Statistical Methods leveraging Uncertainties in Machine Learning

Dissertation

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

eingereicht von
Stefan Michael Stroka
am 12.06.2025

Statistical Methods leveraging Uncertainties in Machine Learning

Erstgutachter:	**Prof. Dr. Christian Heumann**, Institut für Statistik, LMU München
Zweitgutachter:	**Prof. Dr. Volker Schmid**, Institut für Statistik, LMU München
Drittgutachter:	**Prof. Dr. Martin Spieß**, Institut für Psychologie, Universität Hamburg
Vorgelegt von:	**Stefan Stroka**

Tag der mündlichen Prüfung: 14. November 2025

Danksagung

Zu Beginn möchte ich mich herzlichst bei all jenen bedanken, die mich bei meiner persönlichen und fachlichen Weiterentwicklung gefördert und unterstützt haben. Ohne diese Unterstützung wäre die vorliegende Arbeit in dieser Form nicht möglich gewesen.
An erster Stelle möchte ich mich ausdrücklich bei meinem Doktorvater, Prof. Dr. Christian Heumann, bedanken, der sich die Zeit genommen hat, mit mir die komplexen Aufgabenstellungen zu erarbeiten und mich während der gesamten Arbeit stets mit wertvollen Anregungen, seiner fachlichen Expertise und geduldiger Unterstützung begleitet hat.
Mein besonderer Dank gilt meinem Vater Wolfgang Stroka, meinem Bruder Christian Stroka und Janina Wanzke, die mir durch ihre unermüdliche Unterstützung, ihren Zuspruch und ihren festen Glauben an mich maßgeblich ermöglicht haben, die Promotion anzustreben und schließlich diese Arbeit zu vollenden. Darüber hinaus danke ich auch meiner gesamten Familie und meinen Freunden, die mir stets Rückhalt gegeben und mich ermutigt haben. Ihre Geduld, ihr Verständnis und ihre fortwährende Unterstützung haben mir geholfen, auch herausfordernde Zeiten zu meistern.
Mein Dank gilt auch meinen Kollegen bei ams Osram, die mich durch ihre Zusammenarbeit und ihr offenes Ohr stets unterstützt haben. Ihr Feedback und das gemeinsame Arbeiten haben mir in vielen Phasen der Promotion sehr geholfen.
Abschließend danke ich allen, die auf unterschiedliche Weise zum Gelingen dieser Arbeit beigetragen haben.

Abstract

In today's data-driven landscape, machine learning methods are increasingly applied in domains that demand high levels of safety, reliability, and interpretability. However, hidden influencing factors and limited data availability can significantly impair model performance and amplify predictive uncertainty. Recognizing, quantifying, and—where possible—reducing uncertainties such as aleatoric and epistemic uncertainty has therefore become a central concern. This is particularly true in critical fields like autonomous driving, medical diagnostics, finance, weather forecasting, and industrial production, where dependable predictions are not merely advantageous, but essential. Despite its relevance, the broader adoption of uncertainty quantification in practice is often hindered by high computational demands and growing model complexity. Furthermore, aleatoric uncertainty—stemming from noise and imperfections in the data itself—poses a fundamental challenge to the reliability of data-driven models. This dissertation explores multiple strategies for uncertainty quantification across four publications. These contributions examine both the necessity and the practical implementation of probabilistic techniques, while also introducing novel, less computationally intensive methods that reduce model complexity without sacrificing robustness.

Publication 1:

The first study addresses epistemic and aleatoric uncertainties in the preprocessing phase of industrial production modeling. Aleatoric uncertainties are predefined based on expert experience, setting the bounds for acceptable input variation. Given the limited spatial distribution of measurement data, uncertainty-aware interpolation is applied for data augmentation. Probabilistic Gaussian Process Regression is employed to model prediction intervals and serve as a basis for generating synthetic input data. Results using real production data from ams OSRAM show that even with sparse measurements,

highly accurate models can be constructed.

Publication 2:

The second study introduces a novel modeling approach that combines two types of target variables: a continuous regression target and an ordinal classification target. A customized loss function, paired with fuzzy logic, enables the model to optimize regression estimates while simultaneously improving classification performance. The method shows particularly strong performance in imbalanced data scenarios, leading to a significant reduction in latent uncertainty. Applied to housing market data in the United States, the approach yields up to a 17.1% improvement in F1-score.

Publication 3:

The third publication investigates the integration of data-independent, expert-derived knowledge into data-dependent learning models. Unmeasured or unquantified latent uncertainties can reduce model robustness. This approach trains models using both observed data and qualitative expert assessments—without requiring expert input at inference time. Results using synthetic data generated via variational autoencoders (VAEs), based on real-world use cases from ams OSRAM, demonstrate improved optimization even with a marginal increase in mean absolute error (MAE).

Publication 4:

The fourth study addresses the challenge of training with small datasets. A probabilistic modeling approach is presented that estimates the latent distribution of a target variable using ordinal class labels. This enables the generation of additional, reliable input data to support model training. With only 5–10% of the original training data, the method achieves notable improvements: up to 10% in mean squared error (MSE), 5–10% in coefficient of determination (R^2), and approximately 8% in prediction coverage.

Zusammenfassung

In der heutigen datengetriebenen Welt werden Machine-Learning-Methoden zunehmend in sicherheitskritischen und hochzuverlässigen Anwendungsbereichen eingesetzt. Verdeckte Einflussfaktoren sowie begrenzte Datenverfügbarkeit können jedoch die Modellgüte erheblich beeinträchtigen und die Vorhersageunsicherheit erhöhen. Die Erkennung, Quantifizierung und – wo möglich – Reduktion von Unsicherheiten, insbesondere aleatorischer und epistemischer Art, hat daher zentrale Bedeutung gewonnen. Dies gilt insbesondere für Anwendungsfelder wie autonomes Fahren, medizinische Diagnostik, Finanzwesen, Wetterprognose und industrielle Produktion, in denen verlässliche Vorhersagen nicht nur wünschenswert, sondern essenziell sind. Die breite Anwendung von Unsicherheitsquantifizierung scheitert jedoch häufig an hohen Rechenaufwänden und zunehmender Modellkomplexität. Zudem stellt die aleatorische Unsicherheit – verursacht durch zufällige Messfehler und Datenrauschen – eine grundlegende Herausforderung für die Verlässlichkeit datenbasierter Modelle dar. Diese Dissertation untersucht verschiedene Strategien zur Quantifizierung von Unsicherheiten anhand von vier begutachteten Publikationen. Die Beiträge beleuchten sowohl die Notwendigkeit als auch die praktische Umsetzung probabilistischer Verfahren und schlagen darüber hinaus neuartige, rechenökonomische Alternativen vor, die ohne signifikante Steigerung der Modellkomplexität eine robuste Modellierung ermöglichen.

Publikation 1:

Die erste Studie thematisiert epistemische und aleatorische Unsicherheiten in der Vorverarbeitung industrieller Produktionsdaten. Aleatorische Messunsicherheiten werden auf Basis von Erfahrungswerten vordefiniert, um einen Rahmen für die Eingabedatenunsicherheit zu schaffen. Aufgrund der räumlich begrenzten Messpunktverteilung werden Interpolationsverfahren mit Unsicherheitsberücksichtigung zur Datenanreicherung eingesetzt. Eine proba-

bilistische Gaussian-Process-Regression dient zur Modellierung von Prognoseintervallen und zur Generierung zusätzlicher Eingabedaten. Die Anwendung auf Produktionsdaten der Firma ams OSRAM zeigt, dass auch mit wenigen Messwerten präzise Modelle realisierbar sind.

Publikation 2:

Die zweite Studie stellt einen innovativen Modellierungsansatz vor, der zwei Zielgrößen kombiniert: eine kontinuierliche Regressionsgröße und eine ordinale Klassifikationsgröße. Eine angepasste Verlustfunktion in Kombination mit Fuzzy Logic ermöglicht eine gleichzeitige Optimierung beider Zielgrößen. Der Ansatz zeigt insbesondere bei unausgeglichenen Klassenverteilungen signifikante Verbesserungen und reduziert latente Unsicherheiten durch die Kombination beider Ziele. Die Anwendung auf US-Immobiliendaten zeigt eine Verbesserung des F1-Scores um bis zu 17,1 %.

Publikation 3:

Die dritte Veröffentlichung befasst sich mit der Integration von datenneutralem Expertenwissen in datenabhängige Lernmodelle. Nicht gemessene oder nicht quantifizierbare Unsicherheiten können die Modellstabilität gefährden. Der vorgestellte Ansatz kombiniert trainingsseitig beobachtete Daten mit qualitativen Experteneinschätzungen – ohne dass Expertenwissen zur Vorhersagezeit erforderlich ist. Ergebnisse mit synthetischen Daten, erzeugt mittels Variational Autoencoders (VAE) auf Basis realer ams-OSRAM-Anwendungsfälle, zeigen eine verbesserte Modelloptimierung trotz eines geringen Anstiegs des mittleren absoluten Fehlers (MAE).

Publikation 4:

Die vierte Studie widmet sich der Problematik kleiner Stichprobenumfänge. Ein probabilistischer Modellierungsansatz wird vorgestellt, bei dem die Verteilung der Zielgröße über ordinale Klassen abgeschätzt wird. Dadurch lassen sich zusätzliche, zuverlässige Eingabedaten für das Modelltraining erzeugen. Bere-

its mit 5–10 % der ursprünglichen Trainingsdaten lassen sich deutliche Verbesserungen erzielen: eine Reduktion des mittleren quadratischen Fehlers (MSE) um bis zu 10 %, eine Steigerung des Bestimmtheitsmaßes (R^2) um 5–10 % sowie eine um rund 8 % verbesserte Abdeckung der Prognoseintervalle.

Contents

Contributions of the Thesis

This cumulative dissertation consists of the following publication list:

1. Stroka S., Heumann C., Suhrke F., Meindl K. (2023) A probabilistic approach for approximation of optical and opto-electronic properties of an opto-semiconductor wafer under consideration of measuring inaccuracy and model uncertainty *Opto-Electronics Review*

2. Stroka S., Heumann C., Suhrke F. (in Review) Multi-Task Learning for Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes *Operational Research*

3. Stroka S., Heumann C. (in Review) Knowledge-embedded Machine Learning for Production Optimization on Logistical Delivery Grids *The International Journal of Advanced Manufacturing Technology*

4. Stroka S., Heumann C. (2024) Is Anonymization through Discretization reliable? Modeling latent probability distributions for ordinal data as solution for small sample size problem *stats*

These publications are included in the thesis as Chapters 4 to 7. The individual contributions of each author are detailed prior to each publication.

1 Introduction

The digital era, driven by the exponential growth in computational capabilities, has significantly accelerated the evolution of machine learning. The fusion of mathematical and statistical methodologies with computer science has enabled the practical implementation and further development of previously intractable theoretical models [Aggarwal et al., 2022]. In this context, artificial intelligence (AI), particularly machine learning, has experienced rapid advancement and found impactful applications across domains once deemed beyond the reach of automation. Today, AI-powered systems—ranging from language assistants and image recognition to predictive analytics, personalized recommendations, autonomous vehicles, and smart manufacturing within the framework of Industry 4.0—have become integral to everyday operations [Sarker, 2021]. The development of these systems has reached a point where limitations are increasingly defined by the quality and quantity of input data, rather than by model architecture itself [Chen et al., 2021]. Consequently, the demands on trustworthiness and interpretability of AI-driven predictions have intensified [Alam et al., 2023, Bostrom et al., 2024]. As models grow in complexity and autonomy, the risk of reduced transparency and oversight escalates [Holzinger, 2021]. Reliable decision-making therefore hinges not only on model performance but also on the ability to quantify uncertainty in predictions [Begoli et al., 2019, Jalaian et al., 2019]. Uncertainty quantification (UQ) methods, gaining prominence in recent years [Abdar et al., 2021, Psaros et al., 2023], aim to capture both aleatoric and epistemic uncertainty, thus providing a comprehensive framework for assessing predictive reliability. Aleatoric uncertainty pertains to the inherent noise in the input data [Sullivan, 2015]. Despite technological progress, data acquisition

remains vulnerable to both systematic and stochastic errors. In contrast, epistemic uncertainty arises even with reliable data, as models are fundamentally conditioned on the information available during training [Sullivan, 2015]. Uncertainty from unobserved or unmeasurable sources poses a challenge to model trustworthiness. This dissertation is motivated by the imperative to identify effective methodologies for quantifying and incorporating both types of uncertainty into the modeling process. Specifically, it examines potential applications of uncertainty-aware methods, such as Gaussian Process Regression (GPR), and introduces novel techniques aimed at improving data quality and model interpretability.

1.1 Data Augmentation under Uncertainty Using Gaussian Processes

Quantifying uncertainty is a crucial aspect of data modeling, particularly in industrial applications where data may be limited, noisy, or collected asynchronously. GPR is a Bayesian inference method that enables the quantification of epistemic uncertainties while also accounting for aleatoric uncertainty [Schulz et al., 2018a]. Unlike traditional regression techniques that yield only single point estimates, GPR produces full predictive distributions, allowing for the derivation of uncertainty intervals based on prior knowledge or prior distributions. Although GPR is a well-established and extensively studied approach, recent research has shifted its focus toward practical applications—specifically, the generation of data that inherently incorporates uncertainty—rather than on methodological innovation itself [Santoni et al., 2024, Triggiano and Romito, 2024, Tang et al., 2022, Wang et al., 2023]. Within the context of Industry 4.0, GPR has been applied to address imbalances in the availability of input and target data [Gardner et al., 2021]. For

example, in the case study presented in our first paper, a supervised learning model is employed where initial process measurements serve as input to predict final product measurements. Often, these datasets are imbalanced: a limited number of initial input measurements contrasts with a significantly larger volume of final product observations. Although these measurements refer to the same underlying process and product, they are collected at different stages of production and at different points in time. This discrepancy in data availability, coupled with the risk of information loss during preprocessing, motivates the use of approximation techniques capable of generating reliable input data from limited observations. However, to maintain the integrity of the data and avoid introducing ambiguity into the model, the generation process must account for uncertainty in a principled way. The associated publication proposes a novel approach for generating additional input data while explicitly constraining epistemic uncertainty during the augmentation process. This method ensures that augmented data do not dilute the information content of the original inputs, thus preserving the reliability of subsequent model training. By integrating uncertainty directly into the data generation process, the methodology improves both the theoretical soundness and the practical applicability of the model in real-world scenarios.

1.2 Enhancing Regression Models through Classification-Guided Objectives

While traditional regression methods such as Gaussian Process Regression (GPR) have been widely applied, this thesis proposes a novel approach that enhances regression performance by incorporating an additional classification target. To this end, a custom regression loss function [Hastie et al., 2001] is employed to bias the mean estimates, thereby improving the model's ca-

pacity to capture deviations that are better aligned with the classification objective. This loss function is convex and twice differentiable, allowing it to be implemented with any regression technique based on gradient descent optimization [Boyd and Vandenberghe, 2004]. Commonly, regression models assume that the training data are reliable and accurate [Gleser, 1992], neglecting the potential influence of unsupervised factors. However, if aleatoric uncertainty is present in the target data—such as measurement errors—the reliability of the resulting predictions can be substantially diminished [Yazdi et al., 2021]. To address this challenge, the proposed method introduces a second, independent classification target. By jointly optimizing a combined loss function that integrates both the potentially biased regression target and the classification target, the model is able to leverage "weak" supervision signals to enhance overall accuracy. This approach is particularly motivated by practical production scenarios, where data collection can be improved without necessitating fundamental changes to measurement processes. Moreover, non-data-dependent information—such as subjective expert knowledge or other non-quantifiable insights—can be utilized as a secondary classification target. This additional target biases the mean regression estimator and improves classification accuracy by combining two possibly weak targets into a single, optimized model. The model thereby identifies a trade-off that yields the best possible representation of the underlying data. The overarching goal is to integrate both observed input data and non-data-dependent knowledge to capture unobserved and potentially immeasurable factors, thus reducing model and prediction uncertainty. This concept is realized through a novel, customized regression loss function that enables optimization of pointwise regression estimators with respect to a categorical target. A key innovation of this method is the ability to compute the loss between a continuous (metric) target and ordinal classes by leveraging continuous evaluation metrics. This

is achieved through a unique treatment of the ordinal classification target as a pseudo-continuous variable via fuzzy logic. Consequently, continuous metrics can be applied within the classification space, where disjoint ordinal classes partition the continuous value range. Each class is defined by explicit boundaries within this continuous range, allowing clear determination of when a point prediction belongs to a particular class. Unlike conventional optimization strategies, this approach performs optimization iteratively within each model training step, which reduces computational effort—thanks to gradient descent—and decreases model complexity when compared to combined multi-output learning frameworks. Furthermore, the interpretability of each learning step is enhanced relative to black-box methods or multi-output model ensembles, enabling traceable and reliable regression estimates that are specifically optimized for the classification target.

1.3 Reducing Epistemic Uncertainty by Incorporating Aleatoric Uncertainty

Building on the previously introduced approach, the custom loss function methodology is further extended to explicitly incorporate aleatoric uncertainty in the regression target. Known measurement inaccuracies and presumed stochastic fluctuations characterize the quality of the input data [Hüllermeier and Waegeman, 2021]. Comparable methods—such as Bayesian inference [Li et al., 2021] and Monte Carlo simulations [Swiler et al., 2009]—model such uncertainty by means of probability distributions derived from prior knowledge, statistical analysis, or expert input. While these techniques iteratively approximate optimal estimators, they often come at the cost of substantial computational overhead. In contrast, more practical machine learning strategies seek to classify and incorporate sources of uncertainty

directly into the modeling process [Kläs and Vollmer, 2018]. The method proposed in this thesis introduces a deterministic approach based on a "best-guess" estimation to account for aleatoric uncertainty. This allows for the joint modeling of systematic and random errors, while considerably reducing computational complexity. Within this framework, the metric regression target is extended to include a classification target that reflects uncertainty. Specifically, when the continuous value range is divided into disjoint classes with defined boundaries, an overlapping region arises between the metric and classification targets. The presumed uncertainty is then used to augment the point estimate with an uncertainty interval. While the point estimate can only be assigned to a single class, its surrounding interval may span multiple classes. This becomes particularly valuable near class boundaries, where imprecise measurements provide meaningful insight into the confidence of the observed value. As a result, classification based on the point estimate and that based on the uncertainty interval may diverge—creating two contrasting but complementary optimization objectives. Notably, this uncertainty-aware formulation does not require explicit sampling of the distribution itself. The assumed uncertainty is incorporated into the regression model via a tunable hyperparameter, thereby allowing for flexible adjustment. This enhances both the interpretability and adaptability of the model—particularly in settings where measurement accuracy and uncertainty quantification are of central importance.

1.4 Addressing Sample Size Limitations through Distribution-Based Feature Augmentation

Another central challenge addressed in this thesis is the small-sample-size problem, which poses significant difficulties for both deterministic and prob-

abilistic machine learning methods [Chapelle et al., 2002]. Limited datasets often lead to underfitting and poor generalization performance [Pothuganti, Aliferis and Simon, 2024, Montesinos López et al., 2022]. To address this issue, a novel method is proposed that enhances generalization by leveraging statistical probability modeling, even when only sparse input data are available. In many practical applications, generalization is impaired by latent impact factors—that is, influences from partially measured or entirely unobserved features—which introduce hidden variability into the modeling process [Malik, 2020]. Optimized model selection strategies aim to mitigate this effect through established techniques such as Gaussian Process Regression [Ferber et al., 2025], Gaussian Mixture Models (GMMs) [Arora et al., 2021], bootstrapping [Gao et al., 2019], mixtures of experts [Gao et al., 2023], and hierarchical Dirichlet processes [Munro and Ng, 2022, Traunmüller et al., 2015]. These approaches help to model and compensate for uncertainty arising from unobserved data-generating processes. The proposed approach reduces the influence of such unknowns by using probability distributions to interpolate feature spaces and augment input data, with a focus on quantifying uncertainty. It extends latent variable regression models [Burnham et al., 1999] through the integration of modeled probability distributions [Zhou et al., 2014], incorporating Gaussian Mixture Models in combination with Bayesian linear regression to improve the robustness and reliability of predictions. At the core of this method is the reduction of uncertainty in the input data. This is achieved by observing known ordinal classes—combined with available feature information—which segment the continuous value range of the metric target variable. Based on this segmentation, a Gaussian Mixture Model is constructed with a separate normal distribution for each class. The resulting class-conditional distributions allow for interpolation and extrapolation within and between classes, thereby enabling the augmentation of new input

data and improving predictive accuracy.

1.5 Objective of the Dissertation

This dissertation aims to enhance conventional machine learning methodologies by systematically integrating uncertainty modeling through statistical techniques. The overarching goal is to improve the robustness and reliability of predictions by addressing both aleatoric and epistemic uncertainty. The contribution is twofold: (1) leveraging uncertainty intervals to augment and evaluate input data quality, and (2) improving prediction reliability without the explicit estimation of uncertainty via prediction or confidence intervals, which often increase model complexity and computational cost. While the first aspect targets aleatoric uncertainty by refining the input data space, the second focuses on epistemic uncertainty by improving model generalization through structural enhancements.

1.5.1 Primary Goals of the Work

The primary goal of this work is to improve state-of-the-art machine learning models by incorporating statistical methods that explicitly account for aleatoric and epistemic uncertainty. Aleatoric uncertainty—stemming from noise or imprecise measurements—is mitigated through uncertainty-based data generation and augmentation strategies. Epistemic uncertainty—resulting from limited or incomplete information—is addressed by embedding probabilistic models that improve inference under sparse data conditions. Together, these strategies aim to produce more reliable, interpretable, and efficient machine learning models.

1.5.2 Hypotheses and Research Questions

The central hypotheses guiding this dissertation are as follows:

- Uncertainty-aware data generation in the preprocessing phase improves both the quality and quantity of input data, thereby increasing model accuracy.
- The integration of coarser, categorical, or expert-derived information enables an optimization process between regression and classification objectives. This trade-off leads to improved classification accuracy at the cost of a marginal bias in the regression estimate.
- Modeling latent probability distributions using GMMs enhances learning in scenarios with very limited training data.

The key research questions investigated are:

- How reliable are the generated input data, and what are the practical limits of their use in model training?
- To what extent can unobserved or coarsely quantified information be meaningfully incorporated into the learning process?
- How effective is the use of aleatoric uncertainty bounds in constraining subjective or weak knowledge, and what is the optimal balance between introducing bias in the regression estimate and improving classification performance?

1.5.3 Structure of the Dissertation

The remainder of this dissertation is structured as follows: The next section introduces the theoretical background related to uncertainty modeling and outlines its relevance to the application domain. This is followed by an

overview of the individual research papers and a discussion of the original contributions presented in this work.

2 Methodology and General Background

Uncertainty quantification is obviously or latently important for most AI applications nowadays. This applies not only to obvious and highly topical use cases, such as autonomous driving, where even the smallest uncertainty can have fatal consequences, but also to many other, sometimes less conspicuous areas, like production optimization or cost minimization.

2.1 Aleatoric and Epistemic Uncertainty

Uncertainty quantification is divided into irreducible (aleatoric) and reducible (epistemic) uncertainty, which can be further subdivided. In general, a distinction is made between the uncertainty of the input data (aleatoric uncertainty) and the uncertainty of the model and prediction (epistemic uncertainty) [Der Kiureghian and Ditlevsen, 2009, Hüllermeier and Waegeman, 2021]. Figure 2.1 provides a visual illustration of the distinction between aleatoric and epistemic uncertainty in a regression setting.

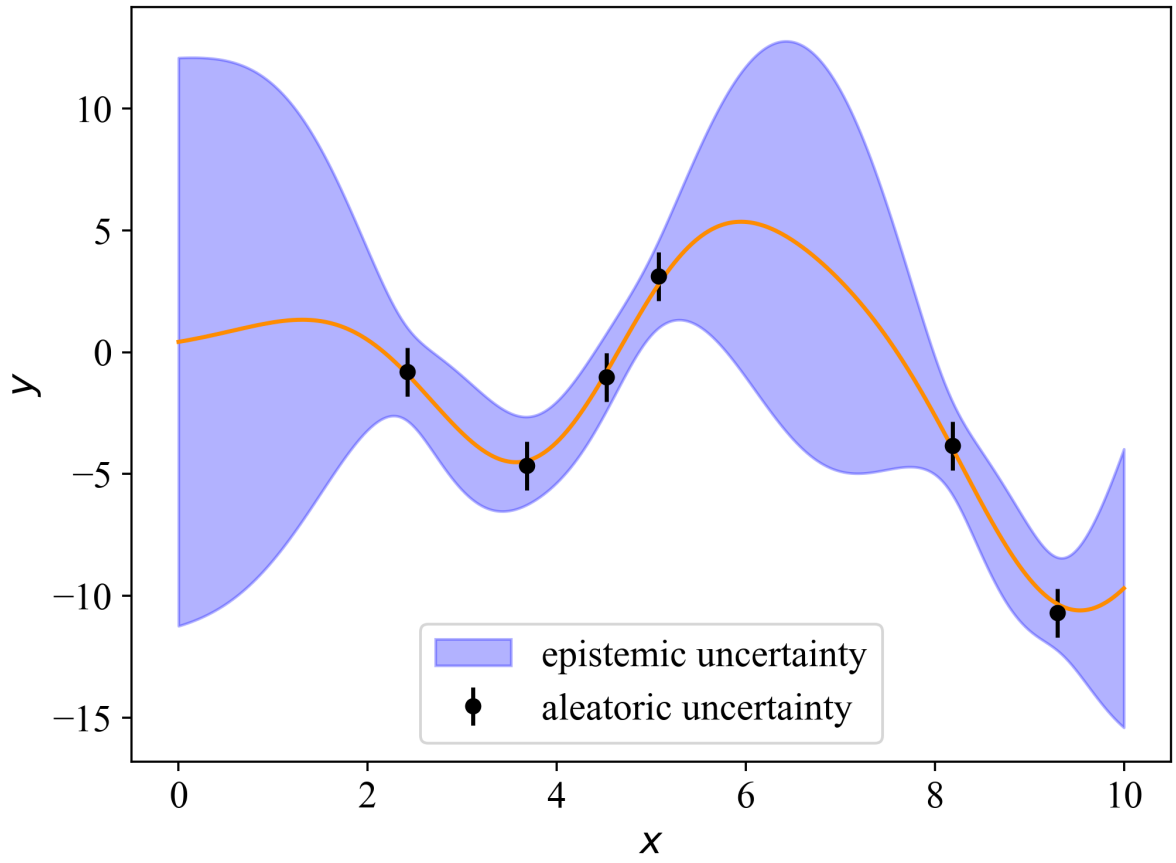


Figure 2.1: Illustration of aleatoric and epistemic uncertainties using a standard Gaussian Process Regression applied to simulated example data with input variability. The prediction intervals (blue bands) represent the model's epistemic uncertainty, incorporating the observation noise depicted by black error bars.

Aleatoric uncertainty can be further subdivided into random and systematic influences [Shaker and Hüllermeier, 2020]. Aleatoric uncertainty is often described simply as random error; however, distinguishing between the two is frequently impractical. In the case of aleatoric uncertainty, one considers systematic measurement errors, which arise from inaccurate measurements or inadequate measurement procedures, as well as stochastic uncertainty. Stochastic uncertainty cannot be mitigated by improved measurement techniques and can arise from various causes such as natural variation, inherent

randomness, or environmental, process, or demographic stochasticity [Glick et al., 2001, Helton, 1997]. Consequently, random (stochastic) errors can only be estimated using statistical or other methods. In contrast, epistemic uncertainty concerns model-dependent uncertainties such as parameter, model, or prediction uncertainties. Epistemic uncertainty is also known as a lack of information [Swiler et al., 2009, Large et al., 2017]. This suggests that improved data quality can result in reduced epistemic uncertainty. However, this uncertainty can also come from subjective judgments. In specific areas where modeling is based on expert opinions or evaluations, uncertainties may arise due to human influences. Therefore, when modeling under uncertainty, the primary focus is on exploring epistemic uncertainty, as the reliability of predictions depends on the level of confidence in this aspect.

2.1.1 Quantification Techniques & Methods for Quantifying and Managing Uncertainty

A wide range of methods are available for quantification [Abdar et al., 2021, Soize, 2017, Kabir et al., 2018, Zhang et al., 2020, Psaros et al., 2023]. The selection of the optimal methodology depends on several factors. These include data availability with measurement errors and variability, model complexity and the number of potential parameters, assumptions and simplifications (which may directly impact the uncertainties), resources and time, and finally the complexity of the uncertainties themselves.

2.1.2 Confidence, Prediction and Certainty Interval

Depending on the prerequisites mentioned and the limitations of uncertainty quantification, the ultimate goal is to determine confidence intervals (CI) [O'Brien and Yi, 2016], prediction intervals (PI) [Khosravi et al., 2011], or credible intervals (CRI) [Whitener, 1990, Eberly and Casella, 2003]. Depend-

ing on the methodology, these intervals can be determined deterministically, frequentistically, or probabilistically and can therefore vary according to the complexity of the model, existing as intervals for point predictions or intervals based on distributions [Berleant et al., 2005]. CI, PI, and CRI should not be confused, as they convey fundamentally different information. CIs are determined by symmetric interval bounds around a central estimator, depending on a α parameter. The interval indicates, with a certain percentage, how likely it is that the central estimator falls within the interval. PI indicates, with a certain percentage, how likely it is that the model predictions will fall within this interval, given the input data and prediction values, as well as an α parameter. Figure 2.2 illustrates the conceptual and quantitative distinction between confidence and prediction intervals in the context of Gaussian Process Regression.

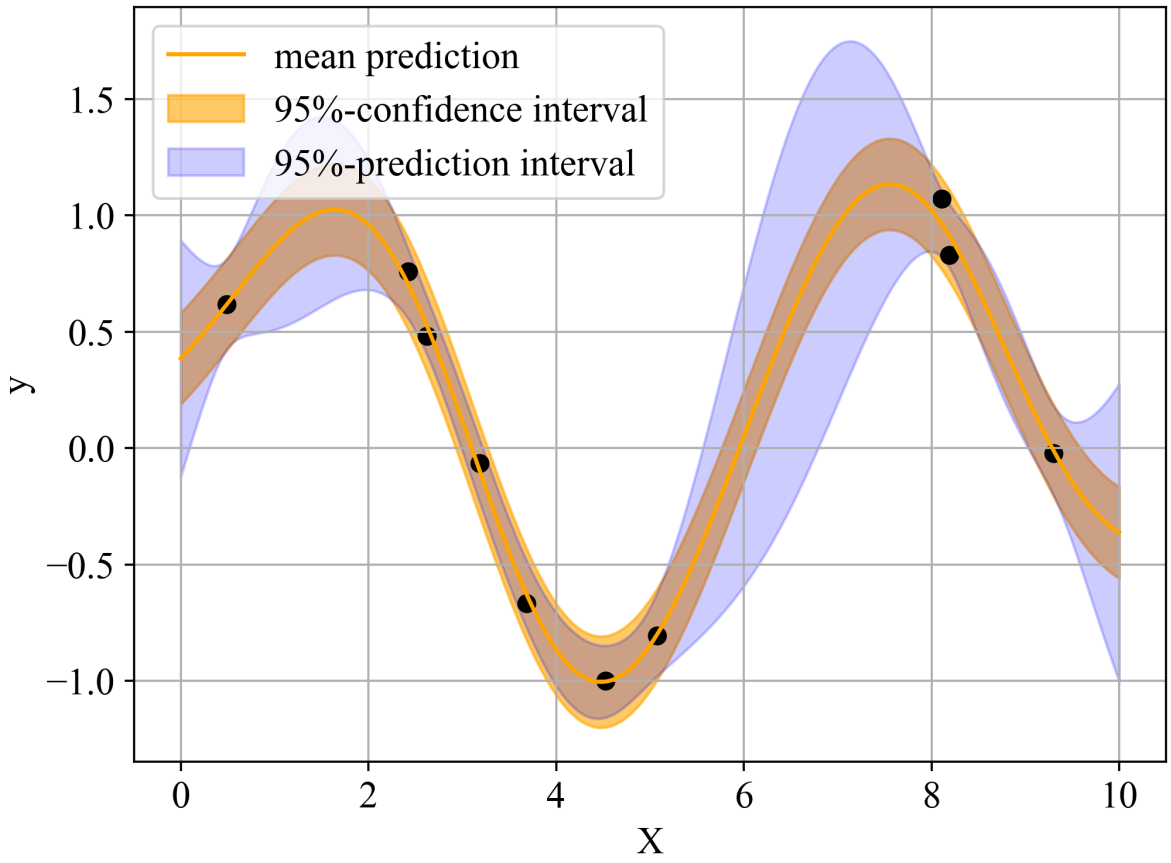


Figure 2.2: Example of Gaussian Process Regression applied to noisy data (black dots). The solid line depicts the mean prediction, the light orange shaded area represents the 95% CI of the model's mean estimate, and the wider light blue shaded area indicates the 95% PI, which accounts for both model uncertainty and observation noise. The PI is consistently wider than the CI, reflecting the added uncertainty from noise in future observations.

The CRI is the Bayesian interval, where the interval represents the uncertainties over the parameters of a model [Whitener, 1990, Eberly and Casella, 2003]. These intervals are directly dependent on the posterior distribution and, therefore, indirectly on the chosen prior distribution. They indicate how likely it is that a parameter value lies within a certain range.

2.2 Gaussian Process Regression

Gaussian Process Regression is a probabilistic regression method based on the Gaussian Process. Through its probabilistic application, a probability distribution of estimation functions is determined, which defines both the mean estimator function and the PI functions [Schulz et al., 2018b, Williams and Rasmussen, 1995].

A Gaussian Process (GP) is equivalent to a stochastic process in probability theory [Pavliotis, 2014]. The finite-dimensional Gaussian distribution consists of a combination of multidimensional normal distributions and is uniquely determined by its mean function and covariance function. Unlike a single Gaussian distribution (also called a normal distribution), the characteristic parameters—mean and standard deviation—are distribution functions in the case of a GP. The distinct feature of a GP is the definition of a probability distribution for all finite-dimensional Gaussian distributions of the GP [Wang, 2023, Chen et al., 2023].

Let $(X_t)_{t\in T}$ be a stochastic process of multidimensional normal distributions on the index set T with $n \in \mathbb{N}$ dimensional normal distributions for each index $t \in T$. Let $X \sim GP(\mu(t), \gamma(s,t))$ be a Gaussian Process, then it is uniquely determined by the mean function $\mu(t) = \mathbb{E}(X_t),\ t \in T$, and the covariance function $\gamma(s,t) = \mathrm{Cov}(X_s, X_t),\ s,t \in T$.

GPR is a regression analysis based on the concept of Gaussian Processes. This Bayesian regression approach combines samples from a kernel function (prior) with given data, resulting in a probability distribution for possible posterior sample functions.

For later applications, we focus on a three-dimensional scenario. Let $D = (X, Y, Z)$ represent the observations, and let k be any kernel function with θ as the hyperparameter vector. Assuming the relationship $z = f(x_i, y_j) + \epsilon_{i,j}$, where $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, we compute the mean vector $\mu((x,y)) = \mathbb{E}(z(x,y))$

and the covariance matrix $K_{ij} = k((x_i, y_i), (x_j, y_j))$. We then randomly sample prior functions dependent on the given data, $p(\theta|z)$, where $z \sim \mathcal{N}(\mu, K)$ in the three-dimensional space.
Assuming $z \sim \mathcal{N}(\mu, \sigma^2)$, the likelihood can be described as the product of the densities of the individual normal distributions for all observations. Thus, the likelihood becomes:

$$p(z \mid X, \theta, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(z_i \mid \mu_i, \sigma^2) \tag{2.1}$$

Using maximum likelihood estimation (MLE), the parameters θ and σ^2 can be found by maximizing the likelihood. This leads to:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} \log p(z \mid X, \theta, \sigma^2) \tag{2.2}$$

as the best estimators. Based on the likelihood $LH(z \mid X, \theta, \sigma^2)$ and the prior functions $p(X \mid \theta)$, the posterior distribution over possible functions can be determined as:

$$p(\theta, \sigma^2 \mid z) \propto LH(z \mid (x, y), \theta, \sigma^2) \times p(\theta, \sigma^2) \tag{2.3}$$

The key advantage of GPR is its high efficiency in hyperparameter optimization through algebraic computation. On the other hand, runtime can be a disadvantage, depending on the number of observations due to the need to invert the kernel matrix, which has a computational complexity of $O(N^3)$ [Tripathy et al., 2016].
The posterior as a GP is defined differently from a univariate normal distribution, using distribution functions for mean and covariance. This allows determining CIs and PIs as functions of these distribution functions. Both intervals depend on a critical value z from the standard normal distribution [Schulz et al., 2018b].

Given $\hat{f}$ as the mean estimate function and σ_post^2 as the variance of the posterior distribution, the CI is given by:

$$\left[\hat{f} - z \cdot \sqrt{\mathrm{Var}(\hat{f})},\ \hat{f} + z \cdot \sqrt{\mathrm{Var}(\hat{f})}\right] \tag{2.4}$$

And for the PI:

$$\left[\hat{f} - z \cdot \sqrt{\mathrm{Var}(\hat{f}) + \sigma^2_{\mathrm{post}}},\ \hat{f} + z \cdot \sqrt{\mathrm{Var}(\hat{f}) + \sigma^2_{\mathrm{post}}}\right] \tag{2.5}$$

Gaussian Process Regression (GPR) is a well-researched algorithm. Further research primarily focuses on reducing runtime. Regarding the research conducted in this work, the focus is not on further investigating GPR itself but rather on its application for data augmentation under uncertainty considerations.

2.3 Reduction of Uncertainties without Explicit Quantification

In addition to conventional methods for parametric or non-parametric quantification of uncertainties, we explore approaches that implicitly address and reduce uncertainty without explicit quantification. Drawing on the Bayesian approach, we use data-independent information as potential prior knowledge and examine a new methodology for incorporating this knowledge into the model training of standard models with minimal additional effort. Data-driven models are ultimately limited by the quality and quantity of the data [Yao, 2021]. These limitations correspond to data and model uncertainties. The approach presented here is therefore motivated by the possibility of reducing these limitations and improving the model with respect to its uncertainties. In practice, data quality or quantity cannot always be easily

improved or increased. High structural costs, such as improved measurement methods or a greater number of measurements, are often not cost-effective in terms of the value generated. On the other hand, there are also stochastic errors or other uncontrollable or unmeasurable influences. To address this lack of data, the proposed methodology allows for the incorporation of data-independent knowledge from, for the model unknown, sources such as statistical methods, expert experience or even physical restrictions into the training process. The use of such knowledge is often hindered by the difficulty of quantifying it. Therefore, the following approach enables the quantification of this knowledge into ordinal classes, which can then be used as classification targets during the training process.

A customized loss function is a self-defined loss function used in supervised learning tasks to influence and control the model training process. In the case of a convex loss function, GD optimization can even be applied [Ebert-Uphoff et al., 2021]. The new customized loss function combines a metric regression loss and a pseudo-classification loss based on fuzzy logic as a linear combination. Both components are weighted by a new hyperparameter, α. The fundamental idea behind this linear combination is to enable the simultaneous training of a supervised regression target and a supervised classification target within a regression model. For the regression loss, we use the Mean Squared Error (MSE), which is a common regression loss that measures the mean squared distance between the metric prediction and the observation (target value). The result, also referred to as pseudo-residuals, is used in the model for prediction evaluation and iterative improvement, depending on the learning rate.

Let matrix X represent the set of n observations and y be a vector of the

newly defined metric target variable. As a regression problem, we assume:

$$y = f(x) + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{2.6}$$

The MSE with $\hat{y}$ as model predictions is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.7}$$

The second component for classification is a pseudo-classification loss based on fuzzy logic [Bothe, 2013]. We first define the implementation of fuzzy logic and then the pseudo-classification function. A classification loss is typically not comparable on a metric scale. To make the two components of the linearly combined custom loss function applicable and comparable, the classification target is quantified through fuzzy logic. Specifically, ordinally scaled and ascendingly ranked classes can be redefined on a metric scale using classification boundaries. Ordinal classes typically have a rank order but do not provide information about the distances between classes. Therefore, classification boundaries are defined based on selected properties to define and make the ordinal classes comparable on a metric scale. These boundaries can be determined either based on natural properties or set randomly or deliberately according to specific objectives. Figure 2.3 illustrates the transformation of a continuous metric space into ordinal classes through predefined classification boundaries.

For example:

Quality	Lower Threshold	Upper Threshold
bad	0	0.25
ok	0.25	0.5
good	0.5	0.75
best	0.75	1

Using class (quality) thresholds as metrical represent for a class:

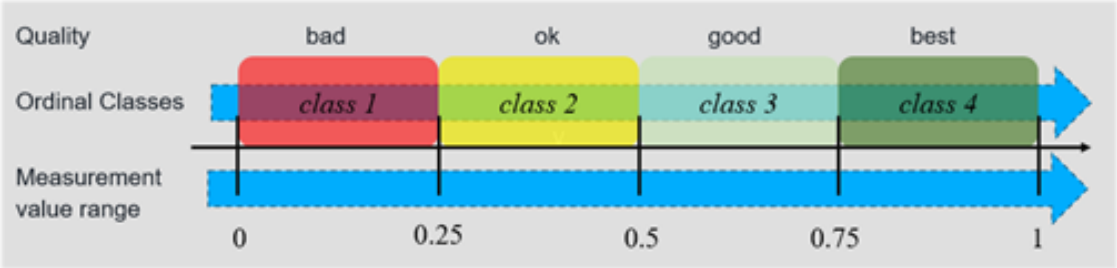


Figure 2.3: The figure illustrates how the metric space is partitioned into ordered classes. The class thresholds divide the continuous space into disjoint ordinal categories.

As visible in the figure, the thresholds of a class can be used as a metric representative for the class and compared with the metric target on a normalized domain. Through the fuzzy logic adjustment, both components of the loss function are thus compatible and computable within a common domain. The piecewise-defined function for the pseudo-classification loss follows, where $\hat{y}$ is the predicted value and $class^{threshold}$ represents the closest threshold of the actual class.

$$\text{dist2trClass}(\hat{y}) = \begin{cases} (\hat{y} - \text{class}_y^{\text{lower b.}}), & \text{class}_{\hat{y}} < \text{class}_y \\ (\hat{y} - \text{class}_y^{\text{upper b.}}), & \text{class}_{\hat{y}} > \text{class}_y \\ 0, & \text{class}_{\hat{y}} = \text{class}_y \end{cases} \tag{2.8}$$

The function, referred to as distance to true class, represents the absolute distance between the metric point estimate or regression prediction and the nearest threshold of the actual class. By minimizing this loss, the model learns to reduce the absolute distance to the true class, eventually reaching the point where $dist2trClass(\hat{y}) = 0$, meaning the regression prediction lies

within the correct class and matches the actual class. For the new custom loss function, the complete form with both components is:

$$custom_{loss} = \alpha * MSE + (1 - \alpha) * dist2trClass(\hat{y})^2 \tag{2.9}$$

The parameter α is used as a hyperparameter in the custom loss function and defines the convex linear combination. Depending on this new hyperparameter, the model is influenced, and the components are weighted accordingly. α is particularly important as a hyperparameter because, like other hyperparameters, it can be tuned based on the training data during the tuning process. The convex combination, due to its twice differentiable nature, allows the application of GD for optimized model training.
Unlike conventional approaches to epistemic uncertainty, this concept does not involve directly quantification. By incorporating data-independent information, which would without this approach be difficult to quantify, a purely data-dependent model is provided with the necessary input to implicitly account for uncertainties. It is assumed that the data-independent information either confirms the uncertainties present in the data or increases the uncertainties when there are significant deviations between the data and the data-independent information. Epistemic uncertainty is equivalent to a lack of information, and the unknown information to the model is intended to reduce this gap. In principle, these new pieces of information could also be introduced into the model as a new feature. However, in model training, this feature—partly generated from subjective information—would not be directly controllable and could have an unpredictable impact on the training process, particularly when using standard regression methods. In contrast, using our concept, the influence is directly controlled by the α hyperparameter, and the weighting of this data-independent information is tuned based on the data. The weighting can also be manually set, allowing direct influence on the train-

ing process. An example of this would be quantifying expert knowledge on output classification based on quality thresholds (epistemic), providing the model with new information on uncertainties. This indirectly reduces model uncertainty without requiring explicit quantification. The optimization of model training with respect to new information and the acceptance of bias in data-dependent regression thus depends on the quality of the new, potentially subjective information. Subjective knowledge from experts or similar sources can vary and, unlike data, may not always be easily comparable. To mitigate this variability, the previous concept is extended in the next step by adding restrictions. Specifically, data-independent information, such as subjective expert assessments, is still used, but its acquisition is now restricted.

As before, the concept involves quantifying uncertain information through the use of fuzzy logic. Previously, it was possible to entirely reclassify any output. For example, in the earlier concept, an output deterministically classified into (quality) class 2 (on a scale from 1 to 5) based on measurements could be reclassified into class 4 based on further data-independent knowledge. This subjectivity presents risks, which the new concept aims to reduce. The new approach not only treats the output as a deterministic result and target of modeling but also incorporates known measurement uncertainties. These uncertainties can be determined by experts or using statistical methods. In this concept, reclassification based on subjective information is restricted by measurement inaccuracy. Specifically, only outputs near the threshold of their classes and whose measurement uncertainty exceeds these threshold regions can be reclassified. Thus, measurement uncertainty is represented, for instance, as a boxplot to show whether it is theoretically possible that the measurement inaccuracy could lead to misclassification, and whether reclassification is justified. Figure 2.4 illustrates this concept of uncertainty-based reclassification, using measurement uncertainty to determine the permissibil-

ity of class boundary transitions.

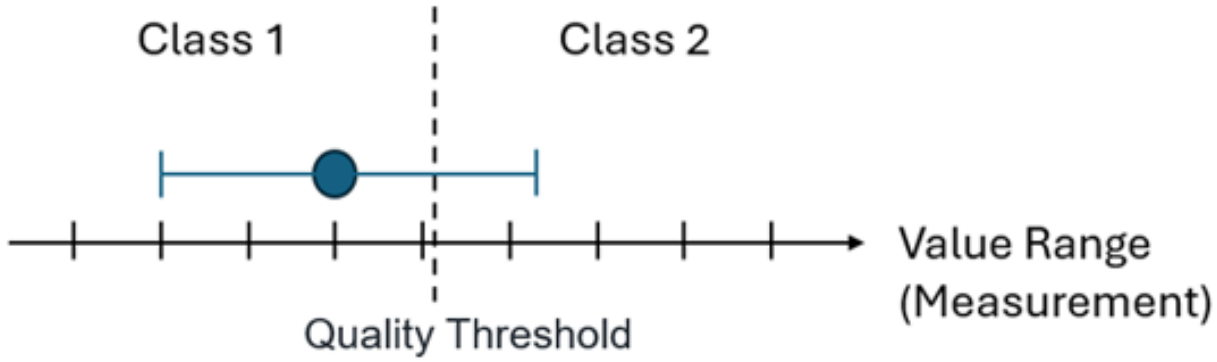


Figure 2.4: The figure shows a one-dimensional continuous measurement range divided into two ordinal classes based on a quality threshold. A measured value lies on the left side, within class 1. Due to measurement uncertainty, the value is represented probabilistically rather than deterministically, illustrated by whiskers of a boxplot extending left and right. The right whisker crosses the threshold, indicating that, probabilistically, the measurement can belong to class 2 with a certain likelihood.

GD is an optimization technique used to find a local minimum of a convex function [Ruder, 2017]. In machine learning, GD is employed to find the parameter values that minimize a given loss function. For optimization using the GD method, the first derivative (gradient) is necessary. Second-order methods, such as Newton's method, additionally require the second derivative (Hessian matrix). GD aims to determine the minimum of a loss function through iterative steps. In each iteration, the current position a is updated using the learning rate γ and the value of the gradient function at a to calculate the next position b. This is expressed mathematically as:

$$b = a - \gamma * \nabla f(a) \tag{2.10}$$

In the specific case of the customized loss function for machine learning, the update step becomes:

$$\theta_{new} = \theta_{old} - \gamma * \nabla custom_{loss}(\theta) \tag{2.11}$$

where θ represents the model parameters, and $custom_{loss}$ is the combined loss function as defined earlier:

$$custom_{loss} = \alpha * MSE + (1 - \alpha) * dist2trClass(\hat{y}) \tag{2.12}$$

The gradient of the customized loss function is computed with respect to the model parameters. This enables the iterative adjustment of the model's weights or parameters to minimize the combined loss function, balancing the contributions of the regression loss (MSE) and the pseudo-classification loss ($dist2trClass$) according to the α hyperparameter. The structure and behavior of the customized loss function are graphically illustrated in Figure 2.5, providing a visual representation of its key characteristics.

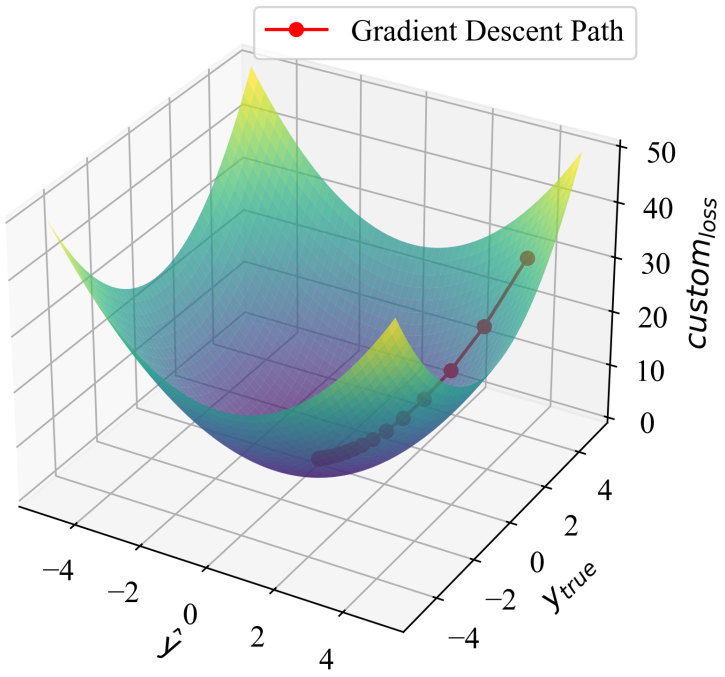


Figure 2.5: 3D plot of an upward-opening paraboloid illustrating a twice-differentiable loss function used for gradient descent optimization in regression training. The loss function is composed as a linear combination of two individual loss components, representing the final objective to minimize. The red curve depicts the optimization path following the gradient descent method along the paraboloid.

The convex custom loss function is designed for use with gradient descent (GD) in a supervised learning regression framework. It supports the application of various regression methods, including linear regression, random forest, and XGBoost. Among these, linear regression is particularly well suited as an illustrative example due to its simplicity and analytical tractability. While a detailed explanation is omitted here, the method serves to demonstrate how the custom loss function integrates with standard regression models during training. Figure 2.6 illustrates how a linear regression model can be applied to a realistic dataset and how continuous target values are discretized into ordinal classes.

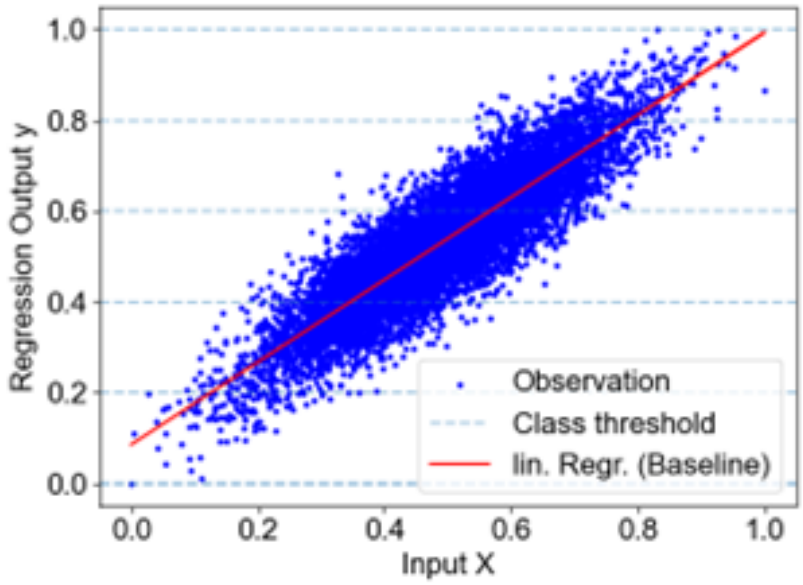


Figure 2.6: Linear regression model fitted to a realistic example dataset representative of the intended application context. The red line shows the estimated regression function, while the dashed horizontal lines illustrate the discretization of continuous target values into ordinal output classes.

Linear regression, therefore, allows for the application of the customized loss function. As described, this directly influences the learning process by biasing the best linear unbiased estimator (BLUE) — represented by the red line —

to achieve higher accuracy on the classification target. This is only feasible when the metric and classification targets are based on different distributions. If the distributions are identical, the classification target merely constitutes a coarse approximation of the metric target, leaving no potential for further optimization. As with standard linear regression, model training with the custom loss is accomplished by successively minimizing the loss. Since both components of the loss are weighted during training depending on α, this enables a direct influence on the biasing of the BLUE estimator. Figure 2.7 visualizes the effect of varying the α parameter in the custom loss function on the regression outcome.

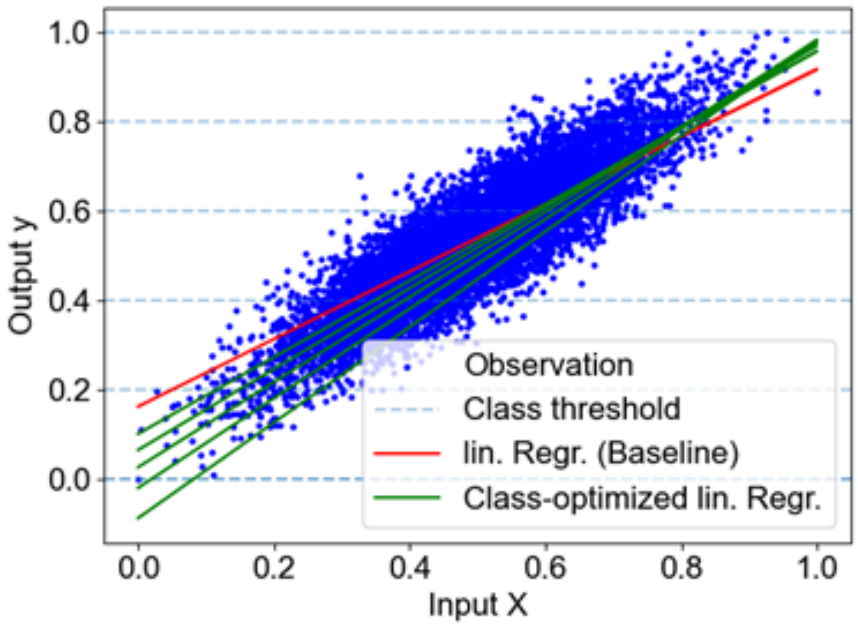


Figure 2.7: A family of regression lines fitted to a realistic example dataset representative of the intended application context, illustrating the optimization process governed by the α parameter in the custom loss function. The green lines represent model fits for different values of $\alpha \in [0, 1)$, while the dashed horizontal lines indicate ordinal class boundaries derived from continuous target values. The red line corresponds to $\alpha = 1$ and represents the standard linear regression fit based solely on the metric target.

2.4 Enhancing Model and Uncertainty Predictions with Latent Probability Distributions for Ordinal Classes of a Metric Target

Having previously described methods that incorporate non-latent influences through ordinal classes in supervised regression learning in conjunction with uncertainty quantification, we now consider latent influences [Borsboom, 2008]. Metric target variables can be categorized and approximated by ordinalization (as a form of sorting metrical values into ordinal classes). This approach offers benefits such as privacy and anonymity, but also drawbacks, such as information loss depending on the number of classes into which the metric information is approximated [Ghinita et al., 2007]. Using these classes (e.g., one-hot encoding) as feature in a regression model may adequately represent the inherent information in case of low complexity. However, with higher complexity, categorical variables may not provide the same value as a metric variable. In cases where a highly relevant variable, potentially with a direct causal relationship to the target, is only available in ordered categories, the model may not optimally utilize this information [Liddell and Kruschke, 2018]. Analyzing data using common machine learning methods often already yields good results and predictions. The limitation of such methods is often the need for a sufficiently large training dataset to make generalized statements [Rajput et al., 2023]. In this approach, we consider the case where only a relatively small, precise sample is available for training (the small-sample-size problem). Depending on the complexity, it can be challenging to identify latent distributions of the metric target variable (i.e., underlying, unobserved patterns) with this very small sample. We distinguish between two use cases. In the first case, in addition to the metric target, the training dataset also includes coarse ordinal classes directly related to the metric target for both

test and training data. These classes, for example, anonymize the target variable. In the second case, no classes are given. These are classified based on the training sample and existing class boundaries to determine the previously provided information retrospectively. The main difference from the first case is that no class information is available for the test data. Therefore, it must be estimated using standard regression methods, such as linear regression. In both cases, this results in a training dataset with categories and metric targets and a test dataset with categories. In the case of a small sample size problem (i.e., challenges arising from limited training data), there is a possibility that the model overgeneralizes and thus underfits [Aliferis and Simon, 2024]. To ensure that the model still considers the latent distribution of the metric target instead of only the ordinal category, we approximate these distributions based on the classes. The proposed methodology extends the input feature to include ordinal classes to represent the probability that the observation or target could belong to each class. Thus, we provide the model with the latent target variable distribution based on classes and the small sample, enhancing the model. Regarding uncertainty quantification, this approach allows us to model the uncertainty of this feature and, consequently, the indirect latent distribution, which may also directly affects the epistemic uncertainty of the prediction.

Since the newly obtained distribution information is incorporated into the model as an input feature, the method can be applied to any supervised regression learning scenario.

The application of this methodology extends to any small-sample-size problems where additional modeling of the latent distribution based on classes is feasible. Essentially, the method was developed to trace anonymized data, back to the metric target using a very small training dataset through approximation. The idea is that it is possible to accurately trace seemingly

anonymized data back to the metric target with very few data points, thereby revealing the anonymity.

In addition to its advantages, the method also has limitations. Machine learning methods are very effective at detecting latent distributions of the target variable, there is a sufficiently large and generalizable dataset [Aliferis and Simon, 2024]. Our method, therefore, offers added value primarily in cases of very small datasets, where additional modeling using ordinal classes is beneficial. However, the method also faces limitations when dealing with extremely small datasets beyond a certain point. For instance, if no meaningful distribution can be approximated due to imbalanced class data.

3 Discussion and Outlook

State-of-the-art machine learning and statistical methods overlap in many areas, but there are fundamental differences in their objectives and approaches. Combining these two fields offers promising synergies and remains an exciting area for future research. In this work, we explored approaches that can improve both the application of machine learning to input data and the reliability of its results.

We first examined how Bayesian methods can assess and improve the quality and quantity of input data, particularly concerning aleatoric uncertainties. Building on this, we developed a new loss function that optimizes for scenarios where the metric and categorical target variables diverge. This optimization is based on the insight that, due to aleatoric uncertainties, target objectives may differ even within the same underlying population. The assumption of differing frequency distributions between a dependent regression target and an ordinal classification target led to an approach where the latent probability distribution is modeled and used as additional input to influence the regression model directly. This also allows for a more comprehensive assessment of predictions and their epistemic uncertainties under these distributional shifts.

The papers included in this dissertation address the impact of aleatoric and epistemic uncertainties on modeling, prediction, and evaluation. The methods we developed follow an end-to-end approach for machine learning applications, focusing on optimization and uncertainty quantification. Starting with the expansion of input data while accounting for uncertainties, improving model fitting when metric and categorical targets diverge, and finally incorporating latent probability distributions to enhance both modeling and uncertainty quantification.

This dissertation contributes to the current literature by combining machine learning with statistical methods. It specifically explores how scale levels and data aggregation impact input data and how statistical probability distributions can improve forecasting and reduce uncertainty.

However, the methods presented also have limitations. For instance, generating input data based on Bayesian approaches should only be applied to a certain extent, as excessive uncertainties may limit the usefulness of the data. When optimizing for two targets with potentially different distributions, these distributional differences must be explicitly present. The approach is therefore only applicable under specific conditions. Moreover, the research on handling multiple distributions due to aleatoric uncertainties shows that prior knowledge about these uncertainties is required or needs to be estimated using other methods.

Each method presented in the papers proves beneficial in its respective use case. Nevertheless, there are several research questions that could be explored further to enhance and generalize the findings. In the first paper, it would be valuable to investigate how universally applicable a prior distribution can be across different problems, and to determine the level of uncertainty at which the generated data begins to degrade model performance. The second paper could benefit from a larger baseline study to evaluate how effectively the new hyperparameter can be tuned alongside others. It might also be useful to introduce a dependency constraint to prevent classification from overshadowing regression in cases of imbalanced class data. The third paper raises the question of whether aleatoric uncertainty, without prior knowledge, can be estimated using statistical methods. Finally, the fourth paper prompts further investigation into whether the use of modeled probability distributions is only effective for small-sample-size problems. Besides the available data, the choice of algorithm also plays a crucial role, so examining different

supervised learning methods across a broader range of datasets would be a valuable next step.

This dissertation was conducted in collaboration with ams Osram, and the research and applications were therefore driven by practical considerations. The results have practical relevance but are constrained by the increased complexity of the methods, which must be carefully weighed in terms of cost-benefit optimization.

Looking forward, the use of probability distributions instead of point estimates holds significant potential. Although the increased complexity and computational requirements remain a challenge, future developments are likely to address these issues, making probabilistic approaches and uncertainty quantification even more relevant. In practice, a prediction is only as good as its reliability.

4 A probabilistic [...] approximation [...] under consideration of measuring inaccuracy and model uncertainty

https://doi.org/10.24425/opelre.2023.145863

Declaration of Author Contributions The conception and design of the study were primarily developed by the first author, with contributions to the ideation phase from Fabian Suhrke and Kathrin Meindl. Data collection, data analysis and interpretation, as well as the literature research, were independently conducted by the first author. The drafting and revision of the manuscript were also carried out by the first author. Throughout all stages of the work, Professor Christian Heumann provided substantial support, close supervision, and critical guidance.

Opto-Electronics Review

journal homepage: https://journals.pan.pl/opelre

A probabilistic approach for approximation of optical and opto-electronic properties of an opto-semiconductor wafer under consideration of measuring inaccuracy and model uncertainty

Stefan M. Stroka[1,2*], Christian Heumann[1], Fabian Suhrke[2], Kathrin Meindl[2]

[1]Department of Statistics, Faculty of Mathematics, Informatics and Statistics, LMU Munich, 80539 Munich, Germany
[2]ams-OSRAM International GmbH, 93055 Regensburg, Germany

Article info

Article history:
Received 24 Oct. 2022
Received in revised form 14 Mar. 2023
Accepted 30 Mar. 2023
Available on-line 12 May 2023

Keywords:
Gaussian process regression;
machine learning;
uncertainty quantification;
photoluminescence;
opto-semiconductor wafer measuring.

Abstract

This paper presents a probabilistic machine learning approach to approximate wavelength values for unmeasured positions on an opto-semiconductor wafer after epitaxy. Insufficient information about optical and opto-electronic properties may lead to undetected specification violations and, consequently, to yield loss or may cause product quality issues. Collection of information is restricted because physical measuring points are expensive and in practice samples are only drawn from 120 specific positions. The purpose of the study is to reduce the risk of uncertainties caused by sampling and measuring inaccuracy and provide reliable approximations. Therefore, a Gaussian process regression is proposed which can determine a point estimation considering measuring inaccuracy and further quantify estimation uncertainty. For evaluation, the proposed method is compared with radial basis function interpolation using wavelength measurement data of 6-inch InGaN wafers. Approximations of these models are evaluated with the root mean square error. Gaussian process regression with radial basis function kernel reaches a root mean square error of 0.814 nm averaged over all wafers. A slight improvement to 0.798 nm could be achieved by using a more complex kernel combination. However, this also leads to a seven times higher computational time. The method further provides probabilistic intervals based on means and dispersions for approximated positions.

1. Introduction

Nowadays, the Bayesian analysis is already being successfully applied in many research areas, like social sciences, ecology, genetics, medicine and more [1]. The particular characteristics of this probabilistic approach is that both observed and unobserved parameters receive a joint probability distribution. The Bayesian approach, which is based on the Bayes' theorem, is thus not only a model based on input data but also extends this with available knowledge about known model parameters [1]. After epitaxy of an opto-semiconductor wafer (wafer), important properties, like brightness or forward voltage, are measured by a destructive method, whereas other properties, for example the wavelength, can be measured non-destructively. The sampled measurements of the destructive process usually cannot completely and accurately reflect the properties of the entire wafer. To reduce the risk due to non-measured wafer positions, an approximation method is proposed. In the context of production, the given data fulfil the properties of spatial data. Many forms of approximations of spatial data have been used in the past. These algorithms are divided into deterministic methods, such as kernel approximation or spline interpolation, and stochastic methods, like spatial structure functions or radial basis functions (RBFs) [2]. Other frequently used algorithms are

*Corresponding author at: stefan.stroka@ams-osram.com

https://doi.org/10.24425/opelre.2023.145863
1896-3757/ Association of Polish Electrical Engineers (SEP) and Polish Academic of Sciences (PAS). Published by PAS

the local neighbourhood approach and the variational approach [3]. Although these established methods usually provide decent results, all methods are only able to provide approximations for certain wafer positions. However, production is affected by uncertainties due to measurement inaccuracy. A pointwise interpolation method does not consider variability in measuring. Thus, only measurement points are used instead of intervals considering input data uncertainty. Gaussian process regression (GPR) as a Bayesian approach can include uncertainties within the observed data and within the model itself through the joint probability distribution and, thus, constitute an approximation considering aleatoric and epistemic uncertainties.

1.1. Application-based GPR for approximation of a wafer

The basis of GPR is the selection of a prior mean and a covariance matrix or a covariance kernel function. This allows subjective knowledge about the wafer measurement to be used as prior information. At this point, it is also possible to consider input uncertainties as normally distributed errors within the prior. Based on the prior probability distribution and the likelihood, the algorithm results in a posterior probability distribution. This posterior follows a multivariate normal distribution which can be used to approximate values for unknown positions on the wafer. The prediction is made by weighting all possible predictions with the posterior distribution. The results are conditionally normally distributed predictions defined by their means and covariances [4].

1.2. Related research

In 2020, Barnes and Henn [5] compared machine learning (ML) algorithms such as RBF interpolation and GPR with a straightforward library lookup method for optical critical dimension (OCD) metrology. In this work, it is described that already 32 training points are sufficient for the ML method to be better than a library search. Schneider *et al*. [6] considered a Bayesian optimisation approach based on a GPR. Numerical simulation is used to reproduce measurement results of periodic micro- or nanostructures. These simulated structures are then described by optimised geometry parameters (geometry reconstruction). An earlier approach from 2015 by Henn *et al*. [7] attempts to obtain reliable estimates for quantitative characteristics of three-dimensional structures and associated realistic uncertainties by optimisable hybrid measurement techniques. A measurement method with a probabilistic prior and an approach with measurement methods combined through regressions are compared. Chen *et al*. [8] use an approach to increase the measurement accuracy of an optical scatterometry by using a fitting error interpolation-based library search method. A fitting error value is used to describe the wafer for a library search. Reference wafers are then those with the minimum difference in fitting error.

1.3. Aim of the paper

Destructive measurement methods can determine the opto-electronic properties of a wafer very well, but they are correspondingly cost-intensive and, therefore, increasing the number is not feasible. In practice, as well as in theory, methods for approximating or improving measurements in the field of opto-semiconductors are already being considered and applied. These approaches are mostly based on point estimators. The complex production of these wafers through the epitaxial process is difficult to control and measurement inaccuracies can also bias these values. A point estimator can deliver decent results compared to the test data but is not able to consider variability in measuring or systematic uncertainty. The aim of this work is, therefore, to get reliable approximations for the wafer measurements based on a probabilistic ML approach. The ML method focused on is the GPR, which is expected to provide robust point estimators and further quantifies uncertainty in the input data, as well as in the model. Comparisons are made with a state-of-the-art baseline method.

1.4. Paper organization

First, the necessity of new approaches in the context of production is shown in section 2. Second, the GPR and the baseline method are described in section 3. The experimental set-up is presented theoretically and practically in section 4. In section 5, results are presented and discussed, before a summary and an outlook for the future steps are given in section 6.

2. Front-end production optimisation with a Bayesian approach

Production at ams-OSRAM aims to manufacture high-quality opto-electronic semiconductors. Therefore, the front end of the production uses an epitaxy process to produce an epitaxial wafer from a substrate wafer as base carrier material. In the following sections, the paper focuses exclusively on a nitride-based process. This production process uses silicon carbide (SiC) as a base substrate or carrier material, respectively and grows gallium nitride (GaN)-based devices on it using metalorganic vapour phase epitaxy (MOVPE) [9]. According to Härle *et al*. [9], it is important for industrial production that, in addition to a stable epitaxy process, also a cost-effective chip technology is developed.

2.1. General idea and description

In the context of the opto-semiconductor process in a front-end production, different measurements are used to monitor the production step and achieve the best possible yield in subsequent further production steps. Measurement procedures are not part of the value chain. Hence, the goal of manufacturers is a maximum information gain with a minimum effort. The fundamental idea is to solve the problem of increasing the amount of information that is making predictions on unmeasured wafer positions, at the lowest possible additional cost using a state-of-the-art ML method and at the same time providing information about the reliability of the predictions based on a Bayesian approach.

2.2. Measuring systems

This paper considers the process steps after epitaxy and the subsequent measuring before further processing. In this regard, several tests are carried out to gain necessary information for further production steps. The main one is the so-called quick test (QT). The QT is a time-consuming and destructive procedure that provides information about the opto-electric properties. Since QT measured points are destroyed during the process, only a few points on the wafer (approximately 120) are tested. Increasing the number is, therefore, often not feasible. A non-destructive method is the photoluminescence (PL) measuring, which measures optical properties, like the wavelength. Here, information is obtained by irradiating the epitaxial surface of the wafer by photoexcitation [10]. PL measuring does not destroy the measured point but is also less accurate compared to the QT measuring.

3. Machine learning algorithms

In terms of application, a multivariate regression is needed to infer information from a higher dimensional space. These dimensions separate in our case into a spatial basis and associated measurements of a wafer. GPR is an approximation algorithm based on spatial dependencies of measurement points [11]. Therefore, the methodology of GPR and its probabilistic properties are described below. Furthermore, a multivariate approximation method based on RBFs is introduced and used as a comparison algorithm [12]. In general, the proposed algorithms can be applied to destructive and non-destructive measurement methods. To evaluate the analysis, the data set with (non-destructive) PL wavelength measurements is chosen.

3.1. Gaussian process regression

GPR is a non-parametric, probabilistic ML approach. The method is determined by Gaussian processes (GP) and uses Gaussian probability distributions. Instead of pointwise estimators, the probabilistic properties result in a distribution for the predictions. This allows to quantify uncertainties [13]. GP are stochastic processes with a finite set of random variables. Each random variable is a linear combination of normally distributed random variables and has, therefore, also a multivariate normal distribution. The paper applies the module Scikit-learn in Python [14], which is based on the presentation of Rasmussen [13]. The goal of GPR is to extract the information inherent in the observation without noise. For this purpose, the GP as multivariate normal distribution is used to model the observation without noise ε. The probabilistic GPR is defined by a posterior probability distribution. According to the Bayes' theorem, the posterior is determined by a prior distribution and the likelihood of actual observations Z. Let $Z = \{Z_i\}_{i\in I} = \{Z_i = (x_i, y_i)\}_{i\epsilon I}$ be the observed data with x_i and y_i as the coordinates for the wafer position, I the associated finite index set and f the GP. Assuming that the observations without inherent noises can be represented by $f(Z)$, it follows:

$$\forall_{i\in I}: w_i = f(Z_i) + \varepsilon_i = f((x_i, y_i)) + \varepsilon_i, \quad (1)$$

with w_i as the wavelength measurement. The prior distribution for each i in the sequence of random variables $\{w_i\}_{i\in I}$ is normal since the error terms ε_i are normal and, therefore, defined by a mean and a variance. Thus, for the multivariate GP, the prior means are given by the expected value function $m_i = E(w_i) = f(Z_i)$ of the observations w_i, while the prior variance must be predefined as a covariance matrix, also called a kernel. For the application, two different kernels are considered with different levels of complexity which are introduced in section 4.1.

3.2. Radial basis function interpolation

RBF interpolation is a method for smoothing or multivariate interpolation of higher-order unstructured data [12]. The algorithm is based on RBFs or, equivalently, radially symmetric basis functions. According to Buhmann [12], a function is radially symmetric if the function value depends solely on the Euclidean distance from the origin. It follows that every function for which $\varphi(x) = \varphi(\|x\|)$ occurs is an RBF. Let Z be again the set of observations and f the inherent function without error ε_i. The aim of the RBF interpolation method is a continuous function s with the property

$$s(Z_i) = f(Z_i)\ \forall_{i\in I} \quad (2)$$

which means that every training point Z_i as support point is met by the interpolation function s and s evaluated at Z_i equals the true value w_i without error ε_i for every Z_i. The algorithm defines the function s as a linear combination of basis functions. Let every φ_i be a basis function, which fulfils the condition of an RBF function, then

$$s(z) = \sum_{i\in I} \lambda_i \varphi_i(z), \quad (3)$$

with the scalar λ_i for every z within the interpolated value range. According to Fasshauer [15], this linear system can be solved uniquely only if the basis functions used are radially symmetrical. For the practical evaluation, the implementation of Scikit-learn [14] in Python is applied.

3.3. Uncertainty quantification

The key aspect of this paper is the quantification of uncertainty. While there are methods like 5-fold cross-validation that allow point estimation algorithms, such as RBF interpolation, to determine prediction uncertainty, these are not feasible in practical applications in the context of the experiment because of data sparsity. Measuring points are expensive and, therefore, only few data points per wafer are available. Hence, this paper focuses solely on the uncertainty quantification by GPR. Uncertainty quantification in a Bayesian approach is divided into two categories. The uncertainty within the data is called aleatoric uncertainty. In a physical approach, this is directly related to the measuring inaccuracy resulting from the measuring process. The second is the model uncertainty, also called systematic uncertainty. This quantifies the lack of knowledge, which is missing from the in theory perfect model [16].

3.3.1. *Confidence interval*

A confidence interval is defined by two bounds which are random variables and depend on a confidence level $\gamma = (1 - \alpha)$ and a population of random samples. The confidence interval states that at a confidence level $\gamma \cdot 100\%$ the unknown parameter θ (e.g., the mean) is covered by the confidence interval at $\gamma \cdot 100\%$ for all repeatedly, randomly drawn samples of this distribution. In practice, the confidence interval depends only on a given population, meaning the training data. It can only quantify the aleatoric uncertainty.

3.3.2. *Prediction interval*

The prediction interval, like the confidence interval, is in this case a symmetric interval around the mean, defined by an upper and lower bound. Unlike the confidence interval, the limits of the prediction interval are determined based on the prediction error. The prediction interval uses the given information to describe which future observations of the same population are covered by the interval with a certain probability $\gamma = (1 - \alpha)$ [13]. In the case of this paper, the prediction interval will be generated by sampling functions from the optimised GP. After fitting the posterior conditional distribution on the training data, it results in a family of not necessarily identical normal distributions equivalent to GP. From this family, an appropriate number of distribution functions are drawn as samples to describe which value ranges are covered by the interval to a fixed probability with the help of percentiles. The interval boundaries are defined by continuous functions, which also provide information about new observations of the same total population beyond the training data. This enables the quantification of uncertainties in both measurement and model accuracy.

4. Experimental setup

In the following, the ML-model setup and the practical application setup are presented before they are applied in section 5.

4.1. *ML model setup*

This section describes the structure of the application in a practical case and which fundamentals must be established for a reasonable implementation. For the practical part, there is a tuple of independent variables, the position data $Z = (x, y)$ on the wafer, and the dependent variable w as the measurement value. In practice, the only task necessary for the application of the GPR is the selection of a prior kernel. The prior represents the subjective view on the dependent variable and, therefore, cannot be unambiguously determined or at least not without very high additional effort. Consequently, two promising kernels were selected for this evaluation. The RBF kernel (squared exponential kernel) k_{RBF} as a standard kernel with an optimizable scalar λ and the length scale l

$$k_{RBF}(Z_i, Z_j) = \lambda * \exp\left(-\frac{d(Z_i, Z_j)^2}{2l^2}\right) \tag{4}$$

is considered at first for the simple model. A linear combination of squared exponential kernel, rational quadratic kernel, and maternal kernel with likewise optimizable scalars

$$k_c(Z_i, Z_j) = a \cdot \exp\left(-\frac{d^2}{2l^2}\right) + b \cdot \left(1 + \frac{d^2}{2\alpha l^2}\right)^{-\alpha} + c \cdot \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}d\right) \tag{5}$$

is considered next for the complex model. Thereby, the hyperparameters to be optimised are the length scale l, the smoothness ν, the scale mixture α and a, b, c as associated scalars. Further applies $d \coloneqq d(Z_i, Z_j)$ as a short form for the Euclidean distance, Γ as the gamma function, and K_ν as the modified Bessel function. The kernel combination can be generated by matrix addition and multiplication since each kernel satisfies the conditions as a covariance matrix [17]. Figure 1 shows the potential sample functions from GP for the respective prior distribution (kernel function). Comparing the GP sample function with a simple and complex kernel, a different degree of the GP functions variability can be seen. For more details regarding the kernels, it is recommended to compare with Rasmussen [13].

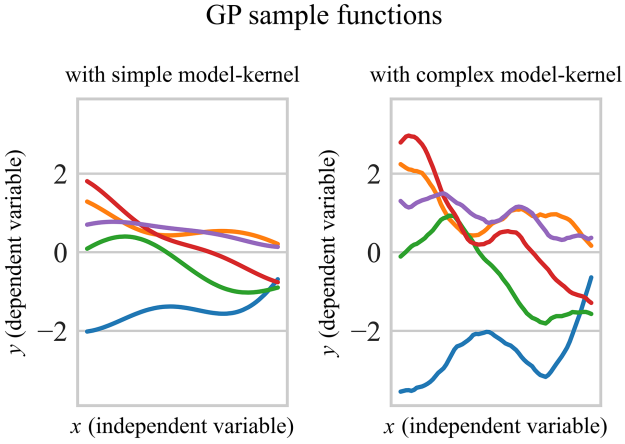


Fig. 1. Sampling y-values from GP with a given prior distribution (kernel). The GP is not yet trained and depends on mean and covariance function (kernel). Figure shows a two-dimensional prior distribution with one independent variable as an example, but more independent variables are also feasible.

4.2. *Practical application setup*

For the practical application, measurement data for a chip type in the blue colour range with nitride-based production processes were selected as an arbitrary prototype for the evaluation. The analysed wafer property in this paper is the wavelength. The QT measurement used in practice cannot be evaluated directly, as all up to 120 measurement points are necessary for training. For this reason, the PL measurement data set is used in the following showcase, as this is larger and thus test data are also given. To recreate the actual use case as realistically as possible, the PL data set is divided into test and training data. The wavelength values are measured for all given wafers of this chip type and used as the statistical population

respectively as training and test data. For each wafer, an equidistant grid is used to declare approximately 120 measurement points of the data set per wafer as training data. This grid is identical to the pattern used for QT measuring. The remaining approximately 3000 measurement points are the test data. Due to the lack of several measured values for the same position, a measurement inaccuracy cannot be estimated directly. However, experts assume a certain uncertainty in the measurement of the wavelength with QT, which is confidential and may not be specified precisely. For transparency, an inaccuracy of 0.5 nm is used throughout the paper. This measurement error is constant and not wafer position-dependent, as the measurement of the edge point is carried out identically to the point within the inner area of the wafer. The used GPR implementation allows to define a specific measuring uncertainty as prior. Hereby, 0.5 nm will be added to the diagonal of the covariance matrix of the GPR. In the process of method application, each GPR is optimised individually for each wafer, resulting in a point estimate (mean vector) and a prediction interval (variance vector as the diagonal of a covariance matrix). The evaluation of the GPR with two different kernels is carried out in comparison to the described interpolation method as the baseline.

5. Results and discussion

Firstly, the results are evaluated based on a single, randomly selected wafer and uncertainties are quantified. Secondly, the evaluation is carried out empirically by considering the data of all wafers.

5.1. Results for an arbitrary wafer

Due to instability of the epitaxial growth, higher fluctuations occur in the edge region. In Table 1, a distinction is made between evaluation on the inner wafer area and evaluation on the complete wafer to represent the performance of the GPR more accurately. The inner wafer area is covered by the equidistant grid consisting of training data. Each measurement point in the complete test data set is part of the inner test data set if it lies within or on the perimeter line passing through the outer points of the equidistant grid. An insight into the results for one arbitrary wafer is given in Table 1.

Table 1.
Results of model fitting and prediction for an arbitrary wafer.

Method	LMLH[a]	RMSE (nm)		CT[b] (s)
		inner area	wafer	
RBF (baseline)	–	0.478	2.001	2.07
GPR (simple kernel)	−159.99	0.844	2.156	7.11
GPR (complex kernel)	−158.15	0.807	2.038	21.07

[a] log marginal likelihood
[b] computational time

Table 1 shows that based on the root mean square error (RMSE); the RBF interpolation model is performing best on both test data sets. The second-best model here is the GPR with a complex kernel. Comparing the two GPRs with different complexity, the high deviation in computational time (CT) is remarkable. Even though the GPR with the complex kernel performs better than the GPR with the simple kernel based on the RMSE and the optimised log marginal likelihood (LMLH) for both data sets, the CT almost triples. Comparing the ratio determined from RMSE divided by the wavelength median of the training data, it becomes clear that the proportional deviation and differences between them are small. The highest difference between proportional deviations is between the baseline and the GPR with simple kernel on the inner area test data set with 0.08%. It is noteworthy that exactly this difference in proportional deviation decreases when evaluating the complete test data set. The difference here is 0.034%. Thus, the GPR seems more stable than the baseline on the more difficult outer wafer area. Regarding the hypothesis, the RBF interpolation as the fastest method with the lowest RMSE can, therefore, dominate the comparison evaluations of point estimators. The disadvantage is obvious when considering uncertainty quantification, as the baseline interpolation method does not consider uncertainties. The lack of uncertainty quantification allows the method to compute the point estimator much faster than the GPR. Generally, GPR should still be preferred for practical purposes, since firstly, the deterioration in RMSE for point-wise regression is relatively small and secondly, a higher computational effort can be justified by a description of the model reliability. The following cross-sections from the wafer surface are used to obtain an insight into the regression and uncertainty quantification. These chosen sections are marked with red (vertical cross-section) and blue (horizontal cross-section) coloured lines within Fig. 2(b) and illustrated within Fig. 3. A cross-section considers the model and its results reduced by one dimension by setting one of the two independent variables x, y to zero. For the following graphical evaluation, the GPR model of the simple kernel is used for demonstration purposes. However, the same evaluation can be done with the complex kernel. Figure 2 shows all given historical data of this selected wafer in Fig. 2(a) and the uncertainty quantification with GPR for the same wafer in Fig. 2(b). The uncertainty is evaluated by using the standard deviation of the respective covariance matrix. A small standard deviation indicates a rather high certainty of the GPR model. Furthermore, those two certain cross-sections of the GPR will now be focused on. In addition to a point estimate, the GPR provides an uncertainty quantification in the form of a covariance matrix related to the wavelength as dependent variable, conditional on the position data. Figure 2(b), therefore, shows the top view of the model uncertainties of the GPR resulting from the regression on a selected wafer. This illustrates that the position dependence, respectively the direct distance to the closest training data point, is decisive for the reliability of the model. The position $(x, y) = (0, 1.1)$, in the outer area in Fig. 2 for example, shows that less training data in the area around the position results in a prediction with much higher prediction uncertainty. This can be seen in Fig. 2 by comparing the standard deviation value between the position $(x_1, y_1) = (-1, 0.59)$ for which measured values are available and the position $(x_2, y_2) = (1, 0.59)$, which is unknown during model training. In terms of practical application, this aspect is crucial, as measurement points

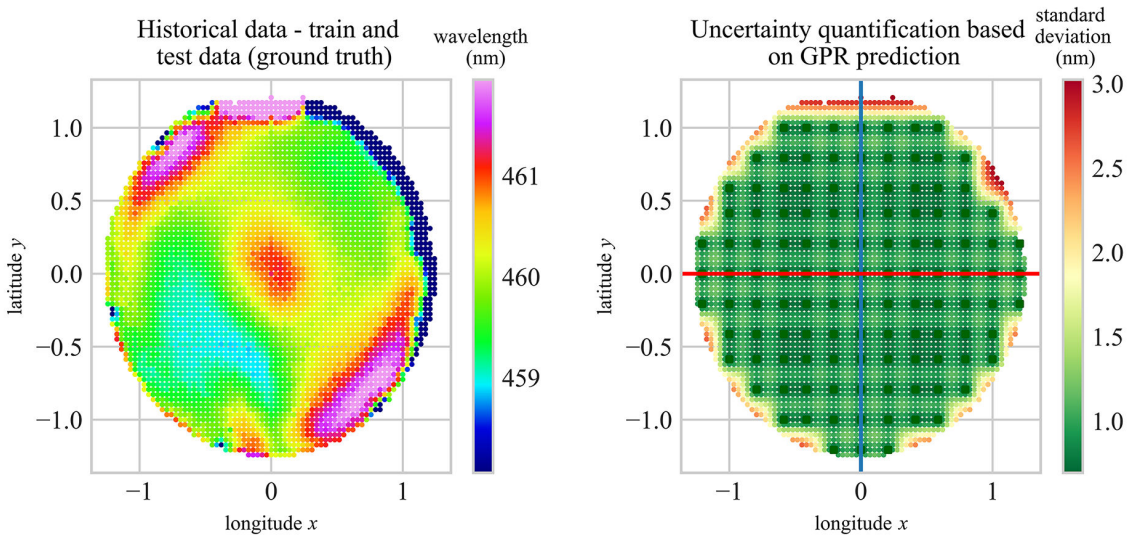


Fig. 2. Train and test data as ground truth (for this arbitrary wafer) (a) and position-based uncertainty quantified by GPR (b) in top view. Figure 2(a) displays the heterogenity of the wavelength for the given observations. GPR provides a standard deviation for every prediction, which indicates the certainty of the model for exactly this predicted position in Fig. 2(b). GPR is based on the PL train data (darkgreen dots).

may be missing in the production and the measurement process for undefined reasons. Information about the reliability of the model at these and surrounding measurement points is therefore essential. Now, the graphical observation is reduced to a section, for this purpose the longitude $x = 0$ in Fig. 3(a) and the latitude $y = 0$ in Fig. 3(b) are set to display the results in a side view. Both models are trained with the complete training data set (not only with data of each cross-section). Figure 3(a) and Fig. 3(b) show the point estimation and the confidence and prediction intervals resulting from the GPR. The graphical comparison of the RBF interpolation curve and the GPR prediction curve for the vertical cross-section in Fig. 3(a) shows that both are almost identical for most of the definition range. Deviation can only be observed in the boundary areas $[-1.21, -1.0]$ and $[1.0, 1.21]$. These deviations become more obvious when comparing these point estimators to the test data. Looking at the model uncertainties, the right border area $[1.0, 1.21]$ has high uncertainties, which results from a missing measurement point at $y = 1.21$. The model, therefore, extrapolates at this point. The same applies to the horizontal cross-section in Fig. 3(b). Both methods are often visually approximately congruent, yet both cannot completely reproduce the test data without deviation. The right border $[0.5, 1.21]$ is noticeable. Here, the RBF interpolation and the GPR prediction diverge strongly in some cases, and yet both fail to recognise the actual trend. This results from the fact that

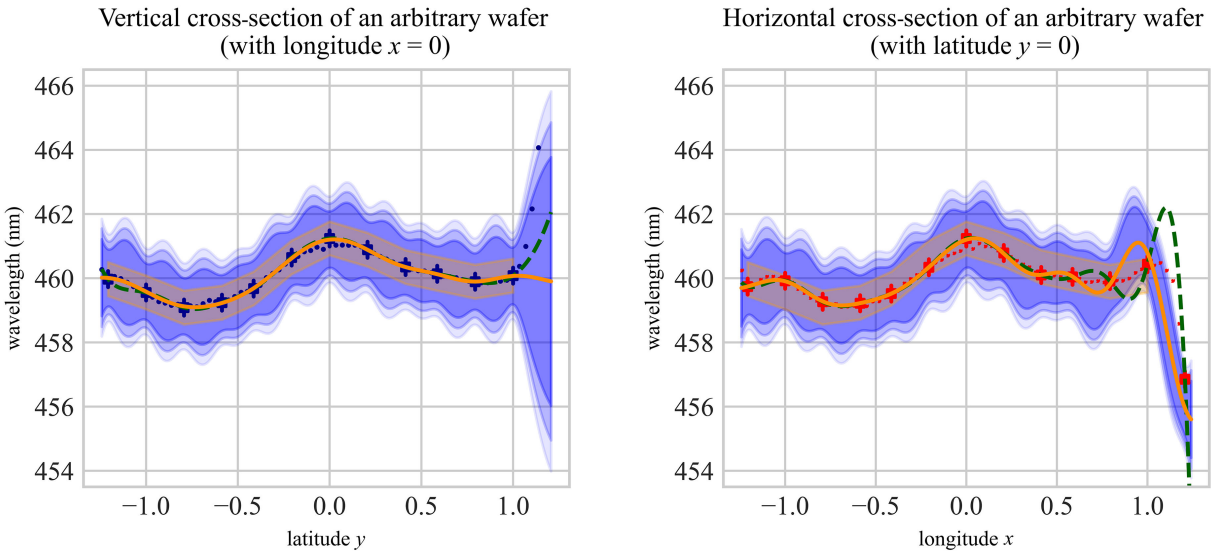


Fig. 3. Two different cross-sections of the same wafer. Both figures show the RBF interpolation (green line), the GPR (orange line), and the real observation, respectively the test data (left: darkblue/right: red dotted line). The results of the green and red line methods are similar for a large part of the range of values where the lines overlap. Each model is trained with the identical training data set. Squared points indicate the PL train data on the cross-section (left: darkblue/right: red squares). Horizontal lines through the training data indicate the fixed measuring inaccuracy (0.5 nm). The narrow interval across the value range (orange) shows the confidence interval for 97.5% probability, whereas the broader intervals (different shades of blue) mark the prediction interval for different probabilities (from the inside out: 90%, 95%, and 97.5%).

both models aim to consider the relatively low wavelength value at $x = 1.21$ from the training data. It is also noteworthy that in Fig. 3(b) in comparison to Fig. 3(a), there is another training point at the right-hand border, from which it follows that the model uncertainty in this area is significantly lower. The relevance of the prediction intervals for the application should be pointed out once again. By method application, there is no longer only one measured value in production that is close to reality only in the optimal case, but an interval range that covers reality with a fixed probability compared to the GPR model. Furthermore, it also becomes graphically clear that the difference between the baseline and the GPR is only small in relative terms and can be neglected regarding the added value due to the prediction interval.

5.2. *Limitation of the ML methods regarding the practical application*

In section 5.1, it is assumed, that approximately 120 measuring points of a wafer are used as model training points. In practice, measurement points can get destroyed during production or measuring. Figure 4 shows that each method needs a certain minimum number of training points to achieve good results.

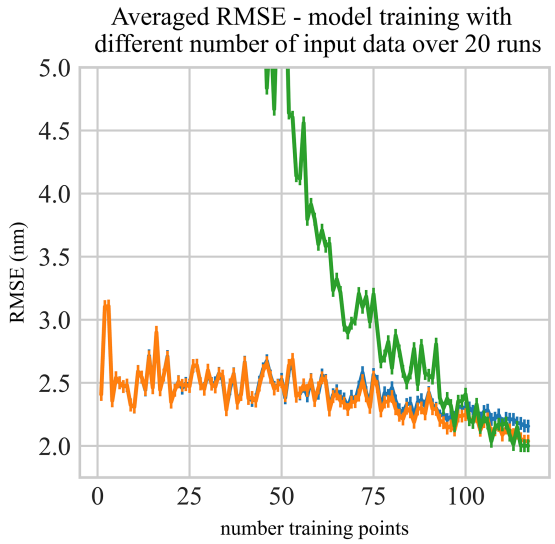


Fig. 4. Evaluation based on averaged RMSE with different numbers of input data. Each model is trained with a fixed number of random sampled training points (x-axis). To obtain robust results, 20 runs are performed for each number of random sampled training points. Sampling is done to simulate the loss of a training point. The evaluation is based on the complete wafer test data set with about 3000 points. The figure shows the RBF interpolation (green line), the GPR with simple kernel (blue line), and the GPR with complex kernel (orange line).

The evaluated models and kernels are trained with different numbers of input data. For this purpose, the respective number of training points is randomly drawn from the training data used in section 5.1. For each fixed number of training points, 20 identical models are trained with different samples and then the mean RMSE is computed. This is necessary because the random sampling of training points for each model has a strong influence on the prediction model. Figure 4 thus shows the tendency of all three approaches to worsen when the number of input data is reduced. In the range above 100 training points, there is hardly any decrease in RMSE value resulting from fewer points. With less than 100 training points, however, a strong deterioration of the RBF interpolation becomes apparent. This worsens up to a maximum RMSE of approx. 14 000 nm, for that reason it cannot be illustrated nicely in Fig. 4 and is therefore truncated. In the comparison of the two GPRs with different kernels, a slight tendency towards deterioration is recognisable in both. It is also noticeable that the more complex kernel cannot deliver an improvement compared to the simple kernel below the minimum number of points. Even though the RMSEs are averaged, the variability of the results increases for smaller numbers of training points. Although this is a limitation for the GPR, the RBF interpolation becomes significantly worse and unreliable at less than 100 training points and below. This is one major advantage of the GPR over the baseline, since in practical application not every single wafer can be checked on its own. GPR provides reliable results even if the number of wafer measurements is exceptionally below 100 points.

5.3. *Empirical results over all wafers*

Table 2 shows the lowest RMSE and subsequently the best overall point estimation achieved by the baseline model. Comparing the GPR models, the smaller LMLH value shows that the model with the complex kernel achieved a much better model fit on train data. However, this is not reflected by the mean of RMSE values related to the test data where both results are similar. When looking at the CT, GPR with the simple kernel takes about 53 times longer than RBF interpolation. The ratio is even higher for the GPR with complex kernel, where it needs about 366 times as much CT as the baseline. It also follows that using the complex kernel instead of the simple kernel takes approximately seven times more CT.

Table 2.
Results of model fitting and prediction for all wafers from the given population.

Method	LMLH[a]	RMSE (nm)		CT[b] (s)
		inner area	wafer	
RBF (baseline)	–	0.224	0.766	0.06
GPR (simple kernel)	−116.65	0.303	0.814	3.18
GPR (complex kernel)	−100.82	0.273	0.798	22.01

[a] log marginal likelihood
[b] computational time

Figure 5 shows the RMSE between measured values and predictions of the models for the inner and the complete wafer test data sets. In relation to the evaluation results of other wafers in the population, the results for the randomly selected wafer from Table 1 are in the upper outlier range.

The overall best performing algorithm in Fig. 5 is the RBF interpolation with the lowest RMSE median for both test data sets. The GPR with a simple kernel and the model with a complex kernel are worse than the baseline at the

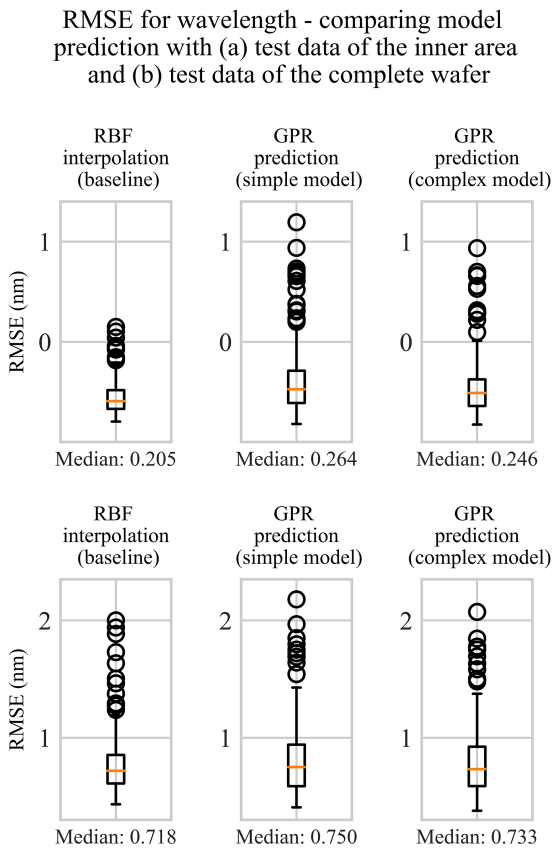


Fig. 5. Evaluation of the empirical results of all models with RMSE for the wavelength for 239 wafers. Each model is trained and tested using the measurements of a specific wafer of the inner area (a) or of the complete wafer (b).

median by 0.059 nm and 0.041 nm for the inner area test data set and 0.032 nm and 0.015 nm for the complete wafer test data. This is further confirmed by comparing the mean values from Table 2. It is noteworthy that the relative difference between the baseline and the GPR models regarding the point-wise prediction error decrease from the application on the inner area to the application on the complete wafer. This indicates a tendency, as in section 5.1, where the GPR is more stable on data with higher variability than the baseline. In summary, the RBF interpolation empirically performs better as a point-wise estimator than the GPR. Based on the results in Table 2, it is evident that the model fitting with a more complex kernel performs better on the training data but hardly represents an improvement compared to the actual test data, especially when considering the longer CT. This indicates an overfitting of the model and, consequently, a possibly unnecessarily high kernel complexity. As already stated in section 5.1, the differences in the mean and median are small in relative terms. These are also significantly lower than the measurement inaccuracy (aleatoric uncertainty). In terms of production, it can be considered not significant. The apparently very high CT compared to the baseline is not particularly noticeable in terms of a real-world application, considering that the regression of a wafer with a simple kernel only takes approximately 3 s. Nevertheless, it must be noted that this modelling was not done using powerful computers and yet a reasonable time was achieved considering the good results for the point estimator and the additional value for production through uncertainty quantification.

6. Conclusions and outlook

Highly complex industrial manufacturing relies on meaningful measurements in the production process to achieve the highest possible and most qualitative yield. Stable and reliable measurements are highly time-consuming and costly, which is why production must get along with sparse measurements and, therefore, accept uncertainties. The proposed probabilistic GPR provides point estimates considering uncertainties of measurement and model. The normally distributed GPR gives a continuous wafer map regarding the measured properties with equally continuous uncertainty quantification for the whole wafer. Although the analysis is carried out for wavelength, any other property can also be investigated in further research. The empirical evaluation for the wavelength shows that a GPR with an RBF kernel as simple kernel is sufficient to achieve an average RMSE of 0.303 nm on the inner area and 0.814 nm on the complete wafer. In relative comparison to the specified measurement accuracy of 5 nm, the fitting error is low and not significant in terms of production. The RBF interpolation as baseline method and the GPR with complex kernel surpass this result only barely with an RMSE of 0.224 nm (inner area) and 0.766 nm (complete wafer), and 0.273 nm (inner area), and 0.798 nm (complete wafer). A point estimator has often little significance when used productively, as it cannot always provide a reliable prediction. This fact poses a great risk to the goals of scrap minimisation, compliance with specification limits, and yield maximisation. A high variability due to the measuring process after epitaxy of a wafer increases this risk significantly. In detail, outliers that exceed the accepted variability limits equivalent to the prediction interval based on measurement and model uncertainties can highlight possible specification failures and thus serve as an alarm system. Measurement and model uncertainties are, therefore, important selection criteria in chip production to estimate dimension of the problem and amount of chips that will be out of specification and the associated yield loss in later production steps. Wafers are selected according to uncertainties for the best possible further processing or also for certain specifications. The GPR, unlike the baseline, has a probabilistic uncertainty quantification. Therefore, GPR directly enables a meaningful uncertainty-based classification of the output to meet the needs of the production. Against the clear advantages stands a higher CT. A GPR with a simple kernel takes on average 53 times longer than baseline interpolation for a complete wafer. Regression with a complex kernel takes even longer, at around 366 times the runtime of the baseline. In productive terms, however, GPR with a RBF kernel can determine a continuous point estimator with uncertainties of a whole wafer in about 3 s, which is why the computing times can be accepted in current applications. Even a GPR with a low-complexity kernel thus offers all the advantages necessary for production, both through exact point estimators and through the determination of uncertainties in measurement and model. Besides the focus on a safety-based categorisation of the output, it is equally important to extend the view to the overall production. The results from GPR are wavelength measurements or intervals of a whole wafer after epitaxy. These are only the results of an intermediate process step.

From an overall production perspective, the GPR results should further be used to optimise the subsequent process steps for chip production. Currently, the approximately 120 measuring points considered in the paper are used to conclude the resulting number and quality of the chips with the help of a regression approach. Viewed holistically, the GPR can thus be seen as a pre-processing step for this regression. Building on the probabilistic approach, a Bayesian regression model, such as a Bayesian neural network can also be used. Possibly, even the prediction intervals respectively to the inherent standard deviation of the GPR model could find further use as a meaningful prior distribution. Furthermore, the significantly higher number of input measurement points resulting from the model should also offer an improvement for any regression. Thus, it can be pointed out that the probabilistic GPR approach provides a solid improvement opportunity for the studied area and offers a clear potential concerning the further process optimisation.

References

[1] van de Schoot, R. *et al*. Bayesian statistics and modelling. *Nat. Rev. Methods Primers* **1**, 1–26 (2021). https://doi.org/10.1038/s43586-020-00001-2

[2] Myers, D. E. Spatial interpolation: an overview. *Geoderma* **62**, 17–28 (1994). https://doi.org/10.1016/0016-7061(94)90025-6

[3] Mitas, L. & Mitasova, H. Spatial Interpolation. in *Geographical information systems: principles, techniques, management and applications (*eds. Longley, P. A., Goodchild, M. F., Maguire, D. J. & Rhind, D. W.) 482–492 (Wiley, 1999).

[4] Rasmussen, C. E. Gaussian Processes in Machine Learning. in *Advanced Lectures on Machine Learning* (eds. Bousquet, O., von Luxburg, U. & Rätsch, G.) 63–71 (Springer, 2004). https://doi.org/10.1007/978-3-540-28650-9_4

[5] Barnes, B. M. & Henn, M.-A. Contrasting conventional and machine learning approaches to optical critical dimension measurements. *Proc. SPIE* **11325**, 222–234 (2020). https://doi.org/10.1117/12.2551504

[6] Schneider, P.-I., Hammerschmidt, M., Zschiedrich, L. & Burger, S. Using Gaussian process regression for efficient parameter reconstruction. *Proc. SPIE* **10959**, 200–207 (2019). https://doi.org/10.1117/12.2513268

[7] Henn, M.-A. *et al*. Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression. *Proc. SPIE* **14**, 044001 (2015). https://doi.org/10.1117/1.JMM.14.4.044001

[8] Chen, X., Liu, S., Zhang, C. & Zhu, J. Improved measurement accuracy in optical scatterometry using fitting error interpolation based library search. *Measurement* **46**, 2638–2646 (2013). https://doi.org/10.1016/j.measurement.2013.04.080

[9] Härle, V. *et al*. GaN-Based LEDs and Lasers on SiC. *Phys. Status Solidi A* **180**, 5–13 (2000). https://doi.org/10.1002/1521-396X(200007)180:1<5::AID-PSSA5>3.0.CO;2-I

[10] Stern, M. L. & Schellenberger, M. Fully convolutional networks for chip-wise defect detection employing photoluminescence images. *J. Intell. Manuf.* **32**, 113–126 (2021). https://doi.org/10.1007/s10845-020-01563-4

[11] Oliver, M. A. & Webster, R. Kriging: a method of interpolation for geographical information systems. *Int. J. Geogr. Inf. Syst.* **4**, 313–332 (1990). https://doi.org/10.1080/02693799008941549

[12] Buhmann, M. D. *Radial Basis Functions: Theory and Implementations*. (Cambridge University Press, 2003).

[13] Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[14] Fasshauer, G. E. *Meshfree Approximation Methods with MATLAB*. (World Scientific, 2007).

[15] Walker, W. E. *et al*. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr. Assess*. **4**, 5–17 (2003). https://doi.org/10.1076/iaij.4.1.5.16466

[16] Patel, J. K. Prediction intervals – a review. *Commun. Stat. – Theory. Methods* **18**, 2393–2465 (1989). https://doi.org/10.1080/03610928908830043

[17] Duvenaud, D. *Automatic Model Construction with Gaussian Processes*. (University of Cambridge, 2014). https://doi.org/10.17863/CAM.14087

5 Multi-Task Learning of Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes

Declaration of Author Contributions The conception and design of the study were primarily developed by the first author, with early support from Fabian Suhrke in shaping the initial research idea and substantial guidance from Professor Christian Heumann during the ideation phase. Data collection, data analysis and interpretation, as well as the literature research, were independently conducted by the first author. The drafting and revision of the manuscript were also carried out by the first author. Throughout all stages of the work, Professor Christian Heumann provided substantial support, close supervision, and critical guidance.

Operational Research

Multi-Task Learning for Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes

--Manuscript Draft--

Manuscript Number:	ORIJ-D-23-00526
Full Title:	Multi-Task Learning for Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes
Article Type:	Original Paper
Corresponding Author:	Stefan Michael Stroka, M.Sc. Ludwig-Maximilians-Universität München: Ludwig-Maximilians-Universitat Munchen Munich, Bavaria GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ludwig-Maximilians-Universität München: Ludwig-Maximilians-Universitat Munchen
Corresponding Author's Secondary Institution:	
First Author:	Stefan Michael Stroka, M.Sc.
First Author Secondary Information:	
Order of Authors:	Stefan Michael Stroka, M.Sc.
	Christian Heumann
	Fabian Suhrke
Order of Authors Secondary Information:	
Funding Information:	ams OSRAM International GmbH — Mr. Stefan Michael Stroka
Abstract:	Achieving perfect models with flawless predictions in real-world applications is impossible. A statistical model with a single target variable can only generalize a complex problem to a limited extent. To enhance generalization, a model must be trained on multiple, often contradictory objectives. Multi-task learning (MTL) enables the joint training of various supervised learning (SL) targets, as separately trained, independent models neglect target correlations. We introduce a novel loss function that allows for the minimization of the trade-off between point estimation and ordinal classification within a shared value range. Combined through an additional hyperparameter, it considers the codependency of the targets and enables a joint training of regression and multiclass classification on ordinal classes. Due to convexity and fuzzy logic, the function is also applicable with the gradient descent (GD) method and avoids the necessity to use methods for imbalanced classes. In contrast to comparable approaches, our methodology not only classifies but also determines class-optimized metrical predictions. The fuzzy logic-based, pseudo-metric consideration of classification allows for the optimization of the metric estimator through ordinal classes without information loss due to discretization of the point estimation, which would be necessary for classification tasks. To demonstrate the usefulness, we evaluate the method using freely available datasets commonly employed for assessing regression or classification tasks. A shared value range among targets is assumed. Out-of-sample evaluation with a focus on maximizing the classification target ($\alpha=0$), compared with a baseline regression, demonstrates that the applied COR loss function can achieve significant improvements in classification results (F1-Score: $+11.1\%$ for Boston Housing, $+17.1\%$ for Ames Housing) despite the challenge of imbalanced classes. This result is possible due to an average two times higher bias of the regression estimator. Minimizing the compromise between both targets results in a much less biased mean absolute error (MAE) ($+35\%$) with a reasonable improvement in classification accuracy (F1-Score: $+9.3\%$ for Boston Housing, $+10.2\%$ for Ames Housing). The increase in computing time remains within reasonable limits ($~1.25-~2$ times).

Suggested Reviewers:	

Title Page

ORJ manuscript No.
(will be inserted by the editor)

Multi-Task Learning for Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes

Stefan M. Stroka[*,1,2], **Christian Heumann**[1], **Fabian Suhrke**[2]

Received: date / Accepted: date

Abstract Achieving perfect models with flawless predictions in real-world applications is impossible. A statistical model with a single target variable can only generalize a complex problem to a limited extent. To enhance generalization, a model must be trained on multiple, often contradictory objectives. Multi-task learning (MTL) enables the joint training of various supervised learning (SL) targets, as separately trained, independent models neglect target correlations. We introduce a novel loss function that allows for the minimization of the trade-off between point estimation and ordinal classification within a shared value range. Combined through an additional hyperparameter, it considers the codependency of the targets and enables a joint training of regression and multiclass classification on ordinal classes. Due to convexity and fuzzy logic, the function is also applicable with the gradient descent (GD) method and avoids the necessity to use methods for imbalanced classes. In contrast to comparable approaches, our methodology not only classifies but also determines class-optimized metrical predictions. The fuzzy logic-based, pseudo-metric consideration of classification allows for the optimization of the metric estimator through ordinal classes without information loss due to discretization of the point estimation, which would be necessary for classification tasks. To demonstrate the usefulness, we evaluate the method using freely available datasets commonly employed for assessing regression or classification tasks. A shared value range among targets is assumed. Out-of-sample evaluation with a focus on maximizing the classification target ($\alpha = 0$), com-

✉ Stefan M. Stroka*
stefan.stroka@ams-osram.com

Christian Heumann
chris@stat.uni-muenchen.de

Fabian Suhrke
fabian.suhrke@ams-osram.com

[1] LMU Munich, 80539 Munich, Germany
[2] ams OSRAM International GmbH, 93055 Regensburg, Germany

pared with a baseline regression, demonstrates that the applied COR loss function can achieve significant improvements in classification results (F1-Score: +11.1% for Boston Housing, +17.1% for Ames Housing) despite the challenge of imbalanced classes. This result is possible due to an average two times higher bias of the regression estimator. Minimizing the compromise between both targets results in a much less biased mean absolute error (MAE) (+35%) with a reasonable improvement in classification accuracy (F1-Score: +9.3% for Boston Housing, +10.2% for Ames Housing). The increase in computing time remains within reasonable limits ($1.25 -$ 2 times).

Keywords Convex and metric loss function · Multi-task learning · Metric regression · Ordinal classification · Imbalanced class data

ORJ manuscript No.
(will be inserted by the editor)

Multi-Task Learning for Regression and Ordinal Classification: A novel loss function avoiding the problem of imbalanced classes

Received: date / Accepted: date

Abstract Achieving perfect models with flawless predictions in real-world applications is impossible. A statistical model with a single target variable can only generalize a complex problem to a limited extent. To enhance generalization, a model must be trained on multiple, often contradictory objectives. Multi-task learning (MTL) enables the joint training of various supervised learning (SL) targets, as separately trained, independent models neglect target correlations. We introduce a novel loss function that allows for the minimization of the trade-off between point estimation and ordinal classification within a shared value range. Combined through an additional hyperparameter, it considers the codependency of the targets and enables a joint training of regression and multiclass classification on ordinal classes. Due to convexity and fuzzy logic, the function is also applicable with the gradient descent (GD) method and avoids the necessity to use methods for imbalanced classes. In contrast to comparable approaches, our methodology not only classifies but also determines class-optimized metrical predictions. The fuzzy logic-based, pseudo-metric consideration of classification allows for the optimization of the metric estimator through ordinal classes without information loss due to discretization of the point estimation, which would be necessary for classification tasks. To demonstrate the usefulness, we evaluate the method using freely available datasets commonly employed for assessing regression or classification tasks. A shared value range among targets is assumed. Out-of-sample evaluation with a focus on maximizing the classification target ($\alpha = 0$), compared with a baseline regression, demonstrates that the applied COR loss function can achieve significant improvements in classification results (F1-Score: +11.1% for Boston Housing, +17.1% for Ames Housing) despite the challenge of imbalanced classes. This result is possible due to an average two times higher bias of the regression estimator. Minimizing the compromise between both targets results in a much less biased mean absolute error (MAE) (+35%)

Address(es) of author(s) should be given

with a reasonable improvement in classification accuracy (F1-Score: +9.3% for Boston Housing, +10.2% for Ames Housing). The increase in computing time remains within reasonable limits ($1.25 - 2$ times).

Keywords Convex and metric loss function · Multi-task learning · Metric regression · Ordinal classification · Imbalanced class data

1 Introduction

Modeling data-dependent correlations between dependent variables to infer multiple targets is crucial to generalize and understand real-world problems. Multiple Machine learning (ML) targets are, therefore, often essential to find suitable solutions for highly complex applications, where combining SL approaches is required. MTL is an approach to jointly model a combination of SL methods, for example, regression and classification, with objectives that are related but not identical [32]. Typical application fields for a combined regression and ordinal classification target include image- or video-based computer vision [31]. For instance, in the context of numerical age estimation based on an image, combining it with a classification into age groups of the same person can lead to improvements. Another application area is disease prediction [31]. For example, heart diseases could be forecasted based on blood pressure readings. Such a prognosis could be enhanced by an ordinal classification of the patient into risk groups for heart diseases, relying on expert assessments (e.g., moderate risk for heart diseases). Application is also possible in the field of fraud prevention [8]. For example, a computer-based continuous risk scoring can potentially be optimized through expert classification of the risk. Additionally, there are several further examples, such as quality control in health treatment [7] or industrial processes [22]. In essence, applications are feasible whenever two less reliable target variables (metric and ordinal) describe the same output based on partly different influences, and their reliability can be enhanced through optimization. In this paper, we focus on the application of a MTL model to optimize the model output based on a continuous regression and a multiclass classification on ordinal classes. This work aims to develop a method that minimizes the trade-off between conflicting objectives and, consequently, to introduce a shared loss function. To be more specific, we determine a point estimate that is optimized conditional on the classification target, thereby accepting a bias in regression estimation for improved classification accuracy. The influence of each target on the prediction is tunable as a hyperparameter. Nevertheless, challenges arise from combining losses of different SL applications, as these fundamentally differ in their definition, purpose, and computation. For instance, loss functions for classification are not designed to be applied to a continuous metric space. These ML approaches target distinct outcomes, yet they are assumed to be codependent, making a shared model training process necessary. Further challenges may emerge due to the data distribution in the case of imbalanced class data, which requires additional model or data processing to achieve accurate and unbiased classification results. The

joint training poses a challenge for most multi-target methods, as often only a random search in a given setting space is conducted. However, ML methods using GD allow simultaneous joint training with progressive improvement under the condition of a convex target function. To effectively train such a regressor for multi-target scenarios with this approach, it is required to use a convex regression loss function that combines the targets on an identical value range. While a standard loss function for a comparable regression is convex and differentiable over the defined, shared value range, a classification model seeks to maximize the class accuracy between predictions and observations, and thus, is often non-differentiable over the same value range. To make these two objectives comparable, we employ fuzzy logic. Fuzzy logic enables the mapping of an ordinal-scaled, linguistic value scale ("low," "medium," "high," "very high") onto a metric value range through the use of characteristic functions [5,30]. This approach allows us to consider the non-convex loss based on the ordinal classes as pseudo-metric and pseudo-continuous, bridging the gap between regression and classification. Following this concept, we introduce our novel loss function. This function enables the simultaneous optimization and combined training of regression and classification on ordinal classes. It is a weighted combination of a metric (L2 loss) and a fuzzy logic-based pseudo-metric loss function (distance-to-nearest-class). While the metric component aims to minimize the empirical distance between estimation and observation, the second component determines the distance of the point estimate to the value of the nearest observed class boundary. By utilizing the weighting parameter as an additional hyperparameter, the optimal model setting with the minimal trade-off can be found through hyperparameter tuning. We apply this loss function and showcase the improvements through experiments conducted with publicly available and real datasets (Ames Housing, Boston Housing), comparing with state-of-the-art baseline models. In addition to a sequentially optimized model training process and improved results compared to baselines, this concept offers a further crucial advantage. By treating ordinal classes as pseudo-metric, class data are also considered metrical. This eliminates the need to process the imbalanced class data. Nevertheless, the baseline classification is processed and optimized on imbalanced class data for a reasonable comparison. In the following sections, we will proceed with a review of related research and alternative approaches. Subsequently, we will explain the methodology in detail, conduct the experiments, and show the results achieved.

2 Related Work

Applying a loss function to model a target aims at minimizing a certain loss value, eventually leading to an optimal fitted model. These functions are not method-specific. In the training process with GD, their gradients are used to calculate pseudo-residuals (deviation errors or classification deviations), which only depend on the ML target and not necessarily the method. In the field of SL, there are several established standard loss functions for regression and

classification. Depending on the task and data distribution, a loss function can be chosen to be more robust for a specific scenario [15]. A detailed overview can be found in [27]. A data-tailored loss function can have various forms, given the condition that it is convex and differentiable when we want to use the GD method for optimization. Customizing a loss function enables a model to adapt more flexibly to data characteristics. Regarding this, we differentiate between symmetric functions and non-symmetric functions, and function combinations with weighting parameters. In scenarios involving asymmetric data distributions, adapting and using asymmetric loss functions can significantly impact the model's quality [12][3]. However, MTL can be addressed using various approaches [33]. A common approach is a weighted linear combination of per-task loss functions [23]. Such an approach as an example includes the custom-log-loss function, which aims to minimize the false alarm rate of a classification by penalizing only directly relevant false positives through a weighting parameter [13]. Another solution aims to optimize the loss function itself within a defined loss function space throughout the training process [25]. Nevertheless, the main challenge of our paper is to achieve an optimized solution between imbalanced data classification and regression without coarsening the metric measurement data by discrete classification. Common methods, like ordinal regression (OR), assign a ranking to a set of ordinal classes to benefit from the order [1]. Another approach, known as chain maximizing ordinal metric learning (CMOML), extends this methodology to establish a metric, allowing for the determination of distances between classes [24]. One drawback is that due to the lack of convexity in the function, a black-box method must be used for optimization. Both ordinal regression and CMOML use the ranking to improve classification and thus do not determine point estimators optimized for ordinal class classification. A similar approach involves sequential execution, where binary classification is performed in the first step, followed by regression with a customized loss in the subsequent step [29]. Further ideas involve a custom loss function based on Mean Squared Error (MSE) as a classification loss within classes. This approach aims to achieve higher accuracy on imbalanced class data using a minibatch logic [21]. A similar approach used previously was the introduction of a novel mean false error loss function [28]. The Mutually equidistant separation loss is another methodology that employs a deep metric learning loss function to enhance model fitting in a highly discriminative feature space. This is achieved by minimizing distance metrics to attain as homogeneous classes and optimal inter-class separation as possible [6]. Further related approaches involve parallel [4][20][17] and sequential execution [36][9], as well as the combination of model estimation with a discriminative classifier as posterior probability [26]. Many of these state-of-the-art ML applications utilize evolutionary algorithms for optimization, where random model settings are experimented with, and the best one with the least compromise between objectives is selected [2]. However, this approach has the drawback of often progressing more slowly because it cannot learn sequentially. Nonetheless, a further advantage of our MTL loss function is its robustness against imbalanced class data. Other approaches tackle this imbalance through a combina-

tion of data processing and stacked generalization [11], minority oversampling [35], an iterative oversampling approach [19], a Bayes cost as loss for ordinal classification of imbalanced data [18], or even a weighted k-nearest neighbor method that employs class membership with quantiles of estimated class probabilities [16].

3 Classification-Optimized Regression Loss (COR Loss)

The introduced COR-loss function serves as a metric loss function to solve certain MTL problems comprising regression and ordinal classification targets. It allows for optimization between codependent point estimation and classification on multiple ordinal-scaled classes, without discretizing the metric data with information loss through coarsening. This is possible due to considering ordinal classes as pseudo-metric using fuzzy logic. It enables the calculation of distances between the pseudo-metric observed class and the regression estimate. Using this metric, we can reduce a combined, weighted loss and approximate true classes which eventually leads to an accurate classification. Furthermore, the convexity of the function satisfies the prerequisites for GD application, enabling sequential optimization. Even though the COR-loss is only usable with a regression model, its specific function definition allows an optimization process between point estimation and classification while modeling a regression. To balance these targets for an optimal result, it is possible to tune the weighting parameter within the COR loss. Thus, the ML algorithm sequentially minimizes the loss, eventually achieving an optimal trade-off between regression and classification prediction.

3.1 Problem Formulation

In this paper, our approach focuses on the joint modeling of regression and classification. These ML targets are codependent, whereby both the ordinal classes (classification target) and the metric values (regression target) depend on the same value range. The latter is no restriction, as, in principle, any metric variable can be transformed to the interval [0,1] and an ordinal variable can be seen as a discretization of a latent metric variable on the interval [0,1]. Real-world application problems are diverse and typically optimization problems. Some examples include maximizing production output though measuring uncertainty while meeting quality thresholds (manufacturing), forecasting accurate disease diagnoses while determining the actual severity (healthcare), or predicting future air quality while considering pollutant limits (environmental science). Summarizing all, these diverse applications are optimization problems where achieving the best possible classification is crucial, while also maintaining a sufficiently accurate point estimation for real observations. The key requirement for decent results in these problems is, therefore, to consider both the target variables of regression and classification, as well as the potentially conflicting objectives. However, independent training of these models can

already result in competitive forecasts for the combined target. Yet, it can also yield entirely contrary outcomes if their objectives are conflicting and codependent. To summarize, a higher complexity of inherent correlations among the target variables requires more than a simple combination and optimization of two independent model outcomes, instead a shared training process is mandatory. When considered separately, regression seeks an optimal unbiased point estimator using a standard L2 loss, whereas, in classification, the focus lies on the accuracy of class predictions. In detail, we want an optimum that allows a biased point estimation to enhance classification. Alternative MTL approaches might, in part, offer competitive forecasts with an optimized compromise between these objectives. However, they have the drawback of exploring random settings for optimization steps, unlike the sequential optimization characteristic of gradient descent (GD). Comparable classification approaches, such as ordinal regression, are also comparable and can provide good classifications, but no regression predictions. In addition, the treatment of imbalanced class data becomes necessary. Imbalanced class data is a general problem anyway, which poses a challenge for most classification algorithms. Hence, a lack of data or model processing leads to bias and improperly weighed models, subsequently resulting in less accurate class predictions.

3.2 Loss Function Definition and Derivatives

To begin with, we now define notations and prerequisites for the theoretical exposition. To optimize a multi-task supervised learning (MTSL) problem, labeled data is required. Let $X = (X)_{n,k}$ represent the input matrix with n observations and k features. The defined MTSL target variable is a tuple consisting of a metric regression target $yreg$ and a classification target y_{clf} on ordinal classes. Both variables are defined on an identical value range. It follows that $y = (y_i^{reg}, y_i^{clf})_{i=1}^n$. The goal of the MTSL solution is to achieve the minimal combined loss for a metric prediction $\hat{y}$, optimized based on the ordinal target variable. The loss function itself is a metric, convex combination of functions. For selected ML applications, it is possible to use non-continuous loss functions, if a given loss value describes the quality of the prediction relative to reality for the target variable. Therefore, obtaining a minimizable value (loss) without sequential optimization is often sufficient. In our case, the minimization of the loss can be complex and computationally intensive. Therefore, we use GD as an iterative optimization approach. In this process, a quadratic approximation of the at least twice-differentiable loss function is conducted. The approximation is represented using the Taylor expansion with the first and second derivatives. For the Taylor expansion follows:

$$L(\hat{y}, y) \approx L(\hat{y}, y) + \nabla L(\hat{y}, y)^T + \frac{1}{2} H_L(\hat{y}, y) \tag{1}$$

with function $L(.)$, gradient $\nabla L(.)$ and Hessian matrix $H(.)$. However, the COR-loss function is a combination of two distinct distance metrics. The first

component (for regression) consists of the L2-loss (equivalent to the MSE) measuring the average squared distance between the estimation $\hat{y}_i$ and the actual value y_i^{reg}. The second component (for classification) measures the mean squared distance between the estimation $\hat{y}$ and its true class y_i^{clf}. In detail, the distance from the prediction to the nearest boundary of the correct classes is thus determined. This distance indicates when the prediction meets its true class, aligning the predicted class with the actual class. Ordinal-scaled classes typically do not comply with a metric system and cannot be metrically optimized. Since both target variables are defined in the same value range, class boundaries, and estimators are metric values and thus comparable. Using fuzzy logic, the classes can be sorted in ascending order of rank and treated in a pseudo-metric manner. Through logical sorting and the metric class boundaries defined on the identical value range as the target variable, the distance of the prediction to the class boundary and thus the class can be determined. The influence of each component on the final target is determined by a weighting parameter α. This additional hyperparameter can be tuned for an optimal solution. Thus, it follows:

$$L(\hat{y}, y) = \alpha L_1^2 + (1 - \alpha) L_2(\hat{y})^2 \tag{2}$$

with

$$L_1^2(\hat{y}, y) = MSE(\hat{y}, y) = (\hat{y} - y)^2 \tag{3}$$

and

$$L_2(\hat{y}) = \begin{cases} (\hat{y} - class_y^{lower\ \ b.}); \ class_{\hat{y}} < class_y \\ (\hat{y} - class_y^{upper\ \ b.}); \ class_{\hat{y}} > class_y \end{cases} \tag{4}$$

as well as the gradient with:

$$\nabla L(\hat{y}, y)^2 = 2\alpha(\hat{y} - y) + (1 - \alpha) L_2(\hat{y}) \tag{5}$$

and the Hessian matrix

$$H(\hat{y}, y) = 2. \tag{6}$$

Finally, Figure 1 shows two examples of how the components of the loss gradient are computed. The calculation example in Figure 1 reveals that in the first scenario (1), no optimization is required. This implies that achieving the regression target automatically satisfies the classification target. In contrast, the second scenario necessitates loss minimization, conditional on the weighting parameter α. In general, regardless of the learning rate (lr) and the pseudo-residuals r, it follows

$$r_{i-1} = -L(\hat{y}_{i-1}, y_{i-1}) = -[2\alpha L_1(\hat{y}_{i-1}, y_{i-1}) + 2(1 - \alpha) L_2(\hat{y}_{i-1})] \tag{7}$$

the updated prediction

$$\hat{y}_i = \hat{y}_{i-1} - r_{i-1} * lr. \tag{8}$$

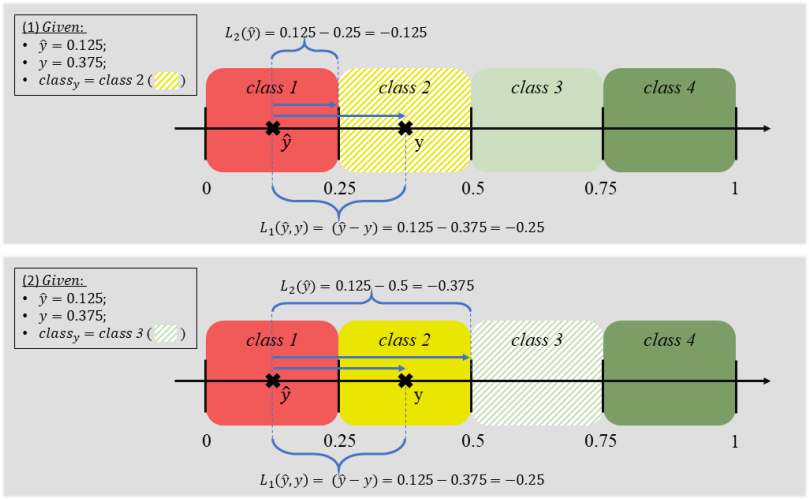


Fig. 1 The shared value range ([0,1]) for the targets is exemplarily separated by four equidistant, ascending classes (e.g. "low", "medium", "high", "very high"). These are chosen randomly for this example and may vary in other cases. We distinguish between two examples. The first case (1) illustrates the scenario in which the observed metric target value y and the observed class $class_y$ align. In the other case (2), contradictory goals are implied, as the observed metric target values y and the observed ordinal class $class_y$ do not align. For both examples, the presented calculations describe how the components L_1 and L_2 are computed.

3.3 Training Process

Following this brief example of loss function calculation, we will now examine the training process in detail. The goal of the training process is to sequentially improve model fit and reach minimal loss, resulting in an optimized prediction for each test data observation. To achieve this, the algorithm provides an initial prediction $\hat{y}_{m=0}$ for all observations, which is the mean of the train data of the target variable. However, the initial estimate is not influenced by the classification target y^{clf}. Instead, it is the mean of the regression target y^{reg} used as the initial prediction. In the case of normalized data, the following applies:

$$\hat{y}_{m=0} = F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = [0.5, ..., 0.5] \tag{9}$$

with γ as predicted values. In each iteration $m \in 1, ..., M$, the pseudo-residuals r_{im} for all observations n are calculated based on the gradient of the loss function $\nabla L(\hat{y}, y)$. It holds that:

$$r_{im} = -\left[\frac{\partial L(\hat{y}_i, y_i)}{\partial \hat{y}_i}\right] = -\nabla L(\hat{y}_i, y_i) = -[2\alpha L_1(\hat{y}_i, y_i) + 2(1-\alpha)L_2(\hat{y}_i)] \tag{10}$$

(compare with equation 5) is the pseudo residual for the observation i in iteration m. Following, terminal regions R_{j_m} (region j in iteration m) are formed by using a chosen base learner. Base learners are typically generated from training data using a base learning algorithm, which can be a decision tree, a neural network, or another type of machine [34]. In the case of decision trees (as base learner), the terminal regions correspond to the leaves. Each region R_{j_m} is defined by the averaged loss γ_{j_m} per region of the pseudo-residuals r_{im} for all leaf-inherent observations $x_i (\in R_{j_m})$. Therefore:

$$\gamma_{j_m} = argmin_\gamma \sum_{x_i \in R_{j_m}} L(y_i, \hat{y}_{m-1} + \gamma) \tag{11}$$

For the updated prediction $\hat{y}_m$, it holds:

$$\hat{y}_m = \hat{y}_{m-1} + \lambda * \sum_{j_m=1}^{J_m} \gamma_{j_m} (x \in R_{j_m}) \tag{12}$$

with $\hat{y}_m = F_m(x)$, learning rate λ and the number of regions J_m in iteration m. Depending on α, the weighted gradient is used to calculate the pseudo-residuals and finally to update the prediction. Thus, the hyperparameter α, similar to learning rate λ, has a significant impact on the model training and the learning process at each algorithm step or decision tree, as it determines the influence of each targets.

4 Experiment

In this section, we present our results using both simulated and real-world data examples. In the first experiment, we apply the loss function in conjunction with linear regression to gain a basic understanding of the optimization process without much model complexity. The second experiment demonstrates the application of the methodology with the more complex XGBoost algorithm on two publicly available datasets. Results are compared with common evaluation metrics of supervised learning.

4.1 Application Scenario

To demonstrate improvement by optimization, we require data with conflicting objectives. Therefore, we use simulated data (for linear regression) and real datasets (for XGBoost). In the first case, we utilize normal distributions for each target variable to sample from. Two similar yet distinct normal distributions are employed to create the desired optimization problem in each case. In the second scenario, real datasets are expanded with a classification target variable for application without altering the real problem or other data. In Detail, the real datasets, presented below, focus on house prices and potential influencing factors. Specifically, the designed use cases are structured to find

an optimum between the seller's desired house price y^{reg} and a rough estimate by a real estate agent y^{clf} represented as a class (with upper and lower limits). Each target variable alone is not reliable, as the price may be set by an untrained person (seller) or estimated cheaply and vaguely by an expert. The aim of the application is to find an optimum by combining both information, enabling the determination of a reasonable selling price without the need for a detailed and expensive house price determination of an expert.

4.2 Data Structure, Preprocessing and Descriptive Analysis

The generated data consists of one input variable $X = (X_1, ..., X_{1000})$ with 1000 observations, a metric, normally distributed target variable y^{reg}, and the classification target variable y^{clf}. y^{reg} is generated using the *make_regression()*-function from the **Python** package *sklearn*. It follows the equation $y^{reg} = f(X) + \epsilon$ (with f normally distributed, and error $\epsilon \sim N(0, 20)$). From this distribution, we draw the samples for $y^{reg} = (y_1^{reg}, ..., y_{1000}^{reg})$. The classification target samples are drawn from a modified distribution ($y^{clf} \sim N(5, 65)$). They are then sorted based on predetermined classes and corresponding thresholds, resulting in the class variable for the classification target. For the real data

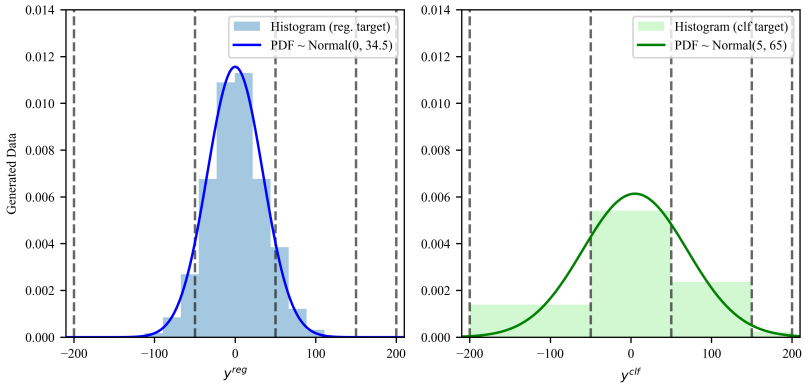


Fig. 2 The left graph displays the density of the normally distributed regression target along with a histogram of 1000 drawn samples. The right graph illustrates the density of the modified, biased normal distribution with a higher variance compared to the left. Drawn samples are sorted into a predetermined grid to represent the frequency distribution across classes (on the right). Grid thresholds are depicted by vertical dashed lines.

application, we use two well-known and publicly available datasets: **Boston Housing (housing)** [14] and **Ames Housing** [10]. These data sets are typically used for validating statistical multivariate regression approaches with

continuous target variables (in Dollars). The described goal of the experiment is the application of the methodology to real data. Since there is no natural classification target in both cases, we design one based on the existing regression target to create realistic optimization examples. To achieve this, classes are generated that depend on the distribution of y^{reg} but are not identical and, therefore, can also be contrary. This is achieved by fitting a two-parameter Gamma distribution (*Gamma 1*) to the metric target variable y^{reg}. Changes in the resulting parameters yield a new Gamma distribution (*Gamma 2*) that biases the reflection of the regression target. Finally, random samples are drawn from the new distribution and classified into predefined classes based on their thresholds. These approaches simulate MTSL problems with codependent regression and generated classification targets. These exam-

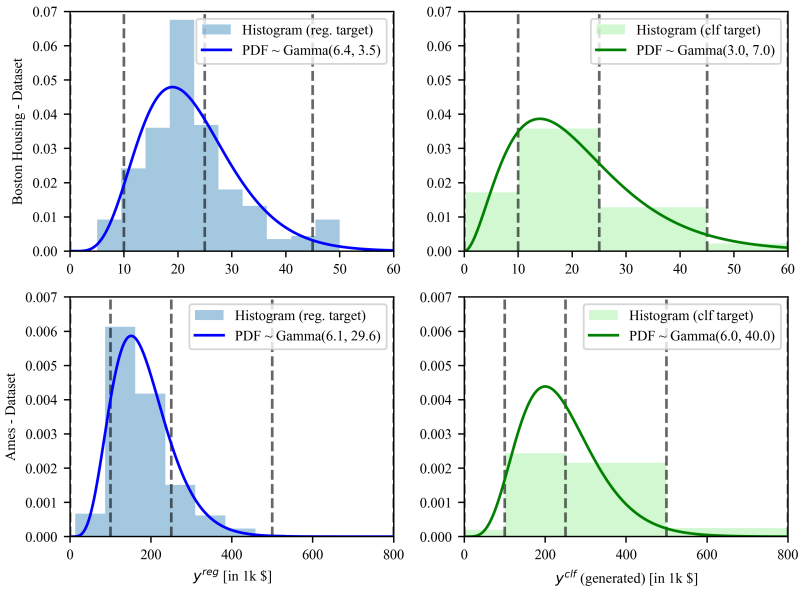


Fig. 3 Illustration of the given data y^{reg} (histogram and density (PDF)) and the generated discrete class data y^{clf}. Every target is displayed with a histogram and density (PDF) for both data sets. Dashed lines (—) show the thresholds of the defined ordinal classes.

ples illustrated in Figure 3 have four defined classes for each data set. These price classes are ordinally scaled and could be described as *low*, *medium*, *high*, and *very high* for instance. The number and width of the classes are again randomly chosen for the experiment and therefore subject to change. Noteworthy, Figure 3 depicts the right-skewed distribution of the regression target and notably highlights a class imbalance for both data sets. This imbalance is

important to demonstrate the side effect of our solution and how it handles imbalanced classification without additional data or model processing.

4.3 Evaluation Metrics

We introduce the used evaluation metrics for regression and classification targets. The Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

serves as the regression metric with $\hat{y}_i$ as predicted value for the unseen i-th observation y_i of all n observations. The classification accuracy is assessed using two metrics based on the multiclass confusion matrix. The multiclass accuracy (ACC) measures the proportion of predicted classes that match the actual classes from the test data across all classes n. Hence

$$ACC = \frac{TP + TN}{TP + TN + FN + TN} = \frac{1}{n}\sum_{i=1}^{n}\chi_{class(y_i)=class(\hat{y}_i)}, \tag{14}$$

with χ as the Indicator function. The $F1$ *score* supports the ACC since the latter can be biased, because of imbalanced class data. The weighted $F1$ *score* of precision and recall per class is given as

$$F1\ score_{weighted} = \frac{1}{C}\sum_{c=1}^{C} w_c * \frac{2}{Recall_c^{-1} + Precision_c^{-1}} \tag{15}$$

and is more robust for imbalanced classes.

4.4 Analysis and Results

In the following section, we examine the results of the optimization using the novel COR-loss function in comparison to defined baselines.

4.4.1 Settings

For the first part, we implemented an object class for a linear regression model, which allows the use of our loss function. For this application, we utilize a 5-fold cross-validation to avoid potential overfitting. Hyperparameter tuning is not required. In the second part, we apply the "eXtrem Gradient Boosting" (XGBoostt) with an *XGBoost Regressor* from the **Python** package *xgboost* in combination with the COR-loss. The implementation also includes preprocessing with the generating of ordinal classes based on the drawn sample values, model training with 5-fold cross-validation and Bayesian hyperparameter tuning (*hyperopt*), as well as the final class-optimized regression predictions.

As imbalanced data often presents a classification challenge, all train, test, and validation splits are performed stratified. To effectively avoid overfitting with XGBoost, an early-stopping approach with an evaluation loss function additional to the hyperparameter tuning is applied. Our focus is on classification accuracy rather than unbiased point estimations. Therefore, we use the $F1$ $score_{weighted}$ as an evaluation metric for early-stopping and hyperparameter tuning. We could also use our COR-loss or MAE as an evaluation function, which focuses on different targets and hence results in a slightly different outcome. To ensure stable and valid results, a test and training data split (out-of-sample ratio: 20/80) is implemented, along with a 5-fold cross-validation (CV) on the remaining training data (in-sample data). Tunable hyperparameters are for each CV pair separately optimized using a Bayesian hyperparameter optimization approach (python: *Hyperopt*) to achieve the best results. The top resulting hyperparameter settings per CV are then used to conduct an out-of-sample evaluation on the test data with the entire training data. The results of these five optimized models provide a stable mean (μ) as well as a standard deviation (σ), offering a comprehensive view. Seeds are set to ensure reproducibility.

4.4.2 Basic Experiment with Linear Regression

Linear regression, as a fundamental statistical methodology, includes fewer tunable hyperparameters compared to other methods (e.g., XGBoost). Therefore, variabilities introduced by the model can be minimized to represent the loss function comprehensibly. The implemented approach is thus limited to three tunable hyperparameters: the learning rate (λ), the number of iterations (n_iters), and the new weighting parameter (α).

Improvement by varying the number of iteration (n_iters)

In the initial analysis, we focus on the impact of the number of iterations while keeping the other two hyperparameters fixed. Therefore we refer to Figure 4. With a higher number of iterations, the model achieved better results in out-of-sample predictions. The baseline (standard linear regression with ordinary least squares with variable number of iterations) improved in terms of the F1-score by an average of 10% (at iteration 500), coupled with a significant reduction in MAE. In comparison, our custom loss function achieved an improvement of approximately 20%. As intended, this enhancement in classification comes at the expense of a biased point estimate, resulting in the MAE being 17% (at iteration 500) worse than the baseline's MAE. The forecast uncertainties of the models (see prediction interval (PI) in Figure 4) are comparable for the MAE, whereas, in the F1-score, the baseline has a much narrower PI.

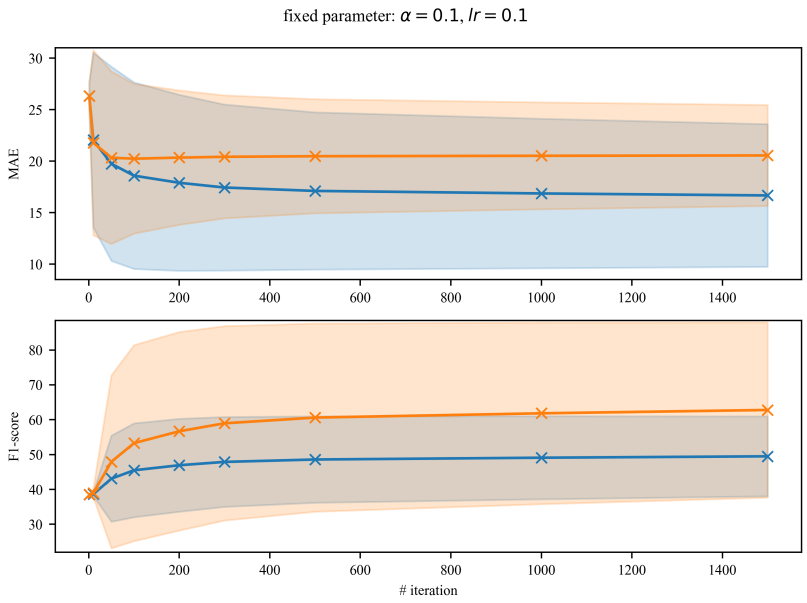


Fig. 4 The figure illustrates the evaluation of out-of-sample predictions with a baseline model (color: 'blue') and a model using the custom loss function. The analysis is conducted on the generated example data, with the hyperparameters $\alpha = 0.1$ and $lr = 0.1$ held constant. Both plots display the mean and the 95 % prediction interval.

Importance of Tuning the Learning Rate (lr)

To illustrate the influence of the lr, we now present a similar model setting with a changed but again fixed lr (with fixed $\alpha = 0.1$). The lr, as a hyperparameter, plays a crucial role in the learning process and can lead to overfitting if chosen incorrectly. It determines the size of learning steps and, consequently, the extent of prediction changes in each iteration. In Figure 5, we can see that the use of the COR loss function with a high learning rate (lr) can lead to significantly poorer point estimation. The low α value significantly weights classification over regression. Through the combination with a high lr, classification is learned in larger steps, and the point estimate is nearly neglected. Nevertheless, the initial estimate in iteration 0 is comparatively good because optimization with our loss function only starts in the first iteration and is not used for the initial estimation.

Influence of α Values on Modeling

In this analysis, we focus on the weighting parameter α. We again fix the other two parameters. Figure 6 leads to the conclusion that changing the α-

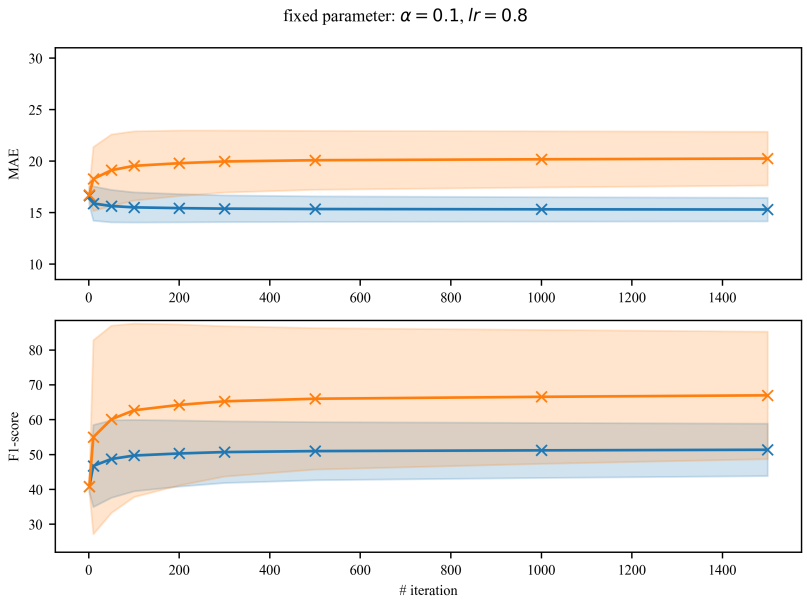


Fig. 5 The figure illustrates the evaluation of out-of-sample predictions with a baseline model (color: 'blue') and a model using the custom loss function. The analysis is conducted on the generated example data, with the hyperparameters $\alpha = 0.1$ and $lr = 0.8$ held constant. Both plots display the mean and the 95% prediction interval.

value allows weighting the optimization towards regression or classification. Hence, the trade-off between improved classification and less bias of regression can be minimized. An optimal α thus results in a metric regression estimator optimized with respect to the contrary class target. However, it is also evident that a decrease in α leads to more uncertain model prediction and accuracy, as indicated by a higher standard deviation and, therefore, an enlargement of the prediction interval.

4.4.3 Advanced Experiment with Decision Tree-based XGBoost Regression

In the following section, we investigate the more complex XGBoost regression application. Since XGBoost has significantly more hyperparameters that need to be tuned for optimal model fitting, we focus on the weighting parameter α and optimize the rest using a Bayesian optimization method.

Analysis of Hyperparameter α

The tuneable, weighting hyperparameter α fundamentally describes the influence of both components in our loss function. In detail, a high α value ($\alpha \approx 1$)

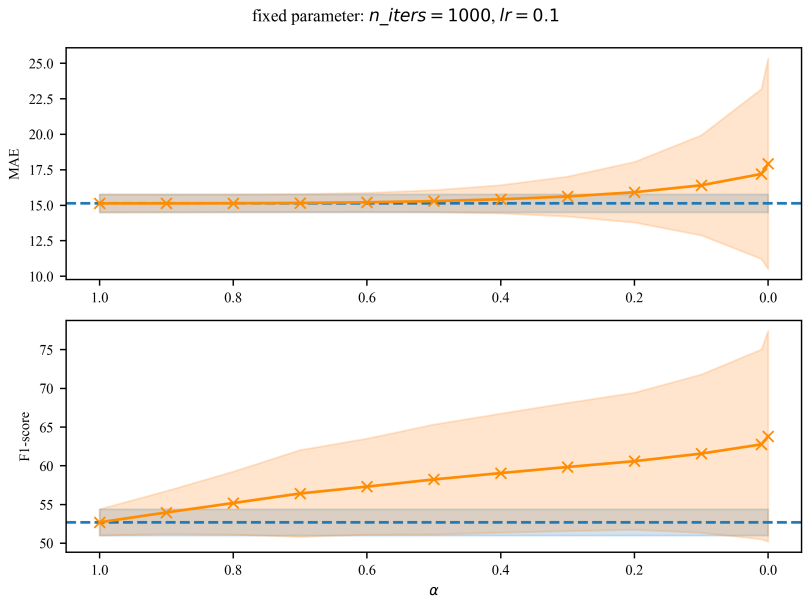


Fig. 6 The figure illustrates the influence of the weighting parameter α for compareable models. A decreasing α-value indicates a higher weighting of the classification component within the loss function. The standard linear regression, as baseline (color: blue) is identical to linear regression with COR loss (color: orange) and $\alpha = 1$ (with the same hyperparameter settings). However, the baseline model is independent of the α-value. The filled areas display the 95% prediction intervals.

mainly uses the L2-loss, approximating regression, and aiming to minimize MAE, whereas a small α value implies the mainly use of the other gradient component for classification. Therefore, a strictly monotonic decreasing α values theoretically imply strictly monotonic increasing metric (e.g., MAE, ACC, or $(1 - F1_{score})$). In practical terms, however, all MTSL models are trained separately and independently for each α value with highly variable decision trees. This implies that other stochastic effects exist, which cannot be fully excluded and only tendencies can be shown. Consequently, the results of practical evaluations are not strictly monotonic (as in 4.4.2) and are subject to natural fluctuations. In both application cases, a significant improvement in the $F1 - score$ with a rising bias of MAE is evident as α decreases. For the Boston Housing dataset, the F1-score increased by approximately 11% when comparing $\alpha = 0$ to $\alpha = 1$. As anticipated, the MAE worsens, increasing by 75%. Visually, we suspect a minimal trade-off at $\alpha = 0.1$, as it exhibits the best classification accuracy with the smallest degradation of MAE. The prediction uncertainty is also minimal in the surrounding range ($\alpha \in 0.3, 0.1, 0.05$). It is noteworthy that the application of the COR-loss function with $\alpha = 0$

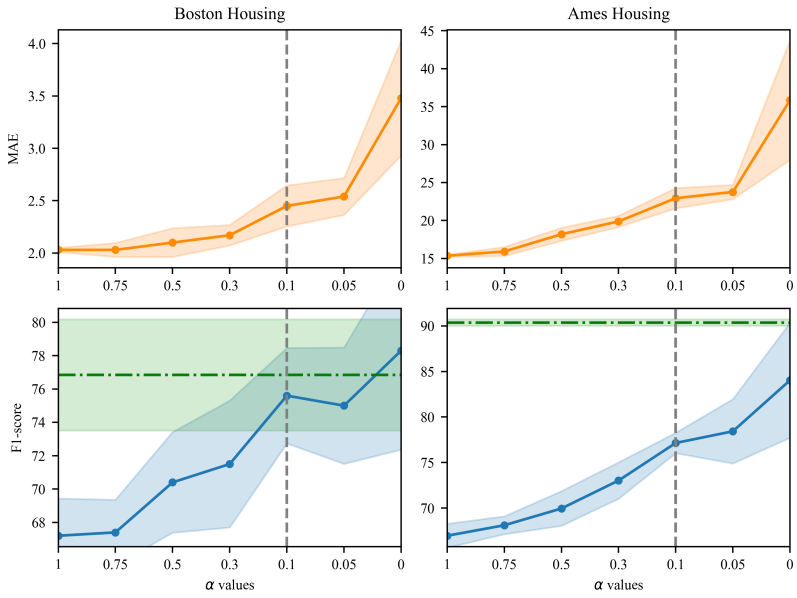


Fig. 7 Out-of-the-sample evaluation for list of discrete α-values illustrating the outcome with MAE and $F1\ score_{weighted}$ for both data sets. Insets show the mean estimation (color: orange, blue, green) and the 95% prediction interval (filled area) for every metric, resulting from evaluating each fold of the 5-fold CV. The horizontal line (and filled area) displays the baseline classification (color: green). The vertical dashed lines (color: grey) display a possible trade-off minimum.

achieved a better F1-score than the baseline. For the Ames Housing dataset, an improvement in the F1-score of approximately 17% was achieved when comparing $\alpha = 1$ to $\alpha = 0$. The regarding MAE increases by a factor of 2.3. The optimum again seems to be near $\alpha = 0.1$. At this point, there is a 49% increase in MAE with an improvement in classification of approximately 6%. The results of the baseline classification were not surpassed here. Further is noteworthy that as α decreases, the forecast uncertainty increases for both point and classification estimators. In addition to the evaluation metrics, Table 1 illustrates differences in computation time (CT). The application of the COR loss function requires, on average, 2 times as much CT for the Boston Housing dataset and 1.1 times for the Ames Housing dataset compared to the baseline regression.

Table 1 Out-of-sample results of XGBoost with 5-fold cross-validation and hyperparameter tuning for selected alpha values. $\alpha = 1$ corresponds to the regression baseline with L2-loss, and clf represents the classification baseline (without our loss function). The mean and standard deviation are derived from results based on the best settings for each fold of the 5-fold CV.

	Metrics		α							clf (b)
			1.0(a)	0.75	0.5	0.3	0.1	0.05	0	
Boston Housing	MAE	μ	**2.03**	2.03	2.10	2.17	2.45	2.54	3.48	-
		(σ)	**(0.01)**	(0.034)	(0.07)	(0.05)	(0.10)	(0.09)	(0.28)	
	ACC	μ	69.46	69.46	72.22	73.17	76.41	75.93	**78.68**	76.77
		(σ)	(1.00)	(0.76)	(1.29)	(1.67)	(1.50)	(1.67)	**(2.92)**	(1.79)
	F1-score	μ	67.26	67.41	70.42	71.51	75.61	75.00	**78.38**	76.84
		(σ)	(1.14)	(1.00)	(1.54)	(1.94)	(1.46)	(1.78)	**(3.03)**	(1.70)
	CT		12.0	27.5	21.6	21.8	24.0	22.3	22.4	**2.9**
Ames Housing	MAE	μ	**15,36**	15,92	18,19	19,88	22,94	23,76	35,84	–
		(σ)	**(0,08)**	(0,30)	(0,44)	(0,38)	(0,68)	(0,48)	(4,01)	
	ACC	μ	70,36	71,29	72,6	75,02	78,61	79,54	84,32	**90,38**
		(σ)	(0,53)	(0,42)	(0,74)	(0,82)	(0,49)	(1,60)	(3,35)	**(0,40)**
	F1-score	μ	66,97	68,12	69,96	73,03	77,15	78,43	84,03	**90,36**
		(σ)	(0,67)	(0,50)	(0,97)	(1,02)	(0,57)	(1,80)	(3,22)	**(0,35)**
	CT	μ	49.8	64.6	63.5	63.8	64.2	52.2	34.9	**29.40**

(a) regression
(b) classification

5 Conclusion

The proposed new loss function serves as an optimized solution for supervised multi-task learning problems. It allows for sequential progressive training (with GD) of regression and ordinal class classification within a shared regression model. It effectively minimizes the trade-off between unbiased point estimation (regression target) and high classification accuracy (classification target). In this paper, we used two publicly available datasets to determine the optimal solution between regression and ordinal classification on a common target value range. Comparisons were performed using standard evaluation metrics and baseline regression and classification models. The models were further hyperparameter-tuned with a Bayesian optimization method and a 5-fold cross-validation. Analyses revealed that our solution, in the case of weighting for an optimal classification, led to a maximum improvement in accuracy of on average 14.1% (Boston Housing: 11.1%, Ames Housing: 17.1%) compared to regression baseline. This is possible due to on average two times higher MAE (Boston Housing: 74%, Ames Housing: 133%). In the case of Boston Housing, the classification result could even surpass the baseline classification. However, in the scenario with a minimal trade-off, the optimum was achieved with an improvement in classification of approximately 9.8% (Boston Housing: 9.3%, Ames Housing: 10.2%) with a significantly lesser biased MAE of approximately 35% (Boston Housing: 21%, Ames Housing: 49%). The increase in complexity due to the new loss function inevitably leads

to higher computational costs. Compared to baseline regression, this means an average increase of approximately 54% (Boston Housing: 94%, Ames Housing: 15%) (depending on hyperparameter tuning). The analysis has shown that despite increased CT, the application can achieve very good results depending on the target weighting. Especially noteworthy is the optimized, joint estimator, which, unlike common methods, allows for both classification and point estimation. Another significant advantage is the joint training of both targets, allowing for sequential learning optimization with GD. The metric consideration of class boundaries also allows for the neglect of imbalanced class data processing as a side effect. The idea that emerged from a practical optimization case will be further developed in future work to demonstrate the potential application in multidimensional multi-task learning problems in real-world applications.

Acknowledgements This work was supported by ams OSRAM group.

References

1. Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.
2. Stamatios-Aggelos N. Alexandropoulos, Christos K. Aridas, Sotiris B. Kotsiantis, and Michael N. Vrahatis. Multi-Objective Evolutionary Optimization Algorithms for Machine Learning: A Recent Survey. In Ioannis C. Demetriou and Panos M. Pardalos, editors, *Approximation and Optimization : Algorithms, Complexity and Applications*, Springer Optimization and Its Applications, pages 35–55. Springer International Publishing, Cham, 2019.
3. Kevin Aretz, Sohnke M Bartram, and Peter F Pope. Asymmetric Loss Functions and the Rationality of Expected Stock Returns.
4. Angelo Bonfitto, Stefano Feraco, Andrea Tonoli, and Nicola Amati. Combined regression and classification artificial neural networks for sideslip angle estimation and road condition identification. *Vehicle System Dynamics*, 58(11):1766–1787, November 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00423114.2019.1645860.
5. Hans-Heinrich Bothe. *Fuzzy Logic: Einführung in Theorie und Anwendungen*. Springer-Verlag, July 2013. Google-Books-ID: LXt_BwAAQBAJ.
6. Yasser Boutaleb, Catherine Soladie, Nam-Duong Duong, Amine Kacete, Jérôme Royan, and Renaud Seguier. MES-Loss: Mutually equidistant separation metric learning loss function. *Pattern Recognition Letters*, 172:58–64, August 2023.
7. Felix J. S. Bragman, Ryutaro Tanno, Zach Eaton-Rosen, Wenqi Li, David J. Hawkes, Sebastien Ourselin, Daniel C. Alexander, Jamie R. McClelland, and M. Jorge Cardoso. Quality control in radiotherapy-treatment planning using multi-task learning and uncertainty estimation. April 2018.
8. Liao Chen, Ning Jia, Hongke Zhao, Yanzhe Kang, Jiang Deng, and Shoufeng Ma. Refined analysis and a hierarchical multi-task learning approach for loan fraud detection. *Journal of Management Science and Engineering*, 7(4):589–607, December 2022.
9. Jui-Sheng Chou and Chih-Fong Tsai. Concrete compressive strength analysis using a combined classification and regression technique. *Automation in Construction*, 24:52–60, July 2012.
10. D. De Cock. Ames housing: The ames iowa housing data, 2011.
11. Marine Desprez, Kyle Zawada, and Daniel Ramp. Overcoming the ordinal imbalanced data problem by combining data processing and stacked generalizations. *Machine Learning with Applications*, 7:100241, March 2022.
12. Jean Dessain. Improving the Prediction of Asset Returns With Machine Learning by Using a Custom Loss Function, September 2022.

13. Yun Gao, Hirokazu Hasegawa, Yukiko Yamaguchi, and Hajime Shimada. Malware Detection Using Gradient Boosting Decision Trees with Customized Log Loss Function. In *2021 International Conference on Information Networking (ICOIN)*, pages 273–278, January 2021. ISSN: 1976-7684.
14. David Jr. Harrison and Daniel L. Rubinfeld. Boston housing (housing).
15. Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2001.
16. Sungil Kim, Heeyoung Kim, and Younghwan Namkoong. Ordinal Classification of Imbalanced Data with Application in Emergency and Disaster Information Services. *IEEE Intelligent Systems*, 31(5):50–56, September 2016. Conference Name: IEEE Intelligent Systems.
17. Na Liu, Fan Zhang, and Fuqing Duan. Facial Age Estimation Using a Multi-Task Network Combining Classification and Regression. *IEEE Access*, 8:92441–92451, 2020. Conference Name: IEEE Access.
18. Marcelino Lázaro and Aníbal R. Figueiras-Vidal. Neural network for ordinal classification of imbalanced data by minimizing a Bayesian cost. *Pattern Recognition*, 137:109303, May 2023.
19. Francisco Marques, Hugo Duarte, João Santos, Inês Domingues, José P. Amorim, and Pedro H. Abreu. An iterative oversampling approach for ordinal classification. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, pages 771–774, New York, NY, USA, April 2019. Association for Computing Machinery.
20. Julia M. H. Noothout, Bob D. De Vos, Jelmer M. Wolterink, Elbrich M. Postma, Paul A. M. Smeets, Richard A. P. Takx, Tim Leiner, Max A. Viergever, and Ivana Išgum. Deep Learning-Based Regression and Classification for Automatic Landmark Localization in Medical Images. *IEEE Transactions on Medical Imaging*, 39(12):4011–4022, December 2020. Conference Name: IEEE Transactions on Medical Imaging.
21. Tri-Cong Pham, Antoine Doucet, Chi-Mai Luong, Cong-Thanh Tran, and Van-Dung Hoang. Improving Skin-Disease Classification Based on Customized Loss Function Combined With Balanced Mini-Batch Logic and Real-Time Image Augmentation. *IEEE Access*, 8:150725–150737, 2020. Conference Name: IEEE Access.
22. Vignesh Sampath, Iñaki Maurtua, Juan José Aguilar Martín, Andoni Rivera, Jorge Molina, and Aitor Gutierrez. Attention-Guided Multitask Learning for Surface Defect Identification. *IEEE Transactions on Industrial Informatics*, 19(9):9713–9721, September 2023. Conference Name: IEEE Transactions on Industrial Informatics.
23. Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
24. Juan Luis Suárez, Salvador García, and Francisco Herrera. Ordinal regression with explainable distance metric learning based on ordered sequences. *Machine Learning*, 110(10):2729–2762, October 2021.
25. Yongxiang Tang, Wentao Bai, Guilin Li, Xialong Liu, and Yu Zhang. CROLoss: Towards a Customizable Loss for Retrieval Models in Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 1916–1924, New York, NY, USA, October 2022. Association for Computing Machinery.
26. Charl van Heerden, Etienne Barnard, Marelie Davel, Christiaan van der Walt, Ewald van Dyk, Michael Feld, and Christian Müller. Combining regression and classification methods for improving automatic speaker age recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5174–5177, March 2010. ISSN: 2379-190X.
27. V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, September 1999. Conference Name: IEEE Transactions on Neural Networks.
28. Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374, July 2016. ISSN: 2161-4407.

29. Peng Ye, Julian Qian, Jieying Chen, Chen-hung Wu, Yitong Zhou, Spencer De Mars, Frank Yang, and Li Zhang. Customized Regression Model for Airbnb Dynamic Pricing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 932–940, New York, NY, USA, July 2018. Association for Computing Machinery.
30. L.A. Zadeh. Fuzzy logic. *Computer*, 21(4):83–93, April 1988. Conference Name: Computer.
31. Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59(2):895–907, January 2012.
32. Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. Publisher: Oxford University Press.
33. Yu Zhang and Qiang Yang. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, December 2022. Conference Name: IEEE Transactions on Knowledge and Data Engineering.
34. Zhi-Hua Zhou. *Ensemble Learning*, pages 270–273. Springer US, Boston, MA, 2009.
35. Tuanfei Zhu, Yaping Lin, Yonghe Liu, Wei Zhang, and Jianming Zhang. Minority oversampling for imbalanced ordinal regression. *Knowledge-Based Systems*, 166:140–155, February 2019.
36. Shaghayegh Zihajehzadeh, Omar Aziz, Chul-Gyu Tae, and Edward J. Park. Combined Regression and Classification Models for Accurate Estimation of Walking Speed Using a Wrist-worn IMU. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3272–3275, July 2018. ISSN: 1558-4615.

6 Knowledge-embedded Machine Learning for Production Optimization on Logistical Delivery Grids

https://doi.org/10.1007/s00170-025-16530-7

Declaration of Author Contributions The conception and design of the study were primarily developed by the first author. Data collection, data analysis and interpretation, as well as the literature research, were independently conducted by the first author. The drafting and revision of the manuscript were also carried out by the first author. Throughout all stages of the work, Professor Christian Heumann provided substantial support, close supervision, and critical guidance.

Applied Intelligence

Knowledge-embedded Machine Learning for Production Optimization on Logistical Delivery Grids

--Manuscript Draft--

Manuscript Number:	APIN-D-24-01080
Full Title:	Knowledge-embedded Machine Learning for Production Optimization on Logistical Delivery Grids
Article Type:	Original Submission
Keywords:	Knowledge-embedded Optimization, Machine Learning, aleatoric uncertainty, logistical grid
Corresponding Author:	Stefan Michael Stroka, M.Sc. Ludwig Maximillians University Munich: Ludwig-Maximilians-Universitat Munchen Munich, Bavaria GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ludwig Maximillians University Munich: Ludwig-Maximilians-Universitat Munchen
Corresponding Author's Secondary Institution:	
First Author:	Stefan Michael Stroka, M.Sc.
First Author Secondary Information:	
Order of Authors:	Stefan Michael Stroka, M.Sc. Christian Heumann, Prof. Dr.
Order of Authors Secondary Information:	
Funding Information:	ams OSRAM Group — Mr. Stefan Michael Stroka
Abstract:	Unobserved measurement errors can significantly affect the reliability of outcome prediction. Biased and highly variable predictions make it challenging to optimize the yield based on set quality thresholds. Although progress has been made in the use of statistical methods to detect and reduce these errors, significant challenges remain. These include the need for multiple measurements with few outliers, and the risk of neglecting systematic errors. A further challenge arises from the input uncertainties caused by random stochastic errors that cannot be eliminated. This is influenced by factors beyond control, such as the measurement methods and technology. In this paper, we propose an approach that leverages aleatoric uncertainties instead of reducing them. This allows for optimized modeling of uncertainty and observations. In production, the yield and quality depend on meeting predefined thresholds for product properties in a logistical delivery grid. The goal is not just an unbiased forecast, but also ensures that the delivered quantity aligns with the ordered quantity within the defined property boundaries. Measurement uncertainties create an optimization problem because the expected value may deviate from the actual delivery. Our method trains a machine learning model using observations and optimizes it with prior knowledge for the best possible delivery. The aim is to achieve unbiased statistical predictions while maximizing the actual yield. Prior information (from expert or external knowledge from other data or models) can be incorporated based on a weighting parameter. To demonstrate the value of the method, we used data generated by a variational autoencoder from measurements in an opto-semiconductor production. Promising results show significant improvements in optimization with respect to data-independent information, and delivery boundaries. Despite the slightly higher normalized Mean Absolute Error (MAE) of +0.0073, an average improvement of approximately +17.78% was achieved across all examples when comparing baseline and optimized models (with).

Applied Intelligence ()
https:/

Knowledge-embedded Machine Learning for Production Optimization on Logistical Delivery Grids

Stefan M. Stroka[1,2,*], Christian Heumann[1]

Received: Accepted: Published online Abstract

Abstract

Unobserved measurement errors can significantly affect the reliability of outcome prediction. Biased and highly variable predictions make it challenging to optimize the yield based on set quality thresholds. Although progress has been made in the use of statistical methods to detect and reduce these errors, significant challenges remain. These include the need for multiple measurements with few outliers, and the risk of neglecting systematic errors. A further challenge arises from the input uncertainties caused by random stochastic errors that cannot be eliminated. This is influenced by factors beyond control, such as the measurement methods and technology. In this paper, we propose an approach that leverages aleatoric uncertainties instead of reducing them. This allows for optimized modeling of uncertainty and observations. In production, the yield and quality depend on meeting predefined thresholds for product properties in a logistical delivery grid. The goal is not just an unbiased forecast, but also ensures that the delivered quantity aligns with the ordered quantity within the defined property boundaries. Measurement uncertainties create an optimization problem because the expected value may deviate from the actual delivery. Our method trains a machine learning model using observations and optimizes it with prior knowledge for the best possible delivery. The aim is to achieve unbiased statistical predictions while maximizing the actual yield. Prior information (from expert or external knowledge from other data or models) can be incorporated based on a weighting parameter. To demonstrate the value of the method, we used data generated by a variational autoencoder from measurements in an opto-semiconductor production. Promising results show significant improvements in optimization with respect to data-independent information, and delivery boundaries. Despite the slightly higher normalized Mean Absolute Error (MAE) of +0.0073, an average improvement of approximately +17.78% was achieved across all examples when comparing baseline and optimized models (with $\alpha = 0$).

Keywords Knowledge-embedded Optimization, Machine Learning, aleatoric uncertainty, logistical grid

Introduction

In recent years, machine learning (ML) has become a valuable tool across various fields, particularly in dealing with complex and voluminous data that is challenging for human understanding. This has led to improvements in production processes and enhanced precision in quality and yield. However, the optimization potential of ML methods is limited by the quality of input data, characterized by aleatoric uncertainty encompassing systematic, reproducible measurement inaccuracies and stochastic errors (Bland & Altman, 1996; Hüllermeier & Waegeman, 2021; Oberkampf & Ferson, 2007; Segalman et al., 2014; Taylor & Thompson, 1982). While systematic errors may be mitigated through technical enhancements, random errors persist, contributing to aleatoric uncertainties (Saris & Revilla, 2016).

Despite advancements in error estimation, complete elimination of random errors using statistical methods remains elusive (Buonaccorsi, 2010; Fuller, 2009; Hamidzadeh & Moradi, 2020; Krippendorff, 1970; Oberski & Satorra, 2013). Probabilistic methods such as Gaussian Process Regression (Schulz et al., 2018; Shi & Choi, 2011; Williams & Rasmussen, 1995) and Bayesian Neural Networks (Izmailov et al., 2021; Lampinen & Vehtari, 2001; Neal, 2012) offer uncertainty estimates, but their high computational complexity hampers practical applications, especially in productive environments (Elishakoff, 2000). Consequently, current research also focuses on developing computationally efficient non-probabilistic approaches to address uncertainties and enhance robust and resilient production optimization (Einbinder et al., 2022;

* Stefan M. Stroka
stefan.stroka@ams-osram.com

1 Department of Statistics,
Faculty of Mathematics, Informatics and Statistics,
LMU Munich, Germany

2 ams OSRAM Group,
93055 Regensburg, Germany

Hammer & Villmann, 2007; Xue & Deng, 2021). Quantifying uncertainties in optimization is a crucial research area, as the reliability of forecasts is exceptionally important in productive applications. Novel approaches (Einbinder et al., 2022; Psaros et al., 2023; Zhao & You, 2019) and method overviews (Y. Chen et al., 2018; Grossmann et al., 2016; Mohammadi & Farsijani, 2023; Ning & You, 2019; Sahinidis, 2004) highlight both benefits and challenges. Even with advanced optimization and error pattern recognition, random errors persist, primarily due to limitations in data input, which often represents only a subset of influential factors. Current research explores augmenting objective data with data-independent information, such as physical correlations (Greis et al., 2023; Jirasek & Hasse, 2023; Karniadakis et al., 2021; Y. Lu et al., 2017; Willard et al., 2022), expert knowledge (Y. Chen & Zhang, 2022; Farbiz et al., 2023; Guevara et al., 2019; Link et al., 2022; Malliaraki & Berditchevskaia, 2023; Wikner et al., 2020; Zhou et al., 2022), or mathematical constraints (Chuang et al., 2020; Kotłowski & Słowiński, 2009; Kurnatowski et al., 2021; Ma et al., 2021; Mangasarian & Wild, 2008; Sideris et al., 2023), to capture unnoticed features.

This paper introduces a novel optimization approach that incorporates data-independent knowledge about measurement uncertainty during training. The goal is to enhance supervised regression modeling and prediction of the target variable, considering the optimal delivery on logistic grids based on experts. Unlike other methods aiming to avoid measurement errors, this approach leverages additional information as prior knowledge to optimize production output. A hyperparameter allows manual weighting of the influence of data or data-independent knowledge.

The paper details the optimization approach and methodologies, emphasizing mathematical principles. The proposed approach assumes that a regression model based on deterministic measurements without considering uncertainties is often insufficient to adequately represent real-world variations. To demonstrate this, the paper presents the application and evaluation of the methodology using real production data from ams Osram International Group, anonymized through de- and encoding with a standard Variational Autoencoder (VAE) in a Python implementation.

Problem Formulation

The goal of an optimization method is to increase production yield while complying with a predefined delivery grid (by quality thresholds), without causing negative impacts on factors like computation time. Optimization involves striking a balance between conflicting objectives. In our case, this involves addressing the discrepancy between deterministic-viewed measurements with potential inaccuracies (metric value) and additional information from experts for assessing aleatoric uncertainties (assessed in ordinal classes). The challenge is to minimize discrepancies between these elements and find the optimal solution for improved production yield.

General Problem

Considering the assumed measurement inaccuracy, which contributes to the observed discrepancy between the delivery class based on the measurement and the true delivered class based on expert assessment, allows us to consider the deterministic measurement in a manner similar to a probability distribution. From a deterministic standpoint, a measured value (refer to Fig. 1) is always assigned to a specific logistic bin, which may deviate from the actual bin if there is considerable measurement uncertainty.

In non-probabilistic methodologies, a measured value (point measurement) is regarded as deterministic. However, the empirical repetition of measurements unveils fluctuations due to measurement errors, thereby exposing an uncertainty interval. In a deterministic perspective of the measured value taken once (without considering uncertainties), the measurement point is classified within the delivery bin ([40;60]). In the statistical view, considering empirically repeated measurements, the measured value can be treated as a probability-distributed random variable. This allows us to infer that other delivery bins ([0;20], [20;40], …, [80;100]) would also be statistically plausible from an empirical standpoint. Therefore, neglecting the measurement error inevitably results in a generalized interpretation of potential outputs. This challenge is prevalent in all statistical, non-probabilistic methods, which provide an expected value estimation without uncertainty quantification as a forecast. Figure 1 illustrates the drawback of a purely deterministic view of the measured value.

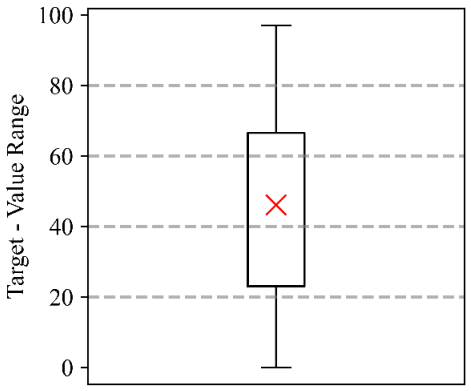


Fig. 1. Repeated measurements (with approximately normally distributed measurement error) are represented as a box plot with the delivery grid and limits indicated by grey dashed lines. The red cross

depicts an example for a deterministic measurement (measured only once).

Application Example in Opto-Semiconductor Production

In opto-semiconductor production, the process begins with the epitaxial growth of prefabricated silicon wafers, which undergo further intricate steps to produce a wafer containing completed LED chips (Härle et al., 2000). An epitaxial wafer can serve as the foundation for various chip types (1:n relationship), distinguished by light color, temperature, and luminosity. However, only one chip type can be produced for each epitaxialized wafer. The choice of further processing (production steering) directly influences chip properties, quantitative yield, and indirectly impacts quality. Consequently, the final product's yield and quality depend on the quality and characteristics of the epitaxial wafer.

Epitaxy, as a form of crystal growth, is inherently not entirely reproducible, despite strict production regulations, and introduces stochastic deviations due to its nature (Kimoto, 2016; Larkin, 1997; Matsunami & Kimoto, 1997). The subsequent processing of wafers is not uniformly standardized across all epitaxial wafers of the same production type. Instead, decisions for each wafer are made based on preliminary measurements or forecasts. To maximize yield and quality, determining the optimal production steering for each individual wafer is crucial.

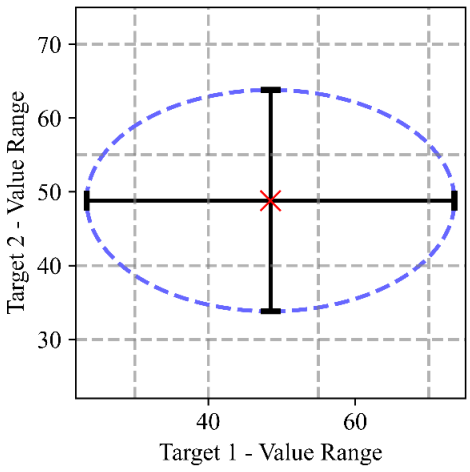


Fig. 2. Repeated measurement of two properties (each with a normally distributed measurement error). The probabilistic view shows the distribution over the logistic bins, while the deterministic view only indicates one possible bin.

The logistic grid, defined by two features—wavelength and brightness, is used to determine the optimal steering. Due to the two-dimensional nature of the grid, the measured value tuple can deviate in 2D directions (refer to Fig. 2). The precise direction and magnitude of the measurement error cannot be easily determined without complex procedures or repeated measurements.

Solution for Application Example

To ensure optimal wafer processing, it is crucial to evaluate outcomes in relation to the input of historical wafers and their corresponding steering. However, an input corrupted by measurement inaccuracies complicates a reliable forecast. One potential solution involves employing precise yet costly and complex measurement methods. Nevertheless, the high costs and time-intensive reorganization outweigh the benefits for a production-wide implementation. Alternatively, we can enhance the limited ability to reduce measurement inaccuracy through statistical methods by incorporating further information (unknown to the model).

Our solution leverages experience gained from previously produced wafers, combining it with data-independent information like expert or other external knowledge to optimize steering for deliveries. Using metrically scaled measured values as the target variable, a pointwise regression can identify inherent data correlations, possibly providing an unbiased expected value estimation. In the context of delivery grid-oriented production, the focus is not solely on the precise forecast of measured values. Instead, it centers on accurate deliveries while adhering to grid-defined thresholds, which define ordinal classification-classes.

Optimal steering for more accurate classification requires finding a balance between potentially conflicting regression and classification targets during the learning process. Therefore, the objective of this paper is to introduce the custom loss (CL) function, as combined regression and classification loss. This function allows regression training for improved classification on logistic delivery bins, providing minimal biased point estimation.

Method and Theoretical Application

The suggested CL function serves as a regression loss applicable to all ML methods facilitating iterative model training through loss minimization. Generally, ML methods utilize a loss function to guide model training by quantifying the disparity between predictions and actual outcomes. For instance, in standard regression, this involves measuring the difference between the estimated expected value and the observed value.

Given the convexity of the loss function, algorithms can seamlessly integrate with iterative optimization techniques like gradient descent. Prominent examples encompass Convolutional Neural Networks (CNN) (Du et al., 2018; Gupta et al., 2018), Recurrent Neural Networks (RNN) (Pascanu et al., 2013), Long Short-Term Memory Networks (LSTM) (Anh et al., 2023), and Reinforcement Learning (Baird, 1999) within the area of deep

learning. Additionally, support vector machines (SVM) (S. Lu & Jin, 2017; Zeyuan et al., 2009), logistic regression (Zou et al., 2019), and gradient boosting (Natekin & Knoll, 2013) are examples from various ML methodologies.
The main emphasis of this paper is on the methodology and its application, particularly in the context of modeling with measurement uncertainty. Our attention is not directed towards exploring potential ML applications. Consequently, we have chosen to exclusively employ eXtreme Gradient Boosting (XGBoost) (T. Chen et al., 2019) as the foundational framework for our analysis.

Fundamental Concept of 2-Target Feature Grid-Steering in Opto-Semiconductor Manufacturing

In the given application scenario, our objective is to meet delivery targets based on (intern) customer specifications. To achieve this, delivery grids are established using thresholds, primarily focused on critical properties such as wavelength and brightness. These grids represent an ordinal classification system for the two continuous measurement value ranges, essentially creating a two-dimensional measurable space.

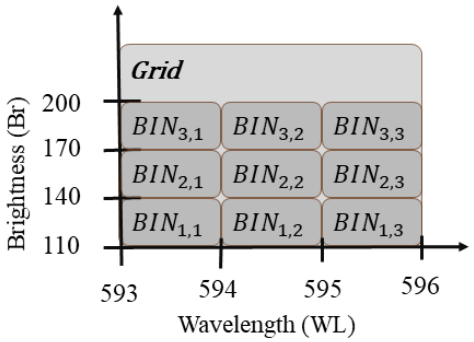


Fig. 3. Example of a logistical two-target delivery bin set. Critical properties are wavelength and brightness. Bins are defined for each target as ordinal classes.

It is assumed that deterministic measurements, (without considering measurement errors) may not accurately capture the true delivery quantity. Point measurements are categorized using the defined bin set, leading to the determination of delivery quantities. However, considering measurements with potential errors suggests that in certain cases, especially at the boundaries, misclassifications may occur, presenting an opportunity for optimizing yield. Since measurement errors cannot be precisely determined, they could be estimated using statistical methods, expert assessment, or other reliable approaches. For instance, a measured value might be reclassified within the range of the measurement error (standard deviation).
Given this consideration, when applied to all provided training data for the targets, a partially new classification within the range of the measurement error might lead to differences between measurement-based classes and actual delivery classes. Consequently, it becomes essential to predict potential deliveries not only pointwise, without considering measurement errors (regression), but also in an optimized manner considering the actual delivery to the grid (classification).

Optimization of 2-Feature Grid Steering

As a result, both measured values and ordinal delivery classes play a crucial role in optimization. The CL function enables training the regression under the condition of delivery classes, ensuring an optimal classification. This means achieving a simultaneous minimally biased expected value estimation for the delivery quantity forecast per bin. As a disclaimer, it is crucial that we need to train an optimized model for each target. Consequently, we do not engage in multi-target optimization, as the ordinality operates solely in one dimension.
The CL as combined loss is formulated as follows:

$$CL = \alpha * Regr_{loss} + (1-\alpha) * Clf_{loss}.$$

The regression loss function is built on the mean-squared-error (MSE), which calculates the mean square distance between the estimated point $\hat{y}$ and the actual measurements y. Specifically,

$$Regr_{loss} = MSE(\hat{y}, y) = E[(\hat{y}-y)^2]$$
$$\approx \frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - y_m)^2 \overset{iid}{\Leftrightarrow} (\hat{y}-y)^2$$

and

$$MSE'(\hat{y}, y) \approx 2 * (\hat{y}-y)$$

where $\dim(y) = \dim(\hat{y}) = (M, 1)$ with M observations. The second part of the loss (Clf_{loss}) is the distance to the true class (*dist2trClass*), determined by the distance between the prediction $\hat{y}$ and the upper or lower threshold of the true class $class_y$ (the threshold depends on the piecewise-defined function) using Fuzzy logic. The class of the prediction, denoted as $class_{\hat{y}}$, corresponds to the class in which the prediction falls. This can be expressed as:

$$Clf_{loss} = dist2trClass(\hat{y}, y)^2$$

with

$$Clf'_{loss} = 2 * dist2trClass'(\hat{y}, y)$$
$$= \begin{cases} 2 * (\hat{y} - class_y^{lower\ t.}); class_{\hat{y}} < class_y \\ 2 * (\hat{y} - class_y^{upper\ t.}); class_{\hat{y}} > class_y \end{cases}$$

In this context, a classification prediction is considered correct when the estimated value surpasses the threshold, placing it in the correct class. Ordinal classes cannot be directly measured using a metric distance measure, but fuzzy logic allows us to treat ordinal classes as pseudo-metric, facilitating distance determination.

In addition to the loss function, the gradient and the Hessian matrix of the same function are fundamental for optimization using the gradient descent optimization method. The focus is not on explicitly minimizing the loss function, but rather its gradient. Hence, the gradient is defined as:

$$CL'(\hat{y}, y) = 2 * \alpha * (\hat{y} - y) + 2 * (1 - \alpha) * dist2trClass'(\hat{y}, y).$$

A metric measurement as well as a delivery classification (class) is therefore a prerequisite for optimization. It is not important whether these targets are contradictory. In the scenario illustrated in Fig. 4, measurements and the delivered bin align, eliminating the need for optimization

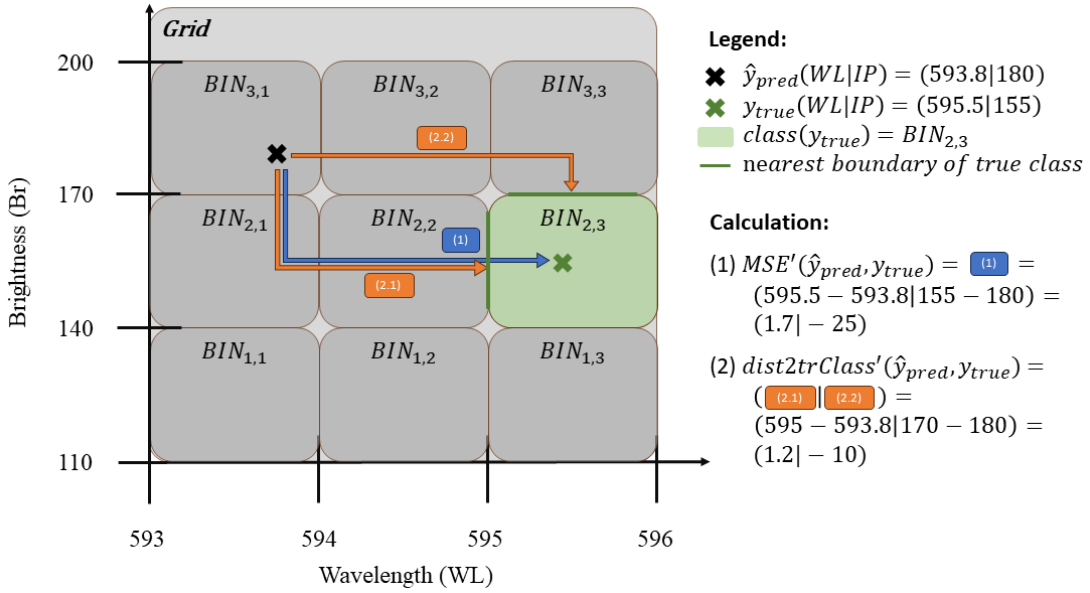


Fig. 4. Example of an iteration in the training process and the calculation of the CL. Measured values and information about true delivery match on the same bin.

In the second case, we examine the scenario depicted in Figure 5. It illustrates a discrepancy arising from measurement inaccuracies between the measurement and the actual delivered bin. This discrepancy is addressed by the hyperparameter α, which plays a crucial role in weighting the CL.

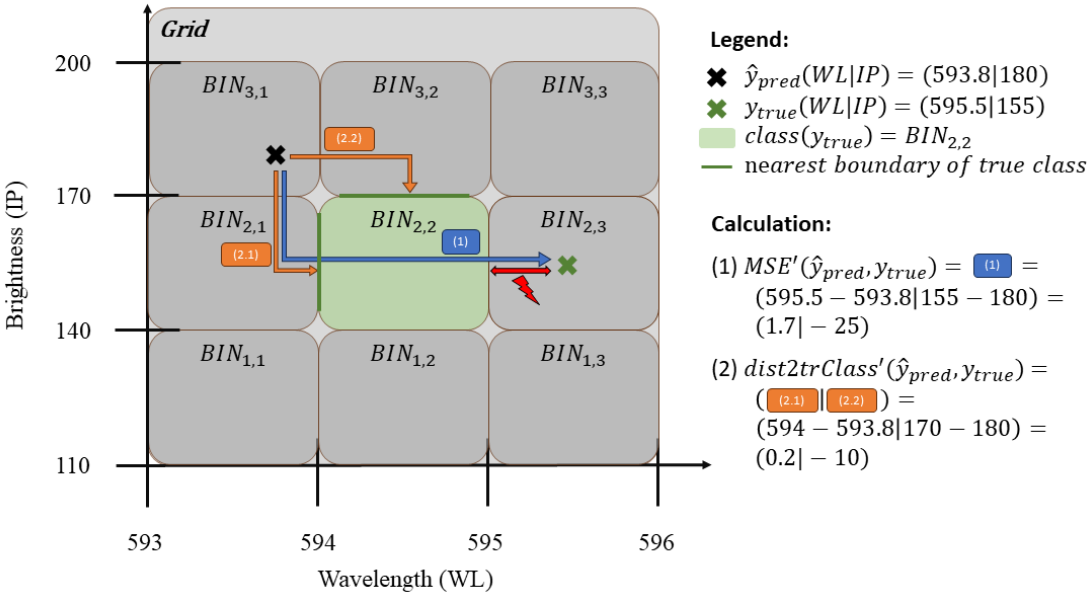


Fig. 5. Example of an iteration in the training process and the calculation of the CL. Measured values and information about true delivery differ.

While the CL function facilitates both classification and point estimation in a joint model training framework, it is designated as a regression loss function. Additionally, the CL function exhibits convexity as a notable property. Unlike optimization methods that rely on repeated random execution, the convex nature of the function allows for the utilization of the gradient descent method to attain optimal results.

ALGORITHM 1: CUSTOM LOSS FUNCTION

Input: *y_true (historical data), y_predicted (model prediction)*
Output: *gradient, hessian matrix*
Get class_true for y_true (per observation)
1 *//Sort continious value y_true into given grid*
Get class_predicted for y_predicted (per obs.)
2 *//Sort continious value y_predicted into given grid*
Calculate delta: *delta = delta(class_true, class_predicted)*
3 *//Difference of integer values of classes*
Compute first gradient part: *grad_1 = (y_pred - y_val)*
4 *//first derivative of the mean_absolut_error (MAE)*
5 ***Calculate second gradient part:*** *assign grad_2 = 0*
6 ***If delta > 0: do*** *(true class is higher than the predicted class)*
assign grad_2 = y_predicted - lower_boundary
7 *//lower boundary value of class_true*
8 ***Else if delta < 0:*** *do //true class is smaller than the predicted class*
assign grad_2 = y_predicted - higher_boundary
9 *//higher boundary value of class_true*
10 ***end***
Compute combined gradient:
*grad = alpha * grad_1 + (1 - alpha) * grad_2*
11 *//alpha as predefined influence constant*
Compute hessian matrix: *hess = d(grad)*
12 *//Derive the derivative of grad*

Gradient Descent

The gradient descent optimization method offers the advantage of sequential progressive improvement based on the gradient function and Hessian matrix. When dealing with a twice-differentiable loss function, expressed as a Taylor approximation, numerical approximation of the complex function becomes feasible. In practical applications, this method employs a convex loss function to assess the deviation between the estimate and the measured value, enhancing the estimate through a fixed learning rate (lr). This iterative process continues until the absolute minimum of the gradient, represented by the first derivative of the loss function, is reached. While the optimization process is primarily designed for convex functions, it is theoretically applicable to non-convex functions. However, in the case of non-convex functions, there is no singular, uniform absolute minimum.
The upper inset in Fig.6 provides a 3D illustration of the CL value range with a fixed $\alpha = 0.5$ (the function's illustration varies depending on the α value). The lower inset presents a cross-section of the CL function, representing the gradient of MSE and illustrating a potential optimization process.

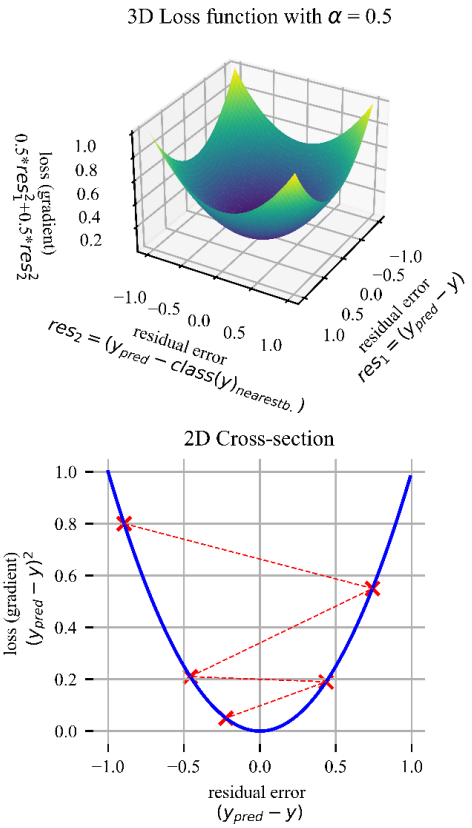


Fig. 6. 3D & 2D visualization of the combined loss function with weighting parameter $\alpha = 0.5$. Upper inset shows the combined loss resulting from both parts of the loss function. The lower inset shows the gradient descent (red line) process, which reduces the loss, resulting in a better model fit for a cross section of the upper inset.

XGBoost

This paper employs the tree-based xGBoost regression as a powerful and computationally efficient methodology. XGBoost is known for its flexibility and stands as a state-of-the-art approach, offering the advantage of comprehensible modeling and intuitive integration of customized loss functions.
XGBoost is a supervised learning method, which servs both classification and regression purposes. The algorithm builds upon an ensemble of independent decision trees, where each tree categorizes input features into nodes and leaves, eventually providing estimations. While a single decision tree (considered a weak learner) might have

limitations in capturing complex data correlations, the ensemble of multiple trees (a strong learner) averages estimates, yielding a more robust and valid result (T. Chen & Guestrin, 2016).

The XGBoost regression is designed to sequentially enhance the initial forecast over a fixed number of iterations. Given its nature as a supervised learning approach, labeled data in tabular form is essential, where dependent target variables and independent input variables are clearly defined.

This paper specifically explores an application involving two target variables. However, for clarity, we focus on describing the boosting methodology within the context of one target variable, with the understanding that the insights gained are applicable across both targets.

In the initial step, the model establishes the mean value of the target variable and uses it as the starting prediction for each observation in the training data. Let X $(\dim(X) = m\,x\,n)$ be the set of m observations with n independent variables and y $(dim(y) = m)$ the measurements of the target variable. It follows that $\hat{y}^{i=0} = mean(y)$ is the first estimate in iteration $i = 0$. As iterations progress, the model attempts to find predictions with smaller deviations from the true values in the training data and eventually minimizes the loss. This involves creating a regression decision tree for each iteration i and calculating pseudo residuals r^i. The pseudo residuals, denoted as $r^i = (r_k^i)_{k\in\{1,\dots,K\}}$, are computed for each k-th end node of the current i-th decision tree. These residuals represent the mean deviation between the current estimate $\hat{y}^i$ and the actual measurement y. Thus, for the i-th decision tree with K end nodes, and considering the set of observations M_k in the k-th end node, the calculation is expressed as follows:

$$r^i = \begin{pmatrix} r_1^i \\ \dots \\ r_K^i \end{pmatrix} = \begin{pmatrix} \frac{1}{|M_1|}\sum_{m\in M_1} \mathrm{CL}'(\hat{y}^i, y^i) \\ \dots \\ \frac{1}{|M_K|}\sum_{m\in M_K} CL'(\hat{y}^i, y^i) \end{pmatrix}.$$

The final step in each iteration i involves updating the estimates using the calculated pseudo residuals r^i and a fixed learning rate lr:

$$\hat{y}^i = \hat{y}^{i-1} + r^i * lr$$

with $lr \in (0,1]$.

Eventually, we achieve the best possible mean expected value, which is optimized by applying the CL function with regard to the actual classes or delivery bins. To ensure the best model fit, avoiding overfitting or underfitting and providing accurate forecasts, the hyperparameters of the models undergo successive tuning via Bayesian optimization. Additionally, 5-fold cross-validation is employed to validate and refine the model's performance across different datasets.

Experiments

In this section, we demonstrate the practical application and validation of our approach (in combination with a XGBoost regressor) within the context of real-world examples.

Experimental Objectives

The upcoming experiment highlights the substantial impact of neglecting measurement errors in the modeling process. It aims to illustrate, through various realistic examples, how incorporating and correcting measurement errors can lead to a noticeable increase in yield with relatively low bias. Additionally, we assess the impact on computational time to provide a comprehensive understanding. The comparison involves a standard regression (baseline) and a regression model optimized specifically for the measurement error classes, evaluated using standard metrics. This investigation allows us to gauge the significance of accounting for measurement errors in improving model performance and overall yield.

Evaluation Metrics

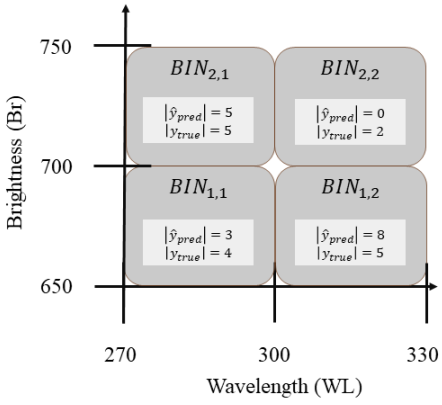
We assess our experiment using standard metrics for both classification and regression. For classification, we use the weighted F1-score due to imbalanced class distribution. In regression, we measure the performance with the mean-absolute-error (MAE), which we normalize with the mean value of the target variable to ensure comparability between the data sets with different value ranges. The MAE is appropriate in this context for several reasons. Firstly, it allows for an absolute comparison, considering bin limits and measurement error values. Secondly, employing MAE avoids potential confusion with the MSE (training metric). The evaluation is carried out independently for each target. In our specific use case, a distinct model is trained for each target (WL and IP). Both models are then evaluated individually using MAE. The final MAE utilized for comparing the combined model output is computed as the mean of the normalized MAE of both models.

Beyond these technical metrics, we also consider a practical performance indicator crucial for the company's operations. This key indicator is the Bin_{KPI}, which evaluates how well our deliveries on the bins align with target requirements in terms of characteristics and quantity. It quantifies the ratio between the predicted and actual deliveries within bins after processing. The Bin_{KPI} does not necessitate an exact match of classes for each observation but instead centers on the overall relationship between the predictions and the actual delivered results. A match of the classes per observation is not explicitly necessary (likewise classification accuracy), the main goal is a quantitative match of predicted and true classes over

all bins (likewise the intersection ratio of two overlapping histograms).
Given a logistic grid represented as g with $dim(g) = k\ x\ l$, the grid consists of $(k\ x\ l)$ −bins. Each bin has two values after classifying the pointwise prediction values into the bins. For the kl −bin, $\widehat{Bin}_{kl}$ represents the count of predicted values within the kl −bin $(\#|\{\forall_{m=1}^{M}: \hat{y}_m \in Bin_{kl}\}|)$ (predicted classes), and $\overline{Bin}_{kl}$ indicates the number of measured values in the same bin $(\#|\{\forall_{m=1}^{M}: y_m \in Bin_{kl}\}|)$ (true classes). Bin_{KPI} is defined as follows:

$$Bin_{KPI} = \frac{\sum_{k,l} \min(\widehat{Bin}_{kl}, \overline{Bin}_{kl})}{\sum_{k,l} \overline{Bin}_{kl}}$$

Fig. 7 illustrates an example for this calculation.



$$Bin_{KPI} = \frac{\widehat{Bin}}{\overline{Bin}} = \frac{\min(5,5) + \min(0,8) + \min(3,4) + \min(8,5)}{16} = \frac{13}{16}$$

Fig. 7. Example of Bin_{KPI} calculation based on a quantity-based delivery in logistical bins. For each logistical bin, the minimum value represents the quantitative accuracy rate. Hence, Bin-KPI is the sum of the quantitative accuracy in relation to the number of given classes.

In the calculation of Bin_{KPI}, the quantitative match value is determined for each bin, represented by the minimum of the predicted and true values within that specific bin. Quantitative agreement implies that the observation and its prediction does not need to match explicitly, but rather the quantity distributed over the grid. The sum of these minima across all bins yields the quantitative agreement in the overall grid. Finally, the quantitative agreement on the grid is put in relation with respect to the total quantity of all observations, providing the percentage of hits.

Data Collection

For our evaluation, we utilize data generated through a Variational Autoencoder (VAE), trained on authentic production data from four distinct opto-semiconductor products. This approach enables the modeling of the production data and inherent structures, preserving data confidentiality. The underlying data comprises measurements of wavelength and brightness before (independent feature) and after processing (target feature).
Measured features are wavelength and brightness for fixed number of sample positions per wafer (for pre and final measurements). The delivery grid, its quality thresholds and the true delivery in Bins are chosen based on expertise. Therefore, each target variable (wavelength or brightness) has a direct target value (feature measurement) and a latent ordinal-scaled target class (delivery bin based on expert assessment). Additional information about the real data and the VAE model can be found in the appendix.

Sample Description (Descriptive Analysis)

In the production of opto-semiconductors, various measurement methods are employed at specific production stages for monitoring and optimizing the process. The data presented in this section originates from authentic production measurements, utilizing a VAE model to simulate real data while ensuring confidentiality.

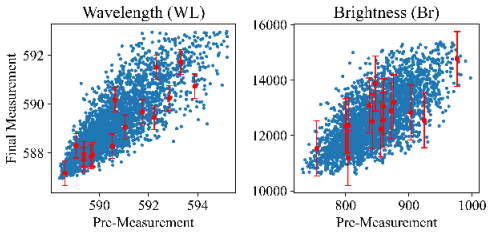


Fig. 8. (In-Output) Scatterplot comparing pre- and final-measurements for wavelength and brightness (depicted in blue). The error bars (in red) represent the assumed measurement uncertainty, particularly for final measurements.

The spread of measurement errors within the value range is depicted in Figure 8, showcasing variations in measurement accuracy based on the method and measured product type. The focus lies on the measurement error of the final stage, given its direct impact on delivery and influence on the model learning process. A high, unnoticed measurement error can lead to a higher error rate in the final product delivery.
A detailed example of the Product 1 dataset (ground truth) is presented in Figure 9. It illustrates the distribution of measured values (1) and the actual delivery represented in bins (4) for both features, as showcased in the insets. The histograms (2,3) visually depict the disparity in classes, revealing instances of incorrect delivery (due to unnoticed impact factors like aleatoric uncertainties). Overlaid on the scatterplot is a 2D histogram, constructed with grid thresholds for the given values. The displayed percentage values represent the ratio within the histogram bins. Inset 4 provides information about

the true delivered classes based on expert assessments, also represented by a 2D histogram.

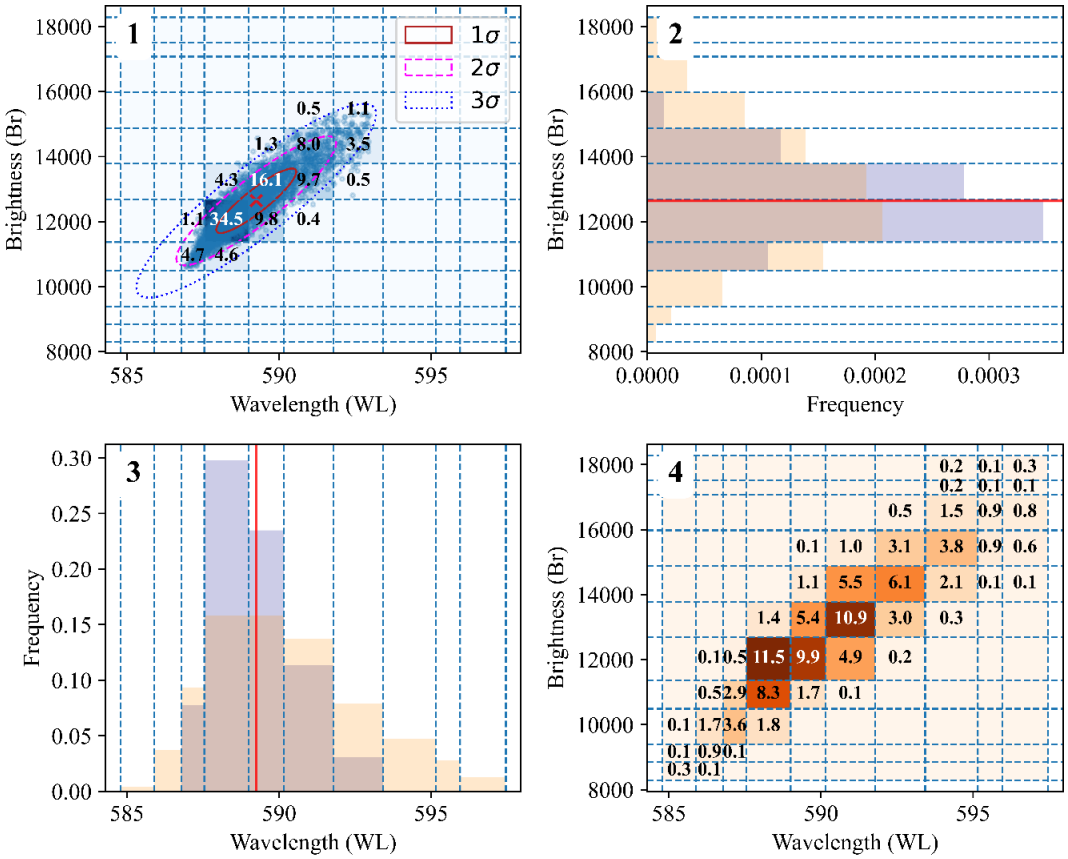


Fig. 9. Descriptive visualization of the model input, incorporating measurement data (depicted in blue) and additional delivery-related information (depicted in orange). The scatterplot (inset 1) illustrates measurements for two targets with possible variance and the percentage of deliveries based on these measurements. Histograms (insets 2 and 3) offer a quantitative comparison between the quantity in bins based on measurements (blue) and expert assessments (orange) for Brightness and Wavelength. Inset 4 presents a confusion matrix of the two targets divided into logistic delivery bins as a percentage, relying solely on expert assessments. The distribution of measured values (1) and actual delivery in bins (4) is depicted in the insets, highlighting discrepancies in classes and, consequently, instances of incorrect delivery.

Model Training, Optimization and Evaluation

We use a tree-based XGBoost for our application, incorporating a custom loss function. The model undergoes training and validation using 5-fold cross-validation on the training dataset, and its performance is evaluated on the out-of-sample test data. The dataset undergoes random partitioning into test (20%) and train (80%) sets. Within the training set, a further division is made into validation (20%) and a smaller train (80%) subsets. The CL function incorporates prior information obtained from expert or external knowledge, representing an ordinal class. In our example, this ordinal class corresponds to a bin in the 2D grid during the training process.

Our method models the correlations between pre-measurements and final measurements (post-wafer processing) and optimizes predictions based on the ordinal delivery class, considering a defined weighting parameter α. Given the multitude of tunable hyperparameters in XGBoost, we conduct Bayesian hyperparameter optimization using the training and validation data. We employ the ***hyperopt*** (Bergstra et al., 2013) package for this purpose. To assess performance, we compare our approach with a baseline model, also tuned for hyperparameters. The baseline model employs XGBoost regression with a mean squared error (MSE) loss function (, which corresponds to the CL function with $\alpha = 1$).

Results and Interpretation

Next, we assess the results for four generated sample datasets. To gain detailed insights, our analysis concentrates on the graphical evaluation of Product 1. Initially, we examine the impact of the new weighting parameter, followed by a comparison of

the optimized results with the baseline. Subsequently, we present a comprehensive table detailing the outcomes for the remaining products, comparing the baseline with the best-performing model.

Results and Evaluation

As a weighting parameter, the new hyperparameter α serves as a weighting parameter that directly influences the model training process. To evaluate its impact, a comprehensive analysis was conducted below, utilizing a list of decreasing α values. While the α value is defined on a continuous range, it must be fixed before training a model. Each point in the analysis represents the outcomes of an independent and optimized model.

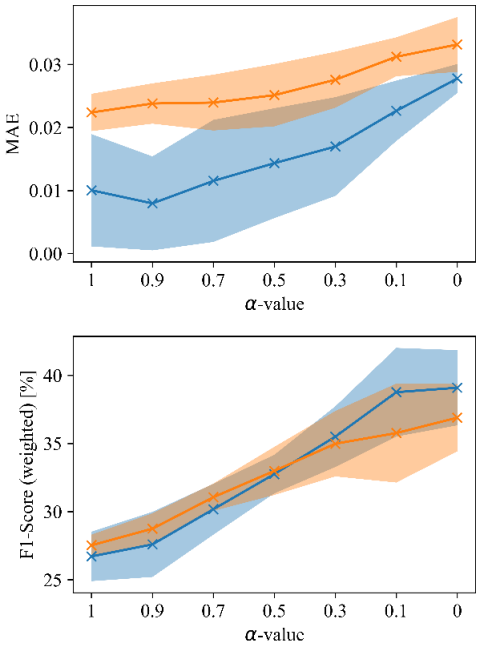


Fig. 10. In-sample (blue) and out-of-sample (orange) evaluation of distinct models for a range of α-values. Results are depicted using mean expectation estimates (line) and standard deviations (colored area).

When examining the models with decreasing α values (see Fig. 10), the normalized Mean Absolute Error (MAE) experiences a relatively modest increase, while the classification results, as measured by the F1-score, exhibit a significant improvement. The comparison between in-sample and out-of-sample data further validates that the optimized models are not overfitted.

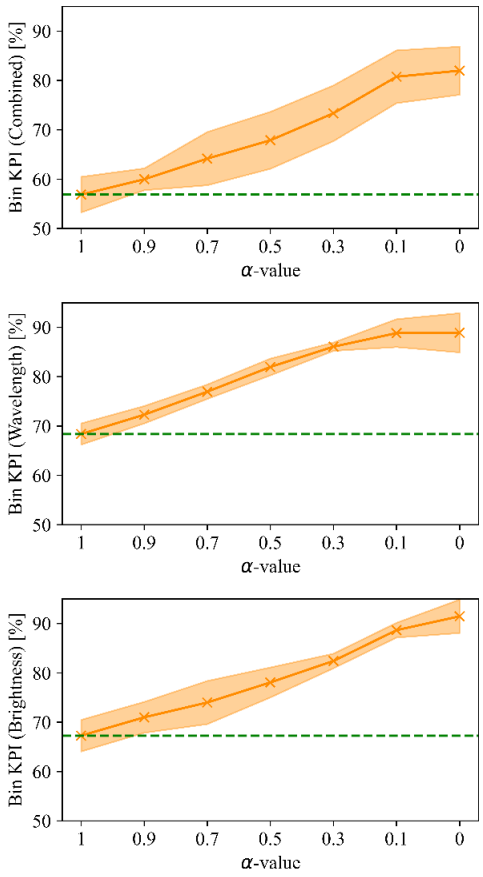


Fig. 11. Out-of-sample evaluation of the actual, quantitative delivery compared to the forecast on the defined logistical delivery grid. Expected values based on the model and α-value, along with standard deviation intervals. Bin-KPI plot for wavelength, brightness, and the mean of both, including the standard regression with $\alpha = 1$ (baseline; green).

In a productive application, the main optimization goal is often on maximizing yield. Figure 11 illustrates how adjusting the α-value can influence the prediction, leading to improved Bin KPI. The strictly monotonic nature of the mean curve demonstrates a maximum improvement in the combined Bin KPI of approximately +26% at the high point ($\alpha = 0$) compared to the baseline ($\alpha = 1$). In this scenario, the normalized MAE increases by 0.0108 (refer to Fig. 10).

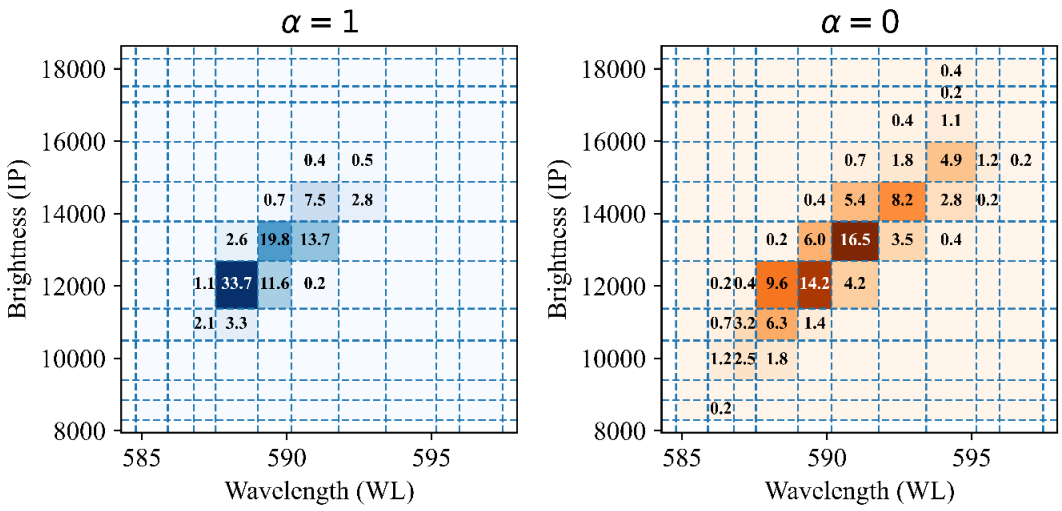


Fig. 12. Percentage delivery quantity forecast in logistical delivery grids for comparable models with $\alpha = 1$ (baseline) and $\alpha = 0$ for the dataset of Product 1.

The considerable improvement in classification results, despite a moderate impact on the expectation estimate, is depicted in Figure 12 through a direct comparison. When comparing $\alpha = 1$ (baseline) with $\alpha = 0$, we observe a more dispersed distribution across logistic bins for the lower alpha. The choice of alpha can be tailored based on the optimization goal and the reliability of data-independent information regarding the quantity of delivery. Finally, Table 1 offers an overview for a direct comparison with other example datasets.

Table 1
Comparison of the baseline and the model with optimal α-values for four selected products (= data input samples). Various metrics used for training/validation and evaluation on in- and out-of-sample are presented. Each metric is reported as the mean and standard deviation from simulation, representing repeated identical training with only hyperparameter optimization varied, reflecting the model uncertainty.

	Metrics		sample	Product 1		Product 2		Product 3		Product 4	
				α = 1	α = 0.0	α = 1	α = 0.0	α = 1	α = 0.0	α = 1	α = 0.0
Training/ Validation	MAE	μ (σ)	in	0.0124 (0.0038)	0.027 (0.001)	0.0241 (0.0068)	0.0271 (0.0053)	0.0281 (0.0096)	0.0492 (0.0027)	0.0095 (0.0022)	0.0096 (0.0022)
			out	0.0216 (0.0012)	0.0346 (0.0022)	0.0373 (0.0018)	0.0384 (0.0025)	0.0438 (0.0034)	0.0565 (0.0022)	0.0166 (0.0007)	0.0188 (0.0015)
	F1-score	μ (σ)	in	26.17 (0.88)	39.92 (1.88)	39.55 (4.97)	37.79 (2.95)	51.83 (2.43)	48.23 (1.06)	24.13 (2.67)	25.2 (3.27)
			out	27.61 (0.3)	35.94 (1.37)	27.66 (0.85)	31.21 (1.34)	43.39 (1.06)	44.5 (1.27)	17.7 (0.48)	18.08 (0.94)
Evaluation	Bin-KPI		out	56.44	83.21	48.39	55.3	58.83	81.02	51.82	67.06
	Bin-KPI WL	μ	out	68.47	89.37	66.83	73.67	80.71	87.41	61.6	73.39
	Bin-KPI IP		out	66.53	92.02	64.29	70.59	89.28	87.55	68.78	80.96

On average, considering all product examples, an enhancement in Bin KPI Accuracy of approximately +17.8% is attained by accepting a bias, with an increased MAE by 0.0073, when comparing the values for $\alpha = 0$ to the baseline.

Interpretation for Application Field

The usage of data-independent information for further evolving or refining a data-based model always involves risks, primarily due to the sometimes untraceable reliability, such as the subjective nature of expert assessments. Nevertheless, in real-life applications, this independent assessment has often proven beneficial and valuable for optimization. Unlike data, expertise is not restricted to a limited sample of the actual ground truth and may account for unnoticed influencing factors. Consequently, it is promising to apply this methodology productively, optimizing for reliability and incorporating biased forecasts.

A deterministic approach to conditions, such as measurements in a variable and uncertain environment like production, may come with the disadvantage of potential loss and reduced yield. Above all, it is frequently observed that purely data-based models possess limited understanding of the true nature of production and may reveal correlations that are biased or misleading from an expert's perspective.

Conclusion

In this paper, we presented an ML-based optimization methodology that extends and improves data-driven models with data-independent information (such as expert knowledge about measurement uncertainties).

In the past, data-based approaches have clearly proven themselves in a productive environment. Due to the purely data-dependent modeling, model quality and improvements are limited by the given data and its quality. Aleatory uncertainties, such as reproducible measurement inaccuracy or stochastic measurement errors, therefore have a significant influence on model reliability and forecast accuracy. Neglecting unobserved influences, such as measurement uncertainties, can lead to a model determining an optimal mean estimator but does not learn the actual inherent spread (variance). In order to model both targets simultaneously and optimized, a combined learning on the training data in combination with data-independent information is necessary.

For this purpose, we used a customized loss function, which biases a metric regression estimation towards an improvement of the classification accuracy on logistic grid. The impact on the regression estimation depends on an additional hyperparameter, which can be optimized. The customized loss combines a metric loss with a classification loss, which can be determined as a distance metric by applying fuzzy logic.

The key advantage of our methodology is therefore that we can incorporate data-independent information, such as expert knowledge about measurement uncertainties, into the model training in order to introduce possibly unobserved influences. introduce possibly unobserved influences to the model. Furthermore, the hyperparameter in the CL function allows the target influences to be weighted in order to determine the trade-off minimum for the most accurate (delivery quantity) classification with the lowest possible bias in the mean forecasts. A further advantage arises from considering the additional information as classes. Metric, non-metric information or even subjective estimates can thus be used for model training.

The CL function is also robust despite the classification of imbalanced data due to the fuzz logic approach.

For the evaluation, we used real production data with two target variables, which were anonymized with a variational autoencoder. Delivery quantities and two-dimensional logistic delivery grids were specified based on experts, which showed, descriptive analysed, a deviations from measurement data. The goal was a data-based regression estimation, which was biased with respect to data-independent information in the form of actual delivery quantities in classes/bins of a logistic grid.

The acceptance of a slightly biased estimation led to a significantly better average classification accuracy of the actual deliveries on the given logistic bins. The optimized models were evaluated in comparison to the standard regression (baseline) with regression (MAE) and classification metrics (weighted F1-score, Bin-KPI). In the out-of-sample evaluation of the data sets of the four sample products, an average improvement of +~17.78 % in the productive metric Bin-KPI was achieved with a deterioration in the MAE of +0.0073.

A practical application, as in the example case at ams Osram International GmbH, arises when there are supposedly measurement errors in the measurement data and a deviating quantity delivery. In practice, the production is expected to deliver the quantity in accordance with predefined quality limits represented by a logistical grid. In reality, products that are at or near the thresholds can be classified differently based on expert assessments or within the standard deviation of measurement inaccuracy in order to cover the requested quantity delivery more optimally.

In this paper, the optimization methodology has focused exclusively on the deviation due to aleatoric uncertainty on the measurement of the target variable. Future research could extend this by considering and incorporating the uncertainties in the input feature data. Also, with regard to epistemic uncertainty, we only consider the simulated model uncertainty through repeated training. In order to incorporate the actual epistemic uncertainty, the optimization could be extended to include common methods for determining the prediction intervals. A Bayesian approach, i.e. a combination of the prior (data-independent information) and the likelihood (measurement-based model), could also provide an approach for developing PI with regard to the optimization procedure shown.

Acknowledgements

This work was supported by ams OSRAM Group.

Data Availability Statement

The production data used for generating the dataset is confidential. The Python script employed for generating the data is accessible upon request.

Appendix

Table 2
VAE generated data for four products using real data from opto-semiconductor production. This includes details about the size and characteristics of the real data, the settings of the VAE model including the loss function, and the size of the generated data sample.

Real Data				VAE Layers				VAE Details			Generated Data
Product	Wafers	Sample Size	Properties	Input	Hidden	Latent	Output	Epochs	Distribution	Loss	Sample size
1	176	7213	approx. symmetric	4	4000	1000	4	30	Normal	(1)	7213
2	212	8688	approx. symmetric	4	4000	1000	4	30	Normal	(1)	8688
3	73	2986	right skewed	4	9000	500	4	20	Gamma	(2)	2986
4	419	17173	right skewed	4	9000	500	4	20	Gamma	(2)	17173

(1) Binary_cross_entropy + kl_divergence (2) Binary_cross_entropy + gamma_entropy + kl_divergence

Declarations

Conflicts of Interest

The authors declare that they have no conflict of interest.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Stefan M. Stroka. The first draft of the manuscript was written by Stefan M. Stroka and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Ethical standards were maintained, and informed consent was obtained throughout the data collection process.

References

Anh, D. T., Thanh, D. V., Le, H. M., Sy, B. T., Tanim, A. H., Pham, Q. B., Dang, T. D., Mai, S. T., & Dang, N. M. (2023). Effect of Gradient Descent Optimizers and Dropout Technique on Deep Learning LSTM Performance in Rainfall-runoff Modeling. *Water Resources Management*, *37*(2), 639–657. https://doi.org/10.1007/s11269-022-03393-w

Baird, L. C. (1999). *Reinforcement learning through gradient descent* [PhD Thesis, Carnegie Mellon University Pittsburgh, PA, USA]. http://reports-archive.adm.cs.cmu.edu/anon/anon/home/ftp/usr/ftp/1999/CMU-CS-99-132.pdf

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *Proceedings of the 12th Python in science conference*, *13*, 20. https://pdfs.semanticscholar.org/d4f4/9717c9adb46137f49606ebbdf17e3598b5a5.pdf

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error. *Bmj*, *312*(7047), 1654.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. CRC Press.

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*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., & Khotilovich, V. (2019). Package 'xgboost'. *R version*, *90*, 1–66.

Chen, Y., Yuan, Z., & Chen, B. (2018). Process optimization with consideration of uncertainties—An overview. *Chinese Journal of Chemical Engineering*, *26*(8), 1700–1706. https://doi.org/10.1016/j.cjche.2017.09.010

Chen, Y., & Zhang, D. (2022). *Integration of knowledge and data in machine learning* (arXiv:2202.10337). arXiv. https://doi.org/10.48550/arXiv.2202.10337

Chuang, H.-C., Chen, C.-C., & Li, S.-T. (2020). Incorporating monotonic domain knowledge in support vector learning for data mining regression problems. *Neural Computing and Applications*, *32*(15), 11791–11805. https://doi.org/10.1007/s00521-019-04661-4

Du, S., Lee, J., Tian, Y., Singh, A., & Poczos, B. (2018). Gradient Descent Learns

One-hidden-layer CNN: Don't be Afraid of Spurious Local Minima. *Proceedings of the 35th International Conference on Machine Learning*, 1339–1348. https://proceedings.mlr.press/v80/du18b.html

Einbinder, B.-S., Romano, Y., Sesia, M., & Zhou, Y. (2022). Training Uncertainty-Aware Classifiers with Conformalized Deep Learning. *Advances in Neural Information Processing Systems*, *35*, 22380–22395.

Elishakoff, I. (2000). *Possible limitations of probabilistic methods in engineering*. https://asmedigitalcollection.asme.org/appliedmechanicsreviews/article-abstract/53/2/19/401531

Farbiz, F., Habibullah, M. S., Hamadicharef, B., Maszczyk, T., & Aggarwal, S. (2023). Knowledge-embedded machine learning and its applications in smart manufacturing. *Journal of Intelligent Manufacturing*, *34*(7), 2889–2906. https://doi.org/10.1007/s10845-022-01973-6

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons. https://books.google.de/books?hl=de&lr=&id=Nalc0DkAJRYC&oi=fnd&pg=PR3&dq=Measurement+error+models&ots=JQy4UxErg6&sig=0_j7EgeTwEwZMr9axzDdl14zxuI

Greis, N. P., Nogueira, M. L., Bhattacharya, S., Spooner, C., & Schmitz, T. (2023). Stability modeling for chatter avoidance in self-aware machining: An application of physics-guided machine learning. *Journal of Intelligent Manufacturing*, *34*(1), 387–413. https://doi.org/10.1007/s10845-022-01999-w

Grossmann, I. E., Apap, R. M., Calfa, B. A., García-Herreros, P., & Zhang, Q. (2016). Recent advances in mathematical programming techniques for the optimization of process systems under uncertainty. *Computers & Chemical Engineering*, *91*, 3–14. https://doi.org/10.1016/j.compchemeng.2016.03.002

Guevara, J., Zadrozny, B., Buoro, A., Lu, L., Tolle, J., Limbeck, J. W., & Hohl, D. (2019). A machine-learning methodology using domain-knowledge constraints for well-data integration and well-production prediction. *SPE reservoir evaluation & engineering*, *22*(04), 1185–1200.

Gupta, H., Jin, K. H., Nguyen, H. Q., McCann, M. T., & Unser, M. (2018). CNN-Based Projected Gradient Descent for Consistent CT Image Reconstruction. *IEEE Transactions on Medical Imaging*, *37*(6), 1440–1453. https://doi.org/10.1109/TMI.2018.2832656

Hamidzadeh, J., & Moradi, M. (2020). Enhancing data analysis: Uncertainty-resistance method for handling incomplete data. *Applied Intelligence*, *50*(1), 74–86. https://doi.org/10.1007/s10489-019-01514-4

Hammer, B., & Villmann, T. (2007). *How to process uncertainty in machine learning?*

Härle, V., Hahn, B., Lugauer, H.-J., Bader, S., Brüderl, G., Baur, J., Eisert, D., Strauss, U., Zehnder, U., Lell, A., & Hiller, N. (2000). GaN-Based LEDs and Lasers on SiC. *Physica Status Solidi (a)*, *180*(1), 5–13. https://doi.org/10.1002/1521-396X(200007)180:1<5::AID-PSSA5>3.0.CO;2-I

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. G. (2021). What Are Bayesian Neural Network Posteriors Really Like? *Proceedings of the 38th International Conference on Machine Learning*, 4629–4640. https://proceedings.mlr.press/v139/izmailov21a.html

Jirasek, F., & Hasse, H. (2023). Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures. *Annual Review of Chemical and Biomolecular Engineering*, *14*(1), 31–51. https://doi.org/10.1146/annurev-chembioeng-092220-025342

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L.

(2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), Article 6. https://doi.org/10.1038/s42254-021-00314-5

Kimoto, T. (2016). Bulk and epitaxial growth of silicon carbide. *Progress in Crystal Growth and Characterization of Materials*, *62*(2), 329–351.

Kotłowski, W., & Słowiński, R. (2009). Rule learning with monotonicity constraints. *Proceedings of the 26th Annual International Conference on Machine Learning*, 537–544. https://doi.org/10.1145/1553374.1553444

Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Kurnatowski, M. von, Schmid, J., Link, P., Zache, R., Morand, L., Kraft, T., Schmidt, I., Schwientek, J., & Stoll, A. (2021). Compensating Data Shortages in Manufacturing with Monotonicity Knowledge. *Algorithms*, *14*(12), Article 12. https://doi.org/10.3390/a14120345

Lampinen, J., & Vehtari, A. (2001). Bayesian approach for neural networks—Review and case studies. *Neural networks*, *14*(3), 257–274.

Larkin, D. J. (1997). An Overview of SiC Epitaxial Growth. *MRS Bulletin*, *22*(3), 36–41. https://doi.org/10.1557/S0883769400032747

Link, P., Poursanidis, M., Schmid, J., Zache, R., von Kurnatowski, M., Teicher, U., & Ihlenfeldt, S. (2022). Capturing and incorporating expert knowledge into machine learning models for quality prediction in manufacturing. *Journal of Intelligent Manufacturing*, *33*(7), 2129–2142. https://doi.org/10.1007/s10845-022-01975-4

Lu, S., & Jin, Z. (2017). Improved Stochastic gradient descent algorithm for SVM. *International Journal of Recent Engineering Science (IJRES)*, *4*(4), 28–31.

Lu, Y., Rajora, M., Zou, P., & Liang, S. Y. (2017). Physics-Embedded Machine Learning: Case Study with Electrochemical Micro-Machining. *Machines*, *5*(1), Article 1. https://doi.org/10.3390/machines5010004

Ma, F., Zhang, F., Ben, S., Qin, S., Zhou, P., Zhou, C., & Xu, F. (2021). *Monotonic Neural Network: Combining Deep Learning with Domain Knowledge for Chiller Plants Energy Optimization* (arXiv:2106.06143). arXiv. https://doi.org/10.48550/arXiv.2106.06143

Malliaraki, E., & Berditchevskaia, A. (2023). *Combining collective and machine intelligence at the knowledge frontier*. OECD. https://doi.org/10.1787/dbbd48a9-en

Mangasarian, O. L., & Wild, E. W. (2008). Nonlinear Knowledge-Based Classification. *IEEE Transactions on Neural Networks*, *19*(10), 1826–1832. https://doi.org/10.1109/TNN.2008.2005188

Matsunami, H., & Kimoto, T. (1997). Step-controlled epitaxial growth of SiC: High quality homoepitaxy. *Materials Science and Engineering: R: Reports*, *20*(3), 125–166.

Mohammadi, R., & Farsijani, H. (2023). Optimization under Uncertainty: Machine Learning Approach. *International Journal of Innovation in Management, Economics and Social Sciences*, *3*(2), Article 2. https://doi.org/10.59615/ijimes.3.2.23

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021

Neal, R. M. (2012). *Bayesian learning for neural networks* (Bd. 118). Springer Science & Business Media. https://books.google.de/books?hl=de&lr=&id=LHHrBwAAQBAJ&oi=fnd&pg=PR3&dq=bayes+neural+network&ots=K6teNScAW6&sig=gO5PNH4k2g1h-AiCIGXlnQUrB3w

Ning, C., & You, F. (2019). Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers & Chemical Engineering*,

125, 434–448. https://doi.org/10.1016/j.compchemeng.2019.03.034

Oberkampf, W. L., & Ferson, S. (2007). *Model Validation Under Both Aleatory and Epistemic Uncertainty.* (SAND2007-7163C). Sandia National Lab. (SNL-NM), Albuquerque, NM (United States). https://www.osti.gov/biblio/1146749

Oberski, D. L., & Satorra, A. (2013). Measurement Error Models With Uncertainty About the Error Variance. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 409–428. https://doi.org/10.1080/10705511.2013.797820

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *Proceedings of the 30th International Conference on Machine Learning*, 1310–1318. https://proceedings.mlr.press/v28/pascanu13.html

Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, *477*, 111902. https://doi.org/10.1016/j.jcp.2022.111902

Sahinidis, N. V. (2004). Optimization under uncertainty: State-of-the-art and opportunities. *Computers & Chemical Engineering*, *28*(6), 971–983. https://doi.org/10.1016/j.compchemeng.2003.09.017

Saris, W. E., & Revilla, M. (2016). Correction for Measurement Errors in Survey Research: Necessary and Possible. *Social Indicators Research*, *127*(3), 1005–1020. https://doi.org/10.1007/s11205-015-1002-x

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Segalman, D. J., Brake, M. R., Bergman, L. A., Vakakis, A. F., & Willner, K. (2014, Februar 12). *Epistemic and Aleatoric Uncertainty in Modeling*. ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2013-13234

Shi, J. Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC press. https://books.google.de/books?hl=de&lr=&id=DkgdN6dRAicC&oi=fnd&pg=PP1&dq=gaussian+process+regression&ots=oXDauUG1-u&sig=bvAt0FGc7rTIOEZbxr5ddaty2d0

Sideris, I., Crivelli, F., & Bambach, M. (2023). GPyro: Uncertainty-aware temperature predictions for additive manufacturing. *Journal of Intelligent Manufacturing*, *34*(1), 243–259. https://doi.org/10.1007/s10845-022-02019-7

Taylor, J. R., & Thompson, W. (1982). *An introduction to error analysis: The study of uncertainties in physical measurements* (Bd. 2). Springer. https://link.springer.com/book/9780935702750

Wikner, A., Pathak, J., Hunt, B., Girvan, M., Arcomano, T., Szunyogh, I., Pomerance, A., & Ott, E. (2020). Combining machine learning with knowledge-based modeling for scalable forecasting and subgrid-scale closure of large, complex, spatiotemporal systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(5). https://pubs.aip.org/aip/cha/article/30/5/053111/1030728

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). *Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems* (arXiv:2003.04919). arXiv. http://arxiv.org/abs/2003.04919

Williams, C., & Rasmussen, C. (1995). Gaussian processes for regression. *Advances in neural information processing systems*, *8*. https://proceedings.neurips.cc/paper_files/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html

Xue, Y., & Deng, Y. (2021). Decision making under measure-based granular uncertainty with intuitionistic fuzzy

sets. *Applied Intelligence*, *51*(8), 6224–6233. https://doi.org/10.1007/s10489-021-02216-6

Zeyuan, A. Z., Weizhu, C., Gang, W., Chenguang, Z., & Zheng, C. (2009). P-packSVM: Parallel primal gradient descent kernel SVM. *2009 Ninth IEEE International Conference on Data Mining*, 677–686. https://ieeexplore.ieee.org/abstract/document/5360294/

Zhao, L., & You, F. (2019). A data-driven approach for industrial utility systems optimization under uncertainty. *Energy*, *182*, 559–569. https://doi.org/10.1016/j.energy.2019.06.086

Zhou, B., Pychynski, T., Reischl, M., Kharlamov, E., & Mikut, R. (2022). Machine learning with domain knowledge for predictive quality monitoring in resistance spot welding. *Journal of Intelligent Manufacturing*, *33*(4), 1139–1163. https://doi.org/10.1007/s10845-021-01892-y

Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic Regression Model Optimization and Case Analysis. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 135–139. https://doi.org/10.1109/ICCSNT47585.2019.8962457

7 Is Anonymization Through Discretization Reliable? Modeling Latent Probability Distributions for Ordinal Data as a Solution to the Small Sample Size Problem

https://doi.org/10.3390/stats7040070

Declaration of Author Contributions The conception and design of the study were primarily developed by the first author, with significant involvement from Professor Christian Heumann in the ideation and conceptualization phases. Data collection, data analysis and interpretation, as well as the literature research, were independently conducted by the first author. The drafting and revision of the manuscript were also carried out by the first author. Throughout all stages of the work, Professor Christian Heumann provided substantial support, close supervision, and critical guidance.

stats

MDPI

Article

Is Anonymization Through Discretization Reliable? Modeling Latent Probability Distributions for Ordinal Data as a Solution to the Small Sample Size Problem

Stefan Michael Stroka * and Christian Heumann

Department of Statistics, Ludwig-Maximilians-University Munich, 80539 Munich, Germany; christian.heumann@lmu.de

* Correspondence: ststroka@gmail.com or st.stroka@campus.lmu.de

Abstract: The growing interest in data privacy and anonymization presents challenges, as traditional methods such as ordinal discretization often result in information loss by coarsening metric data. Current research suggests that modeling the latent distributions of ordinal classes can reduce the effectiveness of anonymization and increase traceability. In fact, combining probability distributions with a small training sample can effectively infer true metric values from discrete information, depending on the model and data complexity. Our method uses metric values and ordinal classes to model latent normal distributions for each discrete class. This approach, applied with both linear and Bayesian linear regression, aims to enhance supervised learning models. Evaluated with synthetic datasets and real-world datasets from UCI and Kaggle, our method shows improved mean point estimation and narrower prediction intervals compared to the baseline. With 5–10% training data randomly split from each dataset population, it achieves an average 10% reduction in *MSE* and a ~5–10% increase in R^2 on out-of-sample test data overall.

Keywords: re-identification; modeling latent class distribution; ordinal class; Bayesian inference; uncertainty quantification; supervised learning regression enhancement

Citation: Stroka, S.M.; Heumann, C. Is Anonymization Through Discretization Reliable? Modeling Latent Probability Distributions for Ordinal Data as a Solution to the Small Sample Size Problem. *Stats* **2024**, *7*, 1189–1208. https://doi.org/10.3390/stats7040070

Academic Editors: Jürgen Pilz, Noelle I. Samia and Dirk Husmeier

Received: 21 August 2024
Revised: 11 October 2024
Accepted: 14 October 2024
Published: 17 October 2024

1. Introduction

Today, both private and business data are incredibly valuable and are targeted by various interest groups at every possible junction. As a result, data privacy has become a highly relevant issue affecting both individuals and organizations. Often, however, data are handled carelessly, with an overreliance on existing anonymization methods, leading to potential risks.

The permanent recording and insufficiently regulated sale of anonymized data present significant risks of re-identification [1], which, according to current research, cannot be completely ruled out despite protective measures [2]. Experts in data governance warn against the false sense of security provided by data anonymization. Advanced machine learning models and statistical techniques, such as modeling probability distributions, are increasingly uncovering methods to re-identify supposedly anonymized data, thereby compromising anonymity [3].

Common challenges in this field include the small sample size problem, whereby the sample size has a significant impact on the generalizability of research findings [4]. According to common recommendations, it should not be less than 10% in pilot studies [5]. Additionally, computational complexity and the reliability of model predictions pose hurdles, as the forecasts depend on the reliability of the given data [6] and the quantification of uncertainties in the predictions [7].

This paper addresses the traceability and reversibility of anonymized metric data or target values through discretization, which may lead to security or information loss. Techniques such as ordinalization [8] and rounding [9] coarsen the metric space into

adjacent ordinal classes, leading to information loss depending on the number of ordinal classes. Previous research has explored the evaluation of data based on discrete classes, considering the uncertainty of data [10], the optimum between data usability and maximum anonymity [11], and data diversity and characteristic details [12]. It has also investigated applying k-anonymization for de-identification [13], the conversion of ordinal classes back to metric values, focusing on assumptions about underlying distributions [14], and an unsupervised learning approach combined with discriminative information [15]. Other approaches have examined how rough set theory, as an example, benefits from discretizing continuous value ranges [16], the conditions under which it makes sense to convert the data [17], and how ordinal classes can be treated as a continuous space [18].

Recent studies have also shown improved methods for enhancing data privacy and anonymity [19–25], where the main focus is on providing a systematic overview of existing anonymization techniques, especially in light of the increasing availability of data from social networks [19], as well as a comparative study of five current techniques for anonymizing collected data, assessing their strengths and weaknesses [20]. Further research has reviewed existing methods such as generalization and bucketization [21], compared suppression with slicing with other common techniques [22], highlighted weaknesses related to the compatibility of independently generalized data [23], and explored anonymization through pseudonymization [24] and a pseudo creation technique [25]. Furthermore, it is interesting to note how the use of latent influencing factors based on ordinal classes improves Bayesian analysis [26] and generates more accurate classifications compared to traditional classification methods [27].

Modeling probabilistic distributions for latent categorical variables suggests that assuming a continuous latent distribution within ordinal classes allows for precise value derivation from limited data using machine learning models. This implies that probability distributions for ordinal classes, even with small datasets, can provide a good distributional model and a reliable approximation of the underlying continuous latent distribution. This paper aims to demonstrate that data coarsening for anonymization can be misleading and not fully reliable. We propose a new approach that enables high prediction accuracy of true metrics values from anonymized data using deterministic or probabilistic supervised learning regression models. The re-anonymized results are also analyzed for uncertainties.

In the following sections, we apply our approach to simulation studies on both low- and high-complexity synthetic data and conduct a benchmarking study with publicly available datasets from various application domains.

2. De-Anonymization of Metric Data

To ensure anonymity in surveys or, in general, in data protection, data discretization is an often-applied approach. This approach involves dividing metric or continuous data into classes, thereby coarsening it into discrete information. In this section, we describe our novel methodology for reliably reverting anonymized information to its true (metric) values, highlighting the issues of this anonymization technique.

2.1. Introduction to the Method

Our methodology is based on the assumption that even a small set of precise, non-discretized information (a very small training dataset with metric values) is sufficient to train models that are capable of de-anonymizing coarsened data and inferring the metric values of out-of-sample data, considering uncertainties. In the process, we model latent distributions (of each class) with normal distributions based on the available precise information and use these to generate probabilities for the discretized observations. The goal is to retrospectively reverse the discretization with minimal bias.

2.2. Process of Discretization

The new approach promises that reliable inferences for the entire discrete ordinal class can be drawn from just a few metric data points per class and their distributions. This

logic describes the partitioning of the distribution of the entire metric space into ordered, discrete classes, each associated with a specific sampling distribution. Consequently, the division and choice of class boundaries significantly influence the sampling distribution within each class.

The grouping of data inevitably leads to a loss of information. However, statistical clustering methods that minimize squared errors can assist in optimally establishing group boundaries, thereby facilitating an optimal classification of normally distributed data concerning the frequency distribution within the class [28]. Other methodologies aim to reduce the complexity of continuous distributions while preserving as much information as possible through Representative Points (RPs). RPs can be generated using techniques such as Monte Carlo sampling, deterministic point selection, or MSE-based clustering, thereby optimizing classification by minimizing a loss function. The commonly used k-means algorithm can also be employed as an approach in this context [29].

In contrast to statistical grouping, there is also the possibility of a predefined, data-independent, or random classification into ordered classes. In practical applications, it may occur that existing classes are utilized by the methodology. Therefore, the objective of this paper is to model existing classes and boundaries that may contradict an optimal statistical clustering.

2.3. Theoretical Formulation

Let $X = \{x_1, x_2, x_3, \ldots x_m\}$ be the feature matrix of m independent variables, where each variable is assumed to be independently and identically distributed (i.i.d.). Let y be the dependent target variable with the value range W_y. We address a regression problem

$$f : \mathbb{R}^{\mathrm{m}} \rightarrow W \subseteq \mathbb{R}$$
$$f : X \rightarrow y$$

with a very small training sample relative to the test data (i.e., small sample size problem). We sloppily define the grid with $K + 2$ subclasses as

$$grid \mid y = \{class_0, class_1, class_2, \ldots, class_K, class_{K+1} \mid y\}, \; |grid| = K + 2$$

comprising disjoint ordinal classes that depend on the target variable y. The *grid* refers to a finite set of ordered, ordinally scaled classes $\{0, 1, 2, \ldots, K + 1\}$. Consequently, the following holds:

$$\{class_0, class_1, class_2, \ldots, class_{K+1}\} = \{0, 1, 2, \ldots, K, K + 1\}.$$

Each class k from the set $\{0, 1, 2, \ldots, K + 1\}$ is defined by two threshold values, a_k and a_{k+1}. These thresholds satisfy the following conditions:

$$-\infty = a_0 < a_1 < \ldots < a_K < a_{K+1} < a_{K+2} = \infty.$$

As a result, the following holds:

$$\forall_{k=1}^{K+1} \forall_{i=1}^{n} : y_i \in class_k \rightarrow y_i \in [a_k, a_{k+1}] \rightarrow a_k \leq y_i < a_{k+1}$$

and for $k = 0$:

$$\forall_{i=1}^{n} : y_i \in class_0 \rightarrow y_i \in [a_0, a_1] \rightarrow -\infty = a_0 < y_i < a_1.$$

$K + 2$ classes are defined by $K + 3$ thresholds. Hence, each class has a lower threshold a_k and an upper threshold a_{k+1}, which partition the continuous value range W of y into discrete, adjacent subgroups. Specifically, this can be expressed as:

$$grid|y = [a_0, a_1, \ldots, a_K, a_{K+1}, a_{K+2}] \subseteq W.$$

In the following application, we focus on a selected finite subset of the classes and their associated thresholds. Consequently, we disregard the outer classes $class_o = [a_0, a_1]$ and $class_{K+1} = [a_{K+1}, a_{K+2}]$. This restriction does not impact the results because:

$$\forall_{i=1}^{n} : y_i \in [a_1, \ldots, a_{K+1}] \subseteq [a_0, a_1, \ldots, a_K, a_{K+1}, a_{K+2}] \subseteq W.$$

In the transformation T, the metric values y are mapped to classes based on these thresholds:

$$T : \mathrm{W} \subseteq \mathbb{R} \rightarrow \{class_1, class_2, \ldots, class_\mathrm{K}\} \subseteq \mathbb{N}_0$$
$$T : y \rightarrow \{1, \ldots, K\}$$

The resulting vector $class^y$ is defined as a new, optionally applicable feature X_{m+1}, which is subsequently used as an ordinal-scaled variable and one-hot encoded.

Following, each class for each observation represents a discretized continuous value y. Now, we split the given data with observations $\{1, \ldots, n\}$ into training data $\{1, \ldots, l\}$ and test data $\{l+1, \ldots, n\}$. For each k-th ordinal class $class_k$, a normal distribution $N\left(\mu_k, \sigma_k^2\right)$ is fitted by estimating its parameters using the mean and standard deviation formulas to approximate the histogram of the training data for $y \in class_k$. With mean μ_k and standard deviation σ_k of the normal distribution for the k-th class from $k = \{1, \ldots, K\}$, it follows that:

$$\left(\forall_{i=1}^{l} : y_i^{train} \in class_k\right) \Big| class_k \sim N\left(\mu_k, \sigma_k^2\right).$$

The parametric modeling of the normal distributions allows for determining the parameters μ and σ for each class, and thus, the densities for each observation can be calculated. While random sampling from the distribution could be used to capture the properties of the density function as an input feature, we aim to provide more precise information for each observation. Therefore, we determine the absolute f_i and relative frequency p_i of y_i based on the density of its class. For each k-th class, we determine the mean and standard deviation as follows:

$$\forall_{k=1}^{K} : \mu(y \in class_k) = \mu_k \wedge \sigma(y \in class_k) = \sigma_k.$$

Then, we calculate the probability density function (PDF) value for the given train data y_i^{train} given the k-th class:

$$f\left(y_i^{train} \Big| class_k\right) = \frac{1}{\sigma_k \sqrt{2\pi}} exp\left(-\frac{\left(y_i^{train} - \mu_k\right)^2}{2\sigma_k^2}\right) = f_{ik}$$

The PDF indicates how densely the random variable y_i^{train} is distributed around a specific value. For the relative frequency density, we have:

$$p_{ik}^{train} = \frac{f_{ik}^{train}}{N_k^{train}}$$

where N_k^{train} is the total number of training samples in class k, and:

$$\sum_{k=1}^{K} N_k = \sum_{k=1}^{K} N_k^{train} + \sum_{k=1}^{K} N_k^{test} = \sum_{k=1}^{K} \left(N_k^{train} + N_k^{test}\right) = n.$$

The relative frequency p_{ik}^{train} thus provides a value that is proportional to the probability density at y_i, indicating how likely it is that y_i lies in a small interval around the given point. By integrating the density over a continuous range

$$\left[b - \frac{\epsilon}{2}, b + \frac{\epsilon}{2}\right]$$

with a sufficient small ϵ, the actual probability can be approximated by:

$$P\Big(b - \frac{\epsilon}{2} \leq y_i \leq b + \frac{\epsilon}{2}\Big|class_k\Big) = \int_{b-\frac{\epsilon}{2}}^{b+\frac{\epsilon}{2}} f(y_i|class_k)dy.$$

Since $f(y_i|class_k)$ is nearly constant within a sufficiently small interval around b, it follows that:

$$P\Big(b - \frac{\epsilon}{2} \leq y_i \leq b + \frac{\epsilon}{2}\Big|class_k\Big) \approx \epsilon * f(b|class_k),$$

where ϵ is independent of the density function f.

To enable this for out-of-sample applications where y-values are not available, we employ a simple linear regression model as a transfer learning method. In the first step, we predict the y^{test}- values for the out-of-sample data. In the next step, we determine the proportional probabilities with PDF-values that the i-th prediction $\hat{y}_i^{test}$ for $i = \{l+1, \ldots, n\}$ falls within the given classes. With the linear prediction model

$$\hat{y}^{test} = X\hat{\beta} + \epsilon, \epsilon \sim N(0, \sigma_\epsilon)$$

and the class distribution based on the training data y^{train} (with μ_k, σ_k), we generate K new input features for the test data for these K classes as follows:

$$f\big(\hat{y}_i^{test}\big|class_k\big) = \frac{1}{\sigma_k\sqrt{2\pi}} exp\left(-\frac{\big(\hat{y}_i^{test} - \mu_k\big)^2}{2\sigma_k^2}\right) = f_{ik}^{test}$$

and

$$p_{ik}^{test} = \frac{f_{ik}^{test}}{N_k}$$

with N_k^{test} as the total number of test samples in class k. Finally, this results in a new feature matrix P with n observations for K classes, and therefore:

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nK} \end{pmatrix} = \begin{pmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{l1} & \cdots & p_{lK} \end{pmatrix} \oplus \begin{pmatrix} p_{(l+1)1} & \cdots & p_{(l+1)K} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nK} \end{pmatrix} = \begin{pmatrix} P^{train} \\ P^{test} \end{pmatrix}$$

with $\oplus$ as the row-wise binding operator.

2.4. Features and Model Definition

The choice of regression model is flexible and independent of our proposed method. For demonstration purposes, this paper uses linear regression, but other supervised learning regression models can also be employed. We differentiate between four models based on the features used. Each of these models is evaluated using both linear regression and Bayesian linear regression approaches.

2.4.1. Linear Regression

Let X^{train} be the $(l \times m)$-Matrix of the training data, $class^{train}\big(= class^{train}\big|y\big)$ as the $(l \times K)$-matrix of ordinal classes dependent on the target variable train data ($class^{train}$ is a one-hot-encoded $(l \times 1) - vector$), and P^{train} as the $(l \times K)$-probability matrix (with the probably for each k-th class of all K for all training data observations l). In the following, we modify the input feature matrix X^* for four different model approaches. For all models, the following applies:

$$y^{train} = X^*\beta + \epsilon, \epsilon \sim N(\mu_\epsilon, \sigma_\epsilon)$$

where X^* $(* \in \{1, \ldots, 4\})$ serves as a placeholder for the respective input feature matrix. It follows that:

Model 1:

$$X^1_{l\times(1+m)} = \begin{pmatrix} 1_{l\times 1} & X^{train}_{l\times m} \end{pmatrix} \Big(\text{with } \beta_{(1+m)\times 1}\Big)$$

Model 2:

$$X^2_{l\times(1+m+K)} = \begin{pmatrix} 1_{l\times 1} & X^{train}_{l\times m} & P^{train}_{l\times K} \end{pmatrix} \Big(\text{with } \beta_{(1+m+K)\times 1}\Big)$$

Model 3:

$$X^3_{l\times(1+m+K)} = \Big(1_{l\times 1} \quad X^{train}_{l\times m} \quad class^{train}_{l\times K}\Big) \Big(\text{with } \beta_{(1+m+K)\times 1}\Big)$$

Model 4:

$$X^4_{l\times(1+m+2K)} = \Big(1_{l\times 1} \quad X^{train}_{l\times m} class^{train}_{l\times K} \quad P^{train}_{l\times K}\Big) \Big(\text{with } \beta_{(1+m+2K)\times 1}\Big).$$

2.4.2. Bayesian Linear Regression

Following a deterministic analysis, it is essential to consider whether uncertainties behave similarly. Therefore, we extend the models to a probabilistic perspective by incorporating prior distributions for β in each model. Therefore, we use standard prior distributions with

$$\beta \sim N\Big(0,\ 10^2 I\Big) \text{ and } \sigma \sim Half - N(1).$$

The linear prediction is then given by

$$\mu = \mathrm{X}^* \beta$$

with the likelihood

$$y|\beta,\sigma \sim N\Big(\mu, \sigma^2\Big)$$

Finally, the posterior distribution is derived as:

$$p(\beta|y) \propto LH * p(\beta) * p(\sigma)$$
$$\text{with } LH:\ p(y \mid \beta,\sigma) = \textstyle\prod_{i=1}^{n} N\big(y_i \mid \beta\check{X}, \sigma^2\big).$$

2.5. Evaluation Metrics

In the following sections, we examine various use cases and compare them based on selected metrics. For deterministic analysis, we use the Mean Squared Error (MSE) and the coefficient of determination (R^2). The evaluation is conducted using a train-test data split, which enables the assessment of the model's performance (in-sample evaluation) and its predictive accuracy (out-of-sample evaluation) using MSE and R^2.

In the probabilistic analysis, we extend these metrics to include the $Coverage_{rate}$, the average width of the prediction interval ($PI\ width$), and the ratio of coverage to width ($ratio$).

The Bayesian linear regression enables sampling M prediction vectors $\hat{y}^{(1)}, \ldots, \hat{y}^{(M)}$ from the posterior predictive distribution, where each draw $\hat{y}^{(m)} = \big(\hat{y}_{l+1}{}^{(m)}, \ldots., \hat{y}_n^{(m)}\big)$ represents an out-of-sample forecast using the same parameter. This results in the following out-of-sample prediction matrix:

$$\hat{y}_{post} = \begin{pmatrix} \hat{y}^{(1)}_{l+1} & \cdots & \hat{y}^{(1)}_{n} \\ \vdots & \ddots & \vdots \\ \hat{y}^{(M)}_{l+1} & \cdots & \hat{y}^{(M)}_{n} \end{pmatrix}^T.$$

From this matrix, we calculate the vector of posterior mean estimates by taking the row-wise mean:

$$mean_{post} = \begin{pmatrix} \overline{y}_{l+1} \\ \vdots \\ \overline{y}_n \end{pmatrix} = \begin{pmatrix} \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{l+1}^{(m)} \\ \vdots \\ \frac{1}{M}\sum_{m=1}^{M} \hat{y}_n^{(m)} \end{pmatrix}$$

The vector of standard deviations is then given by:

$$std_{post} = \begin{pmatrix} std(\hat{y}_{l+1}) \\ \vdots \\ std(\hat{y}_n) \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_{l+1}^{(m)} - \overline{y}_{l+1}\right)^2} \\ \vdots \\ \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_n^{(m)} - \overline{y}_n\right)^2} \end{pmatrix}.$$

For a 95% prediction interval (PI), the interval boundaries are calculated as:

$$\text{PI} : mean_{post} \pm 1.96 * std_{post}$$

Due to the simplicity of the regression problem, using a small M is sufficient for the normal approximation of the PI. However, for more complex data problems, M should be large (e.g., $M \geq 100$) to accurately determine the credibility interval for the 2.5% and 97.5% quantiles of the M values with respect to the model parameters. The coverage rate is calculated as:

$$coverage_{rate} = \frac{\sum_{i=1}^{n} 1_{\{\hat{y} \in PI\}}}{n}$$

Thus, the PI boundaries are based on discrete forecast values. Using these boundaries, the mean distance between them is calculated as follows:

$$width = \frac{\sum_{i=l+1}^{n}\left[\left(mean_{post} + 1.96 * std_{post}\right) - \left(mean_{post} - 1.96 * std_{post}\right)\right]_i}{n - l + 1}$$

Finally, the ratio is given by:

$$ratio = \frac{coverage_{rate}}{width}$$

This ratio represents the number of predictions within the PI divided by the mean width of the PI boundaries.

3. Explanatory Application Example

To gain a deeper understanding of probability distributions and models, we start with a straightforward regression problem. We compare various test-training data ratios for standard and Bayesian linear regression.

3.1. Application Example Design

For the multivariate application problem, 10,000 observations are drawn i.i.d. with $X \sim N(0,1)$ and the target variable is determined using the function:

$$y = f(x) = 4 + 3 * X + \epsilon, \;\; \epsilon \sim N(0, \sigma_\epsilon)$$

Ordinal class boundaries dependent on y are randomly set (*boundaries* = $\{-8.9, 0, 5, 10, 16.8\}$ with $K = 4$ classes), which determine the ordinal class for each observation. The generated population is divided into different ratios of test and training datasets for the analysis. The split is stratified based on the original classes to ensure that the class distribution remains proportional and all classes are represented in the training data. Each ordinal class thus has training data used to estimate the parameters (mean and standard deviation) of a normal distribution for that class.

In the subsequent analysis, X and the ordinal classes for the population are assumed to be given. The goal is to achieve the best possible out-of-sample prediction for the target variable (for the test data), taking uncertainties into account. We compare the four models mentioned with their respective input feature combinations for both standard linear regression and Bayesian linear regression.

3.2. Input Features

The given input features for the population are the observations X and the ordinal class $class^y$. To account for the probability distributions of the classes P, we use the target variable from the training data y^{train} to determine the probability of being in a class. Since the target variable for the test data y^{test} is unknown during model training, we use a simple linear model to obtain an estimate of the target variable and, based on this estimate and the probability distributions, determine the probabilities approximately.

Thus, the prediction is:

$$\hat{y}^{test} = X^{test} * \hat{\beta} \rightarrow \forall_{k=1}^{K} P_k^{test}$$

This results in new input feature variables with probabilities for the test data, depending on the number of classes (and also introducing variability in the prior distribution).

3.3. Models to Compare

We compare standard and Bayesian linear regression using the following feature settings:

Model	**1 (Standard)**	**2 (Baseline)**	**3**	**4**
$X*$	X^1	X^2	X^3	X^4

3.4. Descriptive Statistics

The goal of our new methodology is to provide robust out-of-sample predictions with limited training data. A critical factor in this is the test-training data ratio. Figure 1 illustrates the Mean Squared Error (*MSE*) and the Jensen-Shannon Divergence between the Kernel density estimation (KDE)-modeled distribution of the training data and the distribution of the target variable in the entire population. Various ratios ranging from 0 to 1 are evaluated. The idea is to compare the approximated distribution of the small sample training dataset with the distribution of the ground truth.

Figure 1. Mean squared error (*MSE*) and Jensen-Shannon divergence (*J-S divergence*) as evaluation metrics for comparing KDE-modeled distributions between test data proportions and the population. The train-test split is performed based on the x-range values. The blue line represents the mean result of the evaluation based on repeated train-test splits, while the orange line indicates the corresponding standard deviation.

Figure 1 indicates that, in the case of low complexity, there is no significant increase in discrepancy up to a test proportion of approximately 95%. This suggests that beyond a certain size of the training dataset, the distribution from the training data, despite the small sample size, can probably closely approximate the ground truth. Based on this evaluation, we next examine training data proportions of 5%, 10%, 20%, and 50% from the population.

The following figures display the described data in x-y plots, showing the corresponding probability distributions for each class based on the training-test split with 5/95 (Figure 2) and 80/20 ratios (Figure 3).

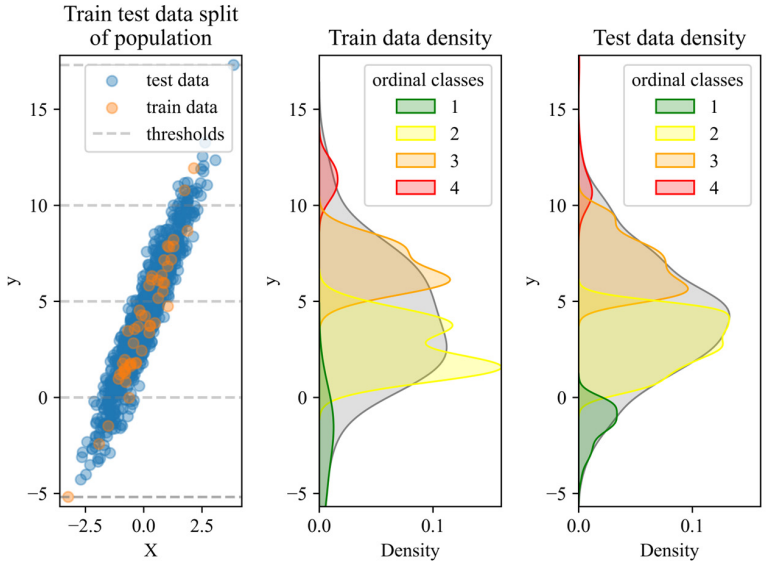


Figure 2. Comparison of KDE-modeled distributions for a 5/95 train-test split of the population (depicted by the gray distribution). The distributions are modeled based on values within the respective thresholds for each ordinal class using training data (middle inset) and test data (right inset). The test and training datasets are split randomly but with class stratification. Despite the visually apparent lower dispersion in the training data, the variability of both datasets is similar.

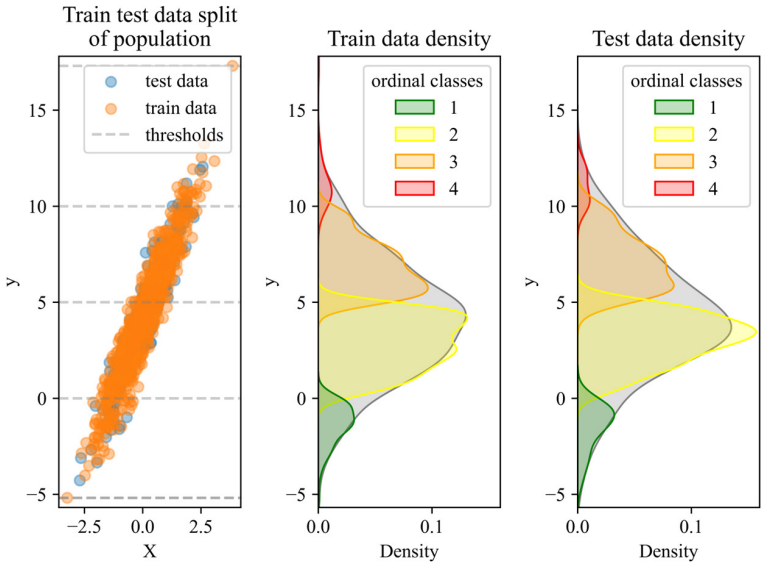


Figure 3. Comparison of KDE-modeled distributions for an 80/20 train-test split (i.e., standard cross validation ratio) of the population (gray distribution). The distributions are modeled based on values within the respective thresholds for each ordinal class, using train data (middle inset) and test data (right inset). The test and training datasets are split randomly but with class stratification.

Figure 2 visually demonstrates that even with just 5% of the training data, the probability distribution is similar to that based on the entire dataset. Increasing the proportion of training data to 80% leads to an even closer approximation of the class distribution from the total data, as shown in Figure 3. However, despite the significantly larger training dataset, there is no exceptionally strong visual improvement in the comparison between Figures 2 and 3. This supports our hypothesis for this application case: with a less complex frequency distribution of y, having as little as 5% of the training data can already yield comparably good results for predicting the true value of y.

3.5. Analysis and Evaluation

In the detailed analysis, we examine how well the models fit depending on the given features. Figure 4 displays the test and training data along with the linear regression predictions for the out-of-sample forecasts across all feature settings. It is important to note that while linear regression can appear non-linear in the subsequent figures, this is due to the graphical representation being limited to a 2D x-y view. Other influencing factors are still incorporated into the predictions. Therefore, several dimensions (influencing factors) used in the analysis are reduced to two for visual clarity.

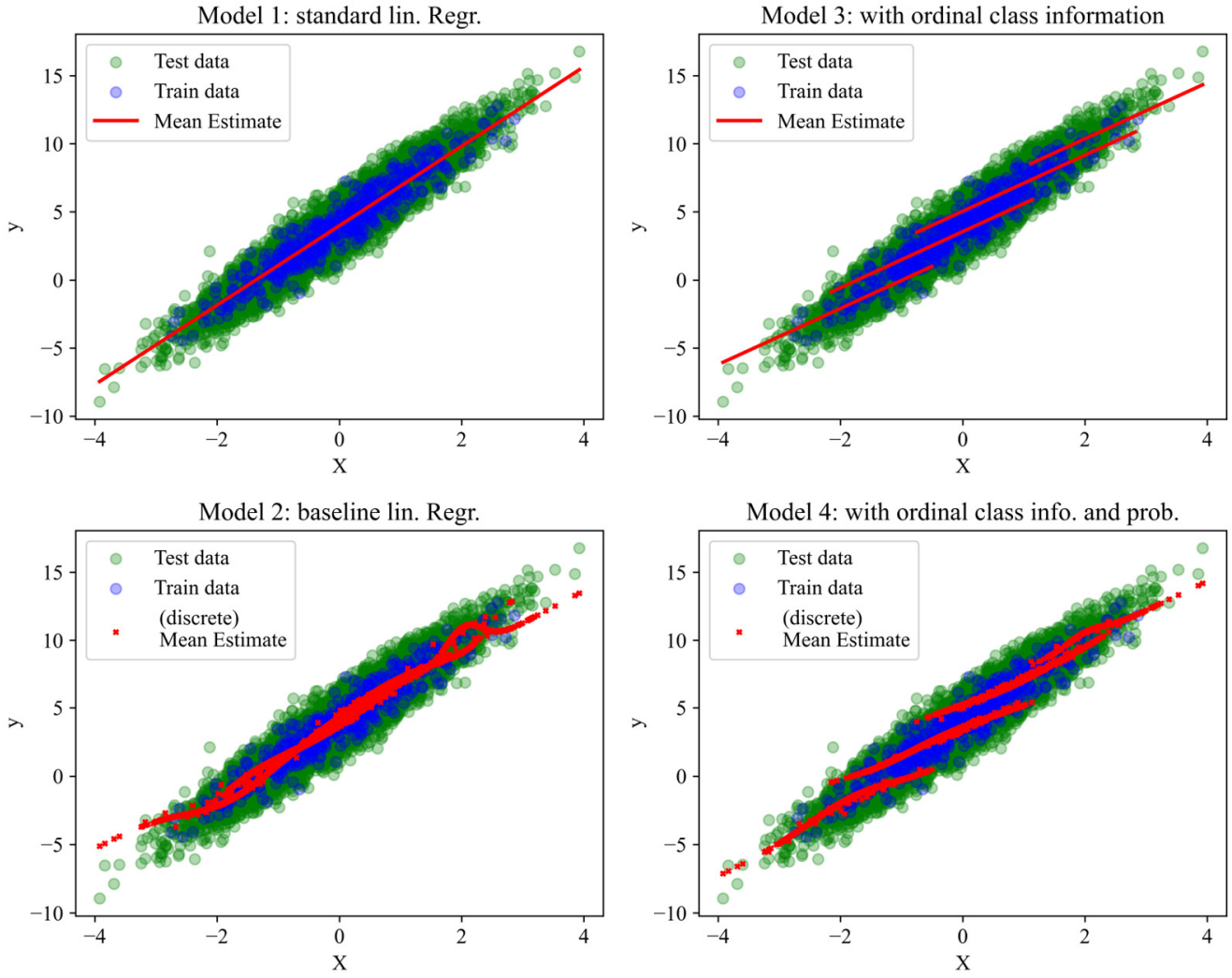


Figure 4. Comparison of four linear regression models based on different input features for a 5/95 train-test split. The insets show a 2D cross-section of the multivariate models, where all features are used. The test and training datasets are split randomly but with class stratification. Despite the visually apparent lower dispersion in the training data, the variability of both datasets is similar.

Figure 4 illustrates the significant impact of including probability features and, especially, class features on model performance. A visual comparison in 2D between (1) and (4) shows that our method allows for a more flexible (and even seemingly nonlinear) modeling compared to standard linear regression. This is because the hyperplane in multivariate standard linear regression is extended by $2K$ additional planes, where K is the number of classes. In the following Table 1, we prove this statement with regression evaluation metrics.

Table 1. Comparison of the different feature input settings (with linear regression) based on the training data proportion for in- and out-of-sample predictions. Evaluation metrics are MSE, R^2, and the computational time (CT). The best results per comparison are highlighted in bold.

Train Data Proportion [in %]	Model	MSE		R^2 [%]		CT [s]
		in *	out **	in *	out **	
5	1	0.876	1.001	89.47	89.39	**1.27**
	2	0.646	0.893	92.24	90.53	1.39
	3	0.499	0.594	94.00	93.71	1.90
	4	**0.427**	**0.527**	**94.86**	**94.41**	1.99
20	1	1.011	0.991	88.41	89.62	**1.50**
	2	0.724	0.873	91.70	90.85	1.66
	3	0.527	0.544	93.95	94.30	1.45
	4	**0.459**	**0.501**	**94.73**	**94.75**	1.47
50	1	1.066	0.921	88.19	90.54	1.22
	2	0.739	0.864	91.82	91.11	1.23
	3	0.465	0.594	94.85	93.89	1.25
	4	**0.426**	**0.547**	**95.28**	**94.38**	**1.14**
80	1	0.998	0.977	89.09	90.52	1.08
	2	0.730	0.987	92.02	90.42	0.95
	3	0.485	0.691	94.69	93.30	**0.87**
	4	**0.467**	**0.616**	**94.89**	**94.02**	0.91

* in-sample. ** out-of-sample.

The out-of-sample results in Table 1 indicate that expanding the model from input feature setting (1) to (2) leads to a significant reduction in MSE (approximately 5–10%) and an average improvement in R^2 of about 1% across all proportions. Notably, as the proportion of training data increases, the benefit from the probability distributions diminishes. This, combined with the in-sample results, suggests that a higher proportion may increase the likelihood of overfitting. However, the new features provide substantial benefits in out-of-sample evaluation, especially when training data are limited (i.e., in small sample size problems). This advantage is also observed in the comparison between settings (3) and (4). The comparison between (1) and (3) is less relevant because the additional categorical variable has a significant impact on the model, making such a comparison less meaningful.

In the next step, we examine the same evaluation considering uncertainties. For Bayesian linear regression, we extend point estimates with a PI. Figure 5 shows, similar to Figure 4, how different feature combinations (1)–(4) affect out-of-sample predictions.

Comparing Figure 5 to Figure 4, the average predictions remain nearly identical. However, when examining the PIs, it becomes apparent that adding probability features increases the uncertainty, as seen in the comparison between models (1) and (2) and between (3) and (4). Nevertheless, the very narrow PI for (1) seems to underestimate the actual uncertainty of the test data.

Regarding Figure 5, Table 2 confirms that while the absolute coverage rate increases in comparisons between (1) and (2) and between (3) and (4), the coverage relative to the width of the PI (Ratio) generally decreases.

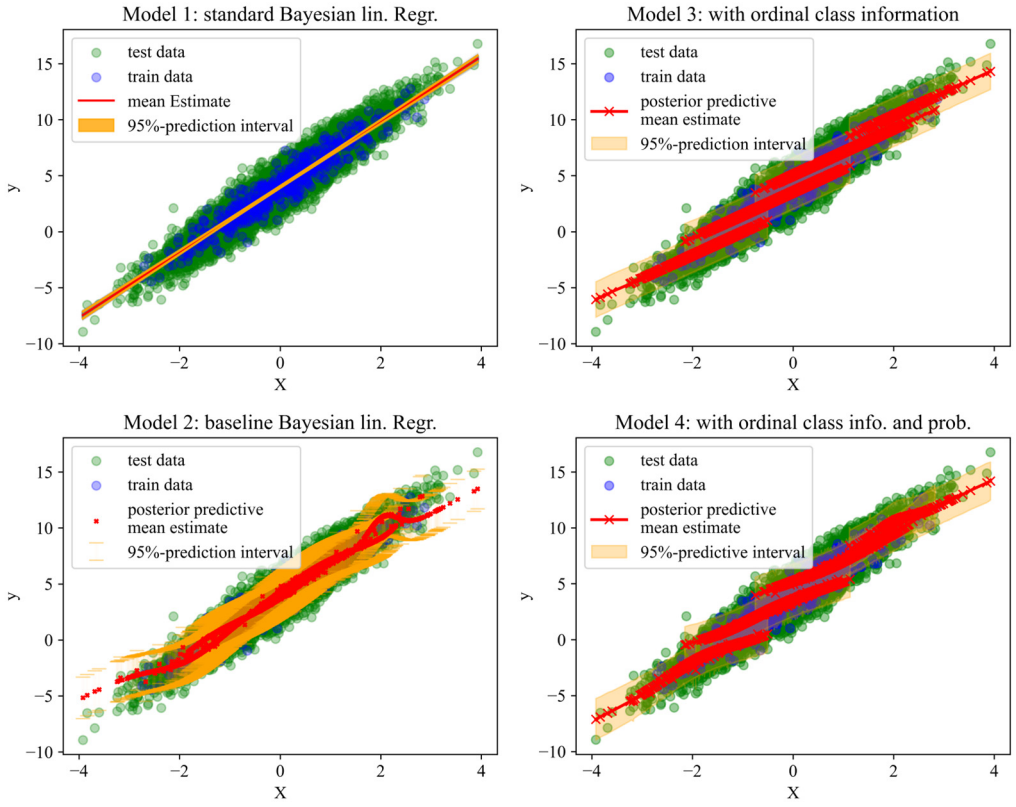


Figure 5. Comparison of four Bayesian linear regression models based on different input features for a 5/95 train-test split. The insets show a 2D cross-section of the multivariate models, where all features are used. The test and training datasets are split randomly but with class stratification. Despite the visually apparent lower dispersion in the training data, the variability of both datasets is similar.

Table 2. Comparison of different feature input settings (using Bayesian linear regression) was conducted by evaluating the coverage of model predictions within the PI (out-of-sample). Evaluation metrics are $coverage_{rate}$, *PI width*, *ratio*, and the computational time (CT). The best results per comparison are highlighted in bold.

Train Data Proportion [in %]	Input Feature Settings	PI Metrics			CT [s]
		Cov. Rate [in %]	PI Width	Ratio [in %]	
5	1	31.68	**1.31**	24.25	**53.94**
	2	53.89	3.75	14.38	56.27
	3	96.21	3.25	**29.61**	78.09
	4	**97.89**	3.43	28.55	97.43
20	1	15.13	**0.56**	27.14	**50.89**
	2	24.63	1.57	15.66	55.21
	3	**96.00**	2.96	32.49	96.06
	4	95.63	2.83	**33.78**	138.68
50	1	9.80	**0.40**	24.79	**50.48**
	2	20.80	0.89	23.45	53.81
	3	**94.00**	2.71	34.66	118.48
	4	93.2	2.62	**35.55**	191.8
80	1	7.50	**0.31**	24.34	**49.81**
	2	15.00	0.62	24.10	54.81
	3	**94.00**	2.75	34.15	154.44
	4	93.50	2.72	**34.37**	215.88

4. Simulation Study

In this section, we examine a more complex simulation study, oriented on a real-world application with respect to variable names and ranges. This approach allows us to gain comprehensive insights into the entire population, unlike using potentially biased or skewed samples. As before, the data are divided into training and test sets.

4.1. Simulation Design

The generated synthetic data include four input features that influence the log-normally distributed target variable. Let X be the feature matrix. This multivariate data problem is designed to simulate a real-world application where, in a street survey, individuals are asked about their age, work experience, education level, weekly hours, and the target variable, salary. A relatively small portion is asked for their exact salary (training data), while the remainder are asked to categorize their salary into ranked (ordinal) classes based on predefined thresholds (test data), such as 'low income' or 'high income'.

For dataset 1, the following equation applies:

$$salary = 0.02 * X_1 + 0.05 * X_2 + 0.1 * X_3 + 0.03 * X_4 + \epsilon_1$$

and for dataset 2:

$$salary = 0.03 * X_1 + 0.06 * X_2 + 0.04 * X_4 + 0.1 * X_5 + \epsilon_2$$

with

$$salary \sim Log - N\left(\mu, \sigma^2\right)$$

and

$$\begin{aligned} X1: \ age\ [in\ a] &\sim U(20, 65) \\ X2: \ experience\ [in\ a] &\sim U(0, 40) \\ X3: \ educational\ level &\sim U(1, 5) \\ X4: \ hours\ per\ week\ [in\ h] &\sim U(20, 60) \\ X5: \ performance\ rating &\sim U(1, 6) \\ \epsilon_1: \ noise &\sim N(0, 0.1) \\ \epsilon_2: \ noise &\sim N(0, 0.2). \end{aligned}$$

The coefficients are chosen randomly. In the simulation, n samples are drawn such that the input matrix $\dim(X) = (n \times 4)$. Samples are also randomly drawn and evaluated for $n = 3000$ and $n = 5000$ for each dataset. For reproducibility, a random seed is used per dataset. Figure 6 shows a simulation example ($n = 3000$) with the correlations between the independent and dependent variables on the left and the approximately log-normally distributed target variable (salary) in detail on the right.

The simulation allows us to apply the methodology to a known and thoroughly understandable ground truth. We compare different train-test split ratios of 5/95, 10/90, and 20/80. This simulates the applicability to small sample size problems to re-anonymize ordinal data. Only for the training data, the exact salary values are provided. In the next step, we evaluate the simulated datasets for the application of the methodology.

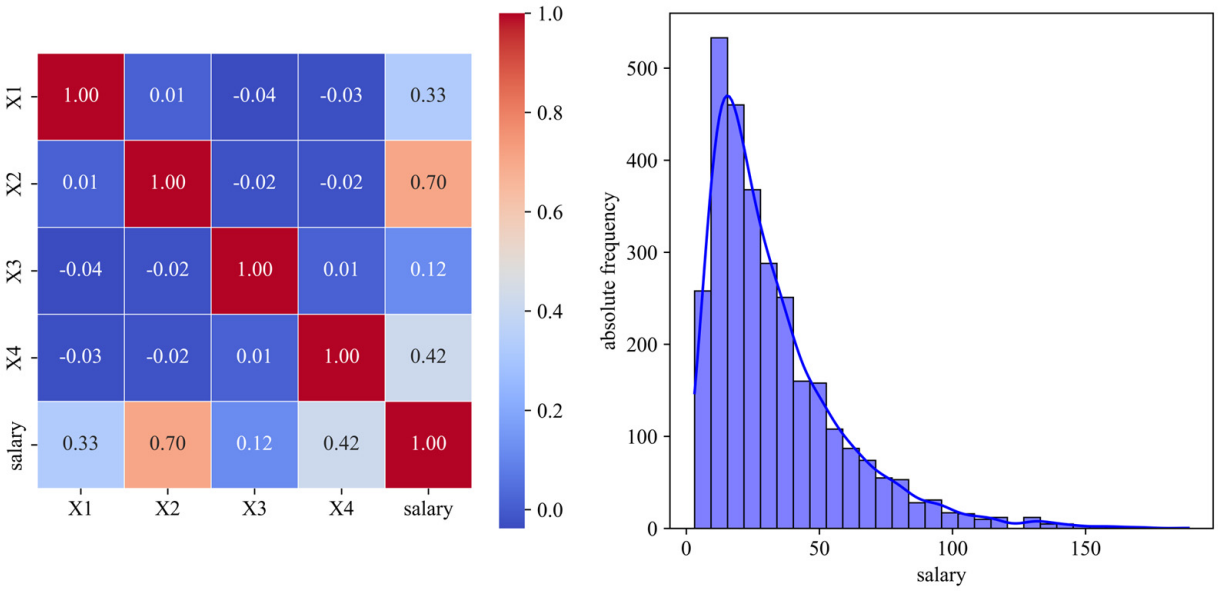


Figure 6. Heatmap and histogram with an approximated log-normal distribution for a simulated example with $n = 3000$.

4.2. Simulation Results

Table 3 shows the evaluation of the six different datasets. Notably, in all cases, using the linear baseline model (2) instead of the standard model (1) provides a significant improvement. This confirms that even without including the ordinal class (3) as a categorical feature, exceptional out-of-sample results can be achieved. In the comparison between models (3) and (4), no substantial improvement was observed. Often, model (4) tends to overfit, indicating that model (3) is sufficiently trained. However, in some cases, such as dataset 3 with 5% training data, an improvement in the out-of-sample R^2 value by 0.5% was achieved despite the already very good result of ~95%. We conclude that, depending on the data complexity and the SL regression model used, further improvements are possible by incorporating probabilities even when using categories as features. Regarding the *MSE*, it is also very clear that using probabilities (2) results in a 3–5 times lower *MSE* compared to standard linear regression (1).

Table 3. Comparison of the different feature input settings (with linear regression) and training data proportion based on the simulation study data for in- and out-of-sample. Evaluation metrics are *MSE* and R^2. The best results per comparison are highlighted in bold.

Dataset	Train Data Proportion	Model	MSE		R^2	
			in *	out **	in *	out **
1 (# 3000)	5	1	0.00246	0.00358	84.97	80.85
		2	0.00050	0.00118	96.97	93.72
		3	0.00055	**0.00113**	**96.65**	93.97
		4	**0.00048**	0.00130	97.06	**93.03**
	10	1	0.00366	0.00352	80.16	81.09
		2	0.00111	**0.00098**	93.98	**94.73**
		3	0.00104	**0.00098**	94.39	94.71
		4	**0.00092**	0.00102	**95.00**	94.51
	20	1	0.00342	0.00354	80.17	81.30
		2	0.00104	0.00108	93.95	94.30
		3	0.00099	0.00099	94.27	94.75
		4	**0.00094**	**0.00093**	**94.56**	**95.07**

Table 3. *Cont.*

Dataset	Train Data Proportion	Model	MSE		R^2	
			in *	out **	in *	out **
2 (# 3000)	5	1	0.00488	0.00528	66.18	65.13
		2	0.00054	0.00121	96.25	92.00
		3	0.00059	0.00114	95.92	92.49
		4	**0.00053**	**0.00113**	**96.32**	**92.54**
	10	1	0.00474	0.00534	66.99	64.89
		2	0.00087	**0.00119**	93.92	92.20
		3	0.00093	**0.00109**	93.55	**92.84**
		4	**0.00084**	0.00110	**94.15**	92.80
	20	1	0.00517	0.00523	66.32	65.31
		2	0.00084	0.00138	94.51	90.84
		3	0.00095	**0.00109**	93.78	**92.75**
		4	**0.00083**	0.00130	**94.59**	91.37
3 (# 5000)	5	1	0.00311	0.00285	80.75	81.05
		2	0.00107	0.00093	93.41	93.86
		3	0.00086	0.00081	94.71	94.59
		4	**0.00073**	**0.00074**	**95.46**	**95.10**
	10	1	0.00297	0.00287	80.54	81.02
		2	0.00085	0.00082	94.40	94.56
		3	0.00090	0.00080	94.13	94.72
		4	**0.00083**	**0.00075**	**94.58**	**95.01**
	20	1	0.00316	0.00278	80.19	81.32
		2	0.00124	0.00080	92.20	94.64
		3	0.00104	0.00074	93.50	95.02
		4	**0.00098**	**0.00072**	**93.88**	**95.19**
4 (# 5000)	5	1	0.00501	0.00497	67.83	66.18
		2	0.00096	0.00107	93.86	92.69
		3	0.00105	**0.00103**	93.24	**92.97**
		4	**0.00095**	0.00107	**93.87**	92.70
	10	1	0.00403	0.00502	70.46	66.24
		2	0.00052	0.00113	96.17	92.40
		3	0.00062	**0.00108**	95.46	**92.76**
		4	**0.00051**	0.00111	**96.28**	92.54
	20	1	0.00480	0.00492	66.17	66.91
		2	0.00117	0.00116	91.76	92.24
		3	0.00095	0.00100	93.32	93.30
		4	**0.00089**	**0.00094**	**93.76**	**93.69**

* in-sample. ** out-of-sample.

5. Benchmarking

To evaluate real-world applicability, we use multiple datasets from the UCI Database and Kaggle to conduct a benchmark study comparing various feature combinations and models.

5.1. Datasets

For a comprehensive analysis, we select datasets that differ in terms of the ratio between the number of observations and the number of features. We examine two datasets for each category: average, low, and high sample/feature ratio. Table 4 shows the settings for class size and thresholds for each dataset, and Table 5 provides additional descriptive information.

Table 4. Descriptive information about class size and thresholds for the multivariate Benchmark datasets.

	Descriptive Analysis of the Target			Class Size	Class Thresholds
	Min	Mean	Max		
AutoMPG	9	23.52	46.6	4	[8, 16, 24, 32.5, 48]
Boston Housing	5	22.53	50	4	[4, 15, 25, 35, 51]
Student Performance	0	11.91	19	4	[−1, 9, 12, 15, 20]
Automobile *	-	-	-	2	[−3.5, 1, 3.5]
California Housing	14,999	206,855	500,001	4	[14,998, 136,249, 257,500, 378,751, 500,002]
Bike Sharing	1	189.46	977	4	[0, 150, 350, 500, 1000]

* Automobile has an ordinal-scaled regression target. The target does not have to be metric scaled to further coarsen into classes.

Table 5. Descriptive information for the normalized multivariate Benchmark datasets.

	Samples Size	Features Size	Sample/ Feature-Ratio	Target	Target Unit	Target Mean	Target IQR
AutoMPG	398	8	49.8	MPG	$\left[\frac{\text{miles}}{\text{gallon}}\right]$	0.3860	0.3059
Boston Housing	506	14	36.1	MEDV	[1k USD]	0.3896	0.1772
Student Performance	649	57	11.4	G3	Points	0.6266	0.2105
Automobile	205	69	3.0	Risk	Level	0.5668	0.4000
California Housing	20,640	14	1474.3	MHV	[USD]	0.3956	0.2992
Bike Sharing	17,379	14	1241.4	Count	Bikes	0.1931	0.2469

5.2. Evaluation

In the evaluation, we use models (1) and (3) as benchmarks for models (2) and (4), respectively, because models (1) and (2) and models (3) and (4) are comparable based on the given input. Looking at Table 6, it is evident that the use of probability distributions as additional input in model (2) often provides a significant improvement compared to standard models (1). This also holds true in comparisons between (1) and (3). However, when comparing models (2) and (3), their performance is often comparable. Specific exceptions include the Boston Housing dataset for 5% and 10% training data, where the use of probability distributions with models (2) and (4) confuses the model, which results in worse results. Comparing models (3) and (4) confirms that overfitting is a concern, particularly with model (4). The datasets for Automobile and Student Performance highlight a key limitation of both linear regression and our new approach: models with a low sample-to-feature ratio struggle to produce reliable results, even when probabilistic features are incorporated. In these cases, our method combined with linear regression reaches its limits. This limitation is particularly evident in the Student Performance dataset, where increasing the proportion of training data leads to improved model results. This suggests that a low sample-to-feature ratio significantly restricts performance. A similar trend is observed in the AutoMPG and Boston Housing datasets, which have a moderately higher sample-to-feature ratio. While we observe improvements from model (1) to (2) and even (3), model (4) appears to overfit, occasionally yielding exceptionally poor results. In contrast, model (4) achieved the best results across all evaluations for the California Housing and Bike Sharing datasets. This demonstrates that, when well calibrated, the use of probabilistic features can marginally improve performance compared to the categorical variable model (3).

Table 6. Comparison of the different feature input settings (with linear regression) and training data proportion based on the Benchmark datasets for in- and out-of-sample. Evaluation metrics are *MSE* and R^2. The best results per comparison are highlighted in bold.

Dataset	Train Data Proportion	Model	MSE		R^2	
			in *	out **	in *	out **
AutoMPG	5	1	0.00273	0.01178	92.92	72.77
		2	0.00079	0.00975	97.96	77.47
		3	0.00087	**0.00507**	97.75	**88.28**
		4	**0.00041**	0.03101	**98.93**	28.36
	10	1	0.00642	0.00938	85.46	78.14
		2	0.00249	0.00398	94.37	90.73
		3	0.00231	**0.00362**	94.76	**91.57**
		4	**0.00202**	0.00528	**95.42**	87.71
	20	1	0.00625	0.00852	86.02	80.04
		2	0.00155	0.00439	96.52	89.71
		3	0.00167	**0.00380**	96.27	**91.10**
		4	**0.00130**	0.00424	**97.10**	90.08
Boston Housing	5	1	0.00318	0.01769	92.32	57.55
		2	0.00216	>1	94.79	$< 1 \times 10^{-5}$
		3	0.00057	**0.01033**	98.63	**75.21**
		4	**0.00037**	0.01878	**99.11**	54.95
	10	1	0.00349	0.02233	88.87	47.85
		2	0.00147	0.01741	95.32	59.35
		3	0.00122	**0.00656**	96.11	**84.69**
		4	**0.00093**	5.82780	**97.03**	$< 1 \times 10^{-5}$
	20	1	0.00653	0.01684	83.81	59.93
		2	0.00258	0.00642	93.59	84.72
		3	0.00185	**0.00329**	95.40	**92.18**
		4	**0.00165**	0.00373	**95.91**	91.12
Automobile	5	1	$< 1 \times 10^{-5}$	0.07078	100.00	−12.98
		2	$< 1 \times 10^{-5}$	0.07211	100.00	−15.10
		3	$< 1 \times 10^{-5}$	**0.06992**	100.00	−11.61
		4	$< 1 \times 10^{-5}$	0.07121	100.00	−13.66
	10	1	$< 1 \times 10^{-5}$	0.08464	100.00	−36.34
		2	$< 1 \times 10^{-5}$	0.08549	100.00	−37.70
		3	$< 1 \times 10^{-5}$	0.05165	100.00	16.80
		4	$< 1 \times 10^{-5}$	**0.05027**	100.00	19.04
	20	1	$< 1 \times 10^{-5}$	0.17308	100.00	−181.87
		2	$< 1 \times 10^{-5}$	0.14112	100.00	−129.83
		3	$< 1 \times 10^{-5}$	0.16912	100.00	−175.43
		4	$< 1 \times 10^{-5}$	**0.14078**	100.00	−129.27
Student Performance	5	1	$< 1 \times 10^{-5}$	0.08539	100.00	−197.44
		2	$< 1 \times 10^{-5}$	0.07496	100.00	−161.09
		3	$< 1 \times 10^{-5}$	**0.02395**	100.00	16.56
		4	$< 1 \times 10^{-5}$	0.02728	100.00	4.99
	10	1	0.00910	0.07507	66.09	−158.08
		2	0.00186	0.03561	93.05	−22.41
		3	0.00151	**0.01194**	94.37	**58.95**
		4	**0.00100**	0.02383	**96.27**	18.07
	20	1	0.01423	0.03304	47.32	−12.67
		2	0.00403	0.01178	85.06	59.82
		3	0.00312	**0.00659**	88.43	**77.53**
		4	**0.00305**	0.00676	**88.70**	76.95

Table 6. *Cont.*

Dataset	Train Data Proportion	Model	MSE		R^2	
			in *	out **	in *	out **
California Housing	5	1	0.01949	0.02040	65.61	63.96
		2	0.00497	0.00532	91.23	90.60
		3	0.00382	**0.00389**	93.27	93.12
		4	**0.00381**	**0.00389**	**93.28**	**93.13**
	10	1	0.01783	0.02049	68.19	63.84
		2	0.00455	0.00518	91.87	90.87
		3	0.00366	**0.00384**	93.48	**93.22**
		4	**0.00363**	0.00386	**93.52**	93.19
	20	1	0.01969	0.02027	64.68	64.32
		2	0.00521	0.00523	90.65	90.80
		3	0.00371	0.00381	93.35	93.29
		4	**0.00369**	**0.00380**	**93.37**	**93.31**
Bike Sharing	5	1	0.02051	0.02143	40.93	37.93
		2	0.00343	0.00339	90.13	90.17
		3	0.00255	0.00297	92.65	91.41
		4	**0.00249**	**0.00296**	**92.84**	**91.43**
	10	1	0.02122	0.02117	39.20	38.63
		2	0.00333	0.00347	90.47	89.94
		3	0.00282	0.00294	91.93	91.48
		4	**0.00280**	**0.00292**	**91.98**	**91.53**
	20	1	0.02141	0.02111	39.42	38.52
		2	0.00411	0.00371	88.36	89.19
		3	0.00295	0.00291	91.65	91.54
		4	**0.00291**	**0.00289**	**91.78**	**91.60**

* in-sample. ** out-of-sample.

6. Conclusions and Discussion

This paper aimed to explore the feasibility of tracing anonymized data from very small sample sizes. We developed a methodology that combines latent probabilistic distributions over ordinal classes—i.e., anonymized data—with a small sample approach. This combination enables significantly improved predictions of actual values using linear regression, applicable to both simple and complex data structures. In the context of increasing data protection concerns, our new method demonstrates how standard anonymization techniques, such as discretizing metric data in street surveys or similar contexts, can be accurately reversed. While it might seem counterintuitive to infer exact values from discrete classes using latent distributions, this approach aligns with existing methodologies. The use of distributions to describe latent relationships within a class provides notable advantages.

Our application results show that even a small training dataset can outperform standard linear regression when using latent probabilistic distributions. However, when comparing models that include ordinal class variables, probabilistic distributions often do not provide substantial additional benefits and may even lead to overfitting. This methodology is particularly versatile for any supervised learning regression task. While data quantity and quality can be limited, the approach remains effective. One limitation is that latent probabilistic distributions require a minimum amount of data, which, in our cases, did not need to be excessively large.

Future research should focus on a more detailed examination of the distribution of latent influencing factors, while considering the potential for optimizing class boundaries. Considering the significant impact of class boundaries, combining the optimal clustering solution and given class boundaries could further improve the methodology.

This paper used normally distributed modeling of ordinal classes and considered these as influencing features. Future work could explore more detailed modeling with Gaussian mixture models or other distributions.

Overall, beyond data protection and anonymization, our approach offers universal applicability and could improve various supervised learning regression problems, particularly when latent probabilistic distributions are not independently or sufficiently recognized by the model.

Author Contributions: Conceptualization, S.M.S. and C.H.; methodology, S.M.S.; software, S.M.S.; validation, S.M.S. and Heumann C.; formal analysis, S.M.S.; investigation, S.M.S.; resources, S.M.S.; data curation, S.M.S.; writing—original draft preparation, S.M.S.; writing—review and editing, S.M.S.; visualization, S.M.S.; supervision, C.H.; project administration, S.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are reproduceable or open access.

Acknowledgments: This work utilized generative artificial intelligence (AI) tools to assist with translation and ensure grammatical correctness.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lubarsky, B. Re-Identification of "Anonymized Data". *Georg. Law Technol. Rev.* **2010**. Available online: https://www.georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017 (accessed on 10 September 2021).
2. Porter, C.C. De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information. *Shidler JL Com. Tech.* **2008**, *5*, 1.
3. Senavirathne, N.; Torra, V. On the Role of Data Anonymization in Machine Learning Privacy. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 29 December 2020–1 January 2021; pp. 664–675.
4. Ercikan, K. Limitations in Sample-to-Population Generalizing. In *Generalizing from Educational Research*; Routledge: Abingdon, UK, 2008; ISBN 978-0-203-88537-6.
5. Hertzog, M.A. Considerations in Determining Sample Size for Pilot Studies. *Res. Nurs. Health* **2008**, *31*, 180–191. [CrossRef]
6. Li, T.; Li, N.; Zhang, J. Modeling and Integrating Background Knowledge in Data Anonymization. In Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 29 March–2 April 2009; pp. 6–17.
7. Stickland, M.; Li, J.D.-Y.; Tarman, T.D.; Swiler, L.P. *Uncertainty Quantification in Cyber Experimentation*; Sandia National Lab. (SNL-NM): Albuquerque, NM, USA, 2021.
8. Oertel, H.; Laurien, E. Diskretisierung. In *Numerische Strömungsmechanik*; Vieweg+Teubner Verlag: Wiesbaden, Germany, 2003; pp. 126–214, ISBN 978-3-528-03936-3.
9. Senavirathne, N.; Torra, V. Rounding Based Continuous Data Discretization for Statistical Disclosure Control. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 15139–15157. [CrossRef]
10. Inan, A.; Kantarcioglu, M.; Bertino, E. Using Anonymized Data for Classification. In Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 29 March–2 April 2009; pp. 429–440.
11. Pors, S.J. Using Discretization and Resampling for Privacy Preserving Data Analysis: An Experimental Evaluation. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2018.
12. Milani, M.; Huang, Y.; Chiang, F. Data Anonymization with Diversity Constraints. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3603–3618. [CrossRef]
13. Bayardo, R.J.; Agrawal, R. Data Privacy through Optimal K-Anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 217–228.
14. Robitzsch, A. Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. *Front. Educ.* **2020**, *5*, 589965. [CrossRef]
15. Zouinina, S.; Bennani, Y.; Rogovschi, N.; Lyhyaoui, A. A Two-Levels Data Anonymization Approach. In *Artificial Intelligence Applications and Innovations*; Maglogiannis, I., Iliadis, L., Pimenidis, E., Eds.; IFIP Advances in Information and Communication Technology; Springer International Publishing: Cham, Switzerland, 2020; Volume 583, pp. 85–95, ISBN 978-3-030-49160-4.

16. Xin, G.; Xiao, Y.; You, H. Discretization of Continuous Interval-Valued Attributes in Rough Set Theory and Its Application. In Proceedings of the 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, China, 19–22 August 2007; Volume 7, pp. 3682–3686.
17. Rhemtulla, M.; Brosseau-Liard, P.É.; Savalei, V. When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods under Suboptimal Conditions. *Psychol. Methods* **2012**, *17*, 354. [CrossRef]
18. Jorgensen, T.D.; Johnson, A.R. How to derive expected values of structural equation model parameters when treating discrete data as continuous. *Struct. Equ. Model. A Multidiscip. J.* **2022**, *29*, 639–650. Available online: https://scholar.google.de/scholar?hl=de&as_sdt=0,5&q=Jorgensen,+T.D.;+Johnson,+A.R.+How+to+Derive+Expected+Values+of+Structural+Equation+Model+Parameters+When+Treating+Discrete+Data+as+Continuous.&btnG= (accessed on 10 October 2024). [CrossRef]
19. Zhou, B.; Pei, J.; Luk, W. A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. *ACM Sigkdd Explor. Newsl.* **2008**, *10*, 12–22. [CrossRef]
20. Murthy, S.; Bakar, A.A.; Rahim, F.A.; Ramli, R. A Comparative Study of Data Anonymization Techniques. In Proceedings of the 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), Washington, DC, USA, 27–29 May 2019; pp. 306–309.
21. Mogre, N.V.; Agarwal, G.; Patil, P. A Review on Data Anonymization Technique for Data Publishing. *Int. J. Eng. Res. Technol. IJERT* **2012**, *1*, 1–5.
22. Kaur, P.C.; Ghorpade, T.; Mane, V. Analysis of Data Security by Using Anonymization Techniques. In Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 287–293.
23. Martinelli, F.; SheikhAlishahi, M. Distributed Data Anonymization. In Proceedings of the 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 5–8 August 2019; pp. 580–586.
24. Marques, J.F.; Bernardino, J. Analysis of Data Anonymization Techniques. In Proceedings of the KEOD 2020—12th International Conference on Knowledge Engineering and Ontology Development, Online Streaming, 2–4 November 2020; pp. 235–241.
25. Abd Razak, S.; Nazari, N.H.M.; Al-Dhaqm, A. Data Anonymization Using Pseudonym System to Preserve Data Privacy. *IEEE Access* **2020**, *8*, 43256–43264. [CrossRef]
26. Muthukumarana, S.; Swartz, T.B. Bayesian Analysis of Ordinal Survey Data Using the Dirichlet Process to Account for Respondent Personality Traits. *Commun. Stat.-Simul. Comput.* **2014**, *43*, 82–98. [CrossRef]
27. Sha, N.; Dechi, B.O. A Bayes Inference for Ordinal Response with Latent Variable Approach. *Stats* **2019**, *2*, 321–331. [CrossRef]
28. Cox, D.R. Note on Grouping. *J. Am. Stat. Assoc.* **1957**, *52*, 543–547. [CrossRef]
29. Fang, K.-T.; Pan, J. A Review of Representative Points of Statistical Distributions and Their Applications. *Mathematics* **2023**, *11*, 2930. [CrossRef]

Bibliography

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, and U. Rajendra Acharya. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021. URL https://www.sciencedirect.com/science/article/pii/S1566253521001081. Publisher: Elsevier.

Karan Aggarwal, Maad M. Mijwil, Abdel-Hameed Al-Mistarehi, Safwan Alomari, Murat Gök, Anas M. Zein Alaabdin, and Safaa H. Abdulrhman. Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, 3(1):115–123, 2022. URL https://www.iasj.net/iasj/download/cefbfd60eb11898a.

Mohammad Nazmul Alam, Mandeep Kaur, and Shahin Kabir. Explainable AI in Healthcare: Enhancing Transparency and Trust upon Legal and Ethical Consideration. *International Research Journal of Engineering and Technology (IRJET)*, 10(06), 2023.

Constantin Aliferis and Gyorgy Simon. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In Gyorgy J. Simon and Constantin Aliferis, editors, *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, pages 477–524. Springer International Publishing, Cham, 2024. ISBN 978-3-031-39355-6. doi: 10.1007/978-3-031-39355-6_10. URL https://doi.org/10.1007/978-3-031-39355-6_10.

Ashish Arora, Niloufar Shoeibi, Vishwani Sati, Alfonso González Briones, Pablo Chamoso, and Emilio Corchado. Data Augmentation Using Gaussian Mixture Model on CSV Files. pages 258–265. January 2021. ISBN 978-3-030-53035-8. doi: 10.1007/978-3-030-53036-5_28.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, January 2019. ISSN 2522-5839. doi: 10.1038/s42256-018-0004-1. URL `https://www.nature.com/articles/s42256-018-0004-1`. Publisher: Nature Publishing Group.

Daniel J. Berleant, Scott Ferson, Vladik Kreinovich, and Weldon A. Lodwick. Combining interval and probabilistic uncertainty: foundations, algorithms, challenges-an overview. In *4th international symposium on imprecise probabilities and their applications*, 2005. URL `https://www.academia.edu/download/46772074/tr05-09.pdf`.

Denny Borsboom. Latent Variable Theory. *Measurement: Interdisciplinary Research & Perspective*, 6(1-2):25–53, May 2008. ISSN 1536-6367, 1536-6359. doi: 10.1080/15366360802035497. URL `http://www.tandfonline.com/doi/abs/10.1080/15366360802035497`.

Ann Bostrom, Julie L. Demuth, Christopher D. Wirz, Mariana G. Cains, Andrea Schumacher, Deianna Madlambayan, Akansha Singh Bansal, Angela Bearth, Randy Chase, Katherine M. Crosman, Imme Ebert-Uphoff, David John Gagne II, Seth Guikema, Robert Hoffman, Branden B. Johnson, Christina Kumler-Bonfanti, John D. Lee, Anna Lowe, Amy McGovern, Vanessa Przybylo, Jacob T. Radford, Emilie Roth, Carly Sutter, Philippe Tissot, Paul Roebber, Jebb Q. Stewart, Miranda White, and John K. Williams. Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, 44(6):

1498–1513, 2024. ISSN 1539-6924. doi: 10.1111/risa.14245. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14245. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.14245.

Hans-Heinrich Bothe. *Fuzzy Logic: Einführung in Theorie und Anwendungen*. Springer-Verlag, 2013. URL https://books.google.de/books?hl=de&lr=&id=LXt_BwAAQBAJ&oi=fnd&pg=PA1&dq=fuzzy+logic&ots=5h9GaaM_iH&sig=JfHxyfZ34RQqQZ0dhHhwvxbjb8E.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. URL https://books.google.de/books?hl=de&lr=&id=IUZdAAAAQBAJ&oi=fnd&pg=PR11&dq=Boyd,+Stephen%3B+Vandenberghe,+Lieven+(2004-03-08).+Convex+Optimization.+Cambridge+University+Press.+doi:10.1017/cbo9780511804441.+ISBN+978-0-521-83378-3.&ots=HPHAdjaGGp&sig=7BWAZJcBeIIHyU763KJDBsCrbLg.

Alison J. Burnham, John F. MacGregor, and Roman Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48(2):167–180, 1999. URL https://www.sciencedirect.com/science/article/pii/S0169743999000180. Publisher: Elsevier.

Olivier Chapelle, Vladimir Vapnik, and Yoshua Bengio. Model Selection for Small Sample Regression. *Machine Learning*, 48(1):9–23, July 2002. ISSN 1573-0565. doi: 10.1023/A:1013943418833. URL https://doi.org/10.1023/A:1013943418833.

Haihua Chen, Jiangping Chen, and Junhua Ding. Data Evaluation and Enhancement for Quality Improvement of Machine Learning. *IEEE Transactions on Reliability*, 70(2):831–847, June 2021. ISSN 1558-1721. doi: 10.

1109/TR.2021.3070863. URL https://ieeexplore.ieee.org/abstract/document/9417095. Conference Name: IEEE Transactions on Reliability.

Zexun Chen, Jun Fan, and Kuo Wang. Multivariate Gaussian processes: definitions, examples and applications. *METRON*, 81(2):181–191, August 2023. ISSN 0026-1424, 2281-695X. doi: 10.1007/s40300-023-00238-3. URL https://link.springer.com/10.1007/s40300-023-00238-3.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112, 2009. URL https://www.sciencedirect.com/science/article/pii/S0167473008000556. Publisher: Elsevier.

Lynn E. Eberly and George Casella. Estimating Bayesian credible intervals. *Journal of statistical planning and inference*, 112(1-2):115–132, 2003. URL https://www.sciencedirect.com/science/article/pii/S0378375802003270. Publisher: Elsevier.

Imme Ebert-Uphoff, Ryan Lagerquist, Kyle Hilburn, Yoonjin Lee, Katherine Haynes, Jason Stock, Christina Kumler, and Jebb Q. Stewart. CIRA Guide to Custom Loss Functions for Neural Networks in Environmental Sciences – Version 1, June 2021. URL http://arxiv.org/abs/2106.09757. arXiv:2106.09757 [cs].

Frédérick Fabre Ferber, Dominique Gay, Jean-Christophe Soulié, Jean Diatta, and Odalric-Ambrym Maillard. Kriging and Gaussian Process Interpolation for Georeferenced Data Augmentation, January 2025. URL http://arxiv.org/abs/2501.07183. arXiv:2501.07183 [cs].

Yuqing Gao, Boyuan Kong, and Khalid M. Mosalam. Deep leaf-bootstrapping generative adversarial network for structural image data augmentation. *Computer-Aided Civil and Infrastructure Engineering*, 34(9):755–

773, 2019. ISSN 1467-8667. doi: 10.1111/mice.12458. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12458. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12458.

Zhitong Gao, Yucong Chen, Chuyu Zhang, and Xuming He. Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Experts, February 2023. URL http://arxiv.org/abs/2212.07328. arXiv:2212.07328 [cs].

P. Gardner, T. J. Rogers, C. Lord, and R. J. Barthorpe. Learning model discrepancy: A Gaussian process and sampling-based approach. *Mechanical Systems and Signal Processing*, 152:107381, May 2021. ISSN 0888-3270. doi: 10.1016/j.ymssp.2020.107381. URL https://www.sciencedirect.com/science/article/pii/S0888327020307676.

Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769, 2007. URL http://www.vldb.org/conf/2007/papers/research/p758-ghinita.pdf.

Leon Jay Gleser. The Importance of Assessing Measurement Reliability in Multivariate Regression. *Journal of the American Statistical Association*, 87(419):696–707, September 1992. ISSN 0162-1459. doi: 10.1080/01621459.1992.10475271. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475271. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1992.10475271.

Henry A Glick, Andrew H Briggs, and Daniel Polsky. Quantifying stochastic uncertainty and presenting results of cost-effectiveness analyses. *Expert Review of Pharmacoeconomics & Outcomes Research*, 1(1):25–36, October

2001. ISSN 1473-7167, 1744-8379. doi: 10.1586/14737167.1.1.25. URL https://www.tandfonline.com/doi/full/10.1586/14737167.1.1.25.

Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2001. ISBN 978-1-4899-0519-2 978-0-387-21606-5. doi: 10.1007/978-0-387-21606-5. URL http://link.springer.com/10.1007/978-0-387-21606-5.

J.C. Helton. Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*, 57(1-4):3–76, April 1997. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949659708811803. URL http://www.tandfonline.com/doi/abs/10.1080/00949659708811803.

Andreas Holzinger. The Next Frontier: AI We Can Really Trust. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 427–440, Cham, 2021. Springer International Publishing. ISBN 978-3-030-93736-2. doi: 10.1007/978-3-030-93736-2_33.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Brian Jalaian, Michael Lee, and Stephen Russell. Uncertain Context: Uncertainty Quantification in Machine Learning. *AI Magazine*, 40(4):40–49, December 2019. ISSN 2371-9621. doi: 10.1609/aimag.v40i4.4812. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/4812. Number: 4.

HM Dipu Kabir, Abbas Khosravi, Mohammad Anwar Hosen, and Saeid Nahavandi. Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE access*, 6:36218–36234, 2018. URL `https://ieeexplore.ieee.org/abstract/document/8371683/`. Publisher: IEEE.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356, 2011. URL `https://ieeexplore.ieee.org/abstract/document/5966350/`. Publisher: IEEE.

Michael Kläs and Anna Maria Vollmer. Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty. In Barbara Gallina, Amund Skavhaug, Erwin Schoitsch, and Friedemann Bitsch, editors, *Computer Safety, Reliability, and Security*, pages 431–438, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99229-7. doi: 10.1007/978-3-319-99229-7_36.

Matthew Large, Cherrie Galletly, Nicholas Myles, Christopher James Ryan, and Hannah Myles. Known unknowns and unknown unknowns in suicide risk assessment: evidence from meta-analyses of aleatory and epistemic uncertainty. *BJPsych bulletin*, 41(3):160–163, 2017. URL `https://www.cambridge.org/core/journals/bjpsych-bulletin/article/known-unknowns-and-unknown-unknowns-in-suicide-risk-assessment-evidence-from-metaanalyses-of-aleatory-and-epistemic-uncertainty/E7888A162A78E24473D1A2F45892FB0B`. Publisher: Cambridge University Press.

Gaoyang Li, Li Yang, Chi-Guhn Lee, Xiaohua Wang, and Mingzhe Rong. A Bayesian Deep Learning RUL Framework Integrating Epistemic and

Aleatoric Uncertainties. *IEEE Transactions on Industrial Electronics*, 68 (9):8829–8841, September 2021. ISSN 1557-9948. doi: 10.1109/TIE.2020.3009593. URL https://ieeexplore.ieee.org/abstract/document/9145803. Conference Name: IEEE Transactions on Industrial Electronics.

Torrin M. Liddell and John K. Kruschke. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348, November 2018. ISSN 0022-1031. doi: 10.1016/j.jesp.2018.08.009. URL https://www.sciencedirect.com/science/article/pii/S0022103117307746.

Momin M. Malik. A Hierarchy of Limitations in Machine Learning, February 2020. URL http://arxiv.org/abs/2002.05193. arXiv:2002.05193 [cs].

Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Overfitting, Model Tuning, and Evaluation of Prediction Performance. In Osval Antonio Montesinos López, Abelardo Montesinos López, and José Crossa, editors, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pages 109–139. Springer International Publishing, Cham, 2022. ISBN 978-3-030-89010-0. doi: 10.1007/978-3-030-89010-0_4. URL https://doi.org/10.1007/978-3-030-89010-0_4.

Evan Munro and Serena Ng. Latent Dirichlet Analysis of Categorical Survey Responses. *Journal of Business & Economic Statistics*, 40(1):256–271, January 2022. ISSN 0735-0015, 1537-2707. doi: 10.1080/07350015.2020.1802285. URL https://www.tandfonline.com/doi/full/10.1080/07350015.2020.1802285.

Sheila F. O'Brien and Qi Long Yi. How do I interpret a confidence interval? *Transfusion*, 56(7):1680–1683, July 2016. ISSN 0041-1132, 1537-2995.

doi: 10.1111/trf.13635. URL https://onlinelibrary.wiley.com/doi/10.1111/trf.13635.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2014. ISBN 978-1-4939-1322-0 978-1-4939-1323-7. doi: 10.1007/978-1-4939-1323-7. URL https://link.springer.com/10.1007/978-1-4939-1323-7.

Swathi Pothuganti. Review on over-fitting and under-fitting problems in Machine Learning and solutions. 7(9).

Apostolos F. Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, 477:111902, March 2023. ISSN 0021-9991. doi: 10.1016/j.jcp.2022.111902. URL https://www.sciencedirect.com/science/article/pii/S0021999122009652.

Daniyal Rajput, Wei-Jen Wang, and Chun-Chuan Chen. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, 24(1):48, February 2023. ISSN 1471-2105. doi: 10.1186/s12859-023-05156-9. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05156-9.

Sebastian Ruder. An overview of gradient descent optimization algorithms, June 2017. URL http://arxiv.org/abs/1609.04747. arXiv:1609.04747 [cs].

Francesco Santoni, Alessio De Angelis, Antonio Moschitta, and Paolo Carbone. Training Gaussian process regression through data augmentation for

battery SOC estimation. *Journal of Energy Storage*, 98:113073, September 2024. ISSN 2352-152X. doi: 10.1016/j.est.2024.113073. URL https://www.sciencedirect.com/science/article/pii/S2352152X24026598.

Iqbal H. Sarker. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3):160, March 2021. ISSN 2661-8907. doi: 10.1007/s42979-021-00592-x. URL https://doi.org/10.1007/s42979-021-00592-x.

E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 2018a.

Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, August 2018b. ISSN 0022-2496. doi: 10.1016/j.jmp.2018.03.001. URL https://www.sciencedirect.com/science/article/pii/S0022249617302158.

Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and Epistemic Uncertainty with Random Forests. In Michael R. Berthold, Ad Feelders, and Georg Krempl, editors, *Advances in Intelligent Data Analysis XVIII*, volume 12080, pages 444–456. Springer International Publishing, Cham, 2020. ISBN 978-3-030-44583-6 978-3-030-44584-3. doi: 10.1007/978-3-030-44584-3_35. URL http://link.springer.com/10.1007/978-3-030-44584-3_35. Series Title: Lecture Notes in Computer Science.

Christian Soize. *Uncertainty Quantification: An Accelerated Course with Advanced Applications in Computational Engineering*, volume 47 of *Interdisciplinary Applied Mathematics*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-54338-3 978-3-319-54339-0. doi: 10.1007/978-3-319-

54339-0. URL http://link.springer.com/10.1007/978-3-319-54339-0.

T. J. Sullivan. *Introduction to Uncertainty Quantification*. Springer, December 2015. ISBN 978-3-319-23395-6. Google-Books-ID: Sik3CwAAQBAJ.

Laura Swiler, Thomas Paez, and Randall Mayes. Epistemic uncertainty quantification tutorial. *Conference Proceedings of the Society for Experimental Mechanics Series*, January 2009.

Zhe Tang, Sihao Li, Kyeong Soo Kim, and Jeremy Smith. Multi-Output Gaussian Process-Based Data Augmentation for Multi-Building and Multi-Floor Indoor Localization. In *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 361–366, May 2022. doi: 10.1109/ICCWorkshops53468.2022.9814616. URL https://ieeexplore.ieee.org/abstract/document/9814616. ISSN: 2694-2941.

Richard Traunmüller, Andreas Murr, and Jeff Gill. Modeling latent information in voting data with Dirichlet process priors. *Political Analysis*, 23(1):1–20, 2015. URL https://www.cambridge.org/core/journals/political-analysis/article/modeling-latent-information-in-voting-data-with-dirichlet-process-priors/338F8FC146746CB76EAA19D1F8517A26. Publisher: Cambridge University Press.

Francesco Triggiano and Marco Romito. Gaussian Processes Based Data Augmentation and Expected Signature for Time Series Classification. *IEEE Access*, 12:80884–80895, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3408712. URL https://ieeexplore.ieee.org/abstract/document/10546274. Conference Name: IEEE Access.

Rohit Tripathy, Ilias Bilionis, and Marcial Gonzalez. Gaussian processes

with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191–223, 2016. URL https://www.sciencedirect.com/science/article/pii/S002199911630184X. Publisher: Elsevier.

Jie Wang. An intuitive tutorial to Gaussian process regression. *Computing in Science & Engineering*, 25(4):4–11, 2023. URL https://ieeexplore.ieee.org/abstract/document/10360364/. Publisher: IEEE.

Yuexi Wang, Nicholas Polson, and Vadim O. Sokolov. Data Augmentation for Bayesian Deep Learning. *Bayesian Analysis*, 18(4):1041–1069, December 2023. ISSN 1936-0975, 1931-6690. doi: 10.1214/22-BA1331. URL https://projecteuclid.org/journals/bayesian-analysis/volume-18/issue-4/Data-Augmentation-for-Bayesian-Deep-Learning/10.1214/22-BA1331.full. Publisher: International Society for Bayesian Analysis.

Ellen M. Whitener. Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, 75(3):315, 1990. URL https://psycnet.apa.org/fulltext/1990-27116-001.html. Publisher: American Psychological Association.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995. URL https://proceedings.neurips.cc/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

Fupin Yao. Machine learning with limited data, January 2021. URL http://arxiv.org/abs/2101.11461. arXiv:2101.11461 [cs].

Mohammad Yazdi, Noorbakhsh Amiri Golilarz, Kehinde Adewale Adesina, and Arman Nedjati. Probabilistic Risk Analysis of Process Systems Con-

sidering Epistemic and Aleatory Uncertainties: A Comparison Study. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 29(02):181–207, April 2021. ISSN 0218-4885. doi: 10.1142/S0218488521500098. URL https://www.worldscientific.com/doi/abs/10.1142/S0218488521500098. Publisher: World Scientific Publishing Co.

Juan Zhang, Junping Yin, and Ruili Wang. Basic Framework and Main Methods of Uncertainty Quantification. *Mathematical Problems in Engineering*, 2020:1–18, August 2020. ISSN 1563-5147, 1024-123X. doi: 10.1155/2020/6068203. URL https://www.hindawi.com/journals/mpe/2020/6068203/.

Le Zhou, Junghui Chen, Zhihuan Song, Zhiqiang Ge, and Aimin Miao. Probabilistic latent variable regression model for process-quality monitoring. *Chemical Engineering Science*, 116:296–305, 2014. URL https://www.sciencedirect.com/science/article/pii/S0009250914002115. Publisher: Elsevier.

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.

Landshut, den 18.November 2025

Stefan M. Stroka