rejection curve obtained for a given uncertainty method and the curve for random rejection, AR_{unc}, is compared to the corresponding area produced by a perfect oracle-based rejection, AR_{orc} (cf. Figure 6). The oracle-based rejection either rejects all incorrectly predicted instances or, in the case of MAE, it successively rejects instances in decreasing order of error magnitude, resulting in a perfect rejection curve:

$$PRR = \frac{AR_{unc}}{AR_{orc}} \,.$$



(a) CMC TU (b) CMC EU (c) CMC AU

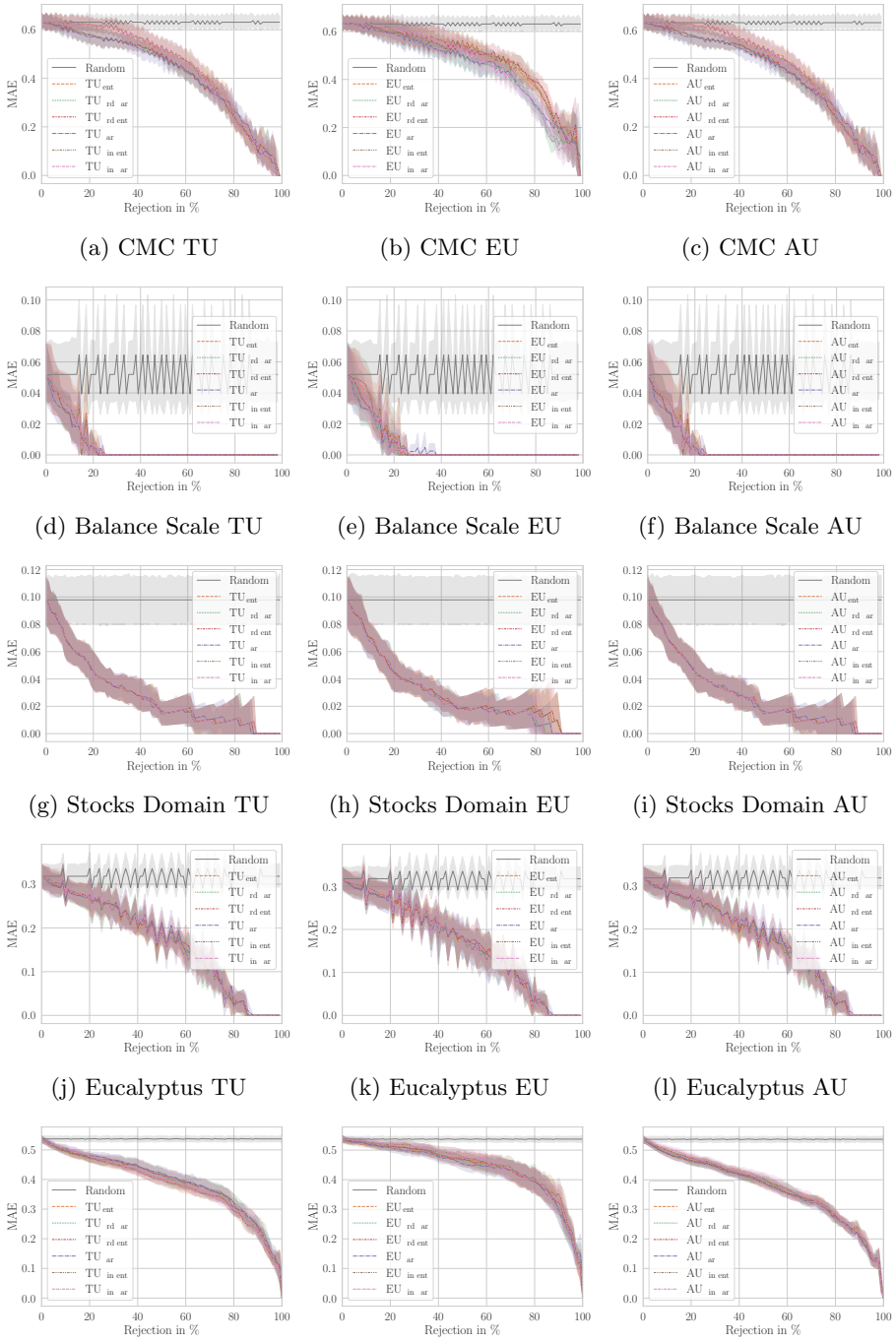
(d) Balance Scale TU (e) Balance Scale EU (f) Balance Scale AU

(g) Stocks Domain TU (h) Stocks Domain EU (i) Stocks Domain AU

(j) Eucalyptus TU (k) Eucalyptus EU (l) Eucalyptus AU

(m) Abalone TU (n) Abalone EU (o) Abalone AU

Fig. 4: Accuracy rejection curves for different datasets, uncertainty types (TU, EU, AU) and measures using an ensemble of GBTs (LightGBM (Ke et al., 2017)). Shaded regions around the mean represent the 95% confidence interval.



(a) CMC TU (b) CMC EU (c) CMC AU

(d) Balance Scale TU (e) Balance Scale EU (f) Balance Scale AU

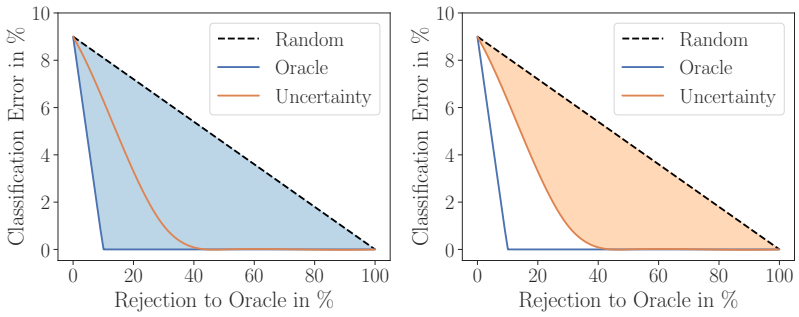
(g) Stocks Domain TU (h) Stocks Domain EU (i) Stocks Domain AU

(j) Eucalyptus TU (k) Eucalyptus EU (l) Eucalyptus AU

(m) Abalone TU (n) Abalone EU (o) Abalone AU

Fig. 5: Mean absolute error rejection curves for different datasets, uncertainty types (TU, EU, AU) and measures using an ensemble of GBTs (LightGBM (Ke et al., 2017)). Shaded regions around the mean represent the 95% confidence interval.

Unlike in accuracy-rejection curves, random rejection will not produce a flat line but a line that decreases linearly in expectation. This is because rejected queries are supposedly delegated to an oracle that will answer queries correctly.



(a) Shaded area is AR_{orc}. (b) Shaded area is AR_{unc}.

Fig. 6: Example Prediction Rejection Curves (Malinin, 2019).

Another property of PRRs is that they assess uncertainty quantification quality independently of classification performance. A PRR value of 1 indicates perfect rejection, consistent with the oracle-based rejection, and therefore represents perfect uncertainty quantification, whereas a value of 0 corresponds to random rejection. The PRR can also become negative, which indicates worse than random uncertainty quantification. To calculate PRR values for uncertainty quantification evaluation, one also needs to select a performance measure, just as for rejection curves. To make all rejection curves go in the same direction, we measure the MCR instead of ACC, as is commonly done (Lahoti, Gummadi, & Weikum, 2023; Malinin et al., 2021), and again MAE.

To compare the PRR values calculated for the different uncertainty methods and datasets, we conduct a Friedman test followed by a Holm-adjusted Wilcoxon signed-rank test (Benavoli, Corani, & Mangili, 2016; Demsar, 2006). The non-parametric Friedman test will first determine whether there is a significant difference in the performance of the uncertainty measures across the datasets overall (with a significance level $p = 0.05$). If it indicates a significant difference, the Wilcoxon signed-rank test will then conduct pairwise comparisons between the uncertainty measures to determine which specific differences are statistically significant (again at a significance level $p = 0.05$). Furthermore, we depict the results by uncertainty type (AU, EU, TU, or All) as well as performance measure (MCR, MAE, or both) and visualize these using critical difference diagrams (CD). The critical difference diagrams show the average ranks of the different measures. If two measures do not differ significantly, they are connected by a horizontal bar or line. The full results of our comparison can be found in the tables presented in Appendix B.

Figure 7 displays the CD diagrams for total uncertainty. Though there is no statistically significant difference among the different measures for total uncertainty,

binary methods outperform standard entropy and variance-based measures, with the variance-based OCS decomposition (ord-var) leading the field when considering MCR and MAE.

When looking at the results for quantifying epistemic uncertainty in Figure 8a, var and ord-var significantly outperform the other measures when simultaneously considering MCR and MAE. With regards to ord-var, the same applies for aleatoric uncertainty (cf. Figure 9a).

Eventually, when looking at the results over all uncertainty types (AU, EU, and TU) in Figure 10, the results become very distinct. Again, ord-var significantly outperforms the other measures (cf. Figure 10a), but also the distinction between measures taking distance into account becomes clear, with ord-var, var, and ord-ent significantly outperforming the other measures on MAE (cf. Figure 10c). Interestingly, for MCR there is no significant difference between the measures (cf. Figure 10b).

In the end, taking distance into account is an important property a measure needs to fulfill in uncertainty quantification for ordinal classification. This becomes also obvious when looking at the overall ranking of measures in Figure 10a for MCR and MAE, with ord-var, ord-ent, and var outperforming ent, bin-ent, and bin-var.

We may conclude that ord-var best represents the inherent trade-off in ordinal classification between exact hit-rate and minimized error distances. This conclusion is based on its qualitative position between variance and ord-ent, as it balances the focus on the extreme bimodal distribution and indicates uncertainty for the uniform distribution (see Figures A1d, A1e, and A1f). The focus of var on the extreme bimodal distribution appears too extreme and results in worse performance on MCR. Conversely, the focus of ord-ent on the extreme bimodal distribution is not strong enough, leading to worse performance on MAE. Overall, ord-var seems to best capture this trade-off.

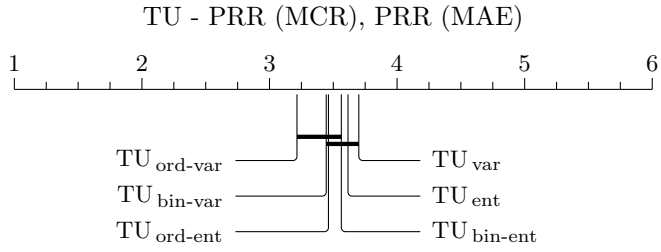
5.4 Out-Of-Distribution (OOD) Detection

A very critical and practically highly relevant challenge for machine learning models is the detection of out-of-distribution (OOD) data, which is data the learner has not seen during training, and which is sampled from a distribution that differs from the distribution of the training data. Think of malicious loan approval requests or unknown clinical conditions. In such cases, the model should ideally be aware of its own incompetency and trigger appropriate fallback scenarios. it is commonly assumed that OOD samples lead to high epistemic uncertainty.

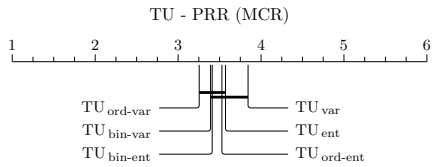
To test this assumption in an OOD evaluation experiment, one typically first trains a model on an in-distribution (ID) dataset and computes uncertainty values on its corresponding test set. Subsequently, the model is exposed to OOD data sampled from out-of-domain datasets. The model, which has not seen this data before, is then expected to exhibit increased epistemic uncertainty. Note that OOD detection for the ordinal case does not inherently differ from the nominal case, as the detection primarily operates in the input space $\mathcal{X}$. The purpose of this experiment is to ensure that our proposed OCS reduction is also competitive when it comes to OOD detection and not only on error detection. Concretely, one can evaluate the quality of OOD detection

[3] https://github.com/mirkobunse/critdd

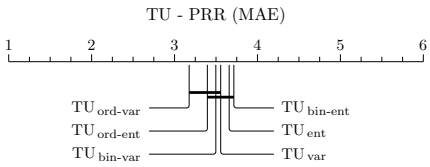
Fig. 7: Critical difference (CD) diagrams[3]for the evaluated Total Uncertainty (TU) measures and performance metrics based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon signed-rank test (Benavoli et al., 2016; Demsar, 2006) using an ensemble of GBTs (LightGBM (Ke et al., 2017)). Groups of uncertainty measures that are not significantly different (at p = 0.05) are connected.



(a) Result for MCR and MAE.

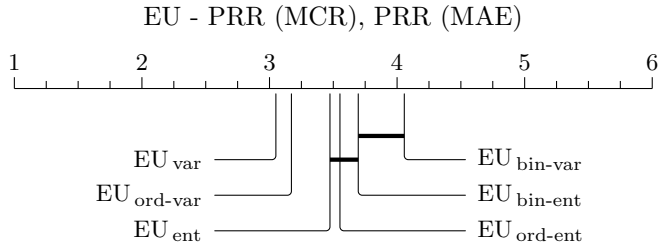


(b) Result only for MCR.

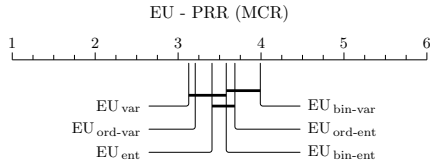


(c) Result only for MAE.

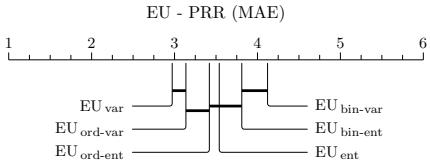
Fig. 8: CD diagrams for Epistemic Uncertainty (EU) using an ensemble of GBTs (LightGBM (Ke et al., 2017)).



(a) Result for MCR and MAE.

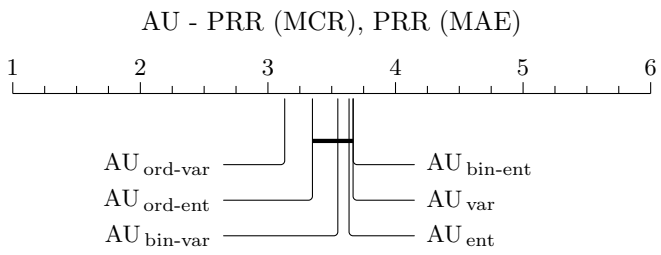


(b) Result only for MCR.

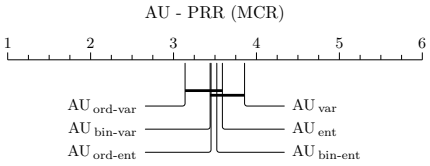


(c) Result only for MAE.

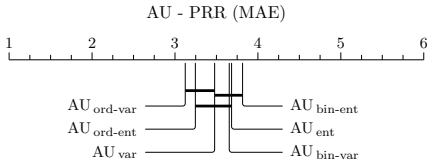
Fig. 9: CD diagrams for Aleatoric Uncertainty (AU) using an ensemble of GBTs (LightGBM (Ke et al., 2017)).



(a) Result for MCR and MAE.

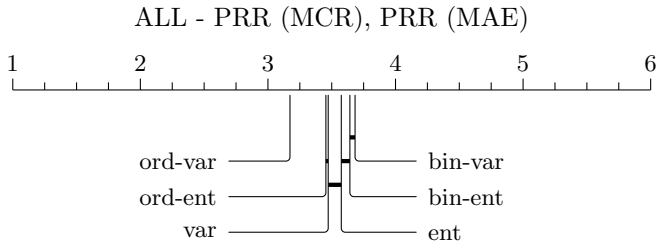


(b) Result only for MCR.

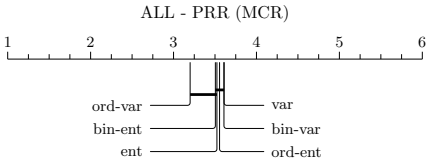


(c) Result only for MAE.

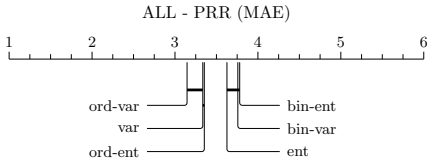
Fig. 10: CD diagrams for all uncertainty types (AU, EU and TU) using an ensemble of GBTs (LightGBM (Ke et al., 2017)).



(a) Result for MCR and MAE.



(b) Result only for MCR.



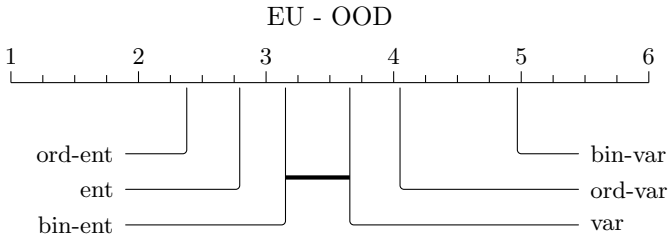
(c) Result only for MAE.

using the computed area under the receiver operating characteristic curve (AUC-ROC), in which OOD and ID data are given binary labels (0 for ID and 1 for OOD). The determined epistemic uncertainty values represent target scores (Hendrycks & Gimpel, 2017).

For our used tabular ordinal datasets, we use the same approach as done by Malinin et al. (2021). For each dataset, we take its test set as ID data. The OOD data of the same size is sampled from the Year MSD dataset (Kelly et al., 2023). All numerical features are normalized by the per-column mean and variance obtained on the ID training data, and categorical features are uniformly sampled at random from the set of all categories of the particular feature.

In the CD diagram in Figure 11, we report the overall result of our OOD experiment based on epistemic uncertainty using AUC-ROC as the OOD performance metric and 10-fold cross-validation over all datasets (see Table F7 for detailed results). It appears as if entropy-based measures have a clear edge over variance-based measures when it comes to OOD detection. This is also underpinned when looking at the results obtained for an ensemble of MLPs (cf. Figure F16). For detailed experimental OOD results, we refer to Appendix F. In general, we can conclude that our proposed OCS decomposition method appears competitive with standard and label-wise uncertainty approaches. However, unlike in error detection, entropy seems more appropriate for OOD detection than variance.

Fig. 11: CD diagram for OOD detection on the basis of epistemic uncertainty quantified by the different measures using an ensemble of GBTs (LightGBM (Ke et al., 2017)).



6 Conclusion and Future Work

In this paper, we have presented and experimentally evaluated several methods for quantifying aleatoric and epistemic uncertainty for probabilistic ordinal classification. Through visualizing accuracy-rejection curves and calculating prediction rejection ratios, we have demonstrated how all explored methods increase predictive performance and improve decision-making, with our OCS decomposition method using variance as base measure (ord-var) overall leading the field. Additionally, the computational complexity of OCS is asymptotically equivalent to that of the baseline methods. Furthermore, we have demonstrated the competitiveness of our ordinal approach

compared to existing methods for OOD detection. Interestingly, for OOD detection, entropy emerges as the superior choice, not only as a base measure for the binary decomposition but also in general.

All in all, we were able to experimentally prove our hypothesis that measures disregarding the ordinal structure, such as entropy and the labelwise approaches, are not ideal candidates for quantifying uncertainty and disentangling aleatoric and epistemic uncertainty in ordinal classification. To this end, one should rather consider our binary decomposition method or variance instead, as minimizing error distances is a crucial factor in ordinal classification. In particular, our binary decomposition method with variance as base measure (ord-var) strikes the best balance between exact hit-rate and reduced error distances, which is crucial in uncertainty quantification for ordinal classification. Furthermore, unlike variance, the binary decomposition method does not assume equal distances between classes, making it theoretically a sound approach for uncertainty quantification in ordinal classification. Nonetheless, despite its assumption of equal distances, variance performs surprisingly well, especially for quantifying epistemic uncertainty, suggesting that this assumption may not be a major limitation in practice, or is at least warranted for many ordinal datasets.

As a general recommendation, we suggest using variance if the predictor tends to commit large distance errors, as variance has a strong focus on the extreme bimodal distribution with all mass equally allocated to the extreme classes. Furthermore, it is also quite effective in detecting in-distribution instances with high epistemic uncertainty. However, if errors are not too widespread, our proposed OCS reduction achieves a favorable balance between exact hit-rate and error distance minimization.

Moreover, in ordinal classification, there exists a significant gap between predictive performance on ordinal metrics, such as QWK or RPS, and uncertainty quantification. While ordinal losses achieve strong results on these metrics, they negatively impact uncertainty quantification due to the strong inductive bias toward unimodal predictive probability distributions. This bias reduces variance, which, as demonstrated, is essential for reliable uncertainty quantification. Addressing this gap should be a key focus for future research. If reliable uncertainty quantification in ordinal classification is crucial, the CE loss appears to be the most suitable choice, as it leads to unbiased predictive probability distributions.

Additionally, future work could investigate uncertainty quantification for ordinal classification from a more theoretical point of view and try to assess or justify specific uncertainty measures axiomatically (Bülte et al., 2025; Wimmer et al., 2023). Furthermore, from an experimental point of view, it might be interesting to explore additional non-tabular ordinal datasets (e.g., image datasets).

Appendix A Comparative Analysis of the Measures

In this section, we conduct a brief comparative analysis of the different measures (cf. Sections 3.1, 3.2, and 4). Figure A1 illustrates the measured (total) uncertainty of the considered uncertainty measures via heatmaps using the probability simplex over $\mathcal{Y} = \{y_1, y_2, y_3\}$. As one can clearly see, the first row of measures is maximized by the uniform distribution, where all probability mass is equally distributed among all three

classes and radiates from the center of the simplex. In contrast, the measures in the second row are maximized by the extreme bimodal distribution, with all probability mass equally concentrated at the extreme classes (in this case at y_1 and y_3). In this case, the uncertainty radiates from the center between y_1 and y_3. While the heatmaps for entropy, label-wise binary entropy, and variance look quite similar, there is a bigger difference between variance (cf. Figure A1d) and the OCS decompositions for variance and entropy (cf. Figures A1e and A1f). Variance appears to have a very strong focus on the extreme bimodal distribution when it comes to uncertainty quantification. For probability mass moving towards the uniform distribution, it significantly measures less uncertainty than the OCS decompositions. Thus, one could conclude that the OCS reductions might strike a better balance and can be seen as standing between uncertainty measures maximized by the uniform distribution and strict dispersion measures like variance.

When comparing the uncertainties of the OCS decompositions, the decomposition with variance as the base measure (cf. Figure A1e) appears slightly less extended toward the uniform distribution than the decomposition with entropy as the base measure (cf. Figure A1f). Since ordinal classification lies between regression and nominal classification, it can be hypothesized that OCS-decomposition-based uncertainty measures are well-suited for uncertainty quantification in this context. These measures could effectively address two key aspects of uncertainty quantification: they may enhance the exact hit rate by indicating uncertainty for uniform distributions, and they may reduce error distances by indicating uncertainty in extreme bimodal cases. In contrast, other measures tend to focus primarily on one of these aspects.

Appendix B Prediction Rejection Ratios (PRR) - Detailed Results

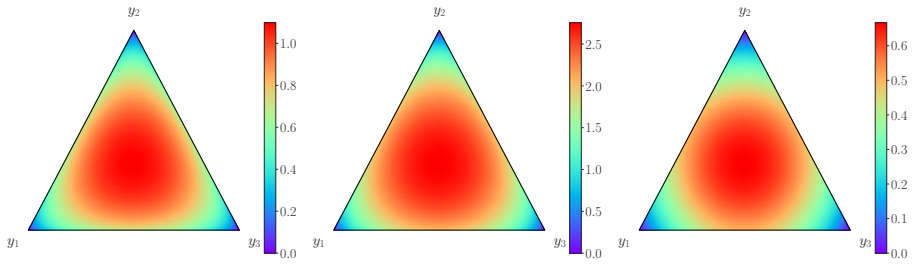
In the following, we present the detailed prediction rejection ratio results for the different tabular ordinal benchmark datasets, uncertainty types, and measures on the basis of an ensemble of GBTs (LightGBM (Ke et al., 2017)) (cf. Subsection 5.3). Table B1 shows results for missclassification rate (MCR) and Table B2 for mean absolute error (MAE).

Table B1: PRRs (MCR) using an ensemble of GBTs (LightGBM (Ke et al., 2017)).

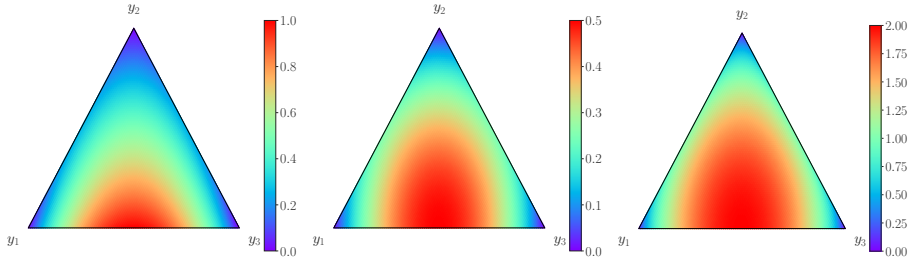
		PRR (MCR) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	0.282±0.06	0.286±0.06	0.289±0.06	0.293±0.06	**0.295**±0.06	**0.295**±0.06
	EU	0.1±0.17	0.096±0.16	0.118±0.15	0.113±0.15	0.121±0.15	**0.128**±0.15
	TU	**0.282**±0.05	0.278±0.05	0.274±0.04	0.251±0.04	0.259±0.04	0.24±0.05
Auto MPG	AU	0.362±0.17	0.383±0.16	0.394±0.15	0.429±0.15	0.436±0.14	**0.446**±0.15
	EU	0.338±0.17	0.349±0.14	0.36±0.16	0.394±0.16	0.396±0.17	**0.424**±0.18
	TU	0.363±0.18	0.383±0.16	0.389±0.16	0.435±0.14	0.435±0.14	**0.451**±0.16
Automobile	AU	0.637±0.47	0.635±0.47	0.625±0.48	0.623±0.47	**0.644**±0.48	0.581±0.51
	EU	0.597±0.47	0.587±0.47	0.609±0.48	0.607±0.47	**0.622**±0.47	0.617±0.56
	TU	0.637±0.47	0.638±0.48	0.634±0.48	0.623±0.47	**0.645**±0.48	0.581±0.51
Balance Scale	AU	0.929±0.06	0.924±0.06	0.913±0.07	**0.96**±0.04	**0.96**±0.04	0.945±0.05
	EU	0.842±0.07	0.798±0.08	0.818±0.08	0.836±0.06	0.879±0.04	**0.911**±0.06

Continued on next page

		PRR (MCR) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
	TU	0.927±0.06	0.921±0.06	0.912±0.07	**0.959**±0.04	**0.959**±0.04	0.946±0.05
Wisconsin Breast Cancer	AU	0.275±0.27	0.29±0.27	**0.294**±0.27	0.255±0.27	0.224±0.29	0.175±0.23
	EU	0.131±0.43	**0.177**±0.41	**0.177**±0.38	0.123±0.41	0.104±0.45	0.1±0.44
	TU	0.27±0.3	0.285±0.28	**0.29**±0.29	0.248±0.28	0.224±0.31	0.181±0.25
ERA	AU	0.151±0.11	0.153±0.12	**0.155**±0.12	0.08±0.08	0.098±0.08	0.043±0.07
	EU	**0.047**±0.1	0.022±0.11	0.025±0.12	-0.013±0.12	0.012±0.13	-0.022±0.13
	TU	0.148±0.11	0.153±0.12	**0.154**±0.12	0.079±0.08	0.099±0.08	0.042±0.06
ESL	AU	0.247±0.13	0.241±0.12	0.235±0.11	0.262±0.1	0.27±0.12	**0.278**±0.11
	EU	0.237±0.13	0.229±0.15	0.223±0.14	0.226±0.14	0.263±0.11	**0.27**±0.11
	TU	0.25±0.13	0.241±0.12	0.241±0.11	0.263±0.09	0.268±0.11	**0.277**±0.11
Eucalyptus	AU	**0.437**±0.1	0.431±0.11	0.424±0.11	0.418±0.12	0.428±0.11	0.414±0.12
	EU	0.44±0.08	0.434±0.08	0.439±0.08	0.437±0.07	**0.449**±0.06	0.447±0.06
	TU	**0.44**±0.1	0.435±0.1	0.428±0.11	0.424±0.11	0.433±0.1	0.416±0.12
Heart (CLE)	AU	0.618±0.16	0.621±0.16	**0.629**±0.16	0.616±0.14	0.602±0.14	0.582±0.14
	EU	0.518±0.19	0.52±0.17	**0.559**±0.17	0.549±0.14	0.548±0.15	0.541±0.15
	TU	0.615±0.16	0.62±0.16	**0.623**±0.16	0.616±0.14	0.6±0.14	0.582±0.13
Boston Housing	AU	**0.406**±0.15	0.397±0.15	0.393±0.15	0.39±0.15	0.398±0.15	0.396±0.14
	EU	0.415±0.15	**0.416**±0.15	0.41±0.15	0.41±0.14	0.414±0.15	0.41±0.14
	TU	**0.412**±0.15	0.404±0.15	0.4±0.16	0.399±0.15	0.408±0.15	0.405±0.14
LEV	AU	0.181±0.1	**0.182**±0.1	0.174±0.1	0.167±0.11	0.178±0.1	0.167±0.1
	EU	0.165±0.12	0.161±0.13	0.175±0.13	0.185±0.14	0.174±0.11	**0.196**±0.11
	TU	**0.181**±0.1	**0.181**±0.1	0.176±0.1	0.166±0.11	0.177±0.1	0.168±0.1
Machine CPU	AU	0.664±0.11	0.692±0.11	0.692±0.1	0.727±0.11	0.734±0.1	**0.742**±0.11
	EU	0.582±0.16	0.619±0.14	0.65±0.13	0.675±0.15	0.649±0.15	**0.702**±0.15
	TU	0.668±0.1	0.684±0.11	0.696±0.1	0.727±0.11	0.732±0.11	**0.742**±0.11
New Thyroid	AU	0.54±0.47	0.54±0.47	0.54±0.47	**0.565**±0.49	**0.565**±0.49	0.56±0.48
	EU	0.545±0.48	0.545±0.48	0.545±0.48	**0.57**±0.49	0.56±0.49	0.565±0.49
	TU	0.54±0.47	0.535±0.47	0.535±0.47	**0.565**±0.49	**0.565**±0.49	0.56±0.48
Pyrimidines	AU	-0.094±0.4	-0.062±0.35	-0.062±0.35	0.038±0.56	**0.099**±0.31	-0.01±0.64
	EU	-0.058±0.49	-0.059±0.37	-0.041±0.39	-0.096±0.52	-0.039±0.5	**0.054**±0.47
	TU	-0.094±0.4	-0.077±0.43	-0.077±0.43	0.056±0.58	**0.099**±0.31	-0.01±0.64
Red Wine	AU	0.432±0.12	0.436±0.12	**0.438**±0.12	0.433±0.12	0.434±0.12	0.429±0.13
	EU	0.438±0.09	0.446±0.08	**0.456**±0.09	**0.456**±0.09	0.443±0.1	0.439±0.11
	TU	0.436±0.11	0.442±0.12	**0.444**±0.12	0.44±0.12	0.437±0.12	0.432±0.13
SWD	AU	0.189±0.09	0.188±0.08	0.182±0.08	0.189±0.08	**0.194**±0.09	0.191±0.09
	EU	0.142±0.11	0.115±0.08	0.136±0.09	0.142±0.09	0.163±0.1	**0.185**±0.11
	TU	0.193±0.09	0.187±0.08	0.182±0.08	0.189±0.08	**0.195**±0.09	0.192±0.09
Stocks Domain	AU	**0.668**±0.06	**0.668**±0.06	**0.668**±0.06	0.666±0.07	0.666±0.06	0.665±0.07
	EU	0.643±0.07	0.627±0.08	0.628±0.08	0.627±0.08	**0.644**±0.07	0.643±0.07
	TU	0.668±0.06	**0.669**±0.06	0.666±0.06	0.664±0.06	0.668±0.06	0.666±0.07
TAE	AU	**0.229**±0.32	0.218±0.34	0.213±0.34	0.164±0.31	0.17±0.31	0.16±0.26
	EU	0.037±0.27	0.001±0.23	0.022±0.27	0.013±0.23	0.094±0.24	**0.108**±0.16
	TU	**0.233**±0.32	0.224±0.33	0.197±0.36	0.157±0.31	0.166±0.3	0.16±0.26
Triazines	AU	**0.308**±0.2	0.284±0.18	0.262±0.18	0.255±0.22	0.259±0.22	0.232±0.24
	EU	0.241±0.25	0.218±0.23	0.236±0.2	0.267±0.23	**0.276**±0.24	0.272±0.23
	TU	**0.315**±0.19	0.287±0.18	0.265±0.18	0.255±0.22	0.262±0.22	0.234±0.24
White Wine	AU	**0.378**±0.05	0.375±0.05	0.368±0.05	0.355±0.05	0.366±0.05	0.345±0.05
	EU	0.436±0.05	0.438±0.05	**0.444**±0.05	0.439±0.04	0.433±0.05	0.419±0.05
	TU	**0.394**±0.05	0.391±0.05	0.383±0.05	0.367±0.05	0.379±0.05	0.355±0.05
CMC	AU	0.359±0.1	**0.36**±0.1	**0.36**±0.1	0.302±0.05	0.305±0.05	0.252±0.05
	EU	**0.285**±0.09	0.236±0.09	0.25±0.09	0.214±0.08	0.262±0.08	0.235±0.07
	TU	0.358±0.1	**0.359**±0.1	**0.359**±0.1	0.3±0.05	0.303±0.05	0.253±0.05
Grub Damage	AU	0.222±0.38	0.249±0.39	0.248±0.4	**0.278**±0.34	0.27±0.33	0.221±0.33
	EU	0.142±0.31	0.174±0.3	**0.223**±0.37	0.174±0.35	0.155±0.37	0.135±0.36
	TU	0.222±0.38	0.242±0.39	0.251±0.4	**0.277**±0.34	**0.277**±0.33	0.229±0.34
Obesity	AU	0.891±0.08	0.892±0.08	0.893±0.08	0.895±0.08	0.895±0.08	**0.898**±0.09
	EU	0.892±0.08	0.892±0.09	0.893±0.09	0.896±0.09	0.895±0.08	**0.899**±0.09
	TU	0.891±0.08	0.892±0.08	0.893±0.09	0.896±0.09	0.895±0.09	**0.899**±0.09



(a) Entropy (b) Label-wise binary entropy (one vs. rest) (c) Label-wise binary variance (one vs. rest)



(d) Variance (e) Ordinal binary variance (f) Ordinal binary entropy

Fig. A1: (Total) uncertainty heatmaps for the different uncertainty measures using the probability simplex over $\mathcal{Y} = \{y_1, y_2, y_3\}$.

Table B2: PRRs (MAE) using an ensemble of GBTs (LightGBM (Ke et al., 2017)).

		PRR (MAE) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	0.299±0.05	0.315±0.05	0.323±0.05	0.334±0.05	0.328±0.05	**0.335**±0.04
	EU	0.113±0.18	0.115±0.18	0.141±0.16	0.136±0.16	0.138±0.16	**0.15**±0.15
	TU	0.304±0.03	0.312±0.03	**0.313**±0.02	0.295±0.02	0.295±0.02	0.282±0.03
Auto MPG	AU	0.332±0.17	0.361±0.16	0.38±0.14	0.47±0.12	0.47±0.12	**0.495**±0.12
	EU	0.332±0.15	0.376±0.12	0.39±0.13	0.437±0.15	0.434±0.14	**0.479**±0.14
	TU	0.328±0.19	0.362±0.16	0.381±0.14	0.475±0.12	0.476±0.11	**0.501**±0.12
Automobile	AU	0.557±0.46	0.556±0.47	0.549±0.47	0.562±0.47	**0.581**±0.47	0.516±0.5
	EU	0.534±0.46	0.528±0.47	0.542±0.49	**0.556**±0.47	0.555±0.45	0.545±0.54
	TU	0.554±0.46	0.561±0.47	0.557±0.47	0.563±0.47	**0.582**±0.47	0.517±0.5
Balance Scale	AU	**0.87**±0.06	0.864±0.06	0.856±0.07	0.848±0.09	0.85±0.08	0.854±0.09
	EU	0.804±0.09	0.766±0.1	0.78±0.09	0.791±0.11	**0.821**±0.11	0.813±0.1
	TU	**0.863**±0.06	0.861±0.06	0.856±0.07	0.847±0.09	0.848±0.09	0.857±0.1
Wisconsin Breast Cancer	AU	0.155±0.33	0.179±0.3	**0.191**±0.28	0.143±0.27	0.123±0.29	0.105±0.27
	EU	-0.004±0.21	0.031±0.21	**0.05**±0.19	0.004±0.26	-0.009±0.24	0.009±0.26
	TU	0.148±0.33	0.181±0.3	**0.188**±0.28	0.13±0.27	0.123±0.29	0.111±0.26
ERA	AU	0.024±0.08	0.013±0.07	0.014±0.07	0.02±0.1	**0.034**±0.09	0.017±0.11
	EU	-0.024±0.13	-0.019±0.12	-0.021±0.12	-0.033±0.12	-0.024±0.13	**-0.013**±0.13
	TU	0.022±0.08	0.011±0.07	0.01±0.07	0.019±0.1	**0.035**±0.09	0.016±0.11
ESL	AU	0.257±0.12	0.255±0.12	0.254±0.11	0.278±0.11	0.285±0.12	**0.295**±0.12
	EU	0.262±0.12	0.255±0.16	0.249±0.16	0.252±0.16	0.288±0.11	**0.293**±0.12
	TU	0.262±0.11	0.257±0.11	0.259±0.11	0.28±0.11	0.284±0.11	**0.294**±0.12
Eucalyptus	AU	**0.42**±0.1	0.417±0.1	0.412±0.11	0.41±0.12	**0.42**±0.11	0.407±0.12
	EU	0.43±0.08	0.429±0.08	0.434±0.08	0.439±0.08	**0.444**±0.07	**0.444**±0.08

Continued on next page

		PRR (MAE) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
	TU	**0.425**±0.1	0.421±0.1	0.416±0.11	0.417±0.11	**0.425**±0.11	0.41±0.12
Heart (CLE)	AU	0.56±0.14	0.565±0.14	0.577±0.13	**0.581**±0.1	0.571±0.1	0.556±0.1
	EU	0.487±0.19	0.491±0.17	0.519±0.17	0.533±0.13	0.534±0.12	**0.535**±0.1
	TU	0.553±0.14	0.566±0.14	0.568±0.14	**0.583**±0.1	0.569±0.09	0.559±0.09
Boston Housing	AU	0.402±0.2	0.394±0.19	0.391±0.2	0.391±0.19	0.398±0.19	**0.403**±0.17
	EU	0.413±0.19	0.42±0.18	0.416±0.18	**0.421**±0.18	0.413±0.19	0.413±0.18
	TU	0.409±0.2	0.401±0.2	0.397±0.2	0.4±0.19	0.407±0.19	**0.412**±0.18
LEV	AU	0.176±0.1	**0.178**±0.1	0.173±0.1	0.168±0.1	0.175±0.09	0.166±0.1
	EU	0.164±0.12	0.167±0.12	0.181±0.13	0.191±0.13	0.174±0.1	**0.193**±0.1
	TU	**0.177**±0.1	**0.177**±0.1	0.175±0.1	0.167±0.1	0.175±0.1	0.166±0.1
Machine CPU	AU	0.643±0.16	0.684±0.16	0.69±0.15	0.733±0.1	0.734±0.09	**0.747**±0.08
	EU	0.571±0.18	0.623±0.16	0.651±0.16	0.695±0.14	0.653±0.16	**0.712**±0.13
	TU	0.648±0.16	0.677±0.16	0.691±0.15	0.731±0.1	0.73±0.1	**0.747**±0.08
New Thyroid	AU	0.54±0.47	0.54±0.47	0.54±0.47	**0.565**±0.49	**0.565**±0.49	0.56±0.48
	EU	0.545±0.48	0.545±0.48	0.545±0.48	**0.57**±0.49	0.56±0.49	0.565±0.49
	TU	0.54±0.47	0.535±0.47	0.535±0.47	**0.565**±0.49	**0.565**±0.49	0.56±0.48
Pyrimidines	AU	0.231±0.45	0.225±0.4	0.192±0.42	0.257±0.34	0.255±0.33	**0.327**±0.43
	EU	**0.219**±0.34	0.04±0.35	0.024±0.34	-0.013±0.51	-0.011±0.51	-0.071±0.53
	TU	0.237±0.42	0.214±0.39	0.216±0.39	0.268±0.36	0.248±0.33	**0.29**±0.49
Red Wine	AU	0.429±0.1	0.435±0.11	**0.44**±0.11	0.436±0.11	0.433±0.11	0.432±0.12
	EU	0.431±0.08	0.437±0.08	**0.446**±0.09	**0.446**±0.09	0.436±0.09	0.431±0.1
	TU	0.434±0.1	0.442±0.11	**0.447**±0.11	0.443±0.11	0.436±0.11	0.435±0.11
SWD	AU	0.136±0.06	0.135±0.07	0.13±0.07	0.14±0.07	0.144±0.07	**0.149**±0.08
	EU	0.129±0.09	0.11±0.07	0.123±0.08	0.122±0.07	0.138±0.07	**0.154**±0.07
	TU	0.138±0.06	0.134±0.07	0.13±0.07	0.141±0.07	0.146±0.07	**0.15**±0.08
Stocks Domain	AU	**0.668**±0.06	**0.668**±0.06	**0.668**±0.06	0.666±0.07	0.666±0.06	0.665±0.07
	EU	0.643±0.07	0.627±0.08	0.628±0.08	0.627±0.08	**0.644**±0.07	0.643±0.07
	TU	0.668±0.06	**0.669**±0.06	0.666±0.06	0.664±0.06	0.668±0.06	0.666±0.07
TAE	AU	0.059±0.3	0.048±0.33	0.044±0.33	0.235±0.22	0.243±0.21	**0.256**±0.2
	EU	0.094±0.13	0.08±0.11	0.091±0.15	0.116±0.16	0.213±0.17	**0.224**±0.22
	TU	0.063±0.3	0.059±0.31	0.041±0.34	0.229±0.21	0.239±0.21	**0.265**±0.2
Triazines	AU	**0.348**±0.21	0.332±0.17	0.314±0.16	0.326±0.21	0.337±0.22	0.299±0.22
	EU	0.304±0.22	0.32±0.16	0.333±0.13	0.362±0.17	0.364±0.18	**0.365**±0.19
	TU	**0.352**±0.2	0.336±0.18	0.316±0.16	0.326±0.21	0.335±0.22	0.301±0.22
White Wine	AU	0.374±0.05	**0.376**±0.04	0.375±0.04	0.365±0.04	0.37±0.04	0.354±0.04
	EU	0.419±0.06	0.422±0.06	**0.433**±0.06	0.427±0.05	0.42±0.06	0.41±0.06
	TU	0.39±0.05	**0.393**±0.04	0.388±0.04	0.376±0.04	0.383±0.04	0.364±0.04
CMC	AU	0.221±0.06	0.219±0.06	0.218±0.06	0.288±0.06	0.289±0.05	**0.294**±0.07
	EU	0.23±0.07	0.203±0.07	0.207±0.07	0.215±0.08	0.253±0.08	**0.267**±0.08
	TU	0.22±0.06	0.219±0.06	0.217±0.06	0.288±0.06	0.289±0.05	**0.295**±0.07
Grub Damage	AU	0.231±0.28	0.251±0.27	0.246±0.28	**0.339**±0.17	0.334±0.16	0.303±0.18
	EU	-0.016±0.27	-0.002±0.27	0.022±0.29	0.01±0.28	0.04±0.26	**0.062**±0.23
	TU	0.221±0.28	0.245±0.29	0.246±0.28	**0.336**±0.17	**0.336**±0.16	0.305±0.17
Obesity	AU	0.895±0.07	0.896±0.08	0.898±0.08	0.899±0.08	0.899±0.08	**0.903**±0.08
	EU	0.896±0.08	0.896±0.08	0.898±0.08	0.9±0.08	0.9±0.08	**0.903**±0.08
	TU	0.895±0.08	0.896±0.08	0.898±0.08	0.9±0.08	0.899±0.08	**0.903**±0.08

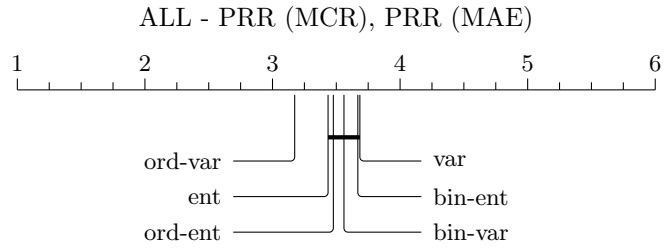
Appendix C Additional Experiments with Ensembles of Gradient Boosted Trees (GBTs)

For the sake of completeness, we also include experimental results for error and OOD detection for the two other popular gradient boosting tree libraries, namely XGBoost (Chen & Guestrin, 2016) and CatBoost (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018). We also create an ensemble of 10 XGBoost and CatBoost GBTs, respectively with subsample rate set to 0.5 to induce stochasticity in the training process. As our primary focus is on uncertainty quantification and not on predictive performance, we leave the parameters with the default values and do not perform any hyperparameter tuning.

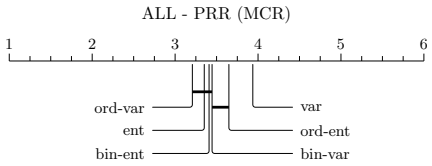
On the one hand, overall error detection results for all uncertainty types (AU, EU, and TU) resemble those obtained for LightGBM (cf. Figure 10), with ord-var significantly outperforming all other measures when considering MCR and MAE simultaneously (cf. Figures C2 and C3). On the other hand, nominal uncertainty measures like ent or bin-ent perform better for XGBoost and CatBoost, which leads to ord-ent and var falling behind overall. Or phrased differently, we can spot the weakness of var when it comes to MCR (cf. Figure C2b) as well as the weaker performance of ord-ent compared to ord-var and var when it comes to MAE. Here it becomes even clearer that, in particular, ord-var captures the inherent trade-off between exact hit rate and minimized error distance for ordinal classification best.

Just like for LightGBM (cf. Figure 11), ord-ent performs best when it comes to OOD detection based on measured epistemic uncertainty (cf. Figures C4 and C5), and the OCS decomposition measures can be considered competitive to the other measures when it comes to OOD detection.

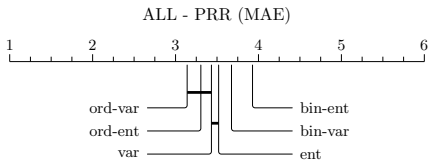
Fig. C2: CD diagrams for all uncertainty types (AU, EU and TU) using an ensemble of GBTs (XGBoost (Chen & Guestrin, 2016)).



(a) Result for MCR and MAE.



(b) Result only for MCR.

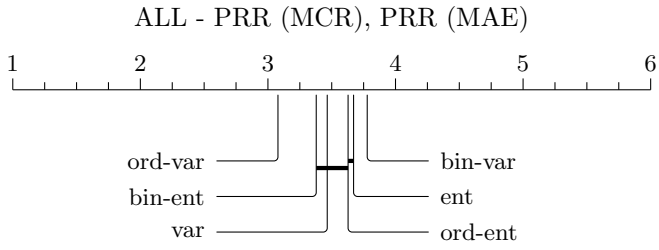


(c) Result only for MAE.

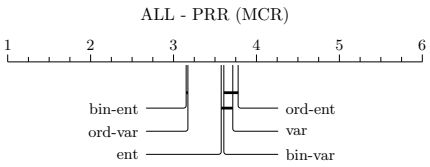
Appendix D Comparison of Predictive Performance and Prediction Rejection Ratios (PRRs)

To complement the error and OOD detection experiments, Table D3 presents the predictive performance of the ensembles utilized, based on their base predictors: MLP with CE loss (Pedregosa et al., 2011), CatBoost (Prokhorenkova et al., 2018), XGBoost

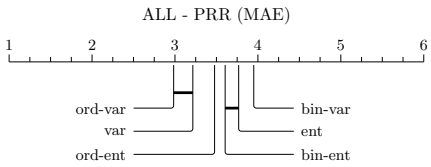
Fig. C3: CD diagrams for all uncertainty types (AU, EU and TU) using an ensemble of GBTs (CatBoost (Prokhorenkova et al., 2018)).



(a) Result for MCR and MAE.



(b) Result only for MCR.



(c) Result only for MAE.

Fig. C4: CD diagram for OOD detection of the different uncertainty measures using an ensemble of GBTs (XGBoost (Chen & Guestrin, 2016)).

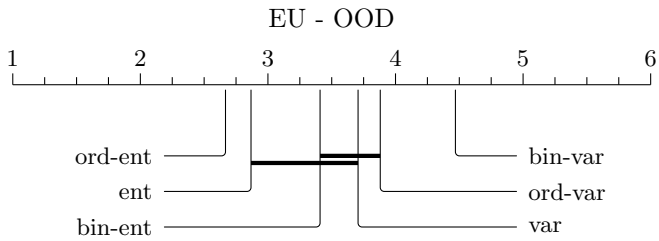
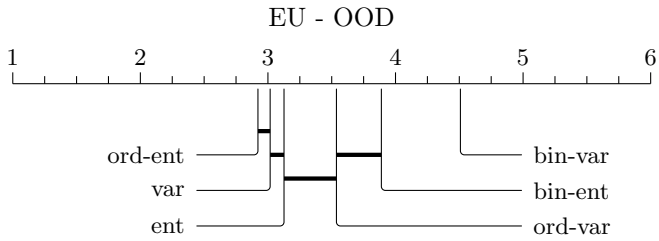


Fig. C5: CD diagram for OOD detection of the different uncertainty measures using an ensemble of GBTs (CatBoost (Prokhorenkova et al., 2018)).



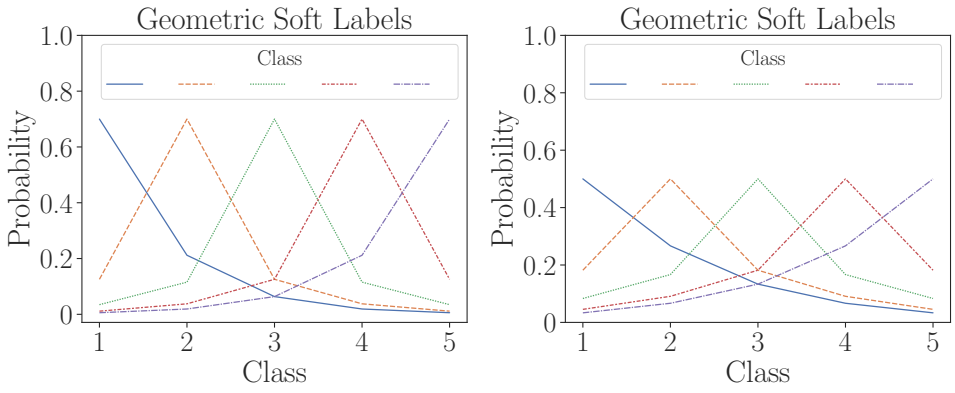
(Chen & Guestrin, 2016), and LightGBM (Ke et al., 2017), averaged across all datasets. Additionally, we include two widely used ordinal losses as alternatives to the CE loss: the squared EMD loss (Hou et al., 2016) and the QWK loss (de La Torre et al., 2018). This comparison highlights the differences in performance between predictors trained with CE loss and those trained with specialized ordinal losses. Moreover, we include a unimodal soft labeling (ULS) approach based on the geometric distribution, using LightGBM as the base learner. This method transforms deterministic one-hot (0/1) encoded labels into soft, unimodal probability distributions (cf. Figure D6 for an illustration) (Haas & Hüllermeier, 2023). The probability distribution is defined as follows:

$$p^{\text{GEO}}(k) = \begin{cases} 1-\alpha, & \text{if } k = c, \\ \frac{1}{G}\,\alpha^{|c-k|+1}(1-\alpha), & \text{if } k \neq c, \end{cases}$$

where α is the smoothing factor, k denotes the k-th class, and c represents the index of the true label in the one-hot (0/1) encoded label vector $\boldsymbol{y}$, with $y_c = 1$ and $y_k = 0$ for all other classes. G serves as a normalizing constant, ensuring that $\sum_{k=1}^{K} p^{\text{GEO}}(k) = 1$. It is defined as:

$$G = \sum_{k \neq c} \alpha^{|c-k|}(1-\alpha).$$

Unimodal soft labeling is a widely used method in ordinal classification that serves two purposes: First, it acts as a regularization technique akin to label smoothing, and second, it converts a standard predictor into an ordinal predictor by enforcing the assumption that adjacent classes are more likely than distant ones. In this approach, instead of subsampling per iteration, the 10 trees in the ensemble are smoothed using various smoothing factors, $\alpha = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$.



(a) $\alpha = 0.2$. (b) $\alpha = 0.5$.

Fig. D6: Unimodal label smoothing for five classes based on the geometric distribution (Haas & Hüllermeier, 2023).

To derive the final probabilistic prediction, we average the predicted probabilities of all ensemble members. The results are obtained using 10-fold cross-validation and

are averaged across all datasets. As our primary focus is on uncertainty quantification rather than predictive performance, we refrain from performing hyperparameter tuning on the GBTs and adhere to the MLP architecture specified in Table E4.

The metrics considered include accuracy (ACC), the accuracy of predictions allowing for errors in adjacent classes (1-OFF) (Bérchez-Moreno et al., 2025), the mean absolute error (MAE), the mean squared error (MSE), the quadratic weighted kappa (QWK), the negative log likelihood (NLL), the Brier score (BS) (Brier, 1950), the ranked probability score (RPS) (Epstein, 1969), and the expected calibration error (ECE) (de Menezes e Silva Filho et al., 2023). To obtain final deterministic predictions, we make a decision that minimizes the expected loss over the predictive probability distributions. For instance, for MAE, we use the l_1 loss, and for MSE, we use the l_2 loss. For ACC and QWK, we choose the class with the maximal probability (arg max), and for 1-OFF, we select the two classes with the highest probabilities, respectively.

CatBoost is the best base predictor when it comes to ACC, 1-OFF, and MAE. Followed by the squared EMD loss and XGBoost. The squared EMD loss leads to the best calibration with the best overall results on RPS, NLL, BS, and ECE. It is also the best predictor on MSE and very competitive on all other metrics. In particular, the RPS is of interest in probabilistic ordinal classification as it is a proper scoring rule for ordinal outcomes (Epstein, 1969; Galdran, 2023):

$$\text{RPS} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K-1}\Big(F_k(\boldsymbol{p}_i) - F_k(\boldsymbol{y}_i)\Big)^2\right),$$

where $\boldsymbol{p}$ is the predicted probability distribution, $\boldsymbol{y}$ is the true one-hot (0/1) encoded probability distribution, and $F_k(\boldsymbol{p})$ and $F_k(\boldsymbol{y})$ are the respective cumulative probability distributions of class k. The RPS assigns lower scores to probabilistic forecasts that allocate high probabilities to classes that are close to the correct class. Since the EMD loss is equivalent to the RPS metric in measuring the squared distance of the cumulative probabilities between predicted and true outcomes, it is plausible that EMD outperforms other predictors on the RPS metric:

$$l_{\text{EMD}}(\boldsymbol{y}, \boldsymbol{p}) = \sum_{k=1}^{K-1}\Big(F_k(\boldsymbol{p}) - F_k(\boldsymbol{y})\Big)^2$$

The same applies to the QWK loss, which leads to the best results on QWK. Remarkably, this is also true for XGBoost with CE loss, although XGBoost exhibits a higher standard deviation on QWK. From this evaluation, we can conclude that ordinal predictors indeed outperform predictors using CE loss on important ordinal metrics like QWK and RPS. However, as noted by Kasa et al. (2024), there appears to be a trade-off between nominal and ordinal performance, as improvements in ordinal metrics are achieved at the expense of nominal metrics.

When comparing the ULS approach with the standard LightGBM ensemble trained using CE loss, we observe that applying unimodal label smoothing effectively transforms a nominal predictor into an ordinal predictor, leading to improvements across

all performance metrics, including the RPS and NLL. However, for metrics such as the BS and ECE, the standard LightGBM ensemble demonstrates superior performance.

Table D3: Predictive performance and calibration of the different ensembles by their base predictors averaged over all datasets.

Predictor	ACC (↑)	1-OFF (↑)	MAE (↓)	MSE (↓)	QWK (↑)	RPS (↓)	NLL (↓)	BS (↓)	ECE (↓)
MLP	0.628±0.199	0.891±0.121	0.503±0.349	0.782±0.756	0.676±0.262	0.387±0.270	1.400±1.028	0.541±0.290	0.067±0.065
CatBoost	**0.639**±0.193	**0.900**±0.113	**0.470**±0.328	0.698±0.658	0.693±0.239	0.345±0.229	0.982±0.543	0.490±0.243	0.048±0.040
XGBoost	0.637±0.200	0.895±0.124	0.477±0.347	0.698±0.667	**0.696**±0.242	0.353±0.238	1.050±0.579	0.503±0.251	0.054±0.042
LightGBM	0.621±0.223	0.886±0.132	0.521±0.435	0.812±1.052	0.649±0.299	0.368±0.283	0.991±0.533	0.488±0.235	0.040±0.026
ULS	0.625±0.197	0.899±0.115	0.483±0.329	0.714±0.667	0.688±0.244	0.353±0.214	0.964±0.407	0.493±0.193	0.045±0.035
EMD	0.635±0.188	0.898±0.116	0.471±0.300	**0.689**±0.611	0.686±0.245	**0.335**±0.210	**0.892**±0.449	**0.467**±0.217	**0.017**±0.010
QWK	0.595±0.196	0.885±0.119	0.537±0.339	0.823±0.725	**0.696**±0.216	0.409±0.245	1.250±0.634	0.568±0.241	0.070±0.057

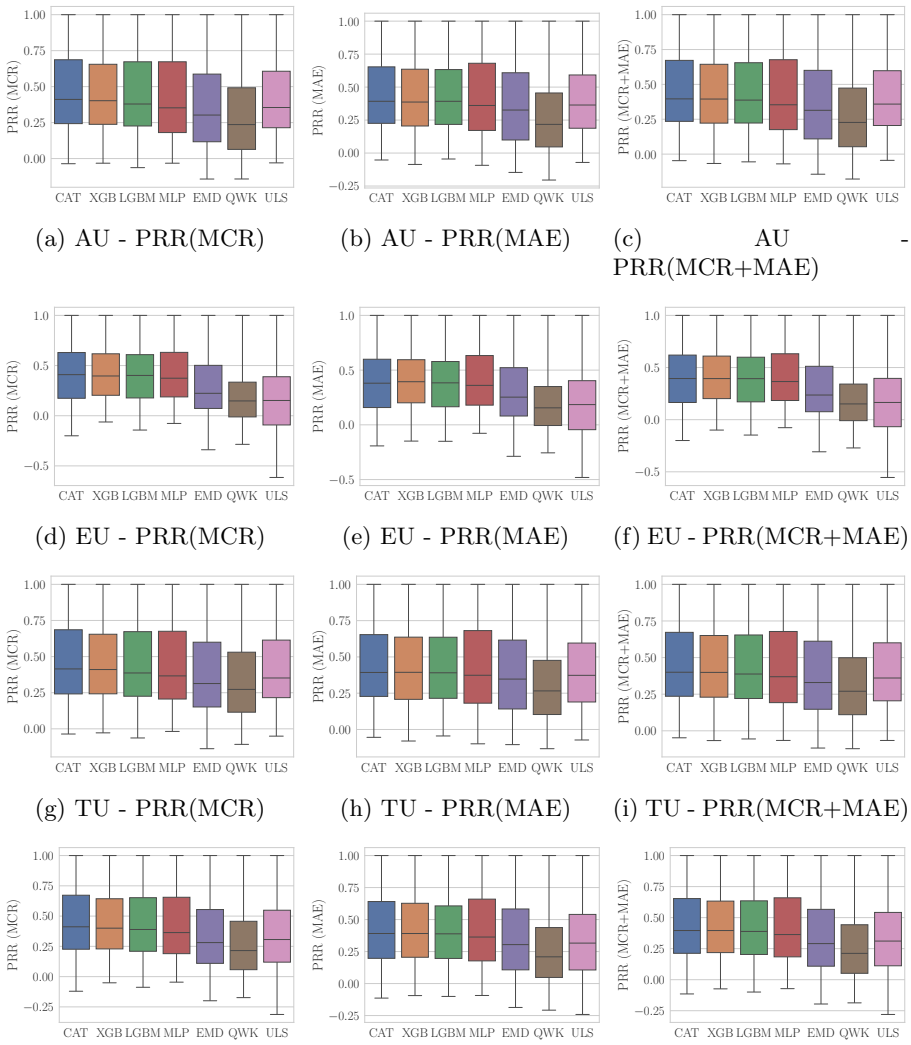
Figure D7 displays the attainable PRRs across all datasets and uncertainty measures, grouped by base predictor. It is important to note that PRRs are independent of predictive performance and solely measure the performance of uncertainty quantification. Specifically, they represent the area between the uncertainty measure-based rejection and random rejection, compared to the area between optimal and random rejection (cf. Figure 6). As one can see, regardless of the uncertainty measure, the base predictors using CE loss (CatBoost, XGBoost, LightGBM, and the MLP with CE loss) achieve higher PRRs compared to those using ordinal losses (EMD, QWK, and ULS).

When examining the CD diagrams in Figure D8, which are based on Wilcoxon signed-rank tests, we observe that these differences are statistically significant in most cases at a significance level of $p = 0.05$ for various uncertainty types and combinations of the considered metrics, namely MCR and MAE. Notably, QWK consistently performs the worst in terms of PRR. This is presumably due to the fact that it penalizes deviations from the true class quadratically, compared to EMD, which does so linearly, as noted by Galdran (2023). Additionally, EMD does not impose a strong penalty in the tails, as CDFs are monotonic; hence, the difference between CDFs will be small in the tails (Kasa et al., 2024). This is particularly evident in the PRRs for MAE, where ULS and EMD may not completely reassign distant class probability mass compared to QWK, thereby allowing for better uncertainty quantification.

Overall, we can conclude that although ordinal losses deliver good ordinal predictions on ordinal metrics, which are even well-calibrated in the case of EMD, they bias the predictor to output unimodal, compressed probability distributions. See Figure D9 for an illustration of this phenomenon, where we compare the predictive probability distributions of the CE loss and the QWK loss. This behavior, as also demonstrated by de La Torre et al. (2018) for the QWK loss, appears to negatively impact uncertainty quantification. Although EMD loss may improve calibration in probabilistic ordinal classification on average, it introduces a bias in the predictive probabilities by enforcing a unimodality assumption that is not universally valid. This inductive bias can obscure information critical for reliable uncertainty quantification in probabilistic ordinal classification. This observation is particularly notable in our experiments, as most datasets exhibit a unimodal prior distribution of class labels, $p(y)$, which might suggest that the predictive distributions, $p(y \mid \boldsymbol{x})$, are also unimodal in most cases, which aligns with standard assumptions in ordinal classification. However, when uncertainty quantification is a priority, CE loss emerges as a more suitable choice, as it provides unbiased predictive probability distributions in ordinal classification.

Appendix E Experiments with Ensemble of Multi-Layer Perceptron (MLP)

This section shows additional experimental results using an ensemble of 10 Multi-Layer Perceptrons (MLPs) (Pedregosa et al., 2011) instead of an ensemble of GBTs (cf. Section 5) to approximate Bayesian inference.



(a) AU - PRR(MCR) (b) AU - PRR(MAE) (c) AU - PRR(MCR+MAE)

(d) EU - PRR(MCR) (e) EU - PRR(MAE) (f) EU - PRR(MCR+MAE)

(g) TU - PRR(MCR) (h) TU - PRR(MAE) (i) TU - PRR(MCR+MAE)

(j) ALL - PRR(MCR) (k) ALL - PRR(MAE) (l) ALL - PRR(MCR+MAE)

Fig. D7: Raw PRRs over all datasets and uncertainty measures by uncertainty type (AU, EU, TU, or All), performance metric (MCR, MAE), and base predictor (CatBoost, XGBoost, LightGBM, MLP, EMD, QWK, or ULS).

E.1 Experimental Setup

For the experiments, the same datasets are used as in Section 5 (cf. Table 1). Besides the preprocessing applied in Section 5, numerical features are also standardized. The parameters of the MLPs are displayed in Table E4. Please note that our focus is not on predictive performance, but on uncertainty quantification, so we deliberately do

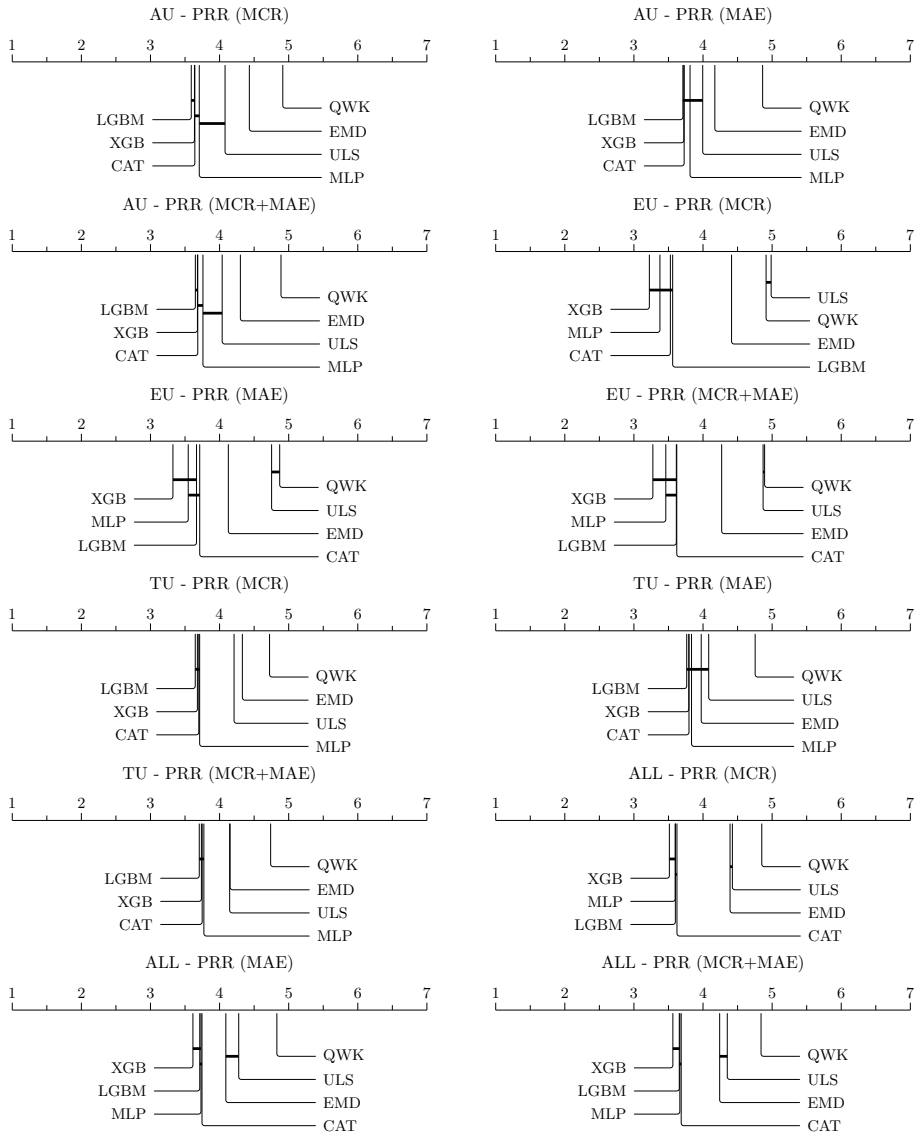


Fig. D8: CD diagrams displaying the ranks of the different base predictors in terms of attainable PRRs per performance measure (MCR, MAE) and uncertainty type (AU, EU, TU, All). Base predictors which are not significantly different at $p = 0.05$ are connected.

not perform extensive hyperparameter tuning. Nonetheless, we selected an architecture for the MLPs that leads to competitive performance compared to LightGBM, or GBTs in general. To create diversity among the MLPs in the ensemble, we provide

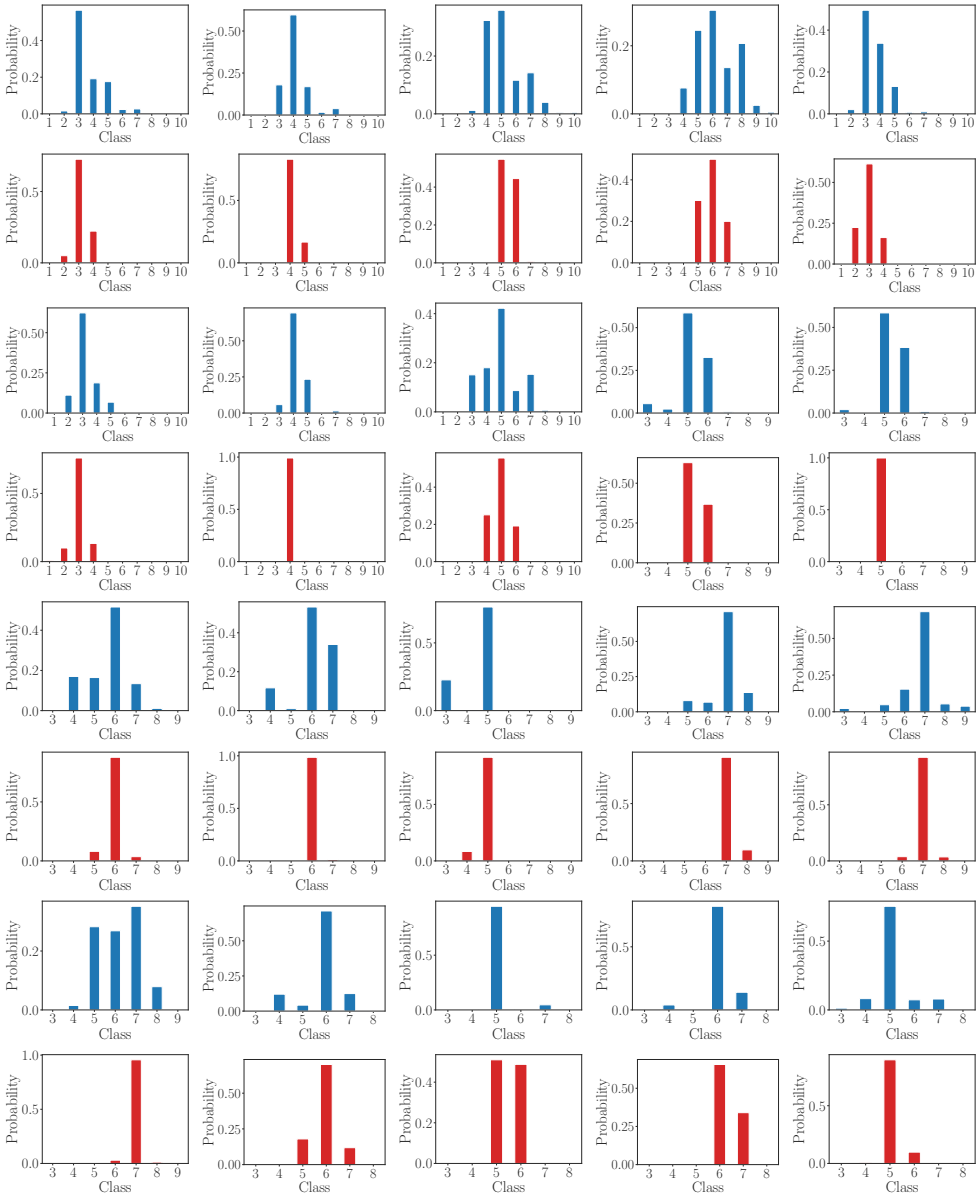


Fig. D9: Comparison of exemplary predictive probability distributions using the MLP with CE loss and the MLP with QWK loss on the Abalone, White Wine, and Red Wine datasets. Odd rows (blue) display the CE loss distributions, while even rows (red) show the corresponding QWK loss distributions.

different seeds, leading to different random number generations for weights and bias initialization.

Parameter	Value
Hidden Layer Sizes	[128,64,32]
Activation Function	ReLU
Solver	Adam
Maximum Epochs	200
Batch Size	200
L2 Regularization (alpha)	1e-4
Learning Rate	1e-3

Table E4: MLP parameters (Pedregosa et al., 2011).

E.2 Rejection Curves

Again, we first display accuracy rejection curves to visually validate the quality of uncertainty quantification for some datasets (cf. Subsection 5.2). Just like for the ensemble of GBTs, the investigated uncertainty methods all appear to be able to quantify epistemic and aleatoric uncertainty, as rejection curves monotonically increase in the case of ACC or decrease in the case of MAE, respectively (cf. Figures E10 and E11).

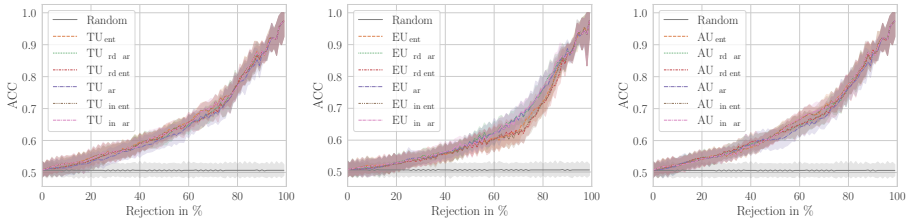
E.3 Prediction-Rejection-Ratios (PRRs)

The following CD diagrams (cf. Figures E12, E13, E14, and E15) show the ranks of the different uncertainty measures according to the obtained PRRs grouped by uncertainty type. The rankings resemble those for GBTs (cf. Subsection 5.3), with measures taking distance into account outperforming nominal measures. The results are even more significant than those for GBTs and underpin the superiority of the OCS decomposition method as well as variance in uncertainty quantification for ordinal classification and the disentanglement of aleatoric and epistemic uncertainty in this context.

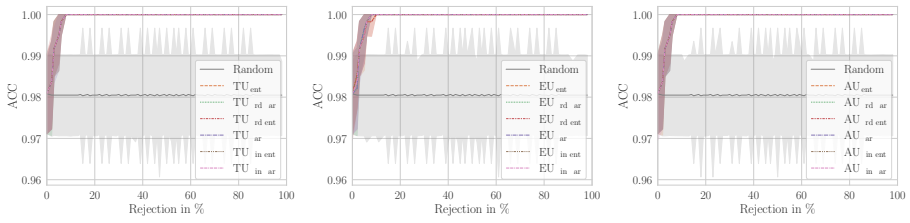
E.4 Prediction Rejection Ratios (PRR) - Detailed Results

In this subsection, we display the detailed prediction rejection ratio results for the different tabular ordinal benchmark datasets, uncertainty types, and measures using an ensemble of MLPs for approximate Bayesian inference. Table E5 shows results for misclassification rate (MCR) and Table E6 for mean absolute error (MAE).

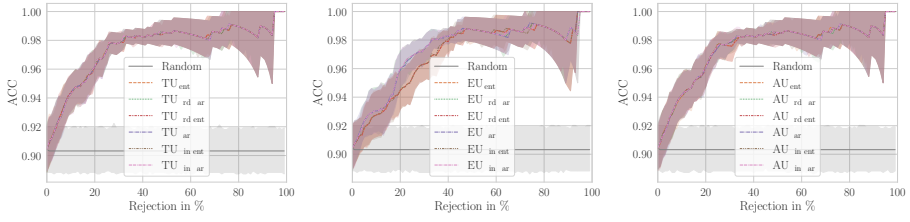
[4] https://github.com/mirkobunse/critdd



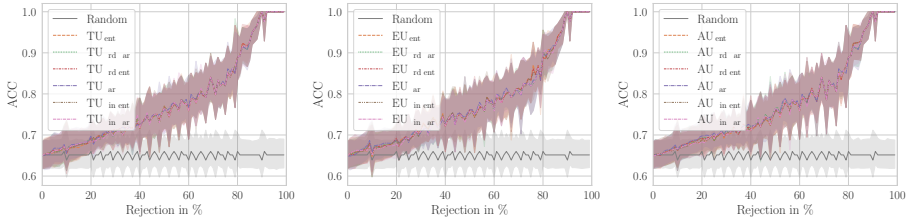
(a) CMC TU (b) CMC EU (c) CMC AU



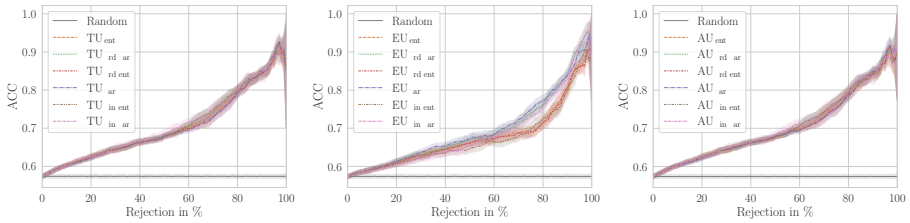
(d) Balance Scale TU (e) Balance Scale EU (f) Balance Scale AU



(g) Stocks Domain TU (h) Stocks Domain EU (i) Stocks Domain AU

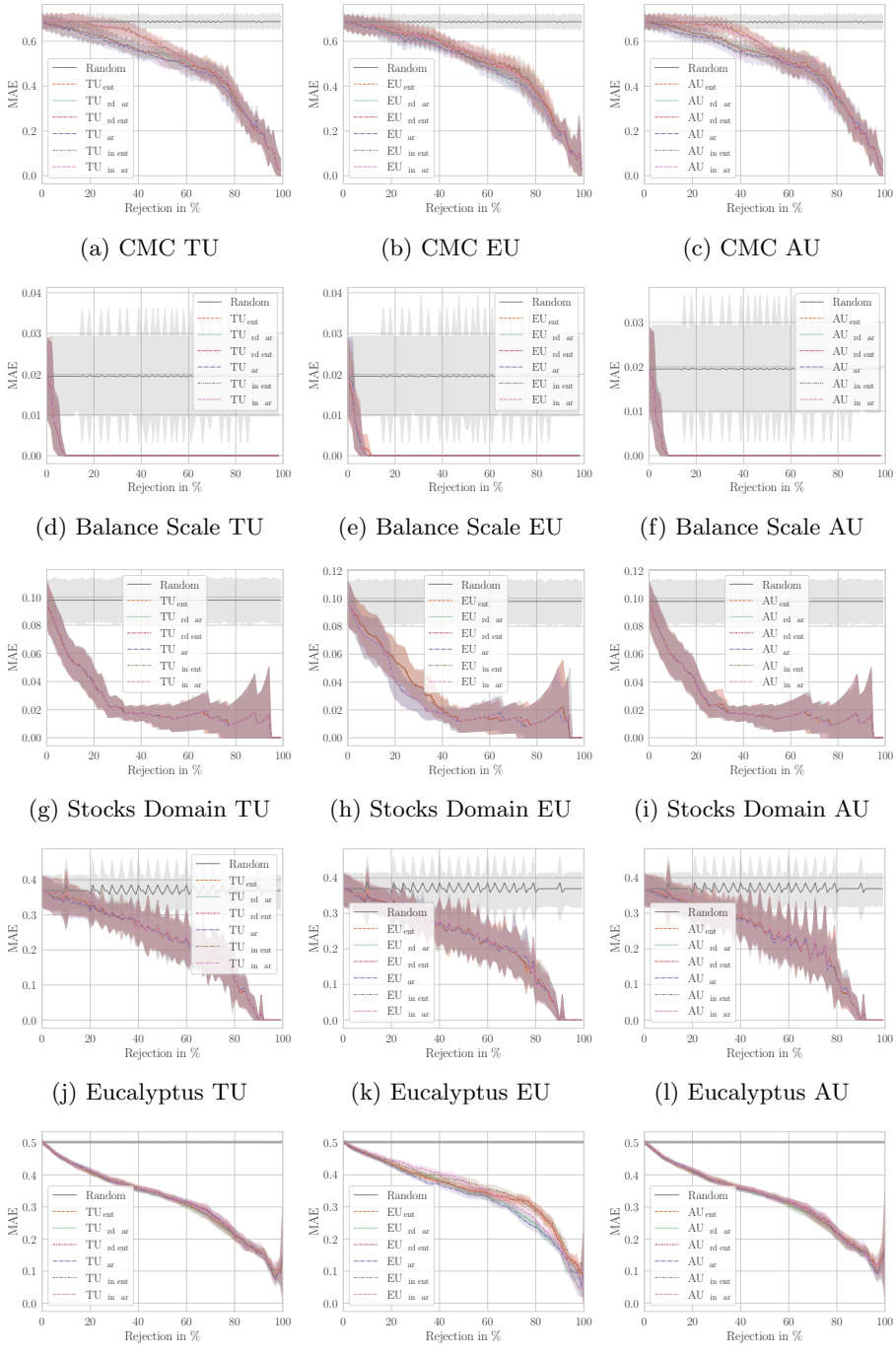


(j) Eucalyptus TU (k) Eucalyptus EU (l) Eucalyptus AU



(m) Abalone TU (n) Abalone EU (o) Abalone AU

Fig. E10: Accuracy rejection curves for different datasets, uncertainty types (TU, EU, AU), and measures using an ensemble of MLPs for approximate Bayesian infernce. Shaded regions around the mean represent the 95% confidence interval.



(a) CMC TU (b) CMC EU (c) CMC AU

(d) Balance Scale TU (e) Balance Scale EU (f) Balance Scale AU

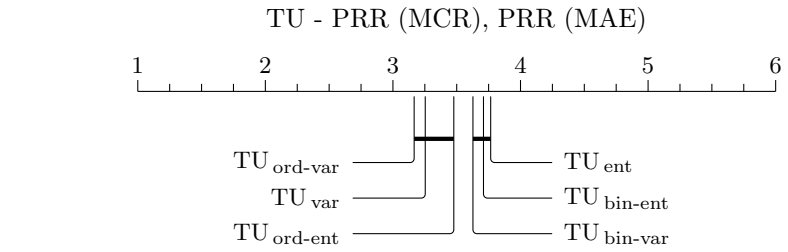
(g) Stocks Domain TU (h) Stocks Domain EU (i) Stocks Domain AU

(j) Eucalyptus TU (k) Eucalyptus EU (l) Eucalyptus AU

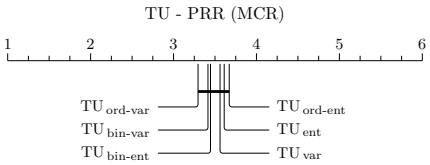
(m) Abalone TU (n) Abalone EU (o) Abalone AU

Fig. E11: Mean absolute error rejection curves for different datasets, uncertainty types (TU, EU, AU), and measures using an ensemble of MLPs for approximate Bayesian inference. Shaded regions around the mean represent the 95% confidence interval.

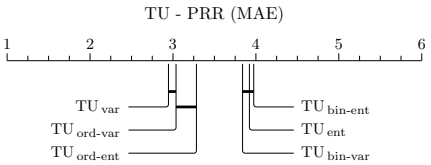
Fig. E12: Critical difference (CD) diagrams[4] for the evaluated Total Uncertainty (TU) measures and performance metrics based on a Friedman test followed by a post-hoc Holm-adjusted Wilcoxon signed-rank test (Benavoli et al., 2016; Demsar, 2006) using an ensemble of MLPs for approximate Bayesian inference. Groups of uncertainty measures that are not significantly different (at p = 0.05) are connected.



(a) Result for MCR and MAE.

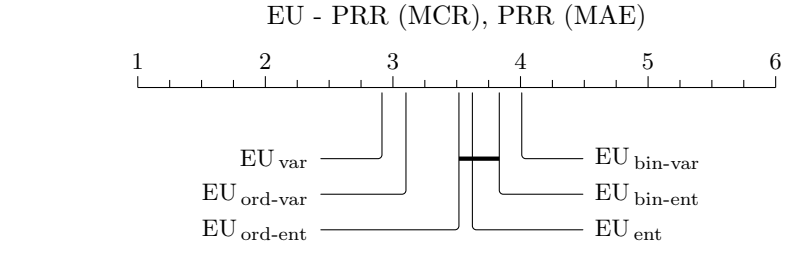


(b) Result only for MCR.

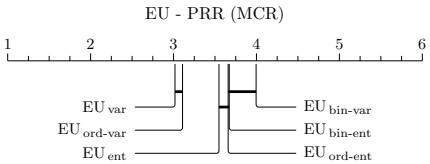


(c) Result only for MAE.

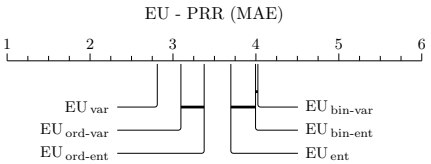
Fig. E13: CD diagrams for Epistemic Uncertainty (EU) using an ensemble of MLPs.



(a) Result for MCR and MAE.

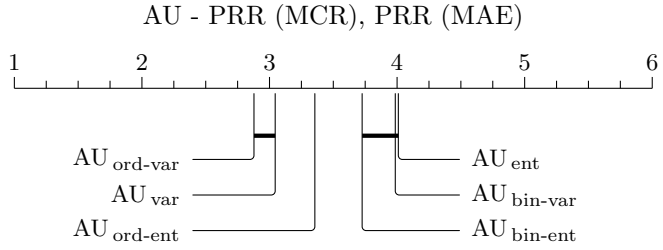


(b) Result only for MCR.

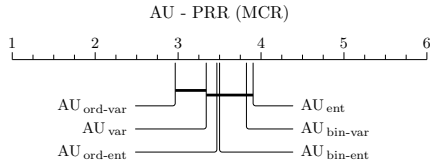


(c) Result only for MAE.

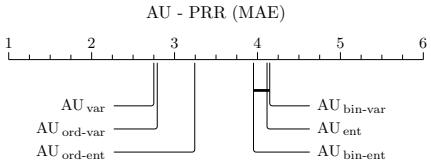
Fig. E14: CD diagrams for Aleatoric Uncertainty (AU) using an ensemble of MLPs.



(a) Result for MCR and MAE.

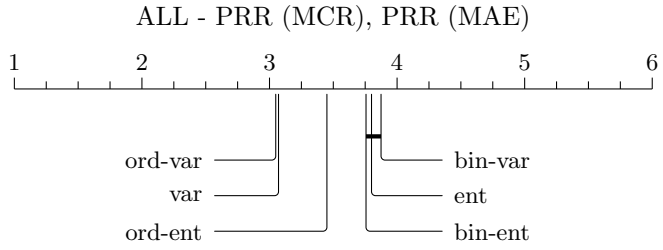


(b) Result only for MCR.

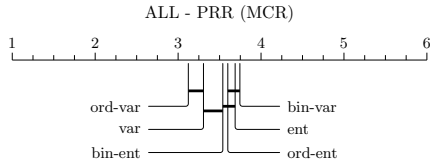


(c) Result only for MAE.

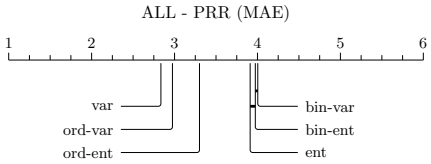
Fig. E15: CD diagrams for all uncertainty types (AU, EU and TU) using an ensemble of MLPs.



(a) Result for MCR and MAE.



(b) Result only for MCR.



(c) Result only for MAE.

Continued on next page

		PRR (MCR) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						

Table E5: PRRs (MCR) using an ensemble of MLPs.

		PRR (MCR) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	**0.34**±0.06	0.334±0.05	0.328±0.05	0.323±0.05	0.333±0.05	0.321±0.05
	EU	0.25±0.07	0.228±0.07	0.251±0.07	0.247±0.07	0.28±0.07	**0.293**±0.07
	TU	**0.344**±0.06	0.338±0.06	0.331±0.05	0.325±0.05	0.337±0.05	0.322±0.05
Auto MPG	AU	0.305±0.17	0.307±0.18	0.31±0.18	0.335±0.18	0.343±0.17	**0.352**±0.17
	EU	0.323±0.13	0.284±0.14	0.3±0.16	0.313±0.17	0.33±0.14	**0.335**±0.16
	TU	0.316±0.18	0.319±0.18	0.318±0.19	0.344±0.17	0.351±0.17	**0.355**±0.17
Automobile	AU	0.664±0.29	0.651±0.31	0.656±0.3	0.666±0.28	0.682±0.27	**0.686**±0.27
	EU	0.694±0.27	0.718±0.25	0.721±0.25	0.721±0.25	0.716±0.27	**0.729**±0.26
	TU	0.681±0.29	0.676±0.29	0.675±0.3	0.696±0.28	0.697±0.29	**0.712**±0.28
Balance Scale	AU	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47
	EU	**0.68**±0.47	0.679±0.47	0.678±0.47	0.678±0.47	0.678±0.47	0.678±0.47
	TU	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47
Wisconsin Breast Cancer	AU	0.213±0.2	0.219±0.19	**0.221**±0.2	0.201±0.25	0.186±0.26	0.206±0.25
	EU	0.2±0.31	0.211±0.31	0.222±0.29	**0.223**±0.32	0.211±0.32	0.207±0.33
	TU	0.228±0.26	0.219±0.25	**0.229**±0.24	0.206±0.28	0.196±0.31	0.208±0.29
ERA	AU	**0.161**±0.06	0.157±0.07	0.152±0.07	0.088±0.1	0.102±0.1	0.027±0.09
	EU	-0.014±0.13	-0.006±0.1	**0.006**±0.1	-0.016±0.1	-0.009±0.11	-0.04±0.11
	TU	**0.161**±0.06	0.154±0.07	0.152±0.07	0.088±0.1	0.101±0.1	0.026±0.09
ESL	AU	**0.238**±0.23	0.228±0.22	0.218±0.21	0.206±0.21	0.236±0.22	0.205±0.21
	EU	0.274±0.23	0.285±0.22	0.272±0.22	0.282±0.22	0.286±0.23	**0.298**±0.21
	TU	**0.242**±0.23	0.238±0.22	0.223±0.21	0.214±0.21	0.241±0.22	0.209±0.21
Eucalyptus	AU	0.325±0.09	0.322±0.09	0.322±0.09	0.33±0.09	0.334±0.09	**0.342**±0.09
	EU	0.329±0.09	0.337±0.09	0.34±0.09	0.34±0.09	0.334±0.09	**0.341**±0.08
	TU	0.337±0.09	0.335±0.09	0.333±0.09	0.34±0.08	0.339±0.08	**0.345**±0.08
Heart (CLE)	AU	0.592±0.15	0.595±0.14	0.595±0.14	0.603±0.13	0.596±0.13	**0.604**±0.12
	EU	0.467±0.22	0.453±0.22	0.476±0.21	0.477±0.23	0.483±0.24	**0.512**±0.22
	TU	0.56±0.18	0.56±0.17	0.562±0.16	0.57±0.15	**0.572**±0.16	0.57±0.16
Boston Housing	AU	0.418±0.24	0.412±0.24	0.414±0.23	0.423±0.24	0.431±0.24	**0.437**±0.25
	EU	0.433±0.19	0.453±0.18	0.455±0.18	**0.467**±0.18	0.448±0.2	0.458±0.2
	TU	0.424±0.22	0.425±0.22	0.431±0.22	0.439±0.23	0.434±0.23	**0.442**±0.24
LEV	AU	0.177±0.08	0.177±0.1	0.175±0.11	0.171±0.11	0.182±0.1	**0.185**±0.11
	EU	0.21±0.06	0.193±0.09	0.183±0.09	0.181±0.09	**0.211**±0.05	0.198±0.06
	TU	0.179±0.08	0.177±0.1	0.174±0.11	0.172±0.11	0.183±0.1	**0.185**±0.11
Machine CPU	AU	0.609±0.2	0.632±0.19	0.649±0.18	0.692±0.15	0.68±0.14	**0.704**±0.15
	EU	0.591±0.13	0.6±0.13	**0.619**±0.12	0.613±0.12	0.61±0.13	0.611±0.13
	TU	0.596±0.19	0.632±0.17	0.647±0.16	0.67±0.14	0.656±0.14	**0.7**±0.15
New Thyroid	AU	**0.479**±0.51	0.474±0.5	**0.479**±0.51	**0.479**±0.51	**0.479**±0.51	**0.479**±0.51
	EU	**0.484**±0.51	**0.484**±0.51	**0.484**±0.51	0.479±0.51	0.479±0.51	0.474±0.5
	TU	0.479±0.51	0.479±0.51	**0.484**±0.51	0.474±0.5	0.474±0.5	0.474±0.5
Pyrimidines	AU	0.181±0.29	**0.189**±0.29	0.167±0.29	0.149±0.36	0.136±0.38	0.05±0.42
	EU	0.029±0.49	0.015±0.47	0.026±0.34	-0.044±0.48	**0.039**±0.49	0.03±0.54
	TU	0.156±0.3	**0.189**±0.24	0.175±0.22	0.089±0.39	0.128±0.4	0.081±0.49
Red Wine	AU	0.316±0.13	0.305±0.12	0.303±0.12	0.311±0.12	0.325±0.12	**0.327**±0.11
	EU	0.423±0.12	0.434±0.12	**0.44**±0.12	0.439±0.13	0.429±0.12	0.423±0.13
	TU	**0.385**±0.13	0.384±0.13	0.381±0.13	0.375±0.13	0.381±0.13	0.375±0.13
SWD	AU	0.184±0.11	0.18±0.11	0.181±0.1	0.19±0.11	0.188±0.11	**0.192**±0.11
	EU	0.129±0.11	0.101±0.08	0.119±0.08	0.117±0.07	0.149±0.11	**0.171**±0.11
	TU	0.184±0.11	0.18±0.11	0.181±0.1	0.19±0.11	0.19±0.11	**0.192**±0.11
Stocks Domain	AU	0.716±0.07	**0.717**±0.07	0.715±0.07	0.714±0.08	0.716±0.07	0.714±0.07
	EU	**0.676**±0.07	0.625±0.08	0.622±0.08	0.623±0.08	**0.676**±0.07	**0.676**±0.07
	TU	0.716±0.07	0.716±0.07	**0.717**±0.07	0.716±0.07	0.716±0.07	0.716±0.07
TAE	AU	0.208±0.38	0.224±0.39	0.231±0.39	0.217±0.45	**0.237**±0.44	0.214±0.45
	EU	**0.36**±0.3	0.27±0.28	0.295±0.29	0.266±0.28	0.332±0.32	0.313±0.39
	TU	0.252±0.36	0.267±0.39	**0.283**±0.4	0.249±0.46	0.256±0.46	0.226±0.48
Triazines	AU	0.194±0.27	0.193±0.27	0.194±0.26	0.226±0.29	0.221±0.28	**0.24**±0.29
	EU	0.273±0.26	0.294±0.28	**0.295**±0.27	0.286±0.28	0.274±0.25	0.266±0.26
	TU	0.214±0.25	0.221±0.25	0.214±0.24	0.236±0.26	0.229±0.26	**0.254**±0.27
White Wine	AU	0.25±0.06	0.246±0.06	0.246±0.06	0.254±0.06	**0.263**±0.06	**0.263**±0.06

Continued on next page

		PRR (MCR) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
	EU	0.363±0.03	0.353±0.04	0.356±0.04	0.356±0.05	**0.37**±0.04	0.361±0.05
	TU	**0.341**±0.05	0.331±0.05	0.323±0.05	0.315±0.05	0.33±0.05	0.311±0.05
CMC	AU	0.316±0.06	0.316±0.06	**0.317**±0.06	0.313±0.07	0.313±0.07	0.293±0.07
	EU	**0.255**±0.07	0.218±0.07	0.229±0.07	0.212±0.07	0.252±0.07	0.246±0.06
	TU	0.328±0.06	0.33±0.05	**0.333**±0.06	0.308±0.07	0.309±0.06	0.284±0.07
Grub Damage	AU	0.152±0.28	0.159±0.27	0.161±0.26	0.182±0.27	0.157±0.27	**0.2**±0.26
	EU	0.152±0.27	0.14±0.29	0.141±0.31	0.167±0.33	**0.181**±0.31	0.162±0.31
	TU	0.201±0.26	0.212±0.26	**0.219**±0.26	0.209±0.28	0.195±0.29	0.21±0.28
Obesity	AU	0.858±0.06	0.853±0.06	0.854±0.06	0.862±0.06	0.864±0.06	**0.87**±0.05
	EU	0.887±0.05	0.891±0.05	0.892±0.05	**0.894**±0.04	0.89±0.05	0.891±0.05
	TU	0.881±0.06	0.882±0.06	0.882±0.06	0.886±0.06	0.884±0.06	**0.887**±0.05

Table E6: PRRs (MAE) using an ensemble of MLPs.

		PRR (MAE) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	0.382±0.05	0.387±0.04	0.387±0.04	0.389±0.04	**0.394**±0.04	0.39±0.04
	EU	0.288±0.05	0.288±0.05	0.317±0.05	0.315±0.05	0.333±0.05	**0.354**±0.04
	TU	0.389±0.05	0.395±0.05	0.393±0.04	0.393±0.04	**0.398**±0.04	0.393±0.04
Auto MPG	AU	0.285±0.14	0.294±0.15	0.302±0.15	0.342±0.15	0.353±0.15	**0.365**±0.16
	EU	0.304±0.16	0.281±0.16	0.298±0.18	0.313±0.18	0.316±0.16	**0.325**±0.16
	TU	0.307±0.16	0.314±0.16	0.317±0.17	0.359±0.17	0.367±0.16	**0.373**±0.17
Automobile	AU	0.657±0.24	0.652±0.25	0.654±0.24	0.688±0.23	0.693±0.22	**0.715**±0.23
	EU	0.692±0.22	0.715±0.2	0.72±0.19	0.729±0.21	0.72±0.21	**0.742**±0.21
	TU	0.672±0.23	0.667±0.23	0.669±0.24	0.711±0.22	0.711±0.23	**0.737**±0.23
Balance Scale	AU	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47
	EU	**0.68**±0.47	0.679±0.47	0.678±0.47	0.678±0.47	0.678±0.47	0.678±0.47
	TU	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47	**0.68**±0.47
Wisconsin Breast Cancer	AU	0.119±0.17	0.139±0.16	0.141±0.16	0.182±0.17	0.164±0.17	**0.189**±0.17
	EU	0.069±0.23	0.08±0.24	0.067±0.21	0.108±0.23	0.112±0.24	**0.131**±0.23
	TU	0.064±0.18	0.07±0.19	0.086±0.17	0.135±0.19	0.115±0.22	**0.154**±0.2
ERA	AU	-0.002±0.09	-0.018±0.09	-0.027±0.08	-0.018±0.13	**0.002**±0.13	-0.022±0.14
	EU	0.014±0.09	0.017±0.12	**0.021**±0.13	0.007±0.12	0.013±0.08	-0.013±0.07
	TU	-0.003±0.09	-0.019±0.08	-0.025±0.08	-0.017±0.13	**0.002**±0.13	-0.022±0.14
ESL	AU	**0.229**±0.25	0.218±0.24	0.208±0.23	0.195±0.23	0.225±0.24	0.192±0.23
	EU	0.266±0.24	**0.286**±0.22	0.273±0.23	0.28±0.23	0.275±0.25	**0.286**±0.23
	TU	**0.231**±0.25	0.229±0.24	0.214±0.24	0.204±0.23	0.23±0.24	0.196±0.23
Eucalyptus	AU	0.321±0.1	0.317±0.1	0.318±0.1	0.334±0.1	0.338±0.11	**0.351**±0.1
	EU	0.339±0.11	0.351±0.11	0.354±0.11	0.356±0.11	0.349±0.11	**0.357**±0.1
	TU	0.339±0.11	0.339±0.11	0.338±0.1	0.351±0.1	0.348±0.1	**0.36**±0.1
Heart (CLE)	AU	0.536±0.16	0.538±0.15	0.538±0.16	0.56±0.12	0.555±0.12	**0.581**±0.09
	EU	0.36±0.2	0.353±0.2	0.376±0.19	0.411±0.18	0.403±0.18	**0.443**±0.15
	TU	0.483±0.18	0.491±0.17	0.496±0.16	0.529±0.11	0.519±0.13	**0.544**±0.1
Boston Housing	AU	0.388±0.26	0.382±0.26	0.385±0.26	0.398±0.26	0.406±0.26	**0.414**±0.27
	EU	0.411±0.22	0.434±0.21	0.436±0.21	**0.451**±0.22	0.426±0.23	0.437±0.23
	TU	0.397±0.25	0.399±0.24	0.405±0.25	0.414±0.25	0.409±0.25	**0.418**±0.26
LEV	AU	0.183±0.09	0.185±0.11	0.185±0.12	0.184±0.12	0.189±0.11	**0.194**±0.12
	EU	0.206±0.1	0.198±0.11	0.197±0.1	0.195±0.1	**0.214**±0.08	0.207±0.07
	TU	0.185±0.09	0.185±0.11	0.185±0.12	0.184±0.12	0.191±0.11	**0.194**±0.12
Machine CPU	AU	0.544±0.3	0.586±0.29	0.613±0.28	0.687±0.22	0.671±0.22	**0.708**±0.21
	EU	0.552±0.2	0.597±0.13	**0.626**±0.14	0.613±0.14	0.595±0.18	0.599±0.19
	TU	0.559±0.29	0.608±0.27	0.626±0.25	0.673±0.21	0.657±0.21	**0.702**±0.21
New Thyroid	AU	0.476±0.5	0.473±0.5	0.48±0.51	**0.487**±0.51	**0.487**±0.51	**0.487**±0.51
	EU	**0.48**±0.51	0.476±0.5	0.476±0.5	0.473±0.5	0.476±0.5	0.473±0.5
	TU	0.476±0.5	0.476±0.5	**0.483**±0.51	0.48±0.51	0.48±0.51	0.48±0.51
Pyrimidines	AU	0.319±0.46	0.29±0.47	0.26±0.45	**0.366**±0.48	0.348±0.48	0.313±0.52
	EU	0.277±0.47	0.263±0.45	0.257±0.43	0.274±0.51	**0.326**±0.49	0.282±0.54
	TU	0.335±0.4	0.309±0.4	0.29±0.37	0.359±0.44	**0.367**±0.45	0.35±0.53
Red Wine	AU	0.306±0.12	0.296±0.11	0.294±0.11	0.306±0.12	0.319±0.12	**0.327**±0.12
	EU	0.407±0.12	0.42±0.12	0.428±0.13	**0.429**±0.13	0.417±0.12	0.416±0.13

Continued on next page

		PRR (MAE) (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
	TU	**0.376**±0.13	0.375±0.12	0.373±0.13	0.371±0.13	**0.376**±0.13	0.373±0.13
SWD	AU	0.121±0.1	0.119±0.1	0.124±0.09	0.135±0.1	0.134±0.1	**0.144**±0.1
	EU	0.143±0.13	0.123±0.13	0.136±0.13	0.125±0.11	0.144±0.12	**0.146**±0.11
	TU	0.122±0.1	0.12±0.1	0.122±0.09	0.136±0.1	0.136±0.1	**0.145**±0.1
Stocks Domain	AU	**0.721**±0.08	**0.721**±0.08	0.719±0.08	0.718±0.08	**0.721**±0.08	0.719±0.08
	EU	**0.679**±0.07	0.628±0.08	0.625±0.08	0.625±0.08	0.678±0.07	0.678±0.07
	TU	0.72±0.07	**0.721**±0.07	**0.721**±0.07	0.72±0.07	0.72±0.07	0.72±0.07
TAE	AU	0.234±0.35	0.251±0.36	0.251±0.36	0.277±0.4	**0.287**±0.39	0.281±0.4
	EU	0.241±0.36	0.143±0.34	0.191±0.35	0.15±0.36	0.233±0.37	**0.249**±0.41
	TU	0.253±0.36	0.266±0.38	0.28±0.39	0.294±0.43	**0.298**±0.43	0.271±0.45
Triazines	AU	0.264±0.26	0.269±0.25	0.27±0.25	0.308±0.25	0.3±0.25	**0.325**±0.24
	EU	0.291±0.24	0.313±0.23	0.32±0.23	**0.341**±0.23	0.328±0.24	0.339±0.23
	TU	0.274±0.23	0.273±0.22	0.272±0.21	0.31±0.23	0.307±0.24	**0.333**±0.22
White Wine	AU	0.235±0.06	0.233±0.06	0.236±0.06	0.25±0.06	0.257±0.06	**0.264**±0.06
	EU	0.354±0.03	0.349±0.03	0.353±0.04	0.353±0.05	**0.363**±0.04	0.356±0.05
	TU	**0.326**±0.05	0.32±0.05	0.314±0.05	0.311±0.05	0.323±0.05	0.31±0.05
CMC	AU	0.172±0.08	0.17±0.08	0.169±0.08	0.246±0.08	0.245±0.08	**0.278**±0.08
	EU	0.222±0.07	0.211±0.07	0.213±0.07	0.231±0.08	0.248±0.07	**0.267**±0.07
	TU	0.181±0.06	0.18±0.06	0.181±0.06	0.27±0.06	0.266±0.06	**0.289**±0.06
Grub Damage	AU	0.127±0.28	0.142±0.26	0.143±0.24	0.175±0.25	0.154±0.27	**0.202**±0.25
	EU	0.072±0.31	0.059±0.32	0.062±0.34	0.087±0.35	0.103±0.34	**0.12**±0.33
	TU	0.128±0.29	0.134±0.3	0.143±0.3	0.185±0.28	0.174±0.3	**0.205**±0.26
Obesity	AU	0.848±0.06	0.843±0.07	0.844±0.07	0.854±0.06	0.857±0.06	**0.865**±0.05
	EU	0.879±0.06	0.884±0.05	0.885±0.05	**0.889**±0.05	0.884±0.05	0.886±0.05
	TU	0.87±0.07	0.872±0.07	0.873±0.07	0.879±0.06	0.877±0.06	**0.881**±0.05

Appendix F Out-Of-Distribution (OOD) Detection - Detailed Results

In this section, we provide detailed results for the OOD detection experiment in Subsection 5.4, as well as additional OOD detection results using an ensemble of MLPs instead of GBTs.

F.1 Ensemble of GBTs - Detailed Results for OOD detection

Table F7 displays detailed results for OOD detection incldluing all uncertainty types. Notably, the best OOD performance is not in all cases achieved by measuring epistemic uncertainty. This is in stark contrast to the results obtained from the ensemble of MLPs, where, as expected, epistemic uncertainty very clearly outperforms TU as well as AU for OOD detection across all datasets (cf. Table F8). Presumably, this is due to the fact that decision trees, unlike neural networks, do not extrapolate their decision function to OOD regions, leading to high-confidence predictions for OOD data with low aleatoric uncertainty. In contrast, in the case of GBTs, OOD data will end up in mixed leaves, also leading to comparably higher aleatoric uncertainty (cf. Figure F17 for an illustration).

Table F7: OOD using an ensemble of GBTs (LGBM (Ke et al., 2017)).

		AUC-ROC (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	0.678±0.08	0.671±0.11	0.675±0.07	0.682±0.05	0.685±0.05	0.693±0.07
	EU	0.671±0.14	0.631±0.14	0.72±0.13	0.795±0.13	0.783±0.14	**0.818**±0.13
	TU	0.751±0.12	0.737±0.15	0.763±0.11	0.808±0.14	0.813±0.13	0.813±0.14
Auto MPG	AU	0.848±0.19	0.803±0.18	0.874±0.18	0.95±0.08	0.933±0.09	**0.967**±0.04
	EU	0.833±0.11	0.731±0.14	0.87±0.13	0.932±0.07	0.863±0.11	0.917±0.08
	TU	0.85±0.19	0.805±0.18	0.879±0.18	0.951±0.07	0.93±0.09	0.964±0.04
Automobile	AU	0.911±0.05	0.903±0.05	0.916±0.05	0.905±0.06	0.9±0.06	0.899±0.06
	EU	0.922±0.05	0.912±0.06	**0.939**±0.04	0.933±0.05	0.921±0.05	0.914±0.05
	TU	0.916±0.04	0.91±0.04	0.92±0.05	0.908±0.06	0.906±0.06	0.901±0.06
Balance Scale	AU	0.872±0.06	0.862±0.06	**0.877**±0.05	0.816±0.07	0.807±0.07	0.789±0.07
	EU	0.843±0.09	0.852±0.08	0.844±0.08	0.833±0.08	0.803±0.07	0.758±0.07
	TU	0.87±0.06	0.859±0.06	0.876±0.05	0.816±0.07	0.807±0.07	0.79±0.07
Boston Housing	AU	0.507±0.13	0.496±0.12	0.528±0.14	0.656±0.13	0.618±0.13	0.773±0.12
	EU	0.704±0.12	0.585±0.14	0.72±0.12	**0.786**±0.11	0.651±0.12	0.729±0.12
	TU	0.52±0.13	0.501±0.12	0.543±0.14	0.67±0.13	0.621±0.13	0.769±0.12
CMC	AU	0.363±0.07	0.362±0.07	0.365±0.06	0.545±0.07	0.54±0.07	0.628±0.06
	EU	0.667±0.09	0.582±0.1	0.675±0.08	**0.702**±0.1	0.647±0.09	0.695±0.09
	TU	0.371±0.07	0.368±0.07	0.375±0.06	0.555±0.08	0.55±0.08	0.634±0.07
ERA	AU	0.218±0.03	0.199±0.02	0.222±0.03	0.205±0.03	0.2±0.03	0.221±0.03
	EU	**0.992**±0.02	0.959±0.03	0.989±0.02	**0.992**±0.02	0.963±0.03	0.925±0.08
	TU	0.258±0.03	0.252±0.02	0.258±0.03	0.221±0.03	0.216±0.03	0.232±0.03
ESL	AU	0.078±0.06	0.084±0.07	0.08±0.06	0.1±0.06	0.103±0.06	0.118±0.07
	EU	0.596±0.2	0.325±0.21	0.58±0.21	**0.607**±0.2	0.357±0.22	0.377±0.23
	TU	0.088±0.07	0.086±0.07	0.087±0.07	0.11±0.07	0.107±0.07	0.128±0.08
Eucalyptus	AU	0.583±0.05	0.568±0.06	0.592±0.05	0.646±0.01	0.646±0.01	0.646±0.01
	EU	0.581±0.07	0.556±0.08	0.601±0.06	0.645±0.03	0.637±0.03	**0.653**±0.02
	TU	0.59±0.05	0.574±0.06	0.599±0.05	0.647±0.01	0.646±0.01	0.646±0.01
Grub Damage	AU	0.522±0.1	0.5±0.09	0.533±0.1	0.63±0.07	0.616±0.07	0.619±0.07
	EU	0.667±0.08	0.62±0.09	0.656±0.09	**0.712**±0.06	0.662±0.07	0.62±0.06
	TU	0.524±0.11	0.506±0.1	0.536±0.1	0.632±0.07	0.617±0.07	0.619±0.07
Heart (CLE)	AU	0.324±0.06	0.316±0.06	0.33±0.05	0.342±0.05	0.332±0.05	0.341±0.06
	EU	**0.405**±0.15	0.368±0.12	0.396±0.13	0.39±0.13	0.359±0.11	0.339±0.1
	TU	0.325±0.06	0.318±0.06	0.325±0.06	0.342±0.05	0.334±0.06	0.341±0.06
LEV	AU	0.129±0.03	0.151±0.04	0.117±0.03	0.12±0.03	0.139±0.04	0.133±0.04
	EU	**0.776**±0.13	0.65±0.09	0.727±0.14	0.727±0.14	0.635±0.09	0.629±0.09
	TU	0.145±0.04	0.16±0.05	0.127±0.03	0.133±0.04	0.147±0.05	0.145±0.05
Machine CPU	AU	0.309±0.13	0.297±0.12	0.33±0.13	0.621±0.14	0.553±0.17	**0.832**±0.07
	EU	0.432±0.15	0.338±0.13	0.459±0.14	0.726±0.13	0.471±0.15	0.701±0.14
	TU	0.316±0.13	0.297±0.12	0.333±0.12	0.626±0.14	0.553±0.17	0.829±0.07
New Thyroid	AU	0.966±0.03	0.957±0.04	**0.97**±0.03	0.954±0.03	0.95±0.04	0.945±0.04
	EU	0.946±0.05	0.946±0.04	0.953±0.04	0.937±0.05	0.942±0.05	0.94±0.04
	TU	0.964±0.03	0.956±0.04	0.966±0.03	0.954±0.03	0.949±0.04	0.947±0.04
Obesity	AU	0.614±0.02	0.617±0.02	0.614±0.02	0.601±0.02	0.607±0.02	0.576±0.02
	EU	**0.687**±0.04	0.66±0.03	0.682±0.04	0.679±0.04	0.661±0.03	0.664±0.03
	TU	0.616±0.02	0.617±0.02	0.615±0.02	0.603±0.02	0.607±0.02	0.576±0.02
Pyrimidines	AU	0.627±0.05	0.627±0.05	0.627±0.05	0.404±0.14	0.443±0.12	0.357±0.17
	EU	0.536±0.26	0.478±0.26	0.559±0.27	0.635±0.26	**0.657**±0.28	0.653±0.25
	TU	0.616±0.05	0.616±0.05	0.616±0.05	0.404±0.14	0.45±0.12	0.377±0.17
Red Wine	AU	0.755±0.15	0.716±0.13	0.793±0.15	0.917±0.09	0.885±0.11	0.962±0.03
	EU	0.95±0.03	0.878±0.08	0.965±0.03	**0.975**±0.02	0.929±0.06	0.961±0.03
	TU	0.829±0.13	0.777±0.13	0.854±0.13	0.936±0.08	0.907±0.1	0.967±0.03
SWD	AU	0.181±0.05	0.192±0.06	0.171±0.05	0.174±0.05	0.184±0.05	0.189±0.04
	EU	**0.648**±0.18	0.494±0.11	0.59±0.15	0.599±0.15	0.492±0.11	0.488±0.1
	TU	0.194±0.06	0.207±0.07	0.186±0.05	0.177±0.05	0.196±0.05	0.194±0.04
Stocks Domain	AU	0.747±0.05	0.744±0.05	0.752±0.05	0.781±0.05	0.772±0.05	0.806±0.05
	EU	0.818±0.1	0.78±0.08	0.826±0.09	**0.836**±0.09	0.79±0.07	0.807±0.08
	TU	0.752±0.06	0.747±0.06	0.759±0.06	0.784±0.06	0.773±0.06	0.807±0.05
TAE	AU	0.832±0.15	0.832±0.15	0.834±0.15	0.603±0.13	0.601±0.13	0.52±0.06
	EU	0.618±0.26	0.641±0.23	0.643±0.25	0.613±0.19	0.635±0.17	0.629±0.14
	TU	0.848±0.15	0.848±0.15	**0.85**±0.15	0.611±0.14	0.612±0.13	0.52±0.06
Triazines	AU	0.567±0.11	0.521±0.12	0.582±0.12	0.697±0.09	0.659±0.09	0.69±0.09
	EU	0.847±0.12	0.759±0.2	0.873±0.08	**0.896**±0.09	0.802±0.19	0.808±0.15
	TU	0.574±0.11	0.519±0.12	0.595±0.12	0.704±0.09	0.678±0.1	0.707±0.1

Continued on next page

		AUC-ROC (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
White Wine	AU	0.644±0.17	0.579±0.16	0.705±0.15	0.846±0.08	0.793±0.09	0.891±0.06
	EU	0.977±0.04	0.937±0.09	**0.986**±0.02	0.981±0.03	0.938±0.08	0.938±0.08
	TU	0.818±0.09	0.763±0.09	0.853±0.09	0.899±0.08	0.844±0.09	0.91±0.07
Wisconsin Breast Cancer	AU	0.873±0.06	0.857±0.06	**0.874**±0.06	0.818±0.08	0.811±0.08	0.805±0.08
	EU	0.335±0.2	0.346±0.2	0.381±0.19	0.381±0.21	0.444±0.23	0.524±0.22
	TU	0.852±0.06	0.839±0.07	0.854±0.07	0.801±0.08	0.797±0.08	0.79±0.1

F.2 Ensemble of MLPs - Detailed Results for OOD detection

In the case of an ensemble of MLPs, EU clearly outperforms TU and AU when it comes to OOD detection. The MLP will output overconfident predictions for OOD data with low to even no aleatoric uncertainty. Just like for GBTs (cf. Figure 11), entropy-based measures outperform variance-based measures (cf. Figure F16), and the proposed OCS decomposition method can be considered competitive with existing standard and labelwise decompositions, though not excelling at OOD.

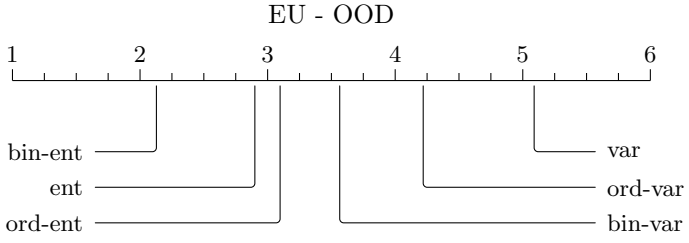
Table F8: OOD using an ensemble of MLPs.

		AUC-ROC (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
Abalone	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.8**±0.35	0.799±0.35	0.799±0.35	0.799±0.35	0.799±0.35	0.795±0.35
	TU	0.104±0.06	0.11±0.06	0.105±0.07	0.186±0.14	0.185±0.14	0.238±0.19
Auto MPG	AU	0.008±0.01	0.015±0.02	0.008±0.01	0.021±0.02	0.029±0.02	0.06±0.03
	EU	**1.0**±0.0	**1.0**±0.0	**1.0**±0.0	**1.0**±0.0	**1.0**±0.0	**1.0**±0.0
	TU	0.647±0.06	0.682±0.08	0.627±0.06	0.87±0.05	0.897±0.05	0.913±0.05
Automobile	AU	0.086±0.07	0.094±0.08	0.085±0.07	0.085±0.07	0.093±0.08	0.09±0.08
	EU	0.569±0.29	0.562±0.29	0.564±0.29	**0.57**±0.29	0.568±0.29	0.562±0.29
	TU	0.481±0.25	0.484±0.25	0.478±0.24	0.5±0.25	0.501±0.25	0.504±0.25
Balance Scale	AU	0.331±0.15	0.334±0.15	0.331±0.15	0.331±0.15	0.334±0.15	0.334±0.15
	EU	**0.573**±0.33	0.569±0.33	**0.573**±0.33	**0.573**±0.33	0.569±0.33	0.569±0.33
	TU	0.556±0.33	0.557±0.33	0.555±0.33	0.555±0.33	0.557±0.33	0.556±0.33
Boston Housing	AU	0.003±0.01	0.003±0.01	0.003±0.01	0.002±0.01	0.003±0.01	0.003±0.01
	EU	**0.731**±0.09	0.728±0.1	0.729±0.1	0.728±0.09	0.726±0.09	0.721±0.08
	TU	0.589±0.12	0.598±0.13	0.58±0.12	0.577±0.12	0.591±0.12	0.582±0.12
CMC	AU	0.001±0.0	0.001±0.0	0.001±0.0	0.001±0.0	0.001±0.0	0.001±0.0
	EU	**0.924**±0.2	0.915±0.2	0.921±0.2	0.921±0.2	0.91±0.2	0.898±0.2
	TU	0.553±0.15	0.548±0.15	0.558±0.15	0.552±0.17	0.544±0.16	0.58±0.18
ERA	AU	0.243±0.02	0.236±0.02	0.244±0.02	0.251±0.03	0.251±0.03	0.264±0.03
	EU	**0.926**±0.03	0.844±0.07	0.903±0.04	0.909±0.04	0.813±0.08	0.752±0.09
	TU	0.259±0.03	0.258±0.03	0.258±0.03	0.26±0.03	0.258±0.03	0.271±0.03
ESL	AU	0.009±0.01	0.012±0.01	0.009±0.01	0.013±0.01	0.017±0.01	0.025±0.01
	EU	0.397±0.28	0.389±0.28	0.396±0.28	**0.398**±0.28	0.389±0.28	0.388±0.28
	TU	0.121±0.11	0.128±0.11	0.114±0.1	0.193±0.18	0.19±0.17	0.25±0.24
Eucalyptus	AU	0.028±0.01	0.03±0.01	0.028±0.01	0.029±0.01	0.03±0.01	0.03±0.01
	EU	**1.0**±0.0	**1.0**±0.0	0.999±0.0	0.994±0.01	0.991±0.01	0.978±0.03
	TU	0.972±0.03	0.97±0.03	0.971±0.03	0.95±0.04	0.952±0.04	0.94±0.04
Grub Damage	AU	0.1±0.04	0.109±0.05	0.099±0.05	0.099±0.04	0.108±0.04	0.115±0.04
	EU	**0.934**±0.13	0.915±0.14	0.929±0.13	0.917±0.15	0.899±0.16	0.868±0.15
	TU	0.745±0.16	0.74±0.16	0.743±0.16	0.723±0.15	0.729±0.15	0.709±0.14
Heart (CLE)	AU	0.016±0.01	0.017±0.01	0.015±0.01	0.016±0.01	0.017±0.01	0.018±0.01
	EU	**0.777**±0.21	0.75±0.21	0.754±0.2	0.738±0.21	0.72±0.23	0.687±0.22
	TU	0.478±0.14	0.503±0.14	0.469±0.13	0.476±0.18	0.492±0.18	0.481±0.18
LEV	AU	0.012±0.0	0.013±0.0	0.011±0.0	0.01±0.0	0.012±0.0	0.011±0.0
	EU	**0.474**±0.23	0.447±0.26	0.469±0.23	0.469±0.23	0.446±0.26	0.445±0.26
	TU	0.035±0.03	0.048±0.03	0.026±0.02	0.024±0.02	0.042±0.03	0.03±0.03
Machine CPU	AU	0.076±0.05	0.083±0.05	0.076±0.05	0.073±0.05	0.081±0.05	0.074±0.04

Continued on next page

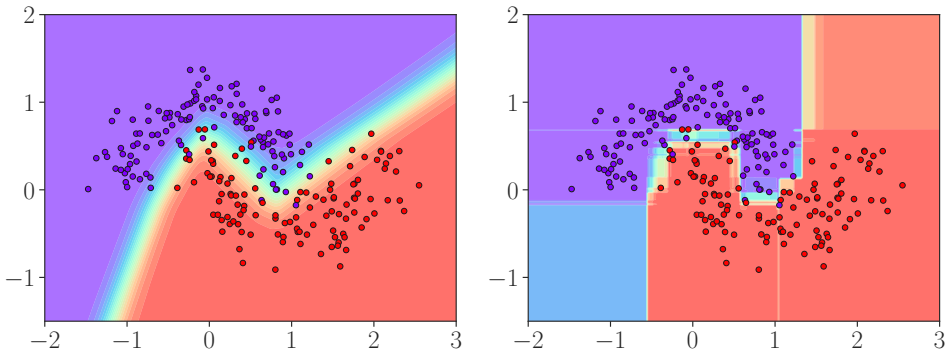
		AUC-ROC (↑)					
	Measure	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
Dataset	**Type**						
	EU	**0.237**±0.09	0.177±0.09	0.227±0.08	0.226±0.08	0.175±0.09	0.171±0.09
	TU	0.092±0.06	0.094±0.06	0.091±0.06	0.087±0.05	0.09±0.06	0.083±0.05
New Thyroid	AU	0.146±0.07	0.15±0.08	0.146±0.07	0.141±0.07	0.145±0.07	0.137±0.07
	EU	**0.175**±0.08	0.17±0.08	0.174±0.08	0.173±0.08	0.168±0.08	0.164±0.08
	TU	0.152±0.08	0.152±0.08	0.152±0.08	0.146±0.07	0.147±0.07	0.14±0.07
Obesity	AU	0.004±0.01	0.004±0.01	0.004±0.01	0.004±0.01	0.004±0.01	0.004±0.01
	EU	**0.92**±0.19	0.911±0.19	0.917±0.19	0.913±0.19	0.909±0.19	0.906±0.19
	TU	0.859±0.17	0.857±0.17	0.859±0.17	0.869±0.17	0.868±0.17	0.872±0.18
Pyrimidines	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.898**±0.22	0.888±0.22	0.888±0.22	0.846±0.2	0.846±0.2	0.814±0.2
	TU	0.692±0.23	0.66±0.21	0.692±0.23	0.715±0.24	0.715±0.25	0.691±0.26
Red Wine	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.998**±0.01	0.99±0.03	0.995±0.01	**0.998**±0.0	0.997±0.0	0.984±0.01
	TU	0.777±0.1	0.802±0.1	0.751±0.1	0.763±0.07	0.799±0.08	0.787±0.08
SWD	AU	0.004±0.0	0.005±0.0	0.004±0.0	0.004±0.0	0.005±0.0	0.005±0.0
	EU	**0.363**±0.3	0.352±0.3	0.361±0.3	0.361±0.3	0.352±0.3	0.351±0.3
	TU	0.044±0.04	0.047±0.04	0.041±0.03	0.04±0.03	0.043±0.04	0.042±0.04
Stocks Domain	AU	0.339±0.08	0.383±0.09	0.339±0.08	0.375±0.08	0.417±0.09	0.46±0.09
	EU	**0.977**±0.05	0.969±0.06	**0.977**±0.05	**0.977**±0.05	0.969±0.06	0.97±0.06
	TU	0.81±0.07	0.804±0.07	0.815±0.08	0.889±0.08	0.875±0.07	0.902±0.08
TAE	AU	0.018±0.02	0.022±0.02	0.018±0.02	0.017±0.02	0.019±0.02	0.017±0.02
	EU	**0.906**±0.17	0.892±0.18	**0.906**±0.17	**0.906**±0.17	0.888±0.18	0.873±0.17
	TU	0.47±0.2	0.475±0.21	0.461±0.2	0.445±0.2	0.455±0.21	0.445±0.21
Triazines	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.609**±0.31	0.594±0.3	0.604±0.31	0.604±0.32	0.581±0.3	0.568±0.3
	TU	0.505±0.27	0.504±0.27	0.509±0.28	0.48±0.26	0.475±0.25	0.47±0.25
White Wine	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.989**±0.01	0.972±0.04	0.977±0.03	0.982±0.01	0.975±0.02	0.951±0.03
	TU	0.489±0.14	0.509±0.13	0.478±0.15	0.605±0.14	0.617±0.13	0.648±0.12
Wisconsin Breast Cancer	AU	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
	EU	**0.907**±0.19	0.878±0.2	0.878±0.2	0.862±0.21	0.839±0.22	0.807±0.24
	TU	0.614±0.18	0.617±0.17	0.601±0.18	0.585±0.22	0.598±0.23	0.577±0.23

Fig. F16: CD diagram for OOD detection of the different uncertainty measures using an ensemble of MLPs.



Appendix G Class Imbalance Analysis

In this section, we examine common class imbalance scenarios in ordinal data and evaluate the previously discussed uncertainty quantification methods in these contexts. Specifically, we generate synthetic datasets that arguably represent the most prevalent imbalance distributions in ordinal data: the extreme bimodal distribution (D1),



(a) The MLP extrapolates smoothly, assigning high-confidence predictions even in OOD regions.

(b) The GBT does not extrapolate outside the training data and assigns more uncertain probabilities instead (closer to 0.5 in the binary case) (Chen & Guestrin, 2016).

Fig. F17: Illustration of different behaviors of MLPs and GBTs when it comes to OOD data. In the case of an MLP, OOD data will be predicted confidently with low aleatoric uncertainty. In contrast, OOD data will lead to high aleatoric uncertainty with GBTs, as GBTs will not extrapolate. If an OOD sample falls outside the learned partitions, it is forced into the nearest known leaf. This results in high aleatoric uncertainty, as the OOD sample may be assigned to a leaf that contains a mix of different labels, leading to a less confident prediction.

extreme right- and left-tailed distributions (D2 and D3), and the extreme unimodal distribution (D4) (see Figure G18).

We assume the existence of a latent standard normal continuous variable $z \sim \mathcal{N}(0,1)$ underlying both the ordinal target y and the features $\boldsymbol{x}$. Ordinal labels y are derived by thresholding z, resulting in ordered but not necessarily equally spaced categories, as the thresholds themselves are not equally spaced:

$$y = k \quad \text{if} \quad t_{k-1} < z \leq t_k \qquad \text{for } k = 1, \ldots, K,$$

where $t_0 = -\infty$, $t_K = +\infty$, and $t_1, \ldots, t_{K-1}$ are thresholds to be determined.

To induce class imbalances according to predefined proportions (see Table G9), we determine thresholds t_k such that

$$P(y = k) = \Phi(t_k) - \Phi(t_{k-1}) = p_k \qquad \text{for } k = 1, \ldots, K,$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution, and p_k is the desired proportion for class k. The exact threshold values can be calculated using the cumulative sums of p_k and the inverse CDF (quantile

function) of the normal distribution:

$$t_k = \Phi^{-1}\left(\sum_{j=1}^{k} p_j\right) \qquad \text{for } k = 1, \ldots, K-1,$$

with $t_0 = -\infty$ and $t_K = +\infty$. The threshold values for the four synthetic datasets are presented in Table G10.

Each feature x_j of the synthetic datasets is generated as a function of z and independent Gaussian noise:

$$\begin{aligned} x_1 &= z + \mathcal{N}(0, 1.0) \\ x_2 &= \sin(z) + \mathcal{N}(0, 1.0) \\ x_3 &= 0.8z + \mathcal{N}(0, 0.2) \\ x_4 &= 0.5z + \mathcal{N}(0, 0.3) \end{aligned}$$

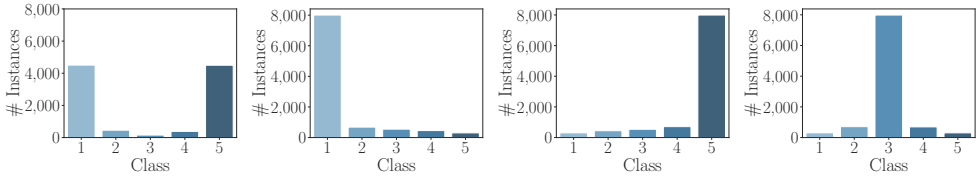
Table G9: Five-class ($K = 5$) distributions used to generate y for the synthetic datasets D1, D2, D3, and D4.

Dataset	$P(y = y_1)$	$P(y = y_2)$	$P(y = y_3)$	$P(y = y_4)$	$P(y = y_5)$
D1	0.45	0.045	0.015	0.04	0.45
D2	0.80	0.07	0.055	0.045	0.03
D3	0.03	0.045	0.055	0.07	0.80
D4	0.03	0.07	0.80	0.07	0.03

Table G10: Threshold values for the synthetic datasets D1, D2, D3, and D4 according to the prior distributions in Table G9

Dataset	t_0	t_1	t_2	t_3	t_4	t_5
D1	$-\infty$	-0.12566135	-0.01253347	0.02506891	0.12566135	$+\infty$
D2	$-\infty$	0.84162123	1.12639113	1.43953147	1.88079361	$+\infty$
D3	$-\infty$	-1.88079361	-1.43953147	-1.12639113	-0.84162123	$+\infty$
D4	$-\infty$	-1.88079361	-1.28155157	1.28155157	1.88079361	$+\infty$

We again make use of an ensemble of 10 GBTs (LightGBM) with a *subsample* rate of 0.5. Figure G19 shows the resulting confusion matrices for the four different datasets. Table G11 displays various performance metrics, including those specifically designed for imbalanced data, like *balanced* ACC (BACC), the *average* MAE (AMAE) (Baccianella et al., 2009) and *maximum* MAE (MMAE) (Cruz-Ramírez et



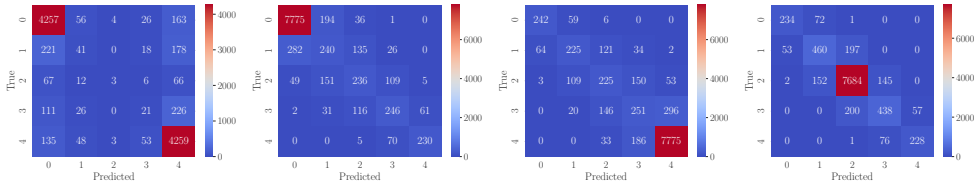
(a) Bimodal class imbalance dataset (D1).

(b) Right tailed class imbalance dataset (D2).

(c) Left tailed class imbalance dataset (D3).

(d) Unimodal class imbalance dataset (D4).

Fig. G18: Common synthetic class imbalances in ordinal data with $n = 10,000$ and generated according to the distributions in Table G9.



(a) Confusion matrix for D1 over 10-folds.

(b) Confusion matrix for D2 over 10-folds.

(c) Confusion matrix for D3 over 10-folds.

(d) Confusion matrix for D4 over 10-folds.

Fig. G19: Confusion matrices for the synthetic datasets.

Table G11: Predictive performance of the GBT ensemble on the four synthetic ordinal datasets.

Dataset	ACC	BACC	MAE	AMAE	MMAE	QWK	NLL	UMOD
D1	0.858±0.010	0.411±0.014	0.322±0.035	1.101±0.053	1.886±0.108	0.877±0.017	0.455±0.025	0.033±0.007
D2	0.873±0.009	0.609±0.038	0.143±0.012	0.438±0.045	0.722±0.063	0.911±0.012	0.319±0.033	0.951±0.008
D3	0.872±0.009	0.607±0.029	0.144±0.008	0.441±0.022	0.726±0.050	0.911±0.006	0.320±0.017	0.951±0.011
D4	0.904±0.009	0.750±0.027	0.096±0.009	0.251±0.027	0.386±0.044	0.871±0.012	0.243±0.027	0.999±0.001

al., 2014). Additionally, we include the degree of unimodality (UMOD) exhibited by the predictor's predictive probabilities, ranging from 0 (none) to 1 (fully unimodal).

To evaluate how the different uncertainty measures perform under class imbalance, we categorize the classes into three groups. For each dataset, the *head* consists of the most frequent class(es), the *tail* comprises the less frequent classes, and the *full* category includes all classes (see Table G12).

Subsequently, we calculate PRRs (Malinin, 2019) for the different class groups and separately for MCR and MAE.

Based on the results, we draw the following conclusions for the full category:

- For MCR (see Figure G20 and Table G13), all measures yield comparable PRRs, except for EU. In this case, ord-var and var achieve the best performance.

Table G12: Definition of head and tail class groupings for each synthetic five-class dataset ($K = 5$).

Dataset	Head	Tail	Full
D1	{1,5}	{2,3,4}	{1,2,3,4,5}
D2	{1}	{2,3,4,5}	{1,2,3,4,5}
D3	{5}	{1,2,3,4}	{1,2,3,4,5}
D4	{3}	{1,2,4,5}	{1,2,3,4,5}

- For MAE (see Figure G21 and Table G14), the differences in PRRs across datasets and uncertainty measures are more pronounced. For D1, D2, and D3, the binary reduction and variance measures provide a clear advantage for all uncertainty types. For D4, the differences are less distinct, which aligns with the assumption that, for centered unimodal distributions, accounting for distance becomes less important.

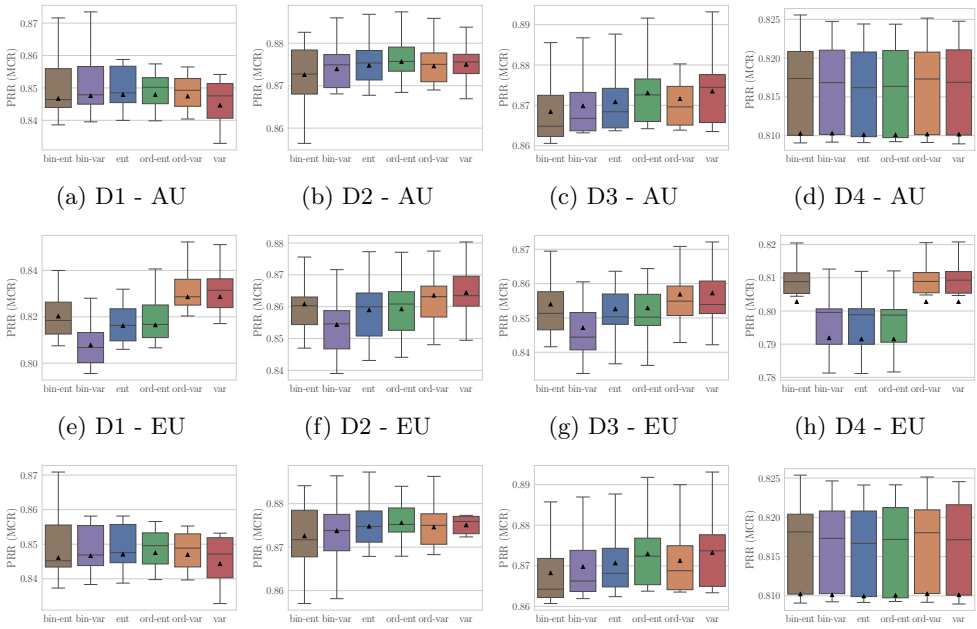
For the head category, containing the most frequent class(es), we draw the following conclusions:

- For MCR (see Figure G22 and Table G15), nominal measures perform slightly better or are at least comparable to ordinal measures.
- For MAE (see Figure G23 and Table G16), the results are similar, with ordinal measures showing a slight improvement and greater overlap between the methods.

For the tail category, which includes the less frequent classes, we draw the following conclusions:

- For MCR (see Figure G24 and Table G17), ordinal measures consistently outperform nominal measures in uncertainty quantification across all uncertainty types. The only exception is D4, where the differences are minimal or overlapping, again supporting the assumption that, for unimodal distributions, accounting for distance becomes less important.
- For MAE (see Figure G25 and Table G18), a similar pattern emerges: ordinal measures outperform nominal measures, with the only exception being D4.

Overall, our findings indicate that ordinal measures, including variance, which performed surprisingly well despite our efforts to avoid introducing equal distances between classes, offer superior uncertainty quantification in ordinal classification for underrepresented classes compared to traditional nominal measures. They remain competitive across all classes and are particularly effective when accounting for distance is important, for example when evaluating performance using the MAE. The strength of this effect also depends on the type of class imbalance. Ordinal measures demonstrate a clear advantage in bimodal, left-tailed, and right-tailed distributions. In contrast, for centered unimodal distributions, incorporating distance does not appear to offer significant benefits.



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU

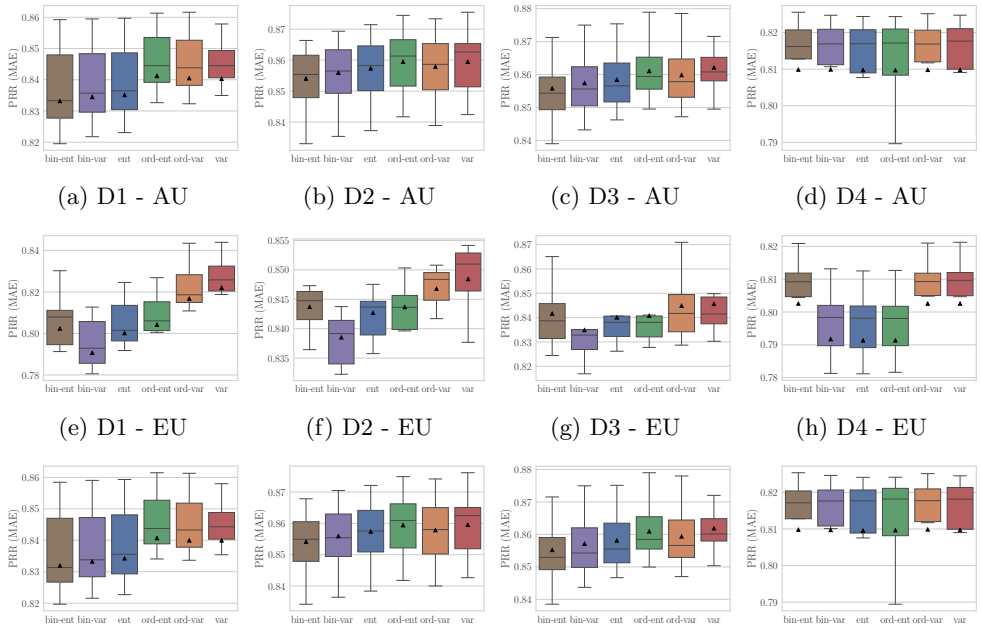
(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

Fig. G20: Full - PRR (MCR)

Table G13: Full - PRR (MCR)

Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	0.847 ±0.017	**0.848** ±0.018	**0.848** ±0.018	**0.848** ±0.017	0.847 ±0.017	0.845 ±0.017
	EU	0.82 ±0.017	0.808 ±0.019	0.816 ±0.019	0.816 ±0.015	**0.829** ±0.015	**0.829** ±0.017
	TU	0.846 ±0.017	**0.847** ±0.018	**0.847** ±0.018	**0.847** ±0.017	**0.847** ±0.017	0.844 ±0.017
D2	AU	0.873 ±0.008	0.874 ±0.008	0.875 ±0.008	**0.876** ±0.008	0.875 ±0.008	0.875 ±0.008
	EU	0.861 ±0.01	0.854 ±0.011	0.859 ±0.011	0.859 ±0.01	**0.864** ±0.01	**0.864** ±0.01
	TU	0.873 ±0.008	0.874 ±0.008	0.875 ±0.008	**0.876** ±0.008	0.875 ±0.008	0.875 ±0.008
D3	AU	0.868 ±0.009	0.87 ±0.008	0.871 ±0.008	**0.873** ±0.009	0.872 ±0.009	**0.873** ±0.009
	EU	0.854 ±0.011	0.847 ±0.011	0.853 ±0.01	0.853 ±0.011	**0.857** ±0.011	**0.857** ±0.011
	TU	0.868 ±0.009	0.87 ±0.008	0.871 ±0.008	**0.873** ±0.009	0.871 ±0.009	**0.873** ±0.009
D4	AU	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.019
	EU	**0.803** ±0.02	0.792 ±0.021	0.791 ±0.021	0.792 ±0.021	**0.803** ±0.02	**0.803** ±0.02
	TU	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.018	**0.81** ±0.019	**0.81** ±0.019



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU

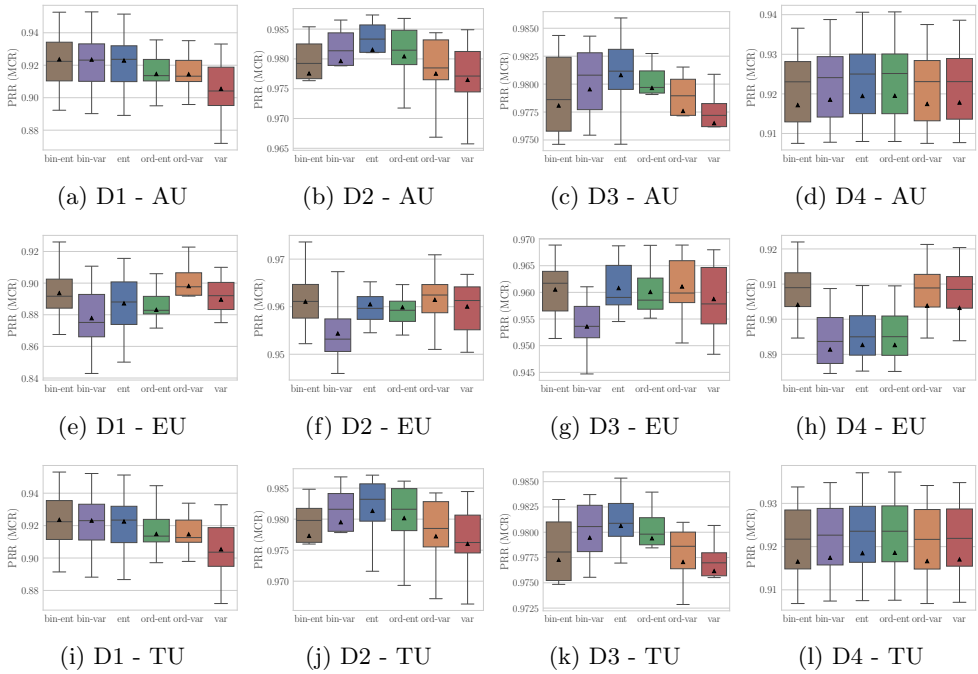
(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

Fig. G21: Full - PRR (MAE)

Table G14: Full - PRR (MAE)

Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	0.833 ±0.021	0.834 ±0.021	0.835 ±0.021	**0.841** ±0.021	0.84 ±0.021	0.84 ±0.02
	EU	0.802 ±0.021	0.791 ±0.022	0.8 ±0.022	0.804 ±0.022	**0.817** ±0.022	**0.822** ±0.022
	TU	0.832 ±0.021	0.833 ±0.022	0.834 ±0.022	**0.841** ±0.021	0.84 ±0.021	0.84 ±0.02
D2	AU	0.854 ±0.011	0.856 ±0.01	0.857 ±0.01	**0.859** ±0.01	0.858 ±0.01	**0.859** ±0.01
	EU	0.844 ±0.009	0.839 ±0.011	0.843 ±0.011	0.844 ±0.011	0.847 ±0.01	**0.848** ±0.01
	TU	0.854 ±0.01	0.856 ±0.01	0.857 ±0.01	0.859 ±0.01	0.858 ±0.01	**0.86** ±0.01
D3	AU	0.856 ±0.011	0.857 ±0.01	0.858 ±0.01	0.861 ±0.009	0.86 ±0.01	**0.862** ±0.009
	EU	0.842 ±0.015	0.835 ±0.015	0.84 ±0.013	0.841 ±0.013	0.845 ±0.014	**0.846** ±0.013
	TU	0.855 ±0.011	0.857 ±0.01	0.858 ±0.01	0.861 ±0.009	0.859 ±0.01	**0.862** ±0.009
D4	AU	**0.81** ±0.019	**0.81** ±0.019	**0.81** ±0.019	**0.81** ±0.019	**0.81** ±0.019	**0.81** ±0.02
	EU	**0.803** ±0.021	0.792 ±0.021	0.791 ±0.021	0.791 ±0.021	**0.803** ±0.021	**0.803** ±0.021
	TU	**0.81** ±0.02	**0.81** ±0.019	**0.81** ±0.019	**0.81** ±0.02	**0.81** ±0.02	**0.81** ±0.02



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU

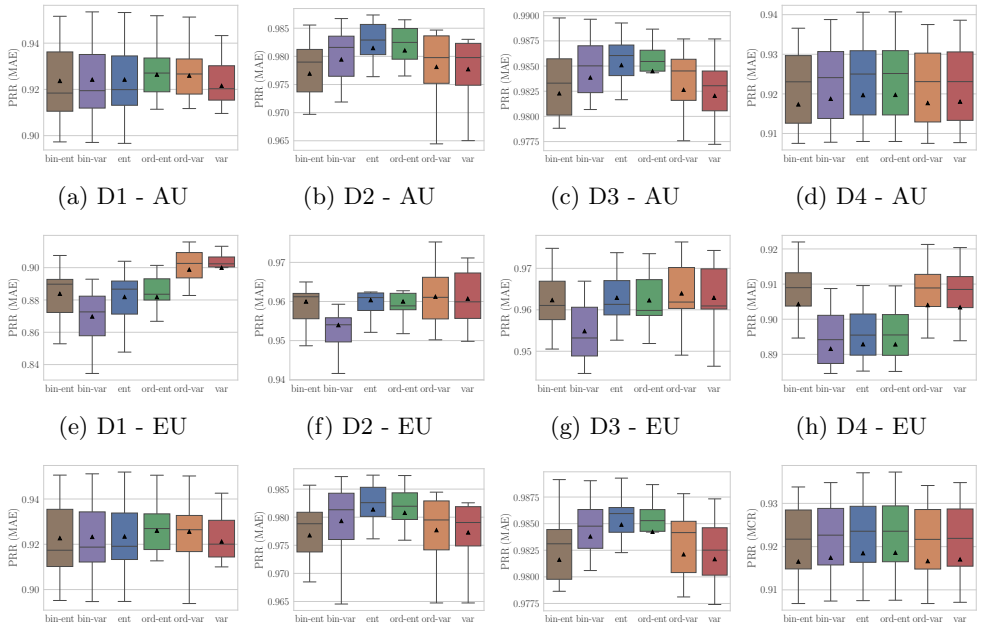
(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

Fig. G22: Head - PRR (MCR)

Table G15: Head - PRR (MCR)

Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	**0.924** ±0.019	0.923 ±0.019	0.923 ±0.019	0.915 ±0.019	0.914 ±0.018	0.905 ±0.019
	EU	0.894 ±0.018	0.878 ±0.022	0.887 ±0.02	0.883 ±0.017	**0.898** ±0.015	0.89 ±0.017
	TU	**0.924** ±0.019	0.923 ±0.019	0.923 ±0.02	0.915 ±0.018	0.915 ±0.018	0.905 ±0.019
D2	AU	0.978 ±0.007	0.98 ±0.007	**0.982** ±0.006	0.98 ±0.006	0.978 ±0.007	0.976 ±0.006
	EU	**0.961** ±0.007	0.954 ±0.006	0.96 ±0.005	0.96 ±0.004	**0.961** ±0.006	0.96 ±0.006
	TU	0.977 ±0.007	0.979 ±0.007	**0.981** ±0.006	0.98 ±0.006	0.977 ±0.007	0.976 ±0.006
D3	AU	0.978 ±0.006	0.98 ±0.005	**0.981** ±0.003	0.98 ±0.003	0.978 ±0.004	0.976 ±0.003
	EU	**0.961** ±0.006	0.954 ±0.005	**0.961** ±0.005	0.96 ±0.005	**0.961** ±0.006	0.959 ±0.007
	TU	0.977 ±0.006	0.979 ±0.004	**0.981** ±0.003	0.979 ±0.003	0.977 ±0.004	0.976 ±0.003
D4	AU	0.917 ±0.021	**0.919** ±0.021	**0.919** ±0.02	**0.919** ±0.02	0.917 ±0.021	0.918 ±0.021
	EU	**0.904** ±0.019	0.891 ±0.017	0.893 ±0.017	0.893 ±0.017	**0.904** ±0.018	0.903 ±0.019
	TU	0.917 ±0.021	0.917 ±0.021	0.918 ±0.02	**0.919** ±0.02	0.917 ±0.021	0.917 ±0.021



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU

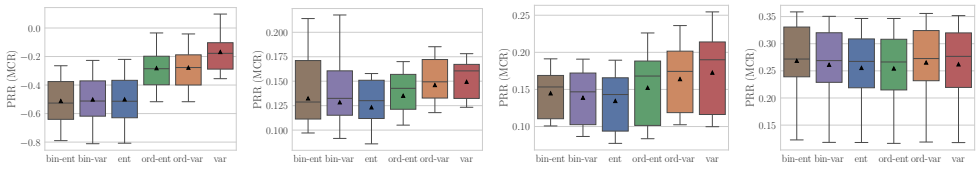
(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

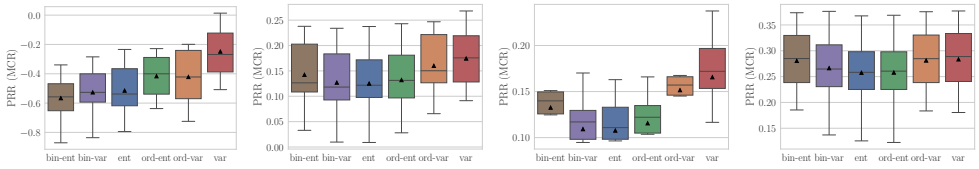
Fig. G23: Head - PRR (MAE)

Table G16: Head - PRR (MAE)

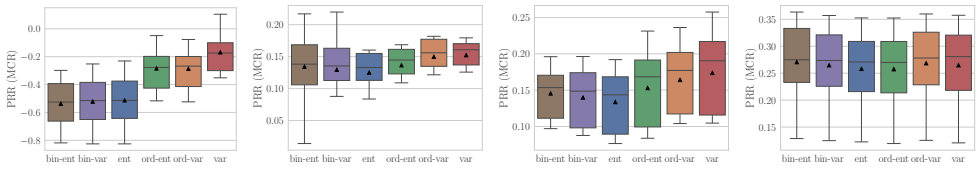
Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	0.924 ±0.019	0.924 ±0.019	0.924 ±0.019	**0.926** ±0.017	**0.926** ±0.017	0.922 ±0.016
	EU	0.884 ±0.016	0.87 ±0.018	0.882 ±0.017	0.882 ±0.017	0.899 ±0.016	**0.9** ±0.015
	TU	0.923 ±0.019	0.923 ±0.019	0.923 ±0.019	**0.926** ±0.017	**0.926** ±0.017	0.921 ±0.016
D2	AU	0.977 ±0.007	0.979 ±0.007	**0.981** ±0.006	**0.981** ±0.006	0.978 ±0.007	0.978 ±0.006
	EU	0.96 ±0.008	0.954 ±0.008	0.96 ±0.006	0.96 ±0.005	**0.961** ±0.008	**0.961** ±0.007
	TU	0.977 ±0.007	0.979 ±0.007	**0.981** ±0.006	**0.981** ±0.006	0.978 ±0.007	0.977 ±0.006
D3	AU	0.982 ±0.006	0.984 ±0.006	**0.985** ±0.004	0.984 ±0.004	0.983 ±0.005	0.982 ±0.004
	EU	0.962 ±0.008	0.955 ±0.008	0.963 ±0.007	0.962 ±0.007	**0.964** ±0.008	0.963 ±0.009
	TU	0.982 ±0.006	0.984 ±0.005	**0.985** ±0.004	0.984 ±0.004	0.982 ±0.006	0.982 ±0.004
D4	AU	0.917 ±0.021	0.919 ±0.021	**0.92** ±0.021	**0.92** ±0.021	0.918 ±0.021	0.918 ±0.021
	EU	**0.904** ±0.019	0.892 ±0.017	0.893 ±0.017	0.893 ±0.017	**0.904** ±0.019	0.903 ±0.019
	TU	0.917 ±0.021	0.918 ±0.021	**0.919** ±0.02	**0.919** ±0.02	0.917 ±0.021	0.917 ±0.021



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU



(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

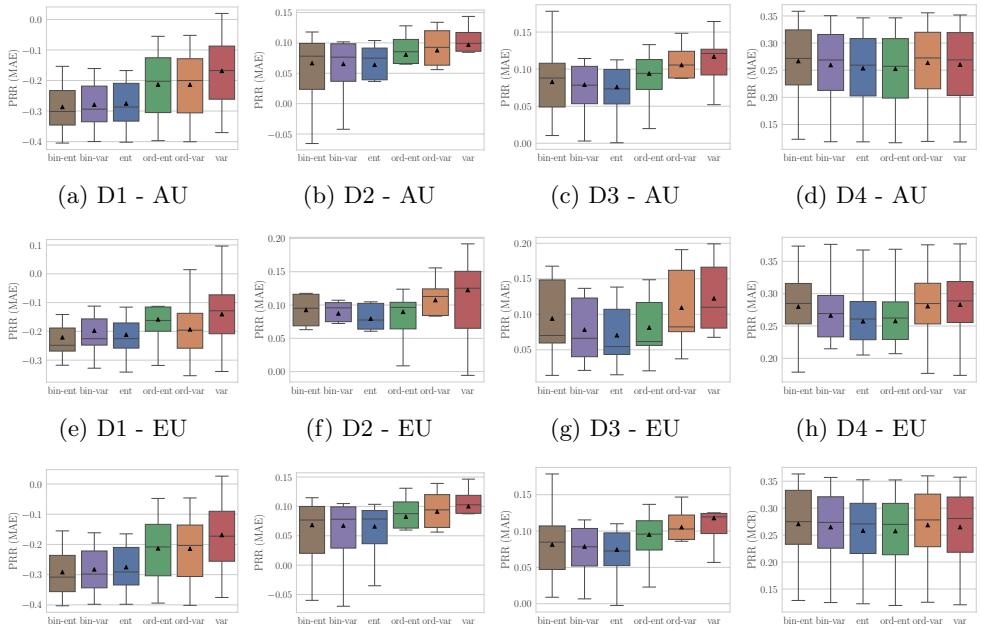


(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

Fig. G24: Tail - PRR (MCR)

Table G17: Tail - PRR (MCR)

Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	-0.512 ±0.185	-0.501 ±0.194	-0.501 ±0.195	-0.282 ±0.155	-0.278 ±0.152	**-0.168** ±0.146
	EU	-0.567 ±0.168	-0.527 ±0.169	-0.515 ±0.184	-0.417 ±0.153	-0.422 ±0.189	**-0.249** ±0.174
	TU	-0.538 ±0.181	-0.523 ±0.187	-0.513 ±0.193	-0.285 ±0.159	-0.287 ±0.151	**-0.169** ±0.15
D2	AU	0.132 ±0.058	0.128 ±0.06	0.123 ±0.061	0.135 ±0.06	0.146 ±0.061	**0.149** ±0.059
	EU	0.142 ±0.07	0.127 ±0.075	0.125 ±0.076	0.132 ±0.071	0.16 ±0.066	**0.174** ±0.06
	TU	0.134 ±0.057	0.129 ±0.06	0.125 ±0.061	0.136 ±0.06	0.15 ±0.058	**0.152** ±0.058
D3	AU	0.145 ±0.034	0.139 ±0.039	0.134 ±0.041	0.152 ±0.05	0.164 ±0.048	**0.173** ±0.056
	EU	0.133 ±0.046	0.109 ±0.045	0.108 ±0.045	0.116 ±0.048	0.151 ±0.05	**0.166** ±0.05
	TU	0.145 ±0.035	0.14 ±0.041	0.134 ±0.044	0.153 ±0.053	0.164 ±0.048	**0.174** ±0.056
D4	AU	**0.268** ±0.079	0.261 ±0.077	0.255 ±0.075	0.254 ±0.076	0.265 ±0.079	0.262 ±0.078
	EU	0.28 ±0.067	0.266 ±0.067	0.257 ±0.067	0.258 ±0.068	0.281 ±0.067	**0.283** ±0.067
	TU	0.27 ±0.077	0.265 ±0.076	0.258 ±0.075	0.257 ±0.075	**0.268** ±0.077	0.265 ±0.078



(a) D1 - AU (b) D2 - AU (c) D3 - AU (d) D4 - AU

(e) D1 - EU (f) D2 - EU (g) D3 - EU (h) D4 - EU

(i) D1 - TU (j) D2 - TU (k) D3 - TU (l) D4 - TU

Fig. G25: Tail - PRR (MAE)

Table G18: Tail - PRR (MAE)

Dataset	Measure Type	bin-ent (↑)	bin-var (↑)	ent (↑)	ord-ent (↑)	ord-var (↑)	var (↑)
D1	AU	-0.287 ±0.082	-0.279 ±0.079	-0.276 ±0.077	-0.214 ±0.111	-0.214 ±0.112	**-0.169** ±0.12
	EU	-0.221 ±0.092	-0.198 ±0.095	-0.212 ±0.088	-0.159 ±0.104	-0.194 ±0.107	**-0.141** ±0.124
	TU	-0.292 ±0.082	-0.283 ±0.078	-0.276 ±0.078	-0.213 ±0.11	-0.215 ±0.112	**-0.169** ±0.121
D2	AU	0.066 ±0.086	0.065 ±0.086	0.064 ±0.087	0.08 ±0.089	0.088 ±0.091	**0.097** ±0.088
	EU	0.092 ±0.081	0.087 ±0.085	0.079 ±0.09	0.09 ±0.088	0.107 ±0.083	**0.122** ±0.082
	TU	0.068 ±0.085	0.067 ±0.086	0.066 ±0.087	0.082 ±0.088	0.091 ±0.09	**0.1** ±0.088
D3	AU	0.083 ±0.048	0.079 ±0.054	0.076 ±0.055	0.094 ±0.061	0.105 ±0.061	**0.117** ±0.06
	EU	0.094 ±0.055	0.078 ±0.046	0.07 ±0.043	0.081 ±0.043	0.109 ±0.054	**0.122** ±0.049
	TU	0.081 ±0.049	0.079 ±0.053	0.074 ±0.055	0.095 ±0.06	0.105 ±0.059	**0.118** ±0.059
D4	AU	**0.267** ±0.076	0.259 ±0.075	0.253 ±0.074	0.252 ±0.074	0.263 ±0.076	0.26 ±0.077
	EU	0.28 ±0.065	0.266 ±0.065	0.257 ±0.066	0.258 ±0.066	0.28 ±0.065	**0.283** ±0.066
	TU	**0.269** ±0.075	0.263 ±0.074	0.256 ±0.073	0.255 ±0.074	0.266 ±0.075	0.263 ±0.077

References

Allwein, E., Schapire, R., Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *The Journal of Machine Learning Research*, *1*, 113–141,

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

Baccianella, S., Esuli, A., Sebastiani, F. (2009). Evaluation measures for ordinal regression. *Ninth international conference on intelligent systems design and applications, ISDA 2009, pisa, italy , november 30-december 2, 2009* (pp. 283–287). IEEE Computer Society. Retrieved from https://doi.org/10.1109/ISDA.2009.230

Benavoli, A., Corani, G., Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, *17*(1), 152–161,

Bérchez-Moreno, F., Ayllón-Gavilán, R., Vargas, V.M., Guijo-Rubio, D., Hervás-Martínez, C., Fernández, J.C., Gutiérrez, P.A. (2025). dlordinal: A python package for deep ordinal classification. *Neurocomputing*, 129305, https://doi.org/doi.org/10.1016/j.neucom.2024.129305

Bischl, B., Casalicchio, G., Das, T., Feurer, M., Fischer, S., Gijsbers, P., ... Wever, M. (2025). Openml: Insights from 10 years and more than a thousand papers. *Patterns*, *6*(7), 101317, https://doi.org/10.1016/J.PATTER.2025.101317 Retrieved from https://doi.org/10.1016/j.patter.2025.101317

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

Bülte, C., Sale, Y., Löhr, T., Hofman, P., Kutyniok, G., Hüllermeier, E. (2025). An axiomatic assessment of entropy-and variance-based uncertainty quantification in regression. *arXiv preprint arXiv:2504.18433*, ,

Cao, W., Mirjalili, V., Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognit. Lett.*, *140*, 325–331, https://doi.org/10.1016/J.PATREC.2020.11.008 Retrieved from https://doi.org/10.1016/j.patrec.2020.11.008

Chang, Y., Chang, K., Wu, G. (2018). Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput.*, *73*, 914–920, https://doi.org/10.1016/J.ASOC.2018.09.029 Retrieved from https://doi.org/10.1016/j.asoc.2018.09.029

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 785–794). ACM. Retrieved from https://doi.org/10.1145/2939672.2939785

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

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., Gutiérrez, P.A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing*, *135*, 21–31, https://doi.org/10.1016/J.NEUCOM.2013.05.058 Retrieved from https://doi.org/10.1016/j.neucom.2013.05.058

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

de Menezes e Silva Filho, T., Song, H., Perelló-Nieto, M., Santos-Rodríguez, R., Kull, M., Flach, P.A. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach. Learn.*, *112*(9), 3211–3260, https://doi.org/10.1007/S10994-023-06336-7 Retrieved from https://doi.org/10.1007/s10994-023-06336-7

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30,

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., Udluft, S. (2018, 10–15 Jul). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 1184–1193). PMLR.

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

Frank, E., & Hall, M.A. (2001). A simple approach to ordinal classification. *Machine learning: EMCL 2001, 12th european conference on machine learning, freiburg, germany, september 5-7, 2001, proceedings* (Vol. 2167, pp. 145–156). Springer. Retrieved from https://doi.org/10.1007/3-540-44795-4_13

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232,

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33nd international conference on machine learning, ICML 2016, new york city, ny, usa, june 19-24, 2016* (Vol. 48, pp. 1050–1059). JMLR.org.

Galdran, A. (2023). Performance metrics for probabilistic ordinal classifiers. *Medical image computing and computer assisted intervention - MICCAI 2023 - 26th international conference, vancouver, bc, canada, october 8-12, 2023, proceedings, part III* (Vol. 14222, pp. 357–366). Springer. Retrieved from https://doi.org/10.1007/978-3-031-43898-1_35

Gaudette, L., & Japkowicz, N. (2009). Evaluation methods for ordinal classification. *Advances in artificial intelligence, 22nd canadian conference on artificial intelligence, canadian AI 2009, kelowna, canada, may 25-27, 2009, proceedings* (Vol. 5549, pp. 207–210). Springer. Retrieved from https://doi.org/10.1007/978-3-642-01818-3_25

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 4878–4887).

Gneiting, T., & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378,

Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, neurips 2022, new orleans, la, usa, november 28 - december 9, 2022.*

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

Haas, S., & Hüllermeier, E. (2022). A prescriptive machine learning approach for assessing goodwill in the automotive domain. *Machine learning and knowledge discovery in databases - european conference, ECML PKDD 2022, grenoble, france, september 19-23, 2022, proceedings, part VI* (Vol. 13718, pp. 170–184). Springer. Retrieved from https://doi.org/10.1007/978-3-031-26422-1_11

Haas, S., & Hüllermeier, E. (2023). Rectifying bias in ordinal observational data using unimodal label smoothing. *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* (Vol. 14174, pp. 3–18). Springer. Retrieved from https://doi.org/10.1007/978-3-031-43427-3_1

Haas, S., & Hüllermeier, E. (2025a). Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests. *Int. J. Data Sci. Anal.*, *20*(3), 2061–2077, https://doi.org/10.1007/S41060-024-00573-2 Retrieved from https://doi.org/10.1007/s41060-024-00573-2

Haas, S., & Hüllermeier, E. (2025b). Uncertainty quantification in ordinal classification: A comparison of measures. *Int. J. Approx. Reason.*, *186*, 109479, https://doi.org/10.1016/J.IJAR.2025.109479 Retrieved from https://doi.org/10.1016/j.ijar.2025.109479

Hendrickx, K., Perini, L., der Plas, D.V., Meert, W., Davis, J. (2024). Machine learning with a reject option: a survey. *Mach. Learn.*, *113*(5), 3073–3110, https://doi.org/10.1007/S10994-024-06534-X Retrieved from https://doi.org/10.1007/s10994-024-06534-x

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.*

Hou, L., Yu, C., Samaras, D. (2016). Squared earth mover's distance-based loss for training deep neural networks. *CoRR*, *abs/1611.05916*, , 1611.05916

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

Huhn, J.C., & Hüllermeier, E. (2009). FR3: A fuzzy rule learner for inducing reliable classifiers. *IEEE Trans. Fuzzy Syst.*, *17*(1), 138–149, https://doi.org/10.1109/TFUZZ.2008.2005490 Retrieved from https://doi.org/10.1109/TFUZZ.2008.2005490

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, *110*(3), 457–506, https://doi.org/10.1007/S10994-021-05946-3 Retrieved from https://doi.org/10.1007/s10994-021-05946-3

Kasa, S.R., Goel, A., Gupta, K., Roychowdhury, S., Priyatam, P., Bhanushali, A., Murthy, P.S. (2024). Exploring ordinality in text classification: A comparative study of explicit and implicit techniques. *Findings of the association for computational linguistics, ACL 2024, bangkok, thailand and virtual meeting, august 11-16, 2024* (pp. 5390–5404). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2024.findings-acl.320

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 3146–3154).

Kelly, M., Longjohn, R., Nottingham, K. (2023). The uci machine learning repository. *URL https://archive. ics. uci. edu*, ,

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 5574–5584).

Lahoti, P., Gummadi, P.K., Weikum, G. (2023). Responsible model deployment via model-agnostic uncertainty learning. *Mach. Learn.*, *112*(3), 939–970, https://doi.org/10.1007/S10994-022-06248-Y Retrieved from https://doi.org/10.1007/s10994-022-06248-y

Lakshminarayanan, B., Pritzel, A., Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 6402–6413).

Li, L., & Lin, H. (2006). Ordinal regression by extended binary classification. *Advances in neural information processing systems 19, proceedings of the twentieth annual conference on neural information processing systems, vancouver, british columbia, canada, december 4-7, 2006* (pp. 865–872). MIT Press.

Li, Q., Wang, J., Yao, Z., Li, Y., Yang, P., Yan, J., ... Pu, S. (2022). Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2022, new orleans, la, usa, june 18-24, 2022* (pp. 20481–20490). IEEE. Retrieved from https://doi.org/10.1109/CVPR52688.2022.01986

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

Löhr, T., Ingrisch, M., Hüllermeier, E. (2024). Towards aleatoric and epistemic uncertainty in medical image classification. *Artificial intelligence in medicine - 22nd international conference, AIME 2024, salt lake city, ut, usa, july 9-12, 2024, proceedings, part II* (Vol. 14845, pp. 145–155). Springer. Retrieved from https://doi.org/10.1007/978-3-031-66535-6_17

Lu, S., Wang, Y., Sheng, L., He, L., Zheng, A., Liang, J. (2025, September). Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Comput. Surv.*, *58*(2), , https://doi.org/10.1145/3760390 Retrieved from https://doi.org/10.1145/3760390

Malinin, A. (2019). *Uncertainty estimation in deep learning with application to spoken language assessment* (PhD thesis). University of Cambridge.

Malinin, A., Prokhorenkova, L., Ustimenko, A. (2021). Uncertainty in gradient boosting via ensembles. *9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021.*

Mobiny, A., Nguyen, H.V., Moulik, S., Garg, N., Wu, C.C. (2021). Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, *11*(1), 5458,

Mucsányi, B., Kirchhof, M., Oh, S.J. (2024). Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in neural information processing systems 38: Annual conference on neural information processing systems 2024, neurips 2024, vancouver, bc, canada, december 10 - 15, 2024.* Retrieved from

http://papers.nips.cc/paper_files/paper/2024/hash/5afa9cb1e917b898ad418216dc726fbd-Abstract-Datasets_and_Benchmarks_Track.html

Nadeem, M.S.A., Zucker, J., Hanczar, B. (2010). Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *Proceedings of the third international workshop on machine learning in systems biology, MLSB 2009, ljubljana, slovenia, september 5-6, 2009* (Vol. 8, pp. 65–81). JMLR.org.

Nguyen, V., Shaker, M.H., Hüllermeier, E. (2022). How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn.*, *111*(1), 89–122, https://doi.org/10.1007/S10994-021-06003-9 Retrieved from https://doi.org/10.1007/s10994-021-06003-9

Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G. (2016). Ordinal regression with multiple output CNN for age estimation. *2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, las vegas, nv, usa, june 27-30, 2016* (pp. 4920–4928). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2016.532

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830,

Prokhorenkova, L.O., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, neurips 2018, december 3-8, 2018, montréal, canada* (pp. 6639–6649).

Rafique, R., Islam, S.R., Kazi, J.U. (2021). Machine learning in the prediction of cancer therapy. *Computational and Structural Biotechnology Journal*, *19*, 4003-4017, https://doi.org/https://doi.org/10.1016/j.csbj.2021.07.003

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

Saberi, N., Shaker, M.H., Duguay, C., Scott, K.A., Hüllermeier, E. (2024). Uncertainty estimation of lake ice cover maps from a random forest classifier using modis toa reflectance data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1-9, https://doi.org/10.1109/JSTARS.2024.3518306

Sale, Y., Hofman, P., Löhr, T., Wimmer, L., Nagler, T., Hüllermeier, E. (2024). Label-wise aleatoric and epistemic uncertainty quantification. *Uncertainty in artificial intelligence, 15-19 july 2024, universitat pompeu fabra, barcelona, spain* (Vol. 244, pp. 3159–3179). PMLR. Retrieved from https://proceedings.mlr.press/v244/sale24a.html

Senge, R., Bösner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inf. Sci.*, *255*, 16–29, https://doi.org/10.1016/J.INS.2013.07.030 Retrieved from https://doi.org/10.1016/j.ins.2013.07.030

Shaker, M.H., & Hüllermeier, E. (2020). Aleatoric and epistemic uncertainty with random forests. *Advances in intelligent data analysis XVIII - 18th international symposium on intelligent data analysis, IDA 2020, konstanz, germany, april 27-29, 2020, proceedings* (Vol. 12080, pp. 444–456). Springer. Retrieved from https://doi.org/10.1007/978-3-030-44584-3_35

Shaker, M.H., & Hüllermeier, E. (2021). Ensemble-based uncertainty quantification: Bayesian versus credal inference. *Proceedings 31. workshop computational intelligence* (Vol. 25, p. 63).

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423,

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Inf. Fusion*, *81*, 84–90, https://doi.org/10.1016/J.INFFUS.2021.11.011 Retrieved from https://doi.org/10.1016/j.inffus.2021.11.011

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., ... Nado, Z. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 13969–13980).

Uddin, N., Uddin Ahamed, M.K., Uddin, M.A., Islam, M.M., Talukder, M.A., Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering*, *4*, 327-339, https://doi.org/https://doi.org/10.1016/j.ijcce.2023.09.001

Vanschoren, J., Van Rijn, J.N., Bischl, B., Torgo, L. (2014). Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, *15*(2), 49–60,

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? *Uncertainty in artificial intelligence, UAI 2023, july 31 - 4 august 2023, pittsburgh, pa, USA* (Vol. 216, pp. 2282–2292). PMLR.

Yıldız, A.Y., & Kalayci, A. (2025). Gradient boosting decision trees on medical diagnosis over tabular data. *2025 ieee international conference on ai and data analytics (icad)* (pp. 1–8).

Yun, V.M.V., Gómez-Orellana, A.M., Guijo-Rubio, D., Bérchez-Moreno, F., Gutiérrez, P.A., Hervás-Martínez, C. (2024). Age estimation using soft labelling ordinal classification approaches. *Advances in artificial intelligence - 20th conference of the spanish association for artificial intelligence, CAEPIA 2024, A coruña, spain, june 19-21, 2024, proceedings* (Vol. 14640, pp. 40–49). Springer. Retrieved from https://doi.org/10.1007/978-3-031-62799-6_5

Zhu, R., Guo, Y., Xue, J. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognit. Lett.*, *133*, 217–223, https://doi.org/10.1016/J.PATREC.2020.03.004 Retrieved from https://doi.org/10.1016/j.patrec.2020.03.004

5 Conclusion, Limitations, and Future Research

This thesis has investigated the challenges associated with automating expert ratings using machine learning, specifically through behavioral cloning. In this learning paradigm, only observed expert decisions are available, whereas consequential outcome data is absent. As a result, it is not possible to predict ground-truth outcomes or estimate causal treatment effects. Instead, the objective is to mimic expert decision-making in a manner that is **consistent**, **auditable**, **uncertainty-aware**, **unbiased**, and **aligned with domain-specific requirements**. The key challenges in automating expert ratings through behavioral cloning include:

- **Human decision biases**, leading to imbalanced data and underutilization of the full rating scale.
- **High data uncertainty**, stemming from the inherently stochastic nature of human judgment.
- **Need for interpretability and auditability**, which is essential in high-stakes contexts to build trust in model decisions.
- **Adaptability of predictors**, requiring models that can learn "appropriate" prescriptive behaviors rather than merely replicating historical decisions.

Guided by a proposed framework specifically tailored to the automation of expert ratings (Section 3.3), this thesis introduces and evaluates several methodological building blocks to address these challenges in the real-world use case of automotive goodwill claim assessment:

- **Ordinal scale-aware cost-sensitive instance weighting** (Section 4.1) and **unimodal label smoothing** (Section 4.4) have proven effective in mitigating decision bias by reducing bias and overconfidence in decisions, while also considering the ordinal structure of rating data. Moreover, these approaches facilitate the integration of domain knowledge and better align learned predictors with business objectives. Resulting in prescriptive models that provide actionable outcomes in different scenarios (e.g., "How should we decide if the budget is limited or customer satisfaction is key?").

- **Explainable AI** has demonstrated its value in examining machine learning-based decisions and addressing organizational resistance to adoption, particularly through global and local feature attribution methods (Section 4.3). Moreover, XAI enables the alignment of expert decisions and organizational objectives by increasing transparency and revealing potential biases.
- **Novel uncertainty quantification methods for ordinal data** have shown substantial improvements over existing techniques for error detection by identifying good trade-offs between exact hit-rate and error distance reduction (Sections 4.5 and 4.6). These advancements help reduce operational risk by enabling selective classification strategies that identify confident and consistent ratings (Section 4.2), also turning a behavioral cloning model into a prescriptive one (e.g., "Can we confidently process this request automatically, or should we abstain and re-evaluate?").

These building blocks have been shown to serve as effective proxy signals, compensating for the absence of true outcome data.

Furthermore, this thesis offers both theoretical and practical contributions to the field of machine learning–based prescriptive analytics, particularly in the context of ordinal classification and uncertainty quantification, by serving as an antidote to the prevalent inductive bias of unimodality in ordinal classification. While this inductive bias may be justified for discretized continuous targets (e.g., age estimation from images), rating data does not necessarily conform to this assumption. The distinction made by Anderson [And84] between "grouped continuous" and "assessed" ordered categorical variables is therefore crucial when selecting the appropriate type of predictor and loss function, especially when uncertainty quantification is a key consideration. As demonstrated by this thesis, "assessed" ordered categorical variables, such as rating data, exhibit strong ties to the social sciences, where such variables are common, for instance, in the form of Likert-scale survey data. This connection makes measures of polarization or consensus [AR22] particularly valuable tools for uncertainty quantification in ordinal classification, and for rating data in particular, surpassing the effectiveness of commonly used uncertainty measures for nominal classification and regression. This thesis has demonstrated that, in ordinal classification applications requiring uncertainty-informed decision-making, it may be more effective to embrace potentially larger errors through unbiased uncertainty representation combined with proper ordinal-aware uncertainty measures, rather than suppressing them by imposing a strong inductive unimodality bias. Here, classic Bayesian decision theory, on the basis of calibrated multi-class predictive probabilities obtained through proper scoring rules and calibration, seems to be the most suitable choice. It has the potential to provide both unbiased uncertainty quantifica-

tion and distance-awareness by minimizing the expected loss of a distance-aware loss function, as detailed in Section 2.2.4.1. Since ordinal targets are prevalent in high-stakes settings such as medicine and finance, these findings have broad and significant implications.

Limitations Despite its contributions, this thesis has certain limitations. One key limitation is its focus on classical supervised machine learning in the form of behavioral cloning, which is not ideal for prescriptive settings [Ker+25]. In particular, for the goodwill claim assessment use case, only information about the input X and the treatment T or decision A, the goodwill contribution, is available, while the outcome Y resulting from the specific treatment, e.g., future service or vehicle purchases triggered by goodwill contributions, remains unknown. Consequently, machine learning models prescribe the most likely treatment or decision, which is treated as a proxy label for the outcome Y despite the lack of information about actual effects on outcomes, a situation deemed unsatisfactory [Fis+24; Ker+25].

This limitation stems from the current unavailability of outcome data for the considered goodwill claim assessment use case. Incorporating such data would enable a deeper understanding of contribution effects and facilitate the development of causal machine learning models better aligned with prescriptive objectives [Feu+24]. In this context, uplift modeling [GG17; DMV18; DBV21; ZLL22] represents a particularly promising practical approach. Its primary goal is to identify individuals who are most likely to respond positively to an intervention (e.g., a marketing campaign), such as remaining loyal to a brand or purchasing a new product. This capability facilitates the more effective allocation of financial resources. Formally, uplift is defined as the difference in outcome probabilities resulting from a treatment [DBV21]:

$$U(\boldsymbol{x}) := P(Y = 1 \mid \boldsymbol{x}, T = 1) - P(Y = 1 \mid \boldsymbol{x}, T = 0),$$

where $Y \in \{0, 1\}$ represents the outcome (e.g., whether a future vehicle purchase occurs), and $T \in \{0, 1\}$ the treatment (e.g., whether a goodwill contribution is made). Essentially, uplift modeling is a practical instantiation of CATE estimation, in which customers are subsequently categorized into four different groups—*sure things*, *lost causes*, *persuadables*, and *do-not-disturbs*. The goal of uplift modeling is to target *persuadables* only, enabling more efficient use of budget, which would be highly valuable in the context of goodwill claim assessment. The standard setting of uplift modeling, however, only considers binary treatments and outcomes. When dealing with rating data, it may be necessary to estimate ordinal treatment effects, as discussed by Chen et al. [Che+18], to identify the appropriate amount or dose of goodwill [Vos+24]. However, the use case of goodwill claim assessment also

illustrates that such outcome information is difficult to obtain in practical scenarios, as it might only become available in the distant future. In the considered example, this corresponds to a subsequent vehicle purchase occurring long after a specific goodwill contribution or treatment was made. Identifying a causal relationship between the contribution and the outcome over such a long time frame, while accounting for confounding external factors, appears to be highly challenging.

Nevertheless, it is debatable whether uplift modeling, or causal machine learning more broadly, is the most appropriate solution for the use case at hand. While such methods may lead to business-optimal decisions, they can also result in outcomes that are perceived as unfair or inconsistent. For instance, two customers experiencing the same vehicle issue might receive different treatments based on factors such as demographics, thereby violating the principle of consistent decision-making. Over time, these inconsistencies could damage the manufacturer's reputation and potentially outweigh any short-term business gains. A combination of behavioral cloning and uplift modeling could offer a more balanced approach by addressing both consistency and business outcome optimization. The choice of which method fits better could be determined on a case-by-case basis.

Another limitation of the approach outlined in this thesis is its exclusive focus on learning decision models purely from observational data. Although the investigated and proposed methods allow for some incorporation of domain knowledge and objectives, through cost-sensitive instance weighting (Section 4.1) and unimodal label smoothing (Section 4.4), this capability remains limited. To better integrate data-driven models with expert insights and nuanced contextual factors that may not be directly captured in the data, it is crucial to address the potential misalignment of the models with real-world decision-making requirements. In particular, the ability to inject ad-hoc decision requirements at a per-instance level could represent a valid and necessary extension.

Future Research An intriguing area for future research in prescriptive analytics, in particular when using behavioral cloning, is the integration of classical supervised machine learning models with foundation models such as *Large Language Models* (LLMs) [Teu+23]. LLMs have the potential to facilitate the combination of probabilistic, data-driven machine learning decisions, including quantified uncertainty and explanations, with user-specified goals and expert knowledge. Specifically, LLMs could act as a fusion and reasoning layer, integrating machine learning predictions with ad-hoc decision requirements specified by experts in user-friendly textual form [Bro+20], such as rule-based decision guardrails or constraints [BBM20; Mur+18]. This approach bridges the gap between traditional prescriptive tech-

nologies, such as expert rules, and data-driven models, leveraging the strengths of both paradigms. This aligns with the principles of neuro-symbolic AI, which aims to integrate symbolic reasoning with data-driven approaches [Yu+23; Bhu+24]. For instance, through the joint optimization of a weighted (λ) learning task loss ($\mathcal{L}_{\text{tsk}}$) and a knowledge base loss ($\mathcal{L}_{\text{kb}}$) [Zha+24a]:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{tsk}} \cdot \mathcal{L}_{\text{tsk}} + \lambda_{\text{kb}} \cdot \mathcal{L}_{\text{kb}}. \tag{5.1}$$

This formulation provides a principled way to balance predictive accuracy on observational data with adherence to organizational rules and constraints. Such an approach could lead to systems that are not only more interpretable, but also more robust and compliant in high-stakes applications.

Last but not least, ordinal classification remains a vital and intriguing area of research due to its prevalence in practical high-stakes applications and the unique challenges it presents. These challenges are often overlooked, with regression or nominal classification techniques being inappropriately applied instead. Moreover, the validity and necessity of the commonly assumed unimodality require further investigation and clarification, particularly with respect to identifying contexts where it may hinder model performance. This thesis has specifically questioned and invalidated this assumption, in the context of uncertainty quantification using probabilistic uncertainty representations. Additionally, the validity of the contiguity assumption in set-based ordinal uncertainty quantification warrants further theoretical and empirical validation. Future research should focus on developing holistic approaches that enhance ordinal classification performance without compromising the quality of uncertainty quantification. In particular, Bayesian decision theory leveraging calibrated probabilities and distance-aware expected loss minimization emerges as a suitable candidate to excel in balancing this trade-off.

References

[Abd+21] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Inf. Fusion* 76 (2021), pp. 243–297.

[AP23] Abhaya Abhaya and Bidyut Kr. Patra. “An efficient method for autoencoder based outlier detection”. In: *Expert Syst. Appl.* 213.Part (2023), p. 118904.

[AR22] Clem Aeppli and Didier Ruedin. *Let’s Measure Agreement, Consensus, and Polarization in Ordinal Data*. Oct. 2022.

[AEV24] Karl Akbari, Markus Eigruber, and Rudolf Vetschera. “Risk attitudes: The central tendency bias”. In: *EURO Journal on Decision Processes* 12 (2024), p. 100042.

[Akt+21] Shahriar Akter, Grace McCarthy, Shahriar Sajib, et al. “Algorithmic bias in data-driven innovation in the age of AI”. In: *Int. J. Inf. Manag.* 60 (2021), p. 102387.

[ACC21] Tomé Albuquerque, Ricardo Cruz, and Jaime S Cardoso. “Ordinal losses for classification of cervical cancer risk”. In: *PeerJ Computer Science* 7 (2021), e457.

[ACC22] Tomé Albuquerque, Ricardo Cruz, and Jaime S Cardoso. “Quasi-unimodal distributions for ordinal classification”. In: *Mathematics* 10.6 (2022), p. 980.

[Alf+24] Michael Alfertshofer, Joanna Kempa, Brian S Biesman, et al. “The “Central Tendency Bias” in the assessment of facial attractiveness in group-based and individual ratings—A survey-based study in 727 volunteers”. In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 92 (2024), pp. 264–275.

[Ali+23] Sajid Ali, Tamer Abuhmed, Shaker H. Ali El-Sappagh, et al. “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence”. In: *Inf. Fusion* 99 (2023), p. 101805.

[All+16] Sarah R Allred, L Elizabeth Crawford, Sean Duffy, and John Smith. “Working memory and spatial judgments: Cognitive load increases the central tendency bias”. In: *Psychonomic bulletin & review* 23 (2016), pp. 1825–1831.

[Ami+20] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. “Deep Evidential Regression”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[and23] Lena Enqvist and. “‘Human oversight’ in the EU artificial intelligence act: what, when and by whom?” In: *Law, Innovation and Technology* 15.2 (2023), pp. 508–535. eprint: `https://doi.org/10.1080/17579961.2023.2245683`.

[And84] John A Anderson. “Regression and ordered categorical variables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.1 (1984), pp. 1–22.

[Ang+22a] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. “Conformal risk control”. In: *arXiv preprint arXiv:2208.02814* (2022).

[AB23] Anastasios N. Angelopoulos and Stephen Bates. “Conformal Prediction: A Gentle Introduction”. In: *Found. Trends Mach. Learn.* 16.4 (2023), pp. 494–591.

[Ang+22b] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. “Machine bias”. In: *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 254–264.

[Ash85] Alison Hubbard Ashton. “Does consensus imply accuracy in accounting studies of decision making?” In: *Accounting Review* (1985), pp. 173–185.

[BES09] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. “Evaluation Measures for Ordinal Regression”. In: *Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009*. IEEE Computer Society, 2009, pp. 283–287.

[BHV14] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2014.

[Bar+21] Javier Barbero-Gómez, Pedro Antonio Gutiérrez, Víctor Manuel Vargas, Juan-Antonio Vallejo-Casas, and César Hervás-Martínez. “An ordinal CNN approach for the assessment of neurological damage in Parkinson’s disease patients”. In: *Expert Syst. Appl.* 182 (2021), p. 115271.

[BP17] 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*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 411–419.

[Ben08] Arie Ben-David. “Rule effectiveness in rule-based systems: A credit scoring case study”. In: *Expert Syst. Appl.* 34.4 (2008), pp. 2783–2788.

[BST09] Arie Ben-David, Leon Sterling, and TriDat Tran. “Adding monotonicity to learning algorithms may impair their accuracy”. In: *Expert Syst. Appl.* 36.3 (2009), pp. 6627–6634.

[Ben+09] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. “Curriculum learning”. In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*. Vol. 382. ACM International Conference Proceeding Series. ACM, 2009, pp. 41–48.

[Bér+25] Francisco Bérchez-Moreno, Rafael Ayllón-Gavilán, Víctor Manuel Vargas Yun, et al. “dlordinal: A Python package for deep ordinal classification”. In: *Neurocomputing* 622 (2025), p. 129305.

[BK20] Dimitris Bertsimas and Nathan Kallus. “From Predictive to Prescriptive Analytics”. In: *Manag. Sci.* 66.3 (2020), pp. 1025–1044.

[Bha+21] Umang Bhatt, Javier Antorán, Yunfeng Zhang, et al. “Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty”. In: *AIES ’21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. ACM, 2021, pp. 401–413.

[Bhu+24] Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and T. P. Singh. “Neuro-symbolic artificial intelligence: a survey”. In: *Neural Comput. Appl.* 36.21 (2024), pp. 12809–12844.

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[Bod+23] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, et al. “Benchmarking and survey of explanation methods for black box models”. In: *Data Mining and Knowledge Discovery* 37.5 (2023), pp. 1719–1778.

[Bol+16] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 2016, pp. 4349–4357.

[BBM20] Andrea Borghesi, Federico Baldo, and Michela Milano. “Improving deep learning models via constraint-based domain knowledge: a brief survey”. In: *arXiv preprint arXiv:2005.10691* (2020).

[BGM22] Leif Brandes, David Godes, and Dina Mayzlin. “Extremity bias in online reviews: The role of attrition”. In: *Journal of Marketing Research* 59.4 (2022), pp. 675–695.

[BUS95] Ivan Bratko, Tanja Urbančič, and Claude Sammut. “Behavioural Cloning: Phenomena, Results and Problems”. In: *IFAC Proceedings Volumes* 28.21 (1995). 5th IFAC Symposium on Automated Systems Based on Human Skill (Joint Design of Technology and Organisation), Berlin, Germany, 26-28 September, pp. 143–149.

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[Bri50] Glenn W Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[Bro+20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[CMR20] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. “Rank consistent ordinal regression for neural networks with application to age estimation”. In: *Pattern Recognit. Lett.* 140 (2020), pp. 325–331.

[CBM22] Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. “Estimating individual treatment effects using non-parametric regression models: A review”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.3 (2022), pp. 1115–1149.

[CMD22] 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*. International Committee on Computational Linguistics, 2022, pp. 4604–4609.

[Cha+02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *J. Artif. Intell. Res.* 16 (2002), pp. 321–357.

[Che+18] Jingxiang Chen, Haoda Fu, Xuanyao He, Michael R Kosorok, and Yufeng Liu. “Estimating individualized treatment rules for ordinal treatments”. In: *Biometrics* 74.3 (2018), pp. 924–933.

[CG16] 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*. ACM, 2016, pp. 785–794.

[Che+23] Valerie Chen, Umang Bhatt, Hoda Heidari, Adrian Weller, and Ameet Talwalkar. “Perspectives on incorporating expert feedback into model updates”. In: *Patterns* 4.7 (2023), p. 100780.

[CWP08] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. “A neural network approach to ordinal regression”. In: *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008*. IEEE, 2008, pp. 1279–1284.

[CW93] Robert T Clemen and Robert L Winkler. “Aggregating point estimates: A flexible modeling approach”. In: *Management Science* 39.4 (1993), pp. 501–515.

[CW99] Robert T Clemen and Robert L Winkler. “Combining probability distributions from experts in risk analysis”. In: *Risk analysis* 19.2 (1999), pp. 187–203.

[Cod+19] Felipe Codevilla, Eder Santana, Antonio M. López, and Adrien Gaidon. “Exploring the Limitations of Behavior Cloning for Autonomous Driving”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 9328–9337.

[Coh60] Jacob Cohen. “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[Coh68] Jacob Cohen. “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” In: *Psychological bulletin* 70.4 (1968), p. 213.

[CAC08] Joaquim F. Pinto da Costa, Hugo Alonso, and Jaime S. Cardoso. “The unimodal model for the classification of ordinal data”. In: *Neural Networks* 21.1 (2008), pp. 78–91.

[CC05] Joaquim F. Pinto da Costa and Jaime S. Cardoso. “Classification of Ordinal Data Using Neural Networks”. In: *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*. Vol. 3720. Lecture Notes in Computer Science. Springer, 2005, pp. 690–697.

[Cre+24] Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. “Conformal Prediction Sets Improve Human Decision Making”. In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[Cru+14] Manuel Cruz-Ramírez, César Hervás-Martínez, Javier Sánchez-Monedero, and Pedro Antonio Gutiérrez. “Metrics to guide a multi-objective evolutionary algorithm for ordinal classification”. In: *Neurocomputing* 135 (2014), pp. 21–31.

[DB20] Jessica Dai and Sarah M Brown. “Label bias, label shift: Fair machine learning with unreliable labels”. In: *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*. Vol. 12. 2020.

[Dan+20] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. “Multi-Objective Counterfactual Explanations”. In: *Parallel Problem Solving from Nature - PPSN XVI - 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*. Vol. 12269. Lecture Notes in Computer Science. Springer, 2020, pp. 448–469.

[DKM00] Elizabeth B Davis, S Jane Kennedy, and Laureen A Maines. “The relation between consensus and accuracy in low-to-moderate accuracy tasks: an auditing example”. In: *Auditing: A Journal of Practice & Theory* 19.1 (2000), pp. 101–121.

[Dep+18] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. “Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning”. In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1192–1201.

[DY14] Sébastien Destercke and Gen Yang. “Cautious Ordinal Classification by Binary Decomposition”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*. Vol. 8724. Lecture Notes in Computer Science. Springer, 2014, pp. 323–337.

[DBV21] Floris Devriendt, Jeroen Berrevoets, and Wouter Verbeke. “Why you should stop predicting customer churn and start using uplift models”. In: *Information Sciences* 548 (2021), pp. 497–515.

[DMV18] Floris Devriendt, Darie Moldovan, and Wouter Verbeke. “A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics”. In: *Big data* 6.1 (2018), pp. 13–41.

[DMK23] 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.

[DM19] 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*. Computer Vision Foundation / IEEE, 2019, pp. 4738–4747.

[Dua+20] Tony Duan, Anand Avati, Daisy Yi Ding, et al. “NGBoost: Natural Gradient Boosting for Probabilistic Prediction”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2690–2700.

[Dwi+23] Rudresh Dwivedi, Devam Dave, Het Naik, et al. “Explainable AI (XAI): Core Ideas, Techniques, and Solutions”. In: *ACM Comput. Surv.* 55.9 (2023), 194:1–194:33.

[EW10] Ran El-Yaniv and Yair Wiener. “On the Foundations of Noise-free Selective Classification”. In: *J. Mach. Learn. Res.* 11 (2010), pp. 1605–1641.

[Elk01] Charles Elkan. “The Foundations of Cost-Sensitive Learning”. In: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*. Morgan Kaufmann, 2001, pp. 973–978.

[Eps69] Edward S Epstein. “A scoring system for probability forecasts of ranked categories”. In: *Journal of Applied Meteorology (1962-1982)* 8.6 (1969), pp. 985–987.

[ER94] Joan-Maria Esteban and Debraj Ray. “On the measurement of polarization”. In: *Econometrica: Journal of the Econometric Society* (1994), pp. 819–851.

[Est+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. AAAI Press, 1996, pp. 226–231.

[Eva97] Christopher Evans. “The use of consensus methods and expert panels in pharmacoeconomic studies: practical applications and methodological shortcomings”. In: *Pharmacoeconomics* 12.2 (1997), pp. 121–129.

[Fav+23] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. “How to be fair? a study of label and selection bias”. In: *Machine Learning* 112.12 (2023), pp. 5081–5104.

[Fer+18a] Alberto Fernández, Salvador García, Mikel Galar, et al. “Cost-Sensitive Learning”. In: *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018, pp. 63–78.

[Fer+18b] Alberto Fernández, Salvador García, Mikel Galar, et al. *Learning from imbalanced data sets*. Vol. 10. 2018. Springer, 2018.

[FGH11] Alberto Fernández, Salvador García, and Francisco Herrera. “Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution”. In: *Hybrid Artificial Intelligent Systems - 6th International Conference, HAIS 2011, Wroclaw, Poland, May 23-25, 2011, Proceedings, Part I*. Vol. 6678. Lecture Notes in Computer Science. Springer, 2011, pp. 1–10.

[Feu+24] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, et al. “Causal machine learning for predicting treatment outcomes”. In: *Nature Medicine* 30.4 (2024), pp. 958–968.

[Fis+24] Unai Fischer-Abaigar, Christoph Kern, Noam Barda, and Frauke Kreuter. “Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector”. In: *Government Information Quarterly* 41.4 (2024), p. 101976.

[FH01] Eibe Frank and Mark A. Hall. “A Simple Approach to Ordinal Classification”. In: *Machine Learning: EMCL 2001, 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001, Proceedings*. Vol. 2167. Lecture Notes in Computer Science. Springer, 2001, pp. 145–156.

[FK04] Eibe Frank and Stefan Kramer. “Ensembles of nested dichotomies for multi-class problems”. In: *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*. Vol. 69. ACM International Conference Proceeding Series. ACM, 2004.

[Fra+19] Davide Frazzetto, Thomas Dyhre Nielsen, Torben Bach Pedersen, and Laurynas Siksnys. “Prescriptive analytics: a survey of emerging trends and technologies”. In: *VLDB J.* 28.4 (2019), pp. 575–595.

[FV14] Benoit Frenay and Michel Verleysen. “Classification in the Presence of Label Noise: A Survey”. In: *IEEE Trans. Neural Networks Learn. Syst.* 25.5 (2014), pp. 845–869.

[Fri01] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[Fri02] Jerome H Friedman. “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[GG16] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 1050–1059.

[Gal23] 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*. Vol. 14222. Lecture Notes in Computer Science. Springer, 2023, pp. 357–366.

[GJ09] 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*. Vol. 5549. Lecture Notes in Computer Science. Springer, 2009, pp. 207–210.

[GE17] Yonatan Geifman and Ran El-Yaniv. “Selective Classification for Deep Neural Networks”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017, pp. 4878–4887.

[Giu+14] Andrea Giustina, Philippe Chanson, David Kleinberg, et al. “Expert consensus document: a consensus on the medical treatment of acromegaly”. In: *Nature Reviews Endocrinology* 10.4 (2014), pp. 243–248.

[GR07] Tilmann Gneiting and Adrian E Raftery. “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[GF17] Bryce Goodman and Seth Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. In: *AI magazine* 38.3 (2017), pp. 50–57.

[Gre+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. “A Kernel Two-Sample Test”. In: *J. Mach. Learn. Res.* 13 (2012), pp. 723–773.

[GOV22] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. “Why do tree-based models still outperform deep learning on typical tabular data?” In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. 2022.

[Gro+11] Crina Grosan, Ajith Abraham, Crina Grosan, and Ajith Abraham. “Rule-based expert systems”. In: *Intelligent systems: A modern approach* (2011), pp. 149–185.

[Gui24] Riccardo Guidotti. “Counterfactual explanations and how to find them: literature review and benchmarking”. In: *Data Min. Knowl. Discov.* 38.5 (2024), pp. 2770–2824.

[Gui+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, et al. “Local rule-based explanations of black box decision systems”. In: *arXiv preprint arXiv:1805.10820* (2018).

[Guo+17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1321–1330.

[Gur+19] Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo A. Cecchi, and Charu C. Aggarwal. “Efficient Data Representation by Selecting Prototypes with Importance Weights”. In: *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*. IEEE, 2019, pp. 260–269.

[GG17] Pierre Gutierrez and Jean-Yves Gérardy. “Causal inference and uplift modelling: A review of the literature”. In: *International conference on predictive applications and APIs*. PMLR. 2017, pp. 1–13.

[Gut+16] Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro, and César Hervás-Martínez. “Ordinal Regression Methods: Survey and Experimental Study”. In: *IEEE Trans. Knowl. Data Eng.* 28.1 (2016), pp. 127–146.

[Har+16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. “Strategic Classification”. In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS ’16. Cambridge, Massachusetts, USA: Association for Computing Machinery, 2016, pp. 111–122.

[HW79] John A Hartigan and Manchek A Wong. “Algorithm AS 136: A k-means clustering algorithm”. In: *Journal of the royal statistical society. series c (applied statistics)* 28.1 (1979), pp. 100–108.

[Has+24] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, et al. “Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence”. In: *Cogn. Comput.* 16.1 (2024), pp. 45–74.

[HT87] Trevor Hastie and Robert Tibshirani. “Generalized additive models: some applications”. In: *Journal of the American Statistical Association* 82.398 (1987), pp. 371–386.

[He+08] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008*. IEEE, 2008, pp. 1322–1328.

[Hec79] James J Heckman. “Sample selection bias as a specification error”. In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.

[Heg+23] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, et al. “TabLLM: Few-shot Classification of Tabular Data with Large Language Models”. In: *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*. Vol. 206. Proceedings of Machine Learning Research. PMLR, 2023, pp. 5549–5581.

[Hen+24] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. “Machine learning with a reject option: a survey”. In: *Mach. Learn.* 113.5 (2024), pp. 3073–3110.

[HG17] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[Her+16] Francisco Herrera, Francisco Charte, Antonio J Rivera, and María J Del Jesus. “Multilabel classification”. In: *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Springer, 2016, pp. 17–31.

[Hoc87] Stephen J Hoch. “Perceived consensus and predictive accuracy: The pros and cons of projection.” In: *Journal of personality and social psychology* 53.2 (1987), p. 221.

[Hoh+18] Erik Hohmann, Jefferson C Brand, Michael J Rossi, and James H Lubowitz. *Expert opinion is necessary: Delphi panel methodology facilitates a scientific approach to consensus*. 2018.

[Hol86] Paul W Holland. “Statistics and causal inference”. In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[Hol+23] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. “TabPFN: A transformer that solves small tabular classification problems in a second”. In: *International Conference on Learning Representations 2023*. 2023.

[Hol+25] Noah Hollmann, Samuel Müller, Lennart Purucker, et al. “Accurate predictions on small data with a tabular foundation model”. In: *Nature* (Jan. 2025).

[HYS16] Le Hou, Chen-Ping Yu, and Dimitris Samaras. “Squared Earth Mover’s Distance-based Loss for Training Deep Neural Networks”. In: *CoRR* abs/1611.05916 (2016). arXiv: `1611.05916`.

[Hua+06] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. “Correcting Sample Selection Bias by Unlabeled Data”. In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. MIT Press, 2006, pp. 601–608.

[HH08] Jens C. Huhn and Eyke Hüllermeier. “Is an ordinal class structure useful in classifier learning?” In: *Int. J. Data Min. Model. Manag.* 1.1 (2008), pp. 45–67.

[Hül21] Eyke Hüllermeier. “Prescriptive Machine Learning for Automated Decision Making: Challenges and Opportunities”. In: *CoRR* abs/2112.08268 (2021). arXiv: 2112.08268.

[HC15] Eyke Hüllermeier and Weiwei Cheng. “Superset Learning Based on Generalized Loss Minimization”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II*. Vol. 9285. Lecture Notes in Computer Science. Springer, 2015, pp. 260–275.

[HDS22] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. “Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison”. In: *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*. Vol. 180. Proceedings of Machine Learning Research. PMLR, 2022, pp. 548–557.

[HW21] Eyke Hüllermeier and Willem Waegeman. “Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods”. In: *Mach. Learn.* 110.3 (2021), pp. 457–506.

[Hup22] Andrea C. Hupman. “Cutoff Threshold Decisions for Classification Algorithms with Risk Aversion”. In: *Decis. Anal.* 19.1 (2022), pp. 63–78.

[Hus+17] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. “Imitation learning: A survey of learning methods”. In: *ACM Computing Surveys (CSUR)* 50.2 (2017), pp. 1–35.

[JTB16] Silke Janitza, Gerhard Tutz, and Anne-Laure Boulesteix. “Random forest for ordinal responses: Prediction and variable selection”. In: *Comput. Stat. Data Anal.* 96 (2016), pp. 57–73.

[JN20] Heinrich Jiang and Ofir Nachum. “Identifying and Correcting Label Bias in Machine Learning”. In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 702–712.

[JSD18] Thorsten Joachims, Adith Swaminathan, and Maarten De Rijke. “Deep learning with logged bandit feedback”. In: *International Conference on Learning Representations*. 2018.

[Jor15] Anthony F Jorm. “Using the Delphi expert consensus method in mental health research”. In: *Australian & New Zealand Journal of Psychiatry* 49.10 (2015). PMID: 26296368, pp. 887–897. eprint: https://doi.org/10.1177/0004867415600891.

[JC96] Mohamed N Jouini and Robert T Clemen. “Copula models for aggregating expert opinions”. In: *Operations research* 44.3 (1996), pp. 444–457.

[Kam21] Margot E Kaminski. “The right to explanation, explained”. In: *Research handbook on information law and governance*. Edward Elgar Publishing, 2021, pp. 278–299.

[Kas+24] Siva Rajesh Kasa, Aniket Goel, Karan Gupta, et al. “Exploring Ordinality in Text Classification: A Comparative Study of Explicit and Implicit Techniques”. In: *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 5390–5404.

[KTL23] Yuko Kato, David M. J. Tax, and Marco Loog. “A Review of Nonconformity Measures for Conformal Prediction in Regression”. In: *Conformal and Probabilistic Prediction with Applications, 13-15 September 2023, Limassol, Cyprus*. Vol. 204. Proceedings of Machine Learning Research. PMLR, 2023, pp. 369–383.

[Ke+17] Guolin Ke, Qi Meng, Thomas Finley, et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017, pp. 3146–3154.

[KW89] Kevin Keasey and Robert Watson. “Consensus and accuracy in accounting studies of decision-making: A note on a new measure of consensus”. In: *Accounting, Organizations and Society* 14.4 (1989), pp. 337–345.

[KG71] Julian Keilson and Hans Gerber. “Some results for discrete unimodality”. In: *Journal of the American Statistical Association* 66.334 (1971), pp. 386–389.

[Ker+25] Christoph Kern, Unai Fischer-Abaigar, Jonas Schweisthal, et al. “Algorithms for reliable decision-making need causal reasoning”. In: *Nature Computational Science* (2025), pp. 1–5.

[KKK16] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. “Examples are not enough, learn to criticize! Criticism for Interpretability”. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 2016, pp. 2280–2288.

[KA12] Kyoung-Jae Kim and Hyunchul Ahn. “A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach”. In: *Comput. Oper. Res.* 39.8 (2012), pp. 1800–1811.

[KGV24] Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. “CARTE: Pretraining and Transfer for Tabular Learning”. In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[KL22] Ask Berstad Kolltveit and Jingyue Li. “Operationalizing machine learning models: a systematic literature review”. In: *Proceedings of the 1st Workshop on Software Engineering for Responsible AI, SE4RAI 2022, Pittsburgh, Pennsylvania, 19 May 2022*. ACM, 2022, pp. 1–8.

[KP04] Sotiris B. Kotsiantis and Panayiotis E. Pintelas. “A Cost Sensitive Technique for Ordinal Classification Problems”. In: *Methods and Applications of Artificial Intelligence, Third Helenic Conference on AI, SETN 2004, Samos, Greece, May 5-8, 2004, Proceedings*. Vol. 3025. Lecture Notes in Computer Science. Springer, 2004, pp. 220–229.

[Kra+01] Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer, and Michael de Groeve. “Prediction of Ordinal Classes Using Regression Trees”. In: *Fundam. Informaticae* 47.1-2 (2001), pp. 1–13.

[KKH23] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. “Machine Learning Operations (MLOps): Overview, Definition, and Architecture”. In: *IEEE Access* 11 (2023), pp. 31866–31879.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. 2009.

[Kru+25] Sven Kruschel, Nico Hambauer, Sven Weinzierl, et al. “Challenging the Performance-Interpretability Trade-off: An Evaluation of Interpretable Machine Learning Models”. In: *Business & Information Systems Engineering* (2025), pp. 1–25.

[KM22] Johnson Kuan and Jonas Mueller. “Back to the Basics: Revisiting Out-of-Distribution Detection Baselines”. In: *CoRR* abs/2207.03061 (2022). arXiv: 2207.03061.

[Kul+19] Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, et al. “Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration”. 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*. 2019, pp. 12295–12305.

[Kum+22] Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. “When Should We Prefer Offline Reinforcement Learning Over Behavioral Cloning?” In: *CoRR* abs/2204.05618 (2022). arXiv: 2204.05618.

[LPV18] Jordi de La Torre, Domenec Puig, and Aida Valls. “Weighted kappa loss function for multi-class classification of ordinal data in deep learning”. In: *Pattern Recognit. Lett.* 105 (2018), pp. 144–154.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017, pp. 6402–6413.

[LK77] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174.

[Lan08] Andreas Lanitis. “Comparative evaluation of automatic age-progression methodologies”. In: *EURASIP Journal on Applied Signal Processing* 2008 (2008).

[Lat+24] Paolo Latorre, Héctor A. López-Ospina, Sebastián Maldonado, C. Angelo Guevara, and Juan Pérez. “Designing employee benefits to optimize turnover: A prescriptive analytics approach”. In: *Comput. Ind. Eng.* 197 (2024), p. 110582.

[Lav+22] Alexander Lavin, Ciarán M Gilligan-Lee, Alessya Visnjic, et al. “Technology readiness levels for machine learning systems”. In: *Nature Communications* 13.1 (2022), p. 6039.

[LHA24] Adrien Le-Coz, Stéphane Herbin, and Faouzi Adjed. “Confidence Calibration of Classifiers with Many Classes”. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. 2024.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[Lei+22] Yiming Lei, Haiping Zhu, Junping Zhang, and Hongming Shan. “Meta Ordinal Regression Forest for Medical Image Classification With Ordinal Labels”. In: *IEEE CAA J. Autom. Sinica* 9.7 (2022), pp. 1233–1247.

[Lep+20] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas. “Prescriptive analytics: Literature review and research challenges”. In: *Int. J. Inf. Manag.* 50 (2020), pp. 57–70.

[Lev+20] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. “Offline reinforcement learning: Tutorial, review, and perspectives on open problems”. In: *arXiv preprint arXiv:2005.01643* (2020).

[LL06] Ling Li and Hsuan-Tien Lin. “Ordinal Regression by Extended Binary Classification”. In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. MIT Press, 2006, pp. 865–872.

[Li+22] Qiang Li, Jingjing Wang, Zhaoliang Yao, et al. “Unimodal-Concentrated Loss: Fully Adaptive Label Distribution Learning for Ordinal Regression”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 20481–20490.

[LH21a] Julian Lienen and Eyke Hüllermeier. “From Label Smoothing to Label Relaxation”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 8583–8591.

[LH21b] Julian Lienen and Eyke Hüllermeier. “Instance weighting through data imprecisiation”. In: *Int. J. Approx. Reason.* 134 (2021), pp. 1–14.

[Lik32] Rensis Likert. “A technique for the measurement of attitudes.” In: *Archives of psychology* (1932).

[Lin+22] Zhipeng Lin, Zhi Gao, Hong Ji, et al. “Classification of cervical cells leveraging simultaneous super-resolution and ordinal regression”. In: *Appl. Soft Comput.* 115 (2022), p. 108208.

[LS10] Charles X. Ling and Victor S. Sheng. “Cost-Sensitive Learning”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 231–235.

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest”. In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422.

[LD14] Li-Ping Liu and Thomas G. Dietterich. “Learnability of the Superset Label Learning Problem”. In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, 2014, pp. 1629–1637.

[Liu+20a] Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, et al. “Unimodal regularized neuron stick-breaking for ordinal classification”. In: *Neurocomputing* 388 (2020), pp. 34–44.

[Liu+19] Xiaofeng Liu, Xu Han, Yukai Qiao, et al. “Unimodal-Uniform Constrained Wasserstein Training for Medical Diagnosis”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 332–341.

[Liu+18] Xiaofeng Liu, Yang Zou, Yuhang Song, et al. “Ordinal Regression with Neuron Stick-Breaking for Medical Diagnosis”. In: *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part VI*. Vol. 11134. Lecture Notes in Computer Science. Springer, 2018, pp. 335–344.

[Liu+20b] Yang Liu, Qi Liu, Hongke Zhao, Zhen Pan, and Chuanren Liu. “Adaptive Quantitative Trading: An Imitative Deep Reinforcement Learning Approach”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.02 (Apr. 2020), pp. 2128–2135.

[Lou+13] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. “Accurate intelligible models with pairwise interactions”. In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. ACM, 2013, pp. 623–631.

[LAP22] 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*. Vol. 13438. Lecture Notes in Computer Science. Springer, 2022, pp. 545–554.

[Lu+18] Jie Lu, Anjin Liu, Fan Dong, et al. “Learning under concept drift: A review”. In: *IEEE transactions on knowledge and data engineering* 31.12 (2018), pp. 2346–2363.

[LEL18] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. “Consistent Individualized Feature Attribution for Tree Ensembles”. In: *CoRR* abs/1802.03888 (2018). arXiv: 1802.03888.

[LL17] Scott M. Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017, pp. 4765–4774.

[LA20] Abdoulaye O Ly and Moulay Akhloufi. “Learning to drive by imitation: An overview of deep behavior cloning methods”. In: *IEEE Transactions on Intelligent Vehicles* 6.2 (2020), pp. 195–209.

[MPU21] Andrey Malinin, Liudmila Prokhorenkova, and Aleksei Ustimenko. “Uncertainty in Gradient Boosting via Ensembles”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Mal+21] Nicholas Maltbie, Nan Niu, Matthew Van Doren, and Reese Johnson. “XAI tools in the public sector: a case study on predicting combined sewer overflows”. In: *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*. 2021, pp. 1032–1044.

[Man+20] Georgios Manthoulis, Michalis Doumpos, Constantin Zopounidis, and Emilios Galariotis. “An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for US banks”. In: *Eur. J. Oper. Res.* 282.2 (2020), pp. 786–801.

[MBD24] Marco Marozzi, Mario Bolzan, and Simone Di Zio. “Robust weighted aggregation of expert opinions in futures studies”. In: *Annals of Operations Research* 342.3 (2024), pp. 1471–1493.

[Mat+25] Anton Matsson, Yaochen Rao, Heather J. Litman, and Fredrik D. Johansson. *Pragmatic Policy Development via Interpretable Behavior Cloning*. 2025. arXiv: 2507.17056 [cs.LG].

[MW19] Jing Lei Mauricio Sadinle and Larry Wasserman. “Least Ambiguous Set-Valued Classifiers With Bounded Error Levels”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 223–234. eprint: https://doi.org/10.1080/01621459.2017.1395341.

[McA+21] Thomas McAndrew, Nutcha Wattanachit, Graham C Gibson, and Nicholas G Reich. “Aggregating predictions from experts: A review of statistical methods, experiments, and applications”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.2 (2021), e1514.

[McC80] Peter McCullagh. “Regression models for ordinal data”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[McE+23] Duncan C. McElfresh, Sujay Khandagale, Jonathan Valverde, et al. “When Do Neural Nets Outperform Boosted Trees on Tabular Data?” 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.

[McG+18] Rory McGrath, Luca Costabello, Chan Le Van, et al. “Interpretable Credit Application Predictions With Counterfactual Explanations”. In: *CoRR* abs/1811.05245 (2018). arXiv: 1811.05245.

[Meh+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[MH18] Vitalik Melnikov and Eyke Hüllermeier. “On the effectiveness of heuristics for learning nested dichotomies: an empirical analysis”. In: *Mach. Learn.* 107.8-10 (2018), pp. 1537–1560.

[Men+23] Telmo de Menezes e Silva Filho, Hao Song, Miquel Perelló-Nieto, et al. “Classifier calibration: a survey on how to assess and improve predicted class probabilities”. In: *Mach. Learn.* 112.9 (2023), pp. 3211–3260.

[Mob+19] Aryan Mobiny, Hien Van Nguyen, Supratik Moulik, Naveen Garg, and Carol C. Wu. “DropConnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks”. In: *CoRR* abs/1906.04569 (2019). arXiv: 1906.04569.

[MRA20] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results”. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020, pp. 243–248.

[Mol25] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 3rd ed. 2025.

[Mos+23] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. “Human-in-the-loop machine learning: a state of the art”. In: *Artif. Intell. Rev.* 56.4 (2023), pp. 3005–3054.

[MST20] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations”. In: *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. ACM, 2020, pp. 607–617.

[MKH19] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. “When does label smoothing help?” 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*. 2019, pp. 4696–4705.

[MMM99] Ákos Münnich, Gyula Maksa, and Robert J Mokken. “Collective judgement: combining individual value judgements”. In: *Mathematical Social Sciences* 37.3 (1999), pp. 211–233.

[Mur+18] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. “Incorporating prior domain knowledge into deep neural networks”. In: *2018 IEEE international conference on big data (big data)*. IEEE. 2018, pp. 36–45.

[Mur70] Allan H Murphy. “The ranked probability score and the probability score: A comparison”. In: *Monthly Weather Review* 98.12 (1970), pp. 917–924.

[MW70] Allan H Murphy and Robert L Winkler. “Scoring rules in probability assessment and evaluation”. In: *Acta psychologica* 34 (1970), pp. 273–286.

[Mur22] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[MB05] DN Prabhakar Murthy and Wallace R Blischke. *Warranty management and product manufacture*. Springer Science & Business Media, 2005.

[NZH10] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. “Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option”. In: *Proceedings of the third International Workshop on Machine Learning in Systems Biology, MLSB 2009, Ljubljana, Slovenia, September 5-6, 2009*. Vol. 8. JMLR Proceedings. JMLR.org, 2010, pp. 65–81.

[Nea12] Radford M Neal. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media, 2012.

[NSH22] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. “How to measure uncertainty in uncertainty sampling for active learning”. In: *Mach. Learn.* 111.1 (2022), pp. 89–122.

[NC05] Alexandru Niculescu-Mizil and Rich Caruana. “Predicting good probabilities with supervised learning”. In: *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*. Vol. 119. ACM International Conference Proceeding Series. ACM, 2005, pp. 625–632.

[Niu+16] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. “Ordinal Regression with Multiple Output CNN for Age Estimation”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4920–4928.

[NJC21] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. “Confident Learning: Estimating Uncertainty in Dataset Labels”. In: *J. Artif. Intell. Res.* 70 (2021), pp. 1373–1411.

[Nto+20] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, et al. “Bias in data-driven artificial intelligence systems—An introductory survey”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020), e1356.

[Pap08] Harris Papadopoulos. “Inductive Conformal Prediction: Theory and Application to Neural Networks”. In: *Tools in Artificial Intelligence*. Rijeka: IntechOpen, 2008. Chap. 18.

[Pap+02] 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*. Vol. 2430. Lecture Notes in Computer Science. Springer, 2002, pp. 345–356.

[PK19] Ellie Pavlick and Tom Kwiatkowski. "Inherent Disagreements in Human Textual Inferences". In: *Trans. Assoc. Comput. Linguistics* 7 (2019), pp. 677–694.

[Pes+20] Dana Pessach, Gonen Singer, Dan Avrahami, et al. "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming". In: *Decis. Support Syst.* 134 (2020), p. 113290.

[PH90] Bercedis Peterson and Frank E Harrell Jr. "Partial proportional odds models for ordinal response variables". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 39.2 (1990), pp. 205–217.

[PÇT25] Gorkem Polat, Ümit Mert Çaglar, and Alptekin Temizel. "Class distance weighted cross entropy loss for classification of disease severity". In: *Expert Syst. Appl.* 269 (2025), p. 126372.

[Pol+22] Gorkem Polat, Ilkay Ergenc, Haluk Tarik Kani, et al. "Class Distance Weighted Cross-Entropy Loss for Ulcerative Colitis Severity Estimation". In: *Medical Image Understanding and Analysis - 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27-29, 2022, Proceedings*. Vol. 13413. Lecture Notes in Computer Science. Springer, 2022, pp. 157–171.

[Pro+18] 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*. 2018, pp. 6639–6649.

[PSM20] Sushil Punia, Surya Prakash Singh, and Jitendra K. Madaan. "From predictive to prescriptive analytics: A data-driven multi-item newsvendor model". In: *Decis. Support Syst.* 136 (2020), p. 113340.

[Qu+25] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. "TabICL: A Tabular Foundation Model for In-Context Learning on Large Data". In: *CoRR* abs/2502.05564 (2025). arXiv: 2502.05564.

[RK87] LKPJ Rdusseeun and P Kaufman. "Clustering by means of medoids". In: *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*. Vol. 31. 1987, p. 28.

[RSG16] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. ACM, 2016, pp. 1135–1144.

[RSG18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2018, pp. 1527–1535.

[Riv+23] Marcos Rivera-Gavilán, Víctor Manuel Vargas, Pedro Antonio Gutiérrez, et al. "Ordinal Classification Approach for Donor-Recipient Matching in Liver Transplantation with Circulatory Death Donors". In: *Advances in Computational Intelligence - 17th International Work-Conference on Artificial Neural Networks, IWANN 2023, Ponta Delgada, Portugal, June 19-21, 2023, Proceedings, Part II*. Vol. 14135. Lecture Notes in Computer Science. Springer, 2023, pp. 517–528.

[RPR13] Filipe Rodrigues, Francisco C. Pereira, and Bernardete Ribeiro. "Learning from multiple annotators: Distinguishing good from random labelers". In: *Pattern Recognit. Lett.* 34.12 (2013), pp. 1428–1436.

[RSC20] 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.

[Ros+22] Riccardo Rosati, Luca Romeo, Víctor Manuel Vargas, et al. "A novel deep ordinal classification approach for aesthetic quality control classification". In: *Neural Comput. Appl.* 34.14 (2022), pp. 11625–11639.

[SAR21] Maria Sahakyan, Zeyar Aung, and Talal Rahwan. "Explainable artificial intelligence for tabular data: A survey". In: *IEEE access* 9 (2021), pp. 135392–135422.

[Sak23] Otmane Sakhi. "Offline Contextual Bandit: Theory and Large Scale Applications". PhD thesis. Institut Polytechnique de Paris, 2023.

[SCH23] Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. "Is the volume of a credal set a good measure for epistemic uncertainty?" In: *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*. Vol. 216. Proceedings of Machine Learning Research. PMLR, 2023, pp. 1795–1804.

[Sal+24] Yusuf Sale, Paul Hofman, Timo Löhr, et al. "Label-wise Aleatoric and Epistemic Uncertainty Quantification". In: *Uncertainty in Artificial Intelligence, 15-19 July 2024, Universitat Pompeu Fabra, Barcelona, Spain*. Vol. 244. Proceedings of Machine Learning Research. PMLR, 2024, pp. 3159–3179.

[Sam11] Caude Sammut. "Behavioral cloning". In: *Encyclopedia of machine learning*. Springer, 2011, pp. 93–97.

[Sen+14] Robin Senge, Stefan Bösner, Krzysztof Dembczynski, et al. "Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty". In: *Inf. Sci.* 255 (2014), pp. 16–29.

[SKK18] Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. “Evidential Deep Learning to Quantify Classification Uncertainty”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 2018, pp. 3183–3193.

[SH20] Mohammad Hossein Shaker and Eyke Hüllermeier. “Aleatoric and Epistemic Uncertainty with Random Forests”. In: *Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27-29, 2020, Proceedings*. Vol. 12080. Lecture Notes in Computer Science. Springer, 2020, pp. 444–456.

[SH21] Mohammad Hossein Shaker and Eyke Hüllermeier. “Ensemble-based uncertainty quantification: Bayesian versus credal inference”. In: *Proceedings 31. Workshop Computational Intelligence*. Vol. 25. 2021, p. 63.

[Sha48] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[Sha+53] Lloyd S Shapley et al. “A value for n-person games”. In: (1953).

[Shi+24] Jongmin Shin, Jonghyeon Won, Hyun-Suk Lee, and Jang-Won Lee. “A review on label cleaning techniques for learning with noisy labels”. In: *ICT Express* 10.6 (2024), pp. 1315–1330.

[SA22] Ravid Shwartz-Ziv and Amitai Armon. “Tabular data: Deep learning is not all you need”. In: *Inf. Fusion* 81 (2022), pp. 84–90.

[ŠP18] Laurynas Šikšnys and Torben Bach Pedersen. “Prescriptive Analytics”. In: *Encyclopedia of Database Systems*. New York, NY: Springer New York, 2018, pp. 2792–2793.

[Smi20] Abir Smiti. “A critical overview of outlier detection methods”. In: *Comput. Sci. Rev.* 38 (2020), p. 100306.

[SH89] Janet A Sniezek and Rebecca A Henry. “Accuracy and confidence in group judgment”. In: *Organizational behavior and human decision processes* 43.1 (1989), pp. 1–28.

[Sol+22] Efrain Solares, Víctor De-León-Gómez, Eduardo Fernández, Emmanuel Contreras-Medina, and Orlando Lopez. “Multicriteria ordinal classification to improve strategic planning in the financial sector of the company”. In: *Int. J. Comb. Optim. Probl. Informatics* 13.2 (2022), pp. 47–57.

[Son+23] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. “Learning From Noisy Labels With Deep Neural Networks: A Survey”. In: *IEEE Trans. Neural Networks Learn. Syst.* 34.11 (2023), pp. 8135–8153.

[Sri+14] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1929–1958.

[Ste+24a] **Stefan Haas**, Konstantin Hegestweiler, Michael Rapp, Maximilian Muschalik, and Eyke Hüllermeier. “Stakeholder-centric explanations for black-box decisions: an XAI process model and its application to automotive goodwill assessments”. In: *Frontiers in Artificial Intelligence - AI in Business* 7 (2024).

[SH22] **Stefan Haas** and Eyke Hüllermeier. “A Prescriptive Machine Learning Approach for Assessing Goodwill in the Automotive Domain”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part VI*. Vol. 13718. Lecture Notes in Computer Science. Springer, 2022, pp. 170–184.

[SH26] **Stefan Haas** and Eyke Hüllermeier. “Aleatoric and Epistemic Uncertainty Measures for Ordinal Classification through Binary Reduction”. In: *Machine Learning* (2026).

[SH24] **Stefan Haas** and Eyke Hüllermeier. “Conformalized prescriptive machine learning for uncertainty-aware automated decision making: the case of goodwill requests”. In: *International Journal of Data Science and Analytics* (2024).

[SH23] **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*. Vol. 14174. Lecture Notes in Computer Science. Springer, 2023, pp. 3–18.

[SH25] **Stefan Haas** and Eyke Hüllermeier. “Uncertainty quantification in ordinal classification: A comparison of measures”. In: *Int. J. Approx. Reason.* 186 (2025), p. 109479.

[Ste+24b] Sarah Sterz, Kevin Baum, Sebastian Biewer, et al. “On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives”. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*. ACM, 2024, pp. 2495–2507.

[SR24] Eleni Straitouri and Manuel Gomez Rodriguez. “Designing Decision Support Systems using Counterfactual Prediction Sets”. In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[SXS25] Qizhou Sun, Yufan Xie, and Yain-Whar Si. “Attention-Based Behavioral Cloning for algorithmic trading”. In: *Appl. Intell.* 55.1 (2025), p. 74.

[Sur05] James Surowiecki. *The wisdom of crowds*. Vintage, 2005.

[SB+98] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998.

[SJ15a] Adith Swaminathan and Thorsten Joachims. “Batch learning from logged bandit feedback through counterfactual risk minimization”. In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 1731–1755.

[SJ15b] Adith Swaminathan and Thorsten Joachims. “Counterfactual risk minimization: Learning from logged bandit feedback”. In: *International conference on machine learning*. PMLR. 2015, pp. 814–823.

[Sze+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826.

[Teu+23] Timm Teubner, Christoph M. Flath, Christof Weinhardt, Wil M. P. van der Aalst, and Oliver Hinz. “Welcome to the Era of ChatGPT et al”. In: *Bus. Inf. Syst. Eng.* 65.2 (2023), pp. 95–101.

[Tin02] Kai Ming Ting. “An Instance-Weighting Method to Induce Cost-Sensitive Trees”. In: *IEEE Trans. Knowl. Data Eng.* 14.3 (2002), pp. 659–665.

[TWS18] Faraz Torabi, Garrett Warnell, and Peter Stone. “Behavioral Cloning from Observation”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 4950–4957.

[Tse17] Dimitrios Tsekouras. “The Effect of Rating Scale Design on Extreme Response Tendency in Consumer Product Ratings”. In: *Int. J. Electron. Commer.* 21.2 (2017), pp. 270–296.

[UL24] Shahadat Uddin and Haohui Lu. “Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data”. In: *Plos one* 19.4 (2024), e0301541.

[Var+23a] 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”. In: *Inf. Sci.* 648 (2023), p. 119606.

[Var+24] Víctor Manuel Vargas, Antonio M. Gómez-Orellana, Pedro Antonio Gutiérrez, César Hervás-Martínez, and David Guijo-Rubio. “EBANO: A novel Ensemble BAsed on uNimodal Ordinal classifiers for the prediction of significant wave height”. In: *Knowl. Based Syst.* 300 (2024), p. 112223.

[Var+23b] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Javier Barbero-Gómez, and César Hervás-Martínez. “Soft labelling based on triangular distributions for ordinal classification”. In: *Inf. Fusion* 93 (2023), pp. 258–267.

[VGH19] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás. “Deep Ordinal Classification Based on the Proportional Odds Model”. In: *From Bioinspired Systems and Biomedical Applications to Machine Learning - 8th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2019, Almería, Spain, June 3-7, 2019, Proceedings, Part II*. Vol. 11487. Lecture Notes in Computer Science. Springer, 2019, pp. 441–451.

[VGH20] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. "Cumulative link models for deep ordinal classification". In: *Neurocomputing* 401 (2020), pp. 48–58.

[VGH22] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. "Unimodal regularisation based on beta distribution for deep ordinal regression". In: *Pattern Recognit.* 122 (2022), p. 108310.

[Var+23c] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Riccardo Rosati, et al. "Exponential loss regularisation for encouraging ordinal constraint to shotgun stocks quality assessment". In: *Appl. Soft Comput.* 138 (2023), p. 110191.

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017, pp. 5998–6008.

[VEJ21] Sahil Verma, Michael D. Ernst, and René Just. "Removing biased data to improve fairness and accuracy". In: *CoRR* abs/2102.03054 (2021). arXiv: 2102.03054.

[VFK21] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. "Mitigating bias in machine learning for medicine". In: *Communications medicine* 1.1 (2021), p. 25.

[Vos+24] Simon De Vos, Christopher Bockel-Rickermann, Stefan Lessmann, and Wouter Verbeke. "Uplift modeling with continuous treatments: A predict-then-optimize approach". In: *CoRR* abs/2412.09232 (2024). arXiv: 2412.09232.

[Vov13] Vladimir Vovk. "Conditional validity of inductive conformal predictors". In: *Mach. Learn.* 92.2-3 (2013), pp. 349–376.

[VGS05] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer, 2005.

[WMR17] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". In: *CoRR* abs/1711.00399 (2017). arXiv: 1711.00399.

[Wan+13] Li Wan, Matthew D. Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. "Regularization of Neural Networks using DropConnect". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 1058–1066.

[Wan+24] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. "A Comprehensive Survey of Continual Learning: Theory, Method and Application". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 46.8 (2024), pp. 5362–5383.

[WLZ19] David Widmann, Fredrik Lindsten, and Dave Zachariah. “Calibration tests in multi-class classification: A unifying framework”. 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*. 2019, pp. 12236–12246.

[Wim+23] Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. “Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?” In: *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*. Ed. by Robin J. Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, 2023, pp. 2282–2292.

[WH25] Jie Wu and Mengshu Hou. “An Efficient Retrieval-Based Method for Tabular Prediction with LLM”. In: *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*. Association for Computational Linguistics, 2025, pp. 9917–9925.

[Wu+22] Xingjiao Wu, Luwei Xiao, Yixuan Sun, et al. “A survey of human-in-the-loop for machine learning”. In: *Future Gener. Comput. Syst.* 135 (2022), pp. 364–381.

[Wyn+19] Laure Wynants, Maarten Van Smeden, David J McLernon, et al. “Three myths about risk thresholds for prediction models”. In: *BMC medicine* 17 (2019), pp. 1–7.

[Xu+24] Maochun Xu, Zixun Lan, Zheng Tao, Jiawei Du, and Zongao Ye. “Deep reinforcement learning for quantitative trading”. In: *2024 4th International Conference on Electronics, Circuits and Information Engineering (ECIE)*. IEEE. 2024, pp. 583–589.

[XGW23] 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*. Vol. 216. Proceedings of Machine Learning Research. PMLR, 2023, pp. 2346–2355.

[Yan+14] Yan Yan, Rómer Rosales, Glenn Fung, Subramanian Ramanathan, and Jennifer G. Dy. “Learning from multiple annotators with varying expertise”. In: *Mach. Learn.* 95.3 (2014), pp. 291–327.

[YDM17] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. “Cautious classification with nested dichotomies and imprecise probabilities”. In: *Soft Comput.* 21.24 (2017), pp. 7447–7462.

[Yan+24] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. “Generalized Out-of-Distribution Detection: A Survey”. In: *Int. J. Comput. Vis.* 132.12 (2024), pp. 5635–5662.

[Yao+21] Liuyi Yao, Zhixuan Chu, Sheng Li, et al. “A survey on causal inference”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.5 (2021), pp. 1–46.

[YK24] A. Yarkin Yildiz and Asli Kalayci. “Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data”. In: *CoRR* abs/2410.03705 (2024). arXiv: 2410.03705.

[YD23] Ayfer Ezgi Yilmaz and Haydar Demirhan. “Weighted kappa measures for ordinal multi-class classification performance”. In: *Appl. Soft Comput.* 134 (2023), p. 110020.

[Yon+22] Ching Wai Yong, Kareen Teo, Belinda Pingguan-Murphy, et al. “Knee osteoarthritis severity classification with ordinal regression module”. In: *Multim. Tools Appl.* 81.29 (2022), pp. 41497–41509.

[Yu+23] Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. “A survey on neural-symbolic learning systems”. In: *Neural Networks* 166 (2023), pp. 105–126.

[Yun+24] Víctor Manuel Vargas Yun, Antonio M. Gómez-Orellana, David Guijo-Rubio, et al. “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*. Vol. 14640. Lecture Notes in Computer Science. Springer, 2024, pp. 40–49.

[ZE02] Bianca Zadrozny and Charles Elkan. “Transforming classifier scores into accurate multiclass probability estimates”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’02. Edmonton, Alberta, Canada: Association for Computing Machinery, 2002, pp. 694–699.

[ZLA03] Bianca Zadrozny, John Langford, and Naoki Abe. “Cost-Sensitive Learning by Cost-Proportionate Example Weighting”. In: *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*. IEEE Computer Society, 2003, p. 435.

[Zar+24] Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. “A survey of imitation learning: Algorithms, recent developments, and challenges”. In: *IEEE Transactions on Cybernetics* (2024).

[ZRH22] Wessel van Zetten, GJ Ramackers, and HH Hoos. “Increasing trust and fairness in machine learning applications within the mortgage industry”. In: *Machine Learning with Applications* 10 (2022), p. 100406.

[Zha+25] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, et al. “Data-centric Artificial Intelligence: A Survey”. In: *ACM Comput. Surv.* 57.5 (2025), 129:1–129:42.

[ZLL22] Weijia Zhang, Jiuyong Li, and Lin Liu. “A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling”. In: *ACM Comput. Surv.* 54.8 (2022), 162:1–162:36.

[Zha+24a] Wenyu Zhang, Fayao Liu, Cuong Manh Nguyen, et al. “Training neural networks with classification rules for incorporating domain knowledge”. In: *Knowl. Based Syst.* 294 (2024), p. 111716.

[Zha+24b] Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. “Mitigating Label Bias in Machine Learning: Fairness through Confident Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.15 (Mar. 2024), pp. 16917–16925.

[Zho18] Zhi-Hua Zhou. “A brief introduction to weakly supervised learning”. In: *National science review* 5.1 (2018), pp. 44–53.

List of Figures

List of Tables

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 08.09.2025 _____

Stefan Haas