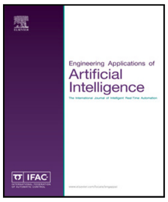
Engineering Applications of Artificial Intelligence 149 (2025) 110363

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Integrating permutation feature importance with conformal prediction for robust Explainable Artificial Intelligence in predictive process monitoring

Nijat Mehdiyev [*], Maxim Majlatow, Peter Fettke

German Research Center for Artificial Intelligence (DFKI) and Saarland University, Saarbrücken, 66123, Germany

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence
Uncertainty quantification
Conformal prediction
Predictive process monitoring

ABSTRACT

As artificial intelligence (AI) systems are increasingly deployed in high-stakes environments, the need for explanations that convey uncertain information has become evident. Conventional explainable AI (XAI) methods often overlook uncertainty, focusing solely on point predictions. To address this gap, we propose using permutation feature importance (PFI) combined with predictive uncertainty evaluation measures. This novel approach examines the significance of features by relating them to the model's confidence in its predictions. By using split conformal prediction (SCP) to quantify predictive uncertainty and integrating the outcomes to PFI, we aim to enhance the robustness and interpretability of machine learning (ML) algorithms. More importantly, we examine three scenarios for conformal prediction-based PFI explanations: permuting feature values in the test data, the calibration data, and both. These scenarios assess the impact of feature permutations from different perspectives, revealing feature sensitivity and the importance of features in various settings. We also perform a series of sensitivity analyses, particularly exploring calibration data size and computational efficiency, to demonstrate the robustness and scalability of our approach for industrial applications. Our comprehensive evaluation offers insights into feature impact on predictions and their associated confidence levels. We validate our proposed approach through a real-world predictive process monitoring use case in manufacturing.

1. Introduction

Explainable Artificial Intelligence (XAI) has emerged as an essential research domain in response to the increasing complexity and opacity of artificial intelligence (AI) systems (Gunning et al., 2019; Emmert-Streib et al., 2020). XAI aims to make AI models more transparent, interpretable, and understandable to human users (Arrieta et al., 2020). This field encompasses various techniques and approaches designed to provide insights into the decision-making processes of AI systems, particularly those based on complex machine learning (ML) models like deep neural networks or ensemble-based approaches (Dwivedi et al., 2023; Adadi and Berrada, 2018). By enhancing the interpretability of AI models, XAI fosters trust and facilitates the integration of AI into critical domains such as healthcare, finance, and criminal justice, where transparency and accountability are paramount (Loh et al., 2022; Guidotti et al., 2018; Ali et al., 2023).

As AI systems are increasingly deployed in high-stakes environments, the need for explanations that also convey uncertain information has become evident (Gawlikowski et al., 2023). Uncertainty quantification (UQ) is crucial in this context, as it provides a measure of confidence or reliability in the outputs of ML models (Abdar et al., 2021). UQ helps to create a more comprehensive picture of an AI system's decision-making process by explaining the predictions and indicating the certainty associated with those predictions (Yang and Yee, 2024). This additional layer of information is vital for assessing the reliability of AI-driven decisions and identifying situations where human intervention might be necessary (Bhatt et al., 2021).

Despite the progress in XAI, integrating uncertainty information into explanations presents significant challenges. Traditional approaches to explainability primarily focus on point predictions without considering the inherent uncertainty in those predictions (Löfström et al., 2024; Mehdiyev et al., 2024b). To address this gap, we propose a novel approach that integrates permutation feature importance (PFI) with conformal prediction (CP), a robust UQ method. Unlike conventional PFI, which measures the importance of features for point predictions, our approach incorporates uncertainty evaluation measures to provide a more holistic understanding of model behavior. In this study, we employ split conformal prediction (SCP), a variant of CP, to quantify uncertainty. SCP offers finite-sample validity and is particularly suitable for scenarios where data is limited (Vovk et al., 2020). By

[*] Corresponding author.
E-mail address: nijat.mehdiyev@dfki.de (N. Mehdiyev).

https://doi.org/10.1016/j.engappai.2025.110363
Received 18 July 2024; Received in revised form 13 January 2025; Accepted 18 February 2025
Available online 15 March 2025

integrating this method with PFI, we aim to enhance the robustness and interpretability of deployed ML algorithms.

We examine in this study three distinct scenarios: permuting feature values in the test data, permuting feature values in the calibration data, and permuting feature values in both calibration and test data. In the first scenario, we permute the feature values in the test data while keeping the calibration data intact. This approach allows us to assess how changes in the test data impact the model's predictive performance and confidence. By isolating the effects of feature permutations in the test data, we can identify which features are crucial for maintaining the model's reliability on new, unseen data. This scenario is particularly relevant for understanding the robustness of the model's predictions under variations in the input data, which is critical for real-world applications where input data can often be noisy or incomplete. We expect this scenario to highlight the sensitivity of the model's predictions and uncertainty estimates to changes in specific features, thereby indicating their importance.

In the second scenario, we permute the feature values in the calibration data while keeping the test data unchanged. The calibration data is used to generate the prediction intervals, and by permuting its feature values, we can observe how this affects the model's uncertainty measures on the test data. This scenario helps us understand the impact of calibration data quality on the reliability of the UQ. It is particularly important for scenarios where the calibration data may be subject to variations or noise, affecting the model's ability to accurately quantify uncertainty. We expect this scenario to reveal how perturbations in the calibration phase influence the model's predictive intervals and highlight the features that are critical for constructing reliable and accurate prediction intervals.

In the third scenario, we permute the feature values in both the calibration and test datasets. This comprehensive approach allows us to examine the combined effect of feature permutations across both stages of the model evaluation process. By analyzing the compounded impact, we can gain insights into the overall robustness of the model's UQ under simultaneous perturbations. This scenario is essential for understanding the interdependencies between calibration and test data features and their joint influence on the model's performance. We expect this scenario to provide a deeper understanding of how simultaneous changes in both datasets affect the prediction intervals and identify the features that are pivotal in maintaining robust and reliable uncertainty estimates.

By exploring these three scenarios, we aim to provide a comprehensive evaluation of feature importance in the context of UQ. This approach not only enhances our understanding of which features are most influential in predicting outcomes but also how they contribute to the confidence we have in those predictions. This dual focus on predictive accuracy and uncertainty is crucial for developing robust and explainable AI systems that can be trusted in high-stakes decision-making environments. Additionally, we conduct a series of sensitivity experiments to investigate how varying the calibration data size affects both coverage and interval accuracy, ensuring the method's robustness for different data availability conditions. We further assess computational efficiency by evaluating runtime performance under various dataset scales, confirming that our approach remains practical even for large industrial scenarios. By uniting sensitivity analyses with considerations of scalability, we offer actionable insights for domains where robust UQ and feature interpretability are paramount. Our proposed approach focuses on the domain of predictive process monitoring, specifically addressing predictive analytics problems related to the duration of production activities based on data from manufacturing execution systems (MES). However, the methodology is transferable to other domains with similar tabular data problems.

The remainder of the paper is organized as follows: Section 2 provides an overview of the background and related work, highlighting key advancements and limitations in the field. In Section 3, we present the details of our proposed method, including its theoretical foundation and implementation. Section 4 describes the experimental setup, datasets, and evaluation metrics used to validate our approach. The results and discussion are presented in Section 5, where we analyze the performance of the proposed model and interpret the model outcomes. Section 6 discusses our findings and outlines potential directions for future research. Finally, Section 7 concludes the paper with a summary of key contributions.

2. Background and related work

2.1. Explainable Artificial Intelligence (XAI)

XAI refers to a set of processes and methods that allow human users to comprehend and trust the results and outputs created by ML algorithms (Ali et al., 2023; Adadi and Berrada, 2018; Langer et al., 2021). XAI aims to make the decision-making process of AI systems transparent and understandable, countering the black-box nature of many AI models (Saeed and Omlin, 2023; Gunning et al., 2019). The concept of XAI has gained significant attention due to the increasing complexity and deployment of AI systems in critical applications such as healthcare, finance, law and autonomous driving, where understanding AI decisions is crucial for safety, fairness, and accountability (Loh et al., 2022; Weber et al., 2024; Dong et al., 2023; Vale et al., 2022).

The necessity for XAI arises from several factors. Firstly, regulatory requirements such as the EU AI Act emphasize the need for transparency and accountability in AI systems to ensure they are used responsibly and ethically (Panigutti et al., 2023). Secondly, XAI is essential for building trust among users and stakeholders, allowing them to understand and validate AI decisions (Arrieta et al., 2020). Thirdly, XAI helps identify and mitigate biases in AI models, ensuring that decisions are fair and unbiased (Nakao et al., 2022). Lastly, explainability is crucial for debugging and improving AI models, as it provides insights into how models operate and where they might be going wrong (Mehdiyev and Fettke, 2021). Several key desiderata have been identified for XAI systems. These include fidelity (the explanation should accurately reflect the model's behavior), comprehensibility (the explanation should be understandable to the intended audience), and actionability (the explanation should provide insights that can be used to improve or act upon the model's decisions) (Mohseni et al., 2021; Liao et al., 2022). Additionally, XAI systems should be able to provide explanations that are consistent across similar inputs and robust to small perturbations in the input data (Baniecki and Biecek, 2024; Chander et al., 2024).

XAI methods and approaches can be broadly categorized into several dimensions. One key distinction is between model-specific and model-agnostic methods (Adadi and Berrada, 2018). Model-specific approaches are tailored to particular types of AI models and can leverage the internal structure of these models to generate explanations. In contrast, model-agnostic methods can be applied to any type of AI model, treating it as a black box and focusing on the relationship between inputs and outputs. Another important categorization is between transparency-based and post-hoc explanation methods (Guidotti et al., 2018). Transparency-based approaches aim to create inherently interpretable models, such as decision trees or linear models. Post-hoc methods, on the other hand, generate explanations for already trained models, often using techniques like feature attribution or example-based explanations. XAI methods can also be classified as global or local (Arrieta et al., 2020). Global methods aim to explain the overall behavior of a model across its entire input space, while local methods focus on explaining individual predictions or decisions. Global methods can provide a high-level understanding of a model's behavior, while local methods offer more detailed insights into specific cases. Despite the progress made in XAI, several challenges remain. One significant issue is the potential trade-off between model performance and explainability, as more complex models that often achieve higher accuracy can be more difficult to explain (Rudin, 2019). Another challenge lies in evaluating the quality and effectiveness of explanations, as there is no universally accepted metric for measuring explanation quality (Mohseni et al., 2021).

A particularly pressing problem in the field of XAI is the integration of UQ. As AI systems are deployed in high-stakes domains, understanding not just the decisions they make but also the confidence or uncertainty associated with those decisions becomes crucial (Slack et al., 2021). The relevance of UQ to XAI lies in its potential to provide a more complete picture of an AI system's decision-making process (Watson et al., 2024). By incorporating measures of uncertainty into explanations, XAI systems can offer more nuanced and informative insights, helping users to assess the reliability of AI-driven decisions better and identify situations where human intervention may be necessary. There is a recent and growing research interest in the intersection of XAI and UQ, and despite this increasing attention, the field remains underexplored, which this study aims to address (Löfström et al., 2024; Hill et al., 2024; Watson et al., 2024; Slack et al., 2021; Chiaburu et al., 2024; Marx et al., 2023; Mehdiyev et al., 2023).

Unlike classical combinatorial feature selection (CFS) methods which aim to identify a minimal set of features that optimally represent the data, PFI focuses on the isolated impact of each feature (Fisher et al., 2019). Different from traditional PFI, which measures feature relevance based on changes in point prediction accuracy, our proposed approach evaluates how individual features contribute to the reliability and informativeness of prediction intervals. Specifically, PFI does not consider feature interactions or subset optimization but instead quantifies how disrupting a single feature affects prediction intervals and UQ metrics like prediction interval coverage probability (PICP), mean prediction interval width (MPIW), mean relative prediction interval width (MRPIW), and the Winkler score. This approach ensures that feature contributions to predictive intervals can be robustly evaluated, highlighting their role in generating transparent and reliable models.

2.2. Uncertainty quantification (UQ)

UQ in predictive modeling refers to the process of estimating and characterizing the uncertainty associated with model predictions (Abdar et al., 2021). It aims to provide a measure of confidence or reliability in the outputs of ML models, going beyond just point predictions to quantify how certain or uncertain those predictions are. Predictive uncertainty arises from various sources within the learning process. This uncertainty is inherent in all ML models and can significantly impact decision-making processes in critical applications such as healthcare, autonomous driving, and financial forecasting (Ghanem et al., 2017). As ML models become increasingly complex and are deployed in high-stakes environments, understanding and quantifying predictive uncertainty has become a crucial area of research and development.

The sources of uncertainty in ML can be broadly categorized into two types: aleatoric and epistemic uncertainty (Hüllermeier and Waegeman, 2021). Aleatoric uncertainty, also known as statistical or data uncertainty, stems from the inherent randomness or noise in the data and is irreducible even with the collection of more data (Kendall and Gal, 2017). This type of uncertainty can be further divided into homoscedastic uncertainty, which remains constant across all inputs, and heteroscedastic uncertainty, which varies depending on the input (Malinin and Gales, 2018). On the other hand, epistemic uncertainty, also referred to as model uncertainty, arises from the model's lack of knowledge or understanding of the underlying data-generating process. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by gathering more data or improving the model architecture (Ovadia et al., 2019).

Various methods have been proposed to quantify uncertainty in ML models, broadly classified into Bayesian and frequentist approaches (Bhatt et al., 2021). These methods aim to provide a comprehensive framework for estimating and understanding the uncertainty associated with model predictions. Bayesian methods, such as Bayesian Neural Networks and Monte Carlo dropout, offer a principled approach to UQ by treating model parameters as random variables and inferring their posterior distributions (MacKay, 1995; Gal and Ghahramani, 2016). These methods are particularly effective at capturing both aleatoric and epistemic uncertainty, making them attractive for applications that require a comprehensive treatment of uncertainty. Bayesian approaches naturally incorporate prior knowledge and provide a probabilistic interpretation of model predictions, which can be valuable in decision-making processes (Graves, 2011).

On the other hand, frequentist methods, while less common, offer alternative approaches to UQ. These include techniques like bootstrap sampling or ensemble methods (Osband et al., 2016; Abdar et al., 2021). Frequentist approaches may not incorporate prior knowledge as explicitly as Bayesian methods, but they can be valuable in scenarios where computational simplicity and interpretability are prioritized. For instance, ensemble methods can provide robust uncertainty estimates by aggregating predictions from multiple models (Lakshminarayanan et al., 2017).

Both Bayesian and frequentist approaches have their strengths and limitations. Bayesian methods often provide more comprehensive uncertainty estimates but can be computationally intensive, especially for large-scale models. Frequentist methods, while potentially less comprehensive, can offer computational advantages and may be easier to implement in certain scenarios. It is worth noting that the choice between Bayesian and frequentist approaches often depends on the specific application, available computational resources, and the desired trade-off between uncertainty estimation quality and computational efficiency. In practice, researchers and practitioners may use a combination of methods or hybrid approaches to leverage the strengths of both paradigms.

2.3. Conformal prediction

Despite the strengths of Bayesian and frequentist methods, there remains a need for approaches that can provide robust uncertainty estimates with minimal assumptions about the underlying data distribution. This is where CP comes into play, which emerged as a powerful framework for UQ that offers distribution-free validity, requiring only the assumption of exchangeability in the data (Vovk et al., 2005; Tibshirani et al., 2019). CP offers several other key advantages over conventional UQ approaches. One of the further strengths of CP is its ability to provide finite-sample validity (Xu and Xie, 2021). Unlike many traditional UQ methods that rely on asymptotic guarantees, CP offers valid prediction intervals even with limited data samples (Tibshirani et al., 2019; Gibbs and Candes, 2021). This property is crucial in practical scenarios where data may be scarce or expensive to obtain. Furthermore, the model-agnostic nature of CP allows it to be applied to various pre-trained models, including complex deep neural networks, without requiring retraining or modification of the underlying model architecture (Lei et al., 2018; Foygel Barber et al., 2021).

There are several types of CP, each with its strengths and weaknesses. Full conformal prediction uses the entire dataset for calibration, providing the most accurate uncertainty estimates but at a high computational cost (Shafer and Vovk, 2008; Angelopoulos et al., 2023). SCP, on the other hand, uses a held-out calibration set, which is faster but potentially less efficient (Papadopoulos et al., 2002; Lei et al., 2015). Inductive conformal prediction is suitable for online learning settings, where the model is continuously updated with new data (Papadopoulos et al., 2002). Mondrian conformal prediction provides conditional coverage guarantees for subgroups, making it useful for applications where the data is heterogeneous (Toccaceli and Gammerman, 2019). Recent work has focused on extending CP to more complex scenarios. Tibshirani et al. (2019) proposed methods for CP under covariate shift, addressing the challenge of distribution mismatch between training and test data (Tibshirani et al., 2019). Angelopoulos and Bates (2023) provide a comprehensive overview of recent developments, including applications to structured outputs, time series data, and scenarios involving outliers or abstaining models (Angelopoulos et al., 2023).

2.4. Predictive process monitoring (PPM)

Predictive Process Monitoring (PPM) is a branch of process mining that aims to forecast the future behavior or outcomes of ongoing business process instances based on historical event logs and current process data (Van Der Aalst, 2012; Breuker et al., 2016). It extends traditional process monitoring by leveraging ML techniques to provide proactive insights and support decision-making during process execution (Evermann et al., 2017; Mehdiyev and Fettke, 2020). PPM enables organizations to anticipate and mitigate potential issues, optimize resource allocation, and improve overall process performance. The field of PPM encompasses several problem types, including predicting the remaining time until process completion, forecasting the next activity in a process instance, estimating the likelihood of specific outcomes (e.g., customer churn or compliance violations), and predicting process performance indicators (Di Francescomarino et al., 2018). These predictions can be made at various stages of process execution, from early predictions based on limited information to more accurate forecasts as more data becomes available.

Numerous ML approaches have been applied to PPM tasks. Early works focused on traditional classification and regression techniques, such as decision trees and support vector machines (Lakshmanan et al., 2011; Maggi et al., 2014; De Koninck et al., 2017). More recently, deep learning methods have gained prominence due to their ability to capture complex temporal dependencies in process data. Long Short-Term Memory (LSTM) networks, in particular, have shown promising results in predicting next activities and remaining time (Evermann et al., 2017). Other approaches include the use of random forests, gradient boosting machines, and ensemble methods that combine multiple predictive models (Di Francescomarino et al., 2018; Teinemaa et al., 2019; Márquez-Chamorro et al., 2017).

As PPM systems become more sophisticated and widely adopted, there is a growing need for XAI techniques to interpret and justify predictions (Mehdiyev and Fettke, 2021). XAI for PPM aims to provide transparent and understandable explanations for the predictions made by complex models, enabling process stakeholders to trust and act upon the insights generated. Techniques such as SHapley Additive exPlanations (SHAP) values, Local Interpretable Model-agnostic Explanations (LIME), and counterfactual explanations have been adapted to explain PPM predictions, offering insights into the most influential factors driving specific forecasts (Mehdiyev and Fettke, 2020; Coma-Puig and Carmona, 2022; Bukhsh et al., 2019; Rizzi et al., 2020; De Koninck et al., 2017; Hsieh et al., 2021).

UQ is another important aspect of PPM that has gained attention in recent years. UQ techniques aim to provide reliable estimates of the confidence or uncertainty associated with process predictions, allowing decision-makers to assess the risk and reliability of forecasts (Weytjens and De Weerdt, 2022). Different methods and approaches have been explored to quantify uncertainty in PPM predictions (Shoush and Dumas, 2022; Bousdekis et al., 2023; Portolani et al., 2022). Incorporating UQ into PPM systems can lead to more robust decision-making processes and help identify cases where additional information or human intervention may be necessary (Mehdiyev et al., 2023). The combination of XAI and UQ techniques with state-of-the-art ML approaches promises to enhance the practical applicability and trustworthiness of PPM systems in real-world business environments (Mehdiyev et al., 2024b).

3. Methodology

Our proposed approach integrated advanced ML techniques with rigorous UQ and XAI practices to create a comprehensive framework for reliable and transparent predictive process monitoring (see Fig. 1). By leveraging data from the examined process-aware information systems, the methodology starts with thorough process data preprocessing to ensure high-quality inputs. Following this, model training focuses on building robust predictive models through hyperparameter optimization and performance evaluation. The UQ stage then introduces SCP to provide reliable prediction intervals, ensuring that the model's predictions are not only accurate but also accompanied by quantified uncertainty. Finally, the XAI component incorporates PFI to shed light on the model's decision-making processes by focusing on the model's confidence. In this regard, we examine the illustrated three scenarios to explain uncertainty from different perspectives.

3.1. Predictive process monitoring: Data preprocessing and model training

In this study, we tackle a specific predictive process monitoring problem to illustrate the relevance and effectiveness of our proposed explainability approach combined with uncertainty estimation. This section provides a thorough mathematical formalization of the process prediction problem, forming the foundation for our subsequent analysis and methodology development. Furthermore, the presented notation is predominantly based on the works of Mehdiyev et al. (2024a, 2023, 2024b), contributing to consistency among relevant literature. We start by describing the crucial elements of the predictive process and then outline the interrelationships among these elements, creating a cohesive framework for process prediction. The formulation of the predictive analytics problem is presented as follows:

Definition 1 (*Event*). An *event* is represented as a tuple
$e = (a, c, t_{\text{start}}, t_{\text{complete}}, v_1, \dots, v_n)$, where:

- $a \in \mathcal{A}$ represents the process activity,
- $c \in \mathcal{C}$ signifies the case identifier,
- $t_{\text{start}} \in \mathcal{T}_{\text{start}}$ is the start timestamp of the event (in Unix epoch time since January 1, 1970),
- $t_{\text{complete}} \in \mathcal{T}_{\text{complete}}$ is the completion timestamp of the event,
- $v_1, \dots, v_n$ are the event-specific attributes, each $v_i \in \mathcal{V}_i$, with $\mathcal{V}_i$ representing the domain of the i-th attribute.

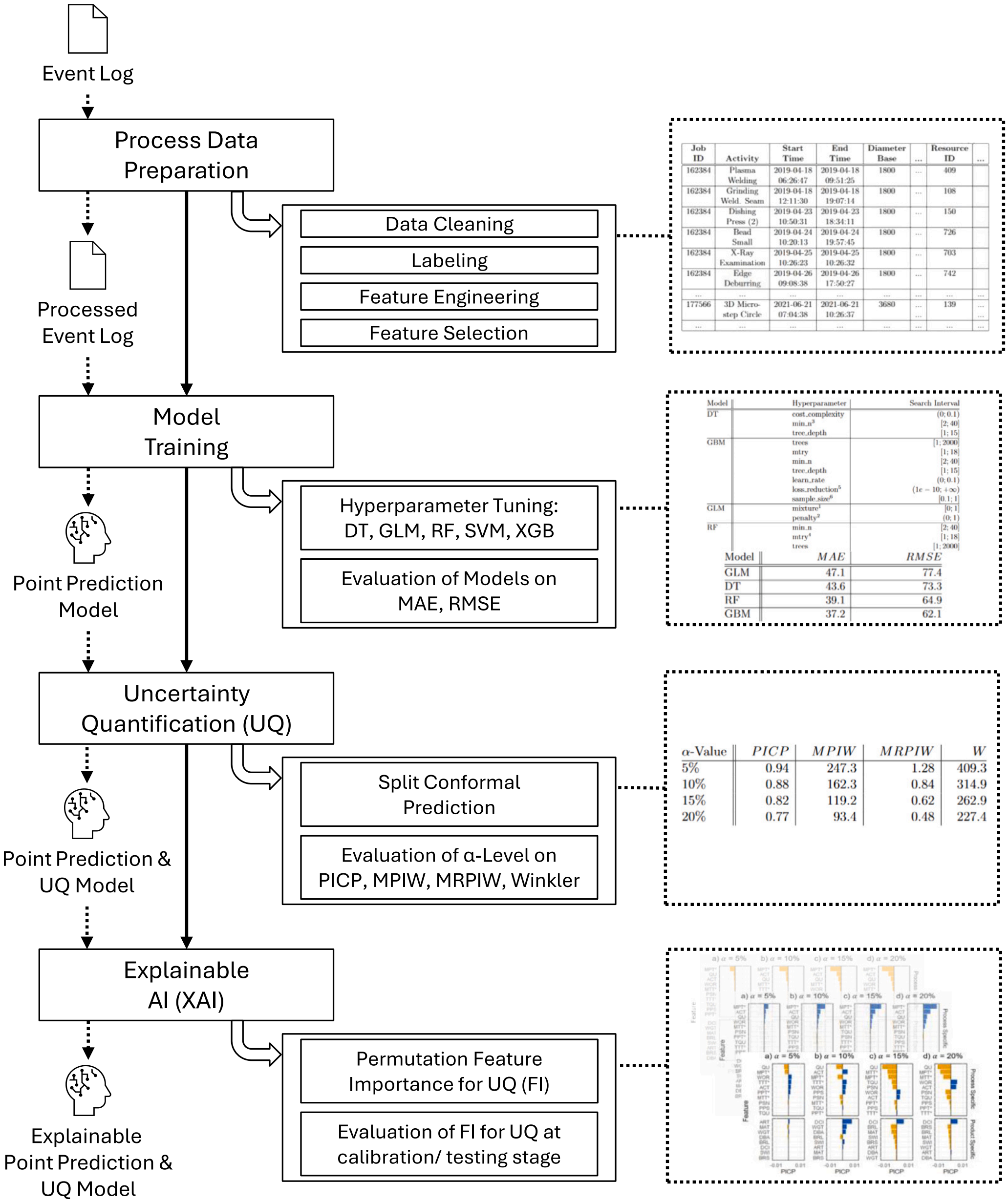
The set of all possible events is $\mathcal{E} = \mathcal{A} \times \mathcal{C} \times \mathcal{T}_{\text{start}} \times \mathcal{T}_{\text{complete}} \times \mathcal{V}_1 \times \cdots \times \mathcal{V}_n$. For an event $e \in \mathcal{E}$, the following projection functions are defined:

- $p_a : \mathcal{E} \to \mathcal{A},\ p_a(e) = a$,
- $p_c : \mathcal{E} \to \mathcal{C},\ p_c(e) = c$,
- $p_{t_{\text{start}}} : \mathcal{E} \to \mathcal{T}_{\text{start}},\ p_{t_{\text{start}}}(e) = t_{\text{start}}$,
- $p_{t_{\text{complete}}} : \mathcal{E} \to \mathcal{T}_{\text{complete}},\ p_{t_{\text{complete}}}(e) = t_{\text{complete}}$,
- $p_{v_i} : \mathcal{E} \to \mathcal{V}_i,\ p_{v_i}(e) = v_i$ for $1 \le i \le n$.

Definition 2 (*Traces and Event Log*). A *trace* $\sigma \in \mathcal{E}^*$ is a finite sequence of events $\sigma_c = \langle e_1, e_2, \dots, e_{|\sigma_c|} \rangle$, where each $e_i \in \sigma$ appears only once and $\forall e_i, e_j \in \sigma$, $p_c(e_i) = p_c(e_j)$ and $p_{t_{\text{start}}}(e_i) \le p_{t_{\text{start}}}(e_j)$ if $1 \le i < j < |\sigma_c|$. The *event log* $\mathcal{E}_C$ is defined as the set of completed traces, $\mathcal{E}_C = \{\sigma_c \mid c \in \mathcal{C}\}$.

Definition 3 (*Partial Traces*). Partial traces are extracted from a full trace σ. Using $hd^i(\sigma_c)$ and $tl^i(\sigma_c)$, prefixes and suffixes are generated:

- Selection operator (.): $\sigma_c(i) = \sigma_i$ for $1 \le i \le n$,
- $hd^i(\sigma_c) = \langle e_1, e_2, \dots, e_{\min(i,n)} \rangle$ for $i \in [1, |\sigma_c|] \subset \mathbb{N}$,
- $tl^i(\sigma_c) = \langle e_w, e_{w+1}, \dots, e_n \rangle$ where $w = \max(n - i + 1, 1)$,
- $|\sigma| = n$ (the length of the trace).


Event Log
Process Data Preparation
Data Cleaning
Labeling
Feature Engineering
Feature Selection
Processed Event Log
Model Training
Hyperparameter Tuning: DT, GLM, RF, SVM, XGB
Evaluation of Models on MAE, RMSE
Point Prediction Model
Uncertainty Quantification (UQ)
Split Conformal Prediction
Evaluation of α-Level on PICP, MPIW, MRPIW, Winkler
Point Prediction & UQ Model
Explainable AI (XAI)
Permutation Feature Importance for UQ (FI)
Evaluation of FI for UQ at calibration/ testing stage
Explainable Point Prediction & UQ Model
a) α = 5%
b) α = 10%
c) α = 15%
d) α = 20%
Feature
Process Specific
Product Specific
PICP

Job ID	Activity	Start Time	End Time	Diameter Base	...	Resource ID	...
162384	Plasma Welding	2019-04-18 06:26:47	2019-04-18 09:51:25	1800	...	409	
162384	Grinding Weld. Seam	2019-04-18 12:11:30	2019-04-18 19:07:14	1800	...	108	
162384	Dishing Press (2)	2019-04-23 10:50:31	2019-04-23 18:34:11	1800	...	150	
162384	Bead Small	2019-04-24 10:20:13	2019-04-24 19:57:45	1800	...	726	
162384	X-Ray Examination	2019-04-25 10:26:23	2019-04-25 10:26:32	1800	...	703	
162384	Edge Deburring	2019-04-26 09:08:38	2019-04-26 17:50:27	1800	...	742	
...	...	...	...	...	...	...	...
177566	3D Micro-step Circle	2021-06-21 07:04:38	2021-06-21 10:26:37	3680	... ...	139	... ...
...	...	...	...	...	...	...	...

Model	Hyperparameter	Search Interval
DT	cost_complexity	(0; 0.1)
	min_n[3]	[2; 40]
	tree_depth	[1; 15]
GBM	trees	[1; 2000]
	mtry	[1; 18]
	min_n	[2; 40]
	tree_depth	[1; 15]
	learn_rate	(0; 0.1)
	loss_reduction[5]	(1e − 10; +∞)
	sample_size[6]	[0.1; 1]
GLM	mixture[1]	[0; 1]
	penalty[2]	(0; 1)
RF	min_n	[2; 40]
	mtry[4]	[1; 18]
	trees	[1; 2000]

Model	MAE	$RMSE$
GLM	47.1	77.4
DT	43.6	73.3
RF	39.1	64.9
GBM	37.2	62.1

α-Value	$PICP$	$MPIW$	$MRPIW$	W
5%	0.94	247.3	1.28	409.3
10%	0.88	162.3	0.84	314.9
15%	0.82	119.2	0.62	262.9
20%	0.77	93.4	0.48	227.4

Fig. 1. Proposed approach: Uncertainty-Related Permutation Feature Importance.

The partial traces produced by $tl^i(\sigma_c)$ enable the construction of a tabular dataset to predict the duration of remaining events in an ongoing trace. Our predictive process monitoring approach starts at the case initiation and updates as events occur. The $tl^i(\sigma_c)$ function is instrumental in generating partial traces for structuring training data. Various process performance indicators (PPIs) are typically used as prediction targets. We focus on the processing time of a specific event, computed as the difference in minutes between its completion and start timestamps.

Definition 4 (*Event Processing Time/Labeling*)**.** For a non-empty trace $\sigma \neq \langle\rangle \in \mathcal{E}^*$, a labeling function resp : $\mathcal{E} \to \mathcal{Y}$ maps an event $e \in \sigma$ to the value of its response variable $\text{resp}(e) \in \mathcal{Y}$. The event processing time is calculated as:

$$\text{resp}(e) = p_{t_{\text{complete}}}(e) - p_{t_{\text{start}}}(e),$$

where $\mathcal{Y} \subset \mathbb{R}^+$.

Definition 5 (*Feature Extraction*)**.** The feature extraction function feat : $\mathcal{E}^* \to \mathcal{X}^*$ extracts features from a non-empty trace $\sigma \neq \langle\rangle \in \mathcal{E}^*$, with $\mathcal{X} \in \mathbb{R}^{\text{dim}}$ representing the feature domain and dim the dimensionality necessary to represent the amount of extracted features. For a trace $\sigma_c = \langle e_1, e_2, \ldots, e_{|\sigma_c|}\rangle$, the function feat produces features $(x_{i,1}, \ldots, x_{i,\text{dim}})$ for each event e_i, including case-specific and event-specific features as well as intra-case features like n-grams.

Definition 6 (*Prediction Task*)**.** Given an incomplete, non-empty trace $\sigma_c = \langle e_1, \ldots, e_i, e_{i+1}, \ldots, e_{|\sigma_c|}\rangle$, we define the prediction task as a supervised learning problem, predicting the processing time of the next upcoming event $\text{resp}(e_{i+1})$ from the suffix based on available data.

3.2. Split conformal prediction

SCP is an advanced statistical framework that strategically divides the original dataset into three distinct subsets: a training set D_{train}, a calibration set D_{cal}, and a test set D_{test}. This division is crucial for the model's ability to generalize effectively beyond the training data and ensures the accuracy of prediction intervals under real-world conditions.

The dataset partitioning is as follows:

- $D_{\text{train}} = \{(X_i, Y_i) : i \in I_{\text{train}}\}$,
- $D_{\text{cal}} = \{(X_i, Y_i) : i \in I_{\text{cal}}\}$,
- $D_{\text{test}} = \{(X_i, Y_i) : i \in I_{\text{test}}\}$,

where I_{train}, I_{cal}, and I_{test} are disjoint subsets of identifiers for rows of the original dataset such that $I_{\text{train}} \cup I_{\text{cal}} \cup I_{\text{test}} = \{1, \dots, n\}$. This ensures comprehensive coverage and unbiased evaluation of the model's predictive capabilities.

A regression model $\hat{f}$ is trained exclusively on D_{train}. The model aims to predict the dependent variable Y based on the explanatory variables X. The effectiveness of $\hat{f}$ depends on its ability to generalize from the training data to unseen data, which is subsequently evaluated using D_{test}.

After training, the non-conformity scores for the calibration set are calculated as:

$$R_i = |Y_i - \hat{f}(X_i)|, \quad \forall i \in I_{\text{cal}}, \tag{1}$$

These scores measure how much each prediction deviates from the actual observed values in D_{cal}, serving as a critical dataset for statistical analysis.

The pivotal step in SCP involves calculating the quantile of the non-conformity scores to determine the width of the prediction intervals. The quantile is calculated as:

$$\hat{q} = \text{Quantile}\left(\left\{ R_i : i \in I_{\text{cal}} \right\}, \ \frac{\left\lceil (1-\alpha)\left(|I_{\text{cal}}| + 1 \right) \right\rceil}{|I_{\text{cal}}|} \right) \tag{2}$$

where α specifies the miscoverage level (thus, the confidence level is $1 - \alpha$), and the quantile calculation includes an adjustment factor $\frac{\lceil (1-\alpha)(, |I_{\text{cal}}| + 1,) \rceil}{|I_{\text{cal}}|}$ to correct for the finite sample size of D_{cal}.

For each test data point X_{n+1} in D_{test}, the prediction interval is constructed as:

$$\hat{C}(X_{n+1}) = \left[\hat{f}(X_{n+1}) - \hat{q}, \hat{f}(X_{n+1}) + \hat{q} \right] \tag{3}$$

The SCP method is designed to guarantee that the prediction interval $\hat{C}(X_{n+1})$ will cover the true response Y_i with a probability of at least $1 - \alpha$. Mathematically, this is expressed as:

$$P(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha \tag{4}$$

This guarantee holds under the assumption that the data points (X_1, Y_1), …, (X_n, Y_n) are exchangeable. Exchangeability means that the joint distribution of the data remains unchanged under permutations, which is a slightly weaker assumption than being independent and identically distributed. The key idea behind the coverage guarantee is the use of the calibration set (D_{cal}) to estimate the distribution of the non-conformity scores. By determining the quantile of these scores from D_{cal}, the method ensures that the constructed prediction interval will cover the true response for new data points with the desired probability, leveraging the empirical distribution of the non-conformity scores for non-parametric, distribution-free predictions.

Regarding the effect of calibration set size on the quality of the prediction intervals yielded by SCP, it is expected that the conformal prediction method guarantees coverage of at least $1 - alpha$ on the test data as long as the calibration data encompasses a sufficient amount of data instances to capture data variability. Furthermore, the variability of conditional coverage follows a Beta distribution with larger calibration sets reducing the fluctuations in coverage, improving reliability (Angelopoulos et al., 2023). The effects of varying the calibration set size and impacts on quality metrics for prediction intervals are examined in Section 5.2.

In summary, SCP leverages empirical data from the calibration set to adaptively size prediction intervals. This approach ensures that the intervals are statistically valid and practically useful, providing a significant advancement in predictive accuracy and reliability across various applications. The method's efficacy is rooted in its structured approach to data partitioning and utilization of the empirical distribution of prediction errors to form robust prediction intervals.

3.3. Permutation feature importance with uncertainty quantification measures

PFI is a widely used technique to assess the importance of individual features in a predictive model (Fisher et al., 2019). It operates by evaluating the effect of randomly shuffling the values of each feature on the model's performance, thereby disrupting the relationship between the feature and the target variable. This method provides insights into which features are most crucial for making accurate predictions. When combined with UQ measures, PFI can offer a nuanced understanding of how features contribute not only to the predictions themselves but also to the confidence in these predictions. This section describes the mathematical formalization of PFI adapted for SCP and UQ measures, exploring three distinct scenarios of feature permutation. These scenarios include: (1) shuffling features only in the test data, (2) shuffling features only in the calibration data, and (3) shuffling features in both calibration and test data. Each scenario is assessed using a generic UQ metric, M, to evaluate the impact of feature permutation on prediction intervals and their quality. UQ measures employed for the underlying dataset are introduced in Section 4.3.

3.3.1. Baseline calculation

The initial step in our methodology involves computing a baseline measure of uncertainty using the test dataset. This baseline serves as a reference point for evaluating the impact of feature permutations on the model's predictive performance and uncertainty estimates. We define the baseline uncertainty measure M_{base} as follows:

$$M_{\text{base}} = M\left(\left\{ \left(\left[\hat{f}(X_i) - \hat{q}, \ \hat{f}(X_i) + \hat{q} \right], \ Y_i \right) \mid i \in I_{\text{test}} \right\}, \ \alpha \right), \tag{5}$$

where $\hat{f}$ is the trained regression model, $\hat{q}$ is the quantile calculated from the calibration set as per the SCP method, and M represents an UQ metric. UQ measures include PICP, MPIW, MRPIW, and the Winkler score, which are introduced in Section 4.3. This baseline measure M_{base} provides insight into the original model's performance and the inherent uncertainty in its predictions without any feature perturbation. It establishes a benchmark against which the effects of permuting individual features can be compared in subsequent analyses.

3.3.2. Scenario 1: Shuffling features only in the test data

In this scenario, we investigate the effect of permuting individual features solely in the test dataset while keeping the training and calibration datasets unchanged. The goal is to assess how sensitive the model's predictions and uncertainty estimates are to each feature when making predictions on new, unseen data. This approach helps us understand the importance of each feature in maintaining the predictive performance and reliability of the model in real-world applications.

For each feature $j \in \{1, \dots, d\}$, we perform the following steps:

1. **Permute the j-th feature in the test set:**
   We create a permuted version of the test dataset, denoted as D^j_{test}, by randomly shuffling the values of the j-th feature among the test samples. Formally, for each test sample $i \in I_{\text{test}}$, we construct a new feature vector X^j_i:

   $$X^j_i = \left(X_{i1}, \dots, X'_{ij}, \dots, X_{id} \right),$$

   where X'_{ij} is the permuted value of the j-th feature for sample i, obtained by randomly reordering the j-th feature values across all test samples. The corresponding permuted test dataset is then:

$$D^j_{\text{test}} = \left\{ (X^j_i, Y_i) \mid i \in I_{\text{test}} \right\}.$$

This permutation breaks the association between the j-th feature and the target variable Y in the test data, simulating the absence of any predictive information from that feature in the test context.

2. **Recompute predictions and prediction intervals for the permuted test data:**
Using the trained regression model $\hat{f}$ (obtained from the training data), we compute new predictions for the permuted test samples:

$$\hat{f}(X^j_i), \quad \forall i \in I_{\text{test}}.$$

We then construct the prediction intervals for each permuted test sample using the quantile $\hat{q}$ computed from the calibration set (which remains unchanged):

$$\hat{C}(X^j_i) = \left[\hat{f}(X^j_i) - \hat{q},\ \hat{f}(X^j_i) + \hat{q} \right], \quad \forall i \in I_{\text{test}}.$$

By recalculating the predictions and intervals with the permuted feature, we can observe how the disruption of the feature's information affects the model's output and uncertainty estimates.

3. **Compute the UQ measure for the permuted test data:**
We evaluate the chosen UQ metric M on the set of prediction intervals and true responses from the permuted test data:

$$M_{\text{perm.test}_j} = M\left(\left\{ \left(\hat{C}(X^j_i),\ Y_i \right) \mid i \in I_{\text{test}} \right\},\ \alpha \right).$$

4. **Compute the feature importance score:**
We quantify the importance of feature j by comparing the UQ measure after permutation to the baseline measure computed without permutation:

$$FI^{\text{perm.test}}_j = M_{\text{perm.test}_j} - M_{\text{base}}.$$

A significant change in the UQ measure indicates that permuting feature j has a substantial impact on the model's predictive uncertainty, suggesting that the feature is important for accurate and reliable predictions on new data.

By shuffling each feature individually in the test data and observing the resulting changes in the UQ measures, the importance scores can be interpreted as follows: A positive $FI^{\text{perm.test}}_j$ indicates that permuting feature j leads to an increase in the UQ measure – such as a decrease in coverage probability or an increase in interval width – implying that the model's uncertainty worsens. This suggests that feature j is critical for making precise and confident predictions. Conversely, a negative or zero $FI^{\text{perm.test}}_j$ implies that the UQ measure remains the same or improves after permutation, indicating that feature j may lack informativeness for the model's predictions on the test data or that the model is robust to variations in that feature.

This scenario specifically examines the effect of feature importance in the context of unseen data, which is critical for evaluating how the model might perform in real-world applications where data distributions may vary. By focusing on the test data, we isolate the impact of each feature on the model's ability to generalize and maintain reliable uncertainty estimates when faced with new instances. It is important to note that the training and calibration datasets remain untouched in this scenario. This means that the model $\hat{f}$ and the quantile $\hat{q}$ are consistent across all permutations, ensuring that any changes in the UQ measures are solely due to the permutation of the test features.

3.3.3. Scenario 2: Shuffling features only in the calibration data

In this scenario, we explore the effect of permuting individual features solely within the calibration dataset while keeping the training and test datasets unchanged. The objective is to assess how sensitive the model's uncertainty estimates are to each feature during the calibration process, which directly influences the construction of prediction intervals. This approach helps us understand the importance of each feature in maintaining the reliability and validity of the model's UQ when making predictions on new, unseen data.

For each feature $j \in \{1, \dots, d\}$, we perform the following steps:

1. **Permute the j-th feature in the calibration set:**
We create a permuted version of the calibration dataset, denoted as D^j_{cal}, by randomly shuffling the values of the j-th feature among the calibration samples. Formally, for each calibration sample $i \in I_{\text{cal}}$, we construct a new feature vector X^j_i:

$$X^j_i = \left(X_{i1}, \dots, X'_{ij}, \dots, X_{id} \right),$$

where X'_{ij} is the permuted value of the j-th feature for sample i, obtained by randomly reordering the j-th feature values across all calibration samples. The corresponding permuted calibration dataset is then:

$$D^j_{\text{cal}} = \left\{ (X^j_i, Y_i) \mid i \in I_{\text{cal}} \right\}.$$

This permutation breaks the association between the j-th feature and the target variable Y in the calibration data, simulating the absence of any informative contribution from that feature during the calibration process.

2. **Recompute the non-conformity scores and recalibrate the quantile:**
Using the trained regression model $\hat{f}$ (obtained from the training data), we compute new non-conformity scores for the permuted calibration samples:

$$R^j_i = \left| Y_i - \hat{f}(X^j_i) \right|, \quad \forall i \in I_{\text{cal}}.$$

We then recalibrate the quantile $\hat{q}^j$ based on these new non-conformity scores:

$$\hat{q}^j = \text{Quantile}\left(\left\{ R^j_i \,:\, i \in I_{\text{cal}} \right\},\ \frac{\lceil (1-\alpha)(|I_{\text{cal}}| + 1) \rceil}{|I_{\text{cal}}|} \right).$$

This recalibration adjusts the width of the prediction intervals to reflect the impact of the permuted feature on the model's prediction errors within the calibration set.

3. **Compute new prediction intervals for the test data:**
Using the recalibrated quantile $\hat{q}^j$, we construct prediction intervals for each test sample $(X_i, Y_i) \in D_{\text{test}}$:

$$\hat{C}^j(X_i) = \left[\hat{f}(X_i) - \hat{q}^j,\ \hat{f}(X_i) + \hat{q}^j \right], \quad \forall i \in I_{\text{test}}.$$

Note that we use the original test features X_i since the test data remains unchanged. The recalibrated prediction intervals reflect how the uncertainty estimates are affected by the permutation of the feature in the calibration data.

4. **Compute the UQ measure for the test data:**
We evaluate the chosen UQ metric M on the set of prediction intervals and true responses from the test data:

$$M_{\text{perm.cal}_j} = M\left(\left\{ \left(\hat{C}^j(X_i),\ Y_i \right) \,:\, i \in I_{\text{test}} \right\},\ \alpha \right).$$

5. **Compute the feature importance score:**
We quantify the importance of feature j by comparing the UQ measure after permutation to the baseline measure computed without permutation. Both measures are evaluated on the test data to assess the impact on the model's predictions for new, unseen instances:

$$FI^{\text{perm.cal}}_j = M_{\text{perm.cal}_j} - M_{\text{base}}.$$

Here, M_{base} is the baseline UQ measure computed on the test data using the original prediction intervals constructed with the original quantile $\hat{q}$.

By shuffling each feature individually in the calibration data and observing the resulting changes in the UQ measures on the test data, the importance scores can be interpreted as follows: A positive $FI^{\text{perm.cal}}_j$

indicates that permuting feature j leads to an increase in the UQ measure – such as a decrease in coverage probability or an increase in interval width – implying that the model's uncertainty estimation has worsened. This suggests that feature j is critical for calibrating precise and confident prediction intervals. Conversely, a negative or zero $FI_j^{\text{perm.cal}}$ implies that the UQ measure remains the same or improves after permutation, indicating that feature j may not significantly influence the model's uncertainty estimates or that the calibration process is robust to changes in that feature.

This scenario specifically examines the role of each feature in the calibration process, which is crucial for ensuring the validity of the uncertainty quantification in the model's predictions. By focusing on the calibration data and evaluating the impact on the test data, we isolate the effect of each feature on the construction of prediction intervals and how they generalize to new, unseen data. It is important to note that the training and test datasets remain untouched in this scenario. This means that the model $\hat{f}$ and the test inputs X_i are consistent across all permutations, ensuring that any changes in the UQ measures are solely due to the permutation of the calibration features. In summary, Scenario 2 helps identify features that are essential for the reliability of the model's uncertainty estimates. By understanding which features significantly affect the calibration of prediction intervals, practitioners can gain insights into the variables that contribute most to the model's confidence in its predictions.

3.3.4. Scenario 3: Shuffling features in both calibration and test data

In this scenario, we investigate the effect of permuting individual features in both the calibration and test datasets while keeping the training dataset unchanged. The goal is to assess how sensitive the model's predictions and uncertainty estimates are to each feature when both the calibration process and the test data are affected by feature perturbations.

For each feature $j \in \{1, \dots, d\}$, we perform the following steps:

1. **Permute the j-th feature in the calibration set:**
   Create a permuted version of the calibration dataset, denoted as D_{cal}^j, by randomly shuffling the values of the j-th feature among the calibration samples. For each calibration sample $i \in I_{\text{cal}}$, the new feature vector $X_i^{\text{cal},j}$ is:
   $$X_i^{\text{cal},j} = \left(X_{i1}, \dots, X'_{ij}, \dots, X_{id}\right),$$
   where X'_{ij} is the permuted value of the j-th feature for sample i in the calibration set. The permuted calibration dataset is:
   $$D_{\text{cal}}^j = \left\{ \left(X_i^{\text{cal},j},\ Y_i\right) \mid i \in I_{\text{cal}} \right\}.$$
2. **Recompute the non-conformity scores and recalibrate the quantile:**
   Using the trained regression model $\hat{f}$, compute the non-conformity scores for the permuted calibration data:
   $$R_i^j = \left| Y_i - \hat{f}\left(X_i^{\text{cal},j}\right) \right|, \quad \forall i \in I_{\text{cal}}.$$
   Recalculate the quantile $\hat{q}^j$ based on these non-conformity scores:
   $$\hat{q}^j = \text{Quantile}\left(\left\{ R_i^j \mid i \in I_{\text{cal}} \right\},\ \frac{\lceil (1-\alpha)(|I_{\text{cal}}|+1) \rceil}{|I_{\text{cal}}|} \right).$$
3. **Permute the j-th feature in the test set:**
   Create a permuted version of the test dataset, denoted as D_{test}^j, by randomly shuffling the j-th feature among the test samples. For each test sample $i \in I_{\text{test}}$, the new feature vector $X_i^{\text{test},j}$ is:
   $$X_i^{\text{test},j} = \left(X_{i1}, \dots, X''_{ij}, \dots, X_{id}\right),$$
   where X''_{ij} is the permuted value of the j-th feature for sample i in the test set. The permuted test dataset is:
   $$D_{\text{test}}^j = \left\{ \left(X_i^{\text{test},j},\ Y_i\right) \mid i \in I_{\text{test}} \right\}.$$
4. **Compute new prediction intervals for the permuted test data:**
   Using the recalibrated quantile $\hat{q}^j$, construct prediction intervals for each permuted test sample:
   $$\hat{C}^j\left(X_i^{\text{test},j}\right) = \left[\hat{f}\left(X_i^{\text{test},j}\right) - \hat{q}^j,\ \hat{f}\left(X_i^{\text{test},j}\right) + \hat{q}^j\right], \quad \forall i \in I_{\text{test}}.$$
5. **Compute the uncertainty quantification (UQ) measure for the permuted test data:**
   Evaluate the chosen UQ metric M on the prediction intervals and true responses from the permuted test data:
   $$M_{\text{perm.cal.test}_j} = M\left(\left\{ \left(\hat{C}^j\left(X_i^{\text{test},j}\right),\ Y_i \right) \mid i \in I_{\text{test}} \right\},\ \alpha \right).$$
6. **Compute the feature importance score:**
   Quantify the importance of feature j by comparing the UQ measure after permutation to the baseline measure computed without permutation:
   $$FI_j^{\text{perm.cal.test}} = M_{\text{perm.cal.test}_j} - M_{\text{base}},$$

A positive $FI_j^{\text{perm.cal.test}}$ indicates that permuting feature j leads to an increase in the UQ measure – such as a decrease in coverage probability or an increase in interval width – implying that the model's uncertainty estimation has worsened. This suggests that feature j is critical for both calibrating precise prediction intervals and making confident predictions on new data. Conversely, a negative or zero $FI_j^{\text{perm.cal.test}}$ implies that the UQ measure remains the same or improves after permutation, indicating that feature j may not significantly influence the model's uncertainty estimates or that the model is robust to changes in that feature. This scenario examines the compounded effect of feature perturbations on both the calibration process and the test data. By considering permutations in both datasets, we assess the overall importance of each feature in the end-to-end predictive performance and uncertainty estimation of the model. This comprehensive analysis helps identify features that are pivotal in maintaining robust prediction intervals.

4. Experiment settings

This section provides a detailed description of the dataset and use case, hyperparameter optimization settings, and evaluation metrics used in our study. It offers insights into the methodology and performance assessment of the ML models employed. These elements ensure that the prediction capabilities of the final ML model are sufficient to reliably measure the effects of feature permutation on the model's UQ. For reproducibility, Appendix A presents experiment settings and results for the open source Production Analysis dataset (Levy, 2014), a similar use case from the field of manufacturing.

4.1. Dataset and use case overview

The used dataset stems from a collaborative research project with a medium-sized German manufacturer specializing in custom and standardized vessel components, consisting of relevant process data regarding the planning and execution of manufacturing tasks (Mehdiyev et al., 2024a, 2023, 2024b). Customer orders, derived from the partner's product catalog, initiate the manufacturing sequence. Each order is assessed for priority, and the sequence of manufacturing steps is determined based on product specifications. These specifications include attributes such as article group identifier, material group identifier, weight, quantity, and other product-specific features. To address the challenge of estimating processing times, which currently relies on expert intuition, a solution leveraging MES has been implemented to capture precise execution details. The manufacturing steps are recorded as events in a structured dataset, with each row containing event- and trace-level information pertaining to the characteristics of the production process. Each event is linked to a specific activity, machine,

Table 1
Extract from the utilized event log data as in Mehdiyev et al. (2024a).

Job ID	Activity	Start Time	End Time	Diameter Base	...	Resource ID	...
162384	Plasma Welding	2019–04–18 06:26:47	2019-04-18 09:51:25	1800	...	409	
162384	Grinding Weld. Seam	2019–04–18 12:11:30	2019-04-18 19:07:14	1800	...	108	
162384	Dishing Press (2)	2019–04–23 10:50:31	2019-04-23 18:34:11	1800	...	150	
162384	Bead Small	2019–04–24 10:20:13	2019-04-24 19:57:45	1800	...	726	
162384	X-ray Examination	2019–04–25 10:26:23	2019-04-25 10:26:32	1800	...	703	
162384	Edge Deburring	2019–04–26 09:08:38	2019-04-26 17:50:27	1800	...	742	
...	...	...	...	...	...	...	...
177566	3D Micro-step Circle	2021–06–21 07:04:38	2021-06-21 10:26:37	3680	... ...	139	... ...
...	...	...	...	...	...	...	...

and qualified worker executing the process step. An excerpt from the dataset is presented in Table 1.

Preprocessing and feature engineering were conducted similarly to Mehdiyev et al. (2024a), although a more relaxed outlier filtering was applied compared to the referenced study. Such an outlier filtering allows for the dataset to contain more aleatoric uncertainty which, in turn, is expected to impact the ML model's performance regarding UQ. Feature engineering was then employed to identify previous activities, and their processing times and integrate statistical data of the current activity.

During iterative feature selection, conducted alongside model training, variables with relative feature importance below 0.05% were categorically excluded, as they were deemed to contribute negligible predictive value. Additionally, variables containing sensitive information, such as data allowing for the personal identification of workers, were omitted to address ethical considerations and safeguard individual privacy. Since the employed data contains only variables with domain-specific relevance regarding the examined production process, all remaining variables were included for the final model training and evaluations.

For hyperparameter tuning, model training, and the calibration of the CP method, the dataset was partitioned into training, calibration, and test sets in an 8:1:1 ratio, resulting in 132,982 events for training, 16,217 events for calibration, and 16,234 events for testing. The splitting was conducted in chronological order, ensuring that only complete traces were allocated to each dataset, thereby maintaining the integrity of each dataset. The product-specific features are detailed in Table 2 and process-specific features in Table 3, while comprehensive information about the datasets is provided in Table 4.

The proposed approach was implemented in *R*, encompassing data processing, hyperparameter optimization, model training, PFI, and result visualization. The *tidyverse* library suite was predominantly utilized for data cleaning, hyperparameter tuning, and model training activities. Visualization tasks were performed using the *ggplot2* library. All computational tasks were executed on a 64-bit system equipped with a 13th Gen Intel(R) Core(TM) i7-13700F processor, running at a clock speed of 2,100 MHz, with 16 cores and 32 GB of RAM.

Table 2
Description of product-specific features.

Feature Name	Abbreviation	Description
Article	ART	The specific article or product being manufactured.
Bend Radius L	BRL	The large bend radius of the sheet metal used in the product.
Bend Radius S	BRS	The small bend radius of the sheet metal used in the product.
Diameter Base	DBA	The diameter of the base component used in the product.
Diameter Circle	DCI	The diameter of the circular component used in the product.
Material	MAT	The material used to manufacture the product.
Sheet Width	SWI	The width of the sheet metal used in the product.
Weight	WGT	The weight of the sheet metal used in the product.

4.2. Machine learning settings

Hyperparameter optimization is a critical process in ML aimed at improving ML model performance by identifying the best set of hyperparameters. Unlike model parameters, which are learned during the training process, hyperparameters are set before training and govern the learning process itself. The objective of hyperparameter tuning is to find the optimal hyperparameters that minimize the loss function, often using techniques such as grid search or random search.

In this study, we conduct hyperparameter optimization for four distinct models: generalized linear model (GLM) employing glmnet (Hastie et al., 2021), decision tree (DT), random forest (RF), and gradient boosting machine (GBM) employing XGBoost (Chen and Guestrin, 2016). The settings for hyperparameter optimization for each model are detailed in Table 5. We employ a grid search via Latin hypercube sampling with a sample size of 30 to comprehensively explore the hyperparameter space, with each model being evaluated using 10-fold cross-validation on the training dataset and the performance being assessed based on mean absolute error (MAE) and root mean squared error (RMSE) (see Section 4.3). By employing this approach, we aim to ensure robust hyperparameter tuning and optimal model performance. The specific settings for hyperparameter optimization for each model are detailed in Table 5.

4.3. Evaluation metrics

This section outlines evaluation methodologies appropriate for assessing the predictive quality of the employed ML model. We focus on evaluation metrics concerning UQ, particularly regarding the quality of prediction intervals. We use PICP, MPIW, MRPIW, and the mean Winkler score to assess this aspect.

Table 3
Description of process-specific features.

Feature Name	Abbreviation	Description
Activity	ACT	The specific activity being performed, including machine equipment type, in the event.
Mean Processing Time	MPT*	The average time expected to complete the current activity. Engineered using historical averages across the corresponding activity of the produced article.
Mean Total Processing Time	MTT*	The average time expected to complete all activities of the current type within trace.
Mean Trace Processing Time	TTT*	The average processing time for the entire event trace. Engineered using historical data.
Planned Processing Steps	PPS	The total number of processing steps planned for the event.
Previous Processing Time	PPT*	The time taken for the previous processing step. Engineered based on historical event logs.
Processing Step Number	PSN	The specific number assigned to each processing step in the event.
Quantity	QU	The quantity of items currently produced.
Total Quantity	TQU	The quantity of items to be produced in total.
Worker	WOR	The worker assigned to the activity in the event.

Table 4
Overview of the employed event log dataset.

	Training	Calibration	Test	Complete Dataset
Number of events	132,982 (80.4%)	16,217 (9.8%)	16,234 (9.8%)	165,433
Number of cases	26,264 (80%)	3,283 (10%)	3,283 (10%)	32,831
Unique activities	32	27	27	32
Mean processing time (min)	102.10	89.55	100.86	100.74
Std. deviation of processing time (min)	118.09	107.90	117.00	117.08
Mean trace length	5.06	4.94	4.94	5.04
Std. deviation of trace length	3.89	4.20	4.39	3.98

PICP measures the proportion of target values from a dataset of length N falling within the prediction intervals and is calculated as follows:

$$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} pic_i, \quad pic_i = \begin{cases} 1, & Y_i \in \left[L_i, U_i\right] \\ 0, & Y_i \notin \left[L_i, U_i\right], \end{cases} \tag{6}$$

with L_i, U_i respectively representing the lower and upper boundaries $\hat{f}(X_i) - \hat{q}$ and $\hat{f}(X_i) + \hat{q}$ of the prediction interval as defined in Eq. (3).

MPIW evaluates the average width of these prediction intervals:

$$MPIW = \frac{1}{N}\sum_{i=1}^{N}(U_i - L_i) \tag{7}$$

To account for the relative scale of prediction intervals, MRPIW normalizes the interval width by the corresponding point predictions:

$$MRPIW = \frac{1}{N}\sum_{i=1}^{N} rWidth_i, \quad rWidth_i = \frac{(U_i - L_i)}{\hat{f}(X_i)} \tag{8}$$

Table 5
Hyperparameter optimization settings for DT, GBM, GLM and RF models.

Model	Hyperparameter	Search Interval
DT	cost_complexity min_n[c] tree_depth	$(0; 0.1)$ $[2; 40]$ $[1; 15]$
GBM	trees mtry min_n tree_depth learn_rate loss_reduction[e] sample_size[f]	$[1; 2000]$ $[1; 18]$ $[2; 40]$ $[1; 15]$ $(0; 0.1)$ $(1e-10; +\infty)$ $[0.1; 1]$
GLM	mixture[a] penalty[b]	$[0; 1]$ $(0; 1)$
RF	min_n mtry[d] trees	$[2; 40]$ $[1; 18]$ $[1; 2000]$

[a] Proportion of pure lasso to ridge regression penalty.
[b] Amount of regularization.
[c] Minimum number of data points in a node for splitting.
[d] Number of predictors randomly sampled at each split.
[e] Reduction in loss function required to split further.
[f] Data set size used for modeling within an iteration of the modeling algorithm.

The MRPIW sets the absolute prediction interval width in the context of the point prediction, capturing the relative proportion of intervals regarding the point forecast.

Additionally, we employ the mean of the Winkler score across single predictions to measure the accuracy and informativeness of prediction intervals, which is calculated as follows:

$$W = \frac{1}{N}\sum_{i=1}^{N} W_i, \tag{9}$$

$$W_i = \begin{cases} \left(U_i - L_i\right) + \frac{2}{\alpha}\left(L_i - Y_i\right) & \text{if } Y_i < L_i \\ \left(U_i - L_i\right) & \text{if } L_i \leq Y_i \leq U_i \\ \left(U_i - L_i\right) + \frac{2}{\alpha}\left(Y_i - U_i\right) & \text{if } Y_i > U_i \end{cases}$$

The Winkler score is useful for UQ as it not only measures the width of the prediction intervals but also penalizes intervals that fail to cover the true values, thereby providing a balanced assessment of both accuracy and informativeness. When aiming at high coverage of ground truths, the alpha levels in the penalty term of the Winkler calculation decrease. However, this also means that the penalty for intervals failing to cover the true values becomes more significant. Consequently, the Winkler score adjusts to discourage overly broad intervals that do not effectively encapsulate the true values, maintaining a critical balance between interval width and coverage accuracy. This property ensures that the prediction model is not only broadening its intervals indiscriminately but is also mindful of maintaining coverage, thus offering a nuanced and robust measure of predictive performance.

5. Results

In this section, we examine the results of hyperparameter optimization and UQ evaluation. First, the optimal hyperparameter settings and performance outcomes for the underlying dataset and use case are presented. Next, the performance regarding point prediction of the optimized ML models was assessed on the test data. The model exhibiting the highest performance was evaluated regarding UQ, providing a comprehensive evaluation of the model's predictive capabilities. This established the necessary foundation for the examination of PFI results, which concludes this section.

Table 6
Optimal settings for GLM, DT, RF and GBM models yielded by hyperparameter tuning and evaluated via 10-fold cross-validation on the training data.

Model	Hyperparameter	Settings	MAE[a]	$RMSE$[a]
GLM	mixture	0.13	46.8	78.4
	penalty	0.846		
DT	cost_complexity	1.3e−07	41.5	72.9
	min_n	35		
	tree_depth	10		
RF	min_n	36	39.5	65.7
	mtry	18		
	trees	979		
GBM	trees	1622	38.2	66.4
	mtry	12		
	min_n	25		
	tree_depth	8		
	learn_rate	0.0356		
	loss_reduction	1.05e−10		
	sample_size	0.864		

[a] In minutes.

Table 7
Evaluation of fitted models on test data.

Model	MAE	$RMSE$
GLM	47.1	77.4
DT	43.6	73.3
RF	39.1	64.9
GBM	37.2	62.1

5.1. Hyperparameter optimization and model evaluation

Hyperparameter tuning was conducted as described in Section 4.2 and evaluated using 10-fold cross-validation on the training data, with a final evaluation on the test data. First, the optimal settings as well as performance results from the 10-fold cross-validation on the training data are being examined (see Table 6).

Both RF and GBM models demonstrate strong predictive capabilities, as evidenced by their performance metrics, followed by the DT model and lastly GLM. The RF model achieved an RMSE of 65.7, slightly lower than the GBM's 66.4, indicating marginally better performance in terms of overall prediction error. However, the GBM model's MAE of 38.2, compared to RF's 39.5, suggests that GBM offers more precise predictions on average. Considering both MAE and RMSE, the GBM model exhibits superior average prediction accuracy, while the RF model demonstrates slightly better performance in minimizing overall prediction error variance. For a comprehensive evaluation, each of the models was trained using the full training dataset, and their performance was examined on the test data (see Table 7).

Regarding computational complexity, a comparative analysis of the computation times for the hyperparameter tuning was conducted, utilizing the parallel processing capabilities of the setup described in Section 4: DT and GLM models yielded the shortest computation times with 5.4 min and 10.2 min respectively, followed by the RF model with 83.0 min and the GBM model with 116 min. This analysis excluded the computation time for the SCP calibration, considering its linear complexity regarding calibration set sizes, its independence from computational costs of the employed point prediction models, and negligible computation times (approximately 2 s) for the maximum calibration set size of 16,217 instances.

When evaluated on the test data, the RF model reported an MAE of 39.1 and an RMSE of 64.9, showing consistent performance across both training and test datasets. This indicates that the RF model generalizes well. However, the GBM model slightly outperformed the RF model, with an MAE of 37.2 and an RMSE of 62.1. Both black-box models showed increased performance on unseen data compared to their transparent counterparts, DT and GLM. Therefore, the GBM model is selected for further analysis and application in this study due to its robustness and higher accuracy in capturing complex data patterns. This selection ensures that the study leverages the model with the best predictive capability, enhancing the reliability of the results.

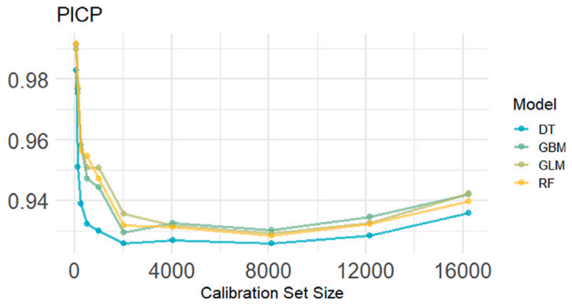


Fig. 2. Evaluation of the fitted models regarding the sensitivity of uncertainty quantification for an α value of 5% on test data across varying sizes of calibration data.

5.2. Uncertainty quantification

To analyze model uncertainty as well as sensitivity, we applied SCP to the fitted models, using unseen data from the calibration dataset to fine-tune prediction intervals and perform analyses on the test data. Figs. 2–5 present the results of the sensitivity analysis for the fitted models at an $alpha$ value of 5% across varying volumes of calibration data regarding the PICP, MPIW, MRPIW, and mean Winkler score metrics respectively. Particularly, the analysis was conducted on calibration set sizes equivalent to 100%, 75%, and 50% of the maximum calibration set size of 16,217 instances, with further evaluation conducted by iteratively halving the set size down to 63 data instances, which is approximately 0.391% of the maximum set size.

While GBM and RF generally outperform the DT and GLM models, all four models demonstrate clear performance gains, evident through high coverage at reduced prediction interval widths and more stable mean Winkler scores when larger calibration datasets (above 12,000 data instances) are employed. When calibration data are extremely limited (below 1,000 data instances), PICP values return to levels comparable to those observed with ample calibration data, suggesting that even small datasets can achieve the same nominal coverage. However, this improvement in coverage arises at the expense of broader intervals, reflected in elevated MPIW and MRPIW scores, which indicates a trade-off between coverage and precision. The trends observed across these models highlight the advantages of using larger calibration datasets to capture the inherent variability of the data more effectively and, consequently, to maintain more precise, and therefore more practical, prediction intervals. For further examination, Appendix B provides Figs. 26–41 for a detailed examination of the sensitivity analysis results across the utilized models, α values and calibration set sizes.

The evaluation of the fitted GBM model on the test data is summarized in Table 8, detailing MPIW, MRPIW, PICP, and mean Winkler score for different significance levels α used for the construction of prediction intervals. These results emphasize the inherent connection of PICP to the value of α, since the width of prediction intervals is calibrated to achieve the desired coverage. A lower α means that the intervals are designed to be more inclusive, reducing the risk of missing the true values. However, this inclusivity comes at the cost of interval width, making the intervals potentially less useful in practical scenarios where narrower intervals are preferred. As α increases, the interval width, and thus the PICP, is expected to decrease, reflecting a relaxation of the desired coverage.

At $\alpha = 5\%$, the PICP is 0.94, indicating that 94% of the true values are within the prediction intervals. This high coverage ensures that the model captures the most true values but results in wide intervals. As α

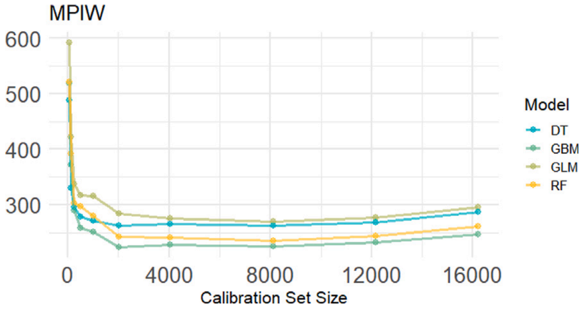


Fig. 3. Evaluation of the fitted models regarding the sensitivity of uncertainty quantification for an α value of 5% on test data across varying sizes of calibration data.

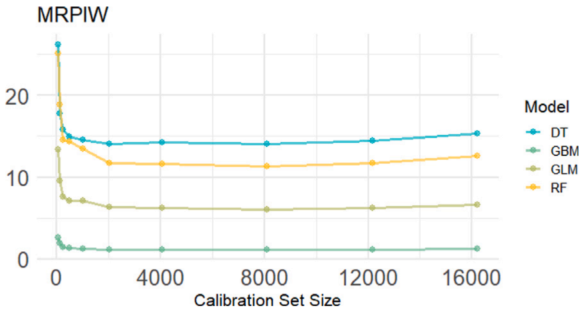


Fig. 4. Evaluation of the fitted models regarding the sensitivity of uncertainty quantification for an α value of 5% on test data across varying sizes of calibration data.

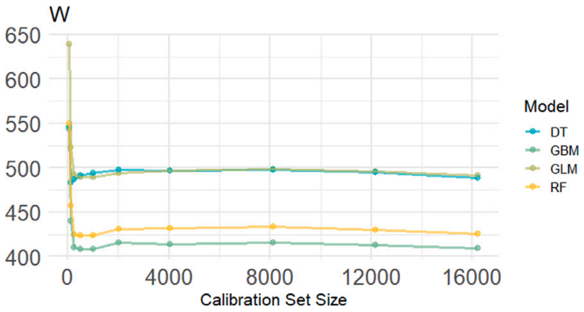


Fig. 5. Evaluation of the fitted models regarding the sensitivity of uncertainty quantification for an α value of 5% on test data across varying sizes of calibration data.

increases to 20%, the PICP decreases to 0.77. The reduction in coverage probability signifies that fewer true values fall within the narrower intervals, indicating the expected trade-off between significance and coverage levels.

The MPIW increase from 93.4 at α = 20% to 247.3 at α = 5%. The decrease in MPIW is particularly pronounced between α = 5% and α = 10%, with a reduction of 85 min. This significant reduction suggests that initially increasing α leads to substantially narrower intervals. As α increases further, the rate of decrease in MPIW continues but at a slower pace, indicating that initial adjustments in α have a more pronounced impact on interval width.

Similarly, MRPIW, which normalizes the interval width relative to the point prediction, increases from 0.48 at α = 20% to 1.28 at α = 5%, with the largest drop in MRPIW occurring between α = 5% and 10%, decreasing the mean relative width by 44%. This trend indicates diminishing returns in the quality of prediction intervals for increases in the desired coverage of ground truths.

The mean Winkler score, which combines both interval width and coverage, increases from 227.4 at α = 20% to 409.3 at α = 5%. Although lower α values result in wider intervals, as indicated by the MPIW and MRPIW metrics, the W tends to increase with higher PICP levels. This increase occurs due to the balance between interval width and the increased penalty for non-coverage. Due to the nature of calculating the Winkler score, increasing coverage by decreasing α levels results in more ground truth values being captured by wider prediction intervals, thus yielding minimal Winkler penalties. However, for instances that were not covered by the prediction intervals, the decrease in α levels results in the penalty term of the Winkler score calculation increasing proportionally. Consequently, uncovered ground truths yield higher penalties for lower α levels, which cannot be compensated by the increase in coverage, resulting in the exhibited trend for the mean Winkler scores.

Table 8
Evaluation of the final GBM model regarding uncertainty quantification on test data.

α-Value	$PICP$	$MPIW$	$MRPIW$	W
5%	0.94	247.3	1.28	409.3
10%	0.88	162.3	0.84	314.9
15%	0.82	119.2	0.62	262.9
20%	0.77	93.4	0.48	227.4

5.3. Uncertainty-related permutation feature importance

The following subsections describe the results of three approaches towards the examination and interpretation of PFI at the calibration and testing stages of UQ evaluation. The analysis commences with the assessment of feature importance regarding UQ metrics at the testing stage, which builds the foundation for the interpretation of subsequent results. Next, feature importance is assessed at the calibration stage and the magnitude of disruption regarding model uncertainty is captured and examined. Lastly, implications of concurrently permuting feature values in the calibration and the test data are investigated in light of the preceding results.

5.3.1. Permutation feature importance on test data

To understand the influence of individual features on a calibrated ML model, we conduct a PFI analysis regarding PICP, MRPIW, and mean Winkler score. The SPC method, which calibrates prediction interval widths for future predictions based on the desired significance level α, inherently keeps the MPIW metric constant. Consequently, we expect the changes in the model's point predictions to subsequently affect UQ metrics, allowing for the examination of the feature importance for the calibrated model.

Figs. 6 (a), (b), (c) and (d) respectively illustrate the PFI results regarding the PICP metric across four different α-levels: 5%, 10%, 15%, and 20%. Each plot segmented the features (see Table 2 and Table 3) into two categories: product-specific and process-specific features. The importance of a feature is denoted by its mean deviation from the baseline (see Table 8) across ten iterations of PFI measurements, and is indicated by an orange bar for negative and a blue bar for positive values.

Across all α-levels, the ranking of most influential features within each category remains consistent: For process-specific features, *MPT** consistently ranks highest, followed by *ACT* and *QU*. This suggests that these features are pivotal in influencing the PICP metric, regardless of the significance level. *DCI* maintains a dominant position across all significance levels, albeit with a minimal absolute deviation from the baseline. Other features like *WGT* and *MAT* show some importance but to a lesser extent compared to *DCI*. The influence of product-specific features on the PICP metric is considerably less pronounced than that of process-specific features, with the engineered feature *MPT** spearheading the importance ranking, suggesting that process-specific features have a more substantial impact on the model predictions.

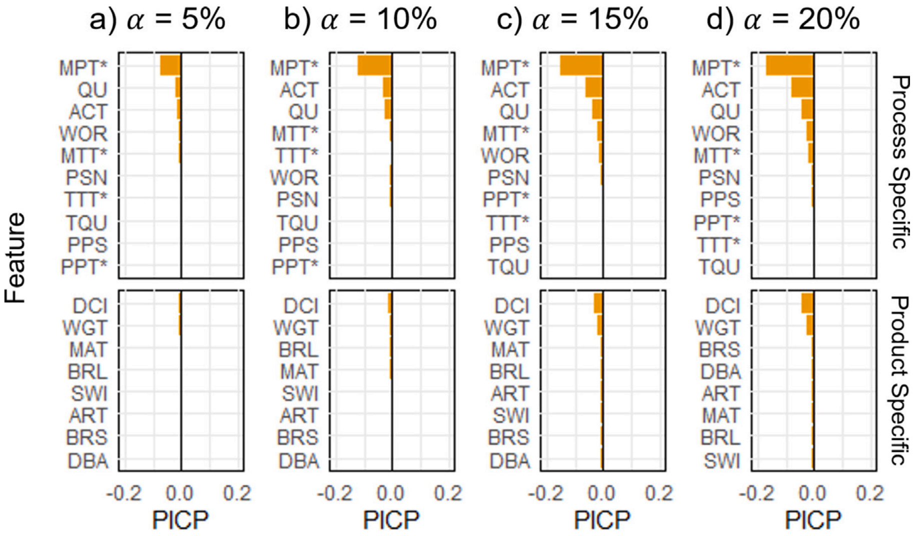


Fig. 6. PICP-related permutation feature importance for the GBM model for permuted test data.

The absolute deviation from the baseline for the most important features diminishes slightly as α levels decrease. Considering that the corresponding baselines increase with decreasing significance levels, this trend indicates that calibration on lower α levels yields increased robustness regarding the coverage of ground truths. This observation can be attributed to higher α levels allowing for narrower prediction interval widths which are more sensitive to changes in the point prediction and, subsequently, perturbances of the input data.

The order of features within each category remains relatively stable across varying α-levels, indicating that the relative importance of features does not drastically change with different levels of significance. Many features across all categories exhibit negligible importance, regardless of α-level. This suggests that only a few key features predominantly influence the PICP metric, while the rest have a minimal impact. These findings highlight the robustness and stability of certain key features across different significance levels, emphasizing their marginal effect on the model prediction.

Regarding the MRPIW metric, Figs. 7 (a), (b), (c) and (d) depict feature importance in similar fashion as for PICP. Across all feature categories, deviations from the baselines exhibit a more balanced distribution when compared to the PICP, with deviations decreasing their magnitude with higher α values.

Across all significance levels, the ranking of the most influential features within the process-specific category remains consistent. *WOR* and *MTT** consistently rank highest, followed by *ACT* and *TQU*, influencing the MRPIW metric regardless of the significance level. For product-specific features, *WGT* maintains a dominant position across all $\alpha \leq 15\%$. Other features like *MAT* and *SWI* also show some importance but to a lesser extent compared to *WGT*. The influence of product-specific features on the MRPIW metric is slightly less pronounced than that of process-specific features, especially with decreasing significance levels.

In the context of SCP, the prediction interval width is determined during the calibration phase and remains constant regardless of model uncertainty. Therefore, the MPIW also remains consistent across significance levels. However, the MRPIW metric differs as it is influenced by changes in the model's point predictions. The MRPIW is calculated by dividing the determined interval width by the model's point prediction. Therefore, increases in MRPIW can be traced back to an average decrease in the predicted target values.

Deviations in the mean Winkler score are presented in Figs. 8 (a), (b), (c) and (d). The stability importance rankings across varying significance levels indicate that the relative impact of these features remains relatively unchanged with different levels of significance. For process-specific features, *MPT** is the most influential feature, followed by *QU* and *ACT*, regardless of the α value. *DCI* and *WGT* are the most significant product specific features. However, the overall impact of product-specific features on the Winkler score is negligible compared to that of process-specific features.

Similar to the observations in Table 8, greater deviations from the baseline can be observed with decreasing significance levels. This behavior stems from the Winkler score's intrinsic connection to the α value. As α decreases, the penalty for uncovered ground truths increases. This increase in penalty is not sufficiently offset by the higher likelihood of the prediction intervals covering the true values, leading to greater deviations from the baseline mean Winkler score with decreasing significance levels.

In summary, the changes in the test dataset result in the model misinterpreting the permuted feature's values, negatively affecting the point forecast and subsequently the quality of UQ. However, the measured effects of the permuted variables indicate that changes in the model's point prediction accuracy may be a relevant contributor. The following experiment settings aim at further isolating the feature impact on UQ.

5.3.2. Permutation feature importance on calibration data

After examining the influence of individual features on a calibrated ML model, we explore and compare the effects of permuting values of individual features during the conformal inference calibration on the final model's UQ performance. Due to these changes, it is expected that the prediction intervals to be affected according to the magnitude of uncertainty introduced during calibration. We evaluate the calibration quality on the test dataset, examining the deviations of PICP, MPIW, MRPIW, and mean Winkler score from the baseline established in Table 8.

Figs. 9 (a), (b), (c) and (d) generally follow the same trends as their respective counterparts in Figs. 6 (a), (b), (c) and (d), although the values deviate in a positive direction from the baseline. With the magnitude of deviations being limited by the number of uncaptured ground truths of the corresponding α value, we observe that the impact of relevant variables drives the ground truth coverage towards its maximum, especially for lower significance levels. Again, *MPT**, *ACT* and *QU* consistently rank highest for process-specific features, with *DCI* and *WGT* showing the highest impact regarding product-specific features. The minimal impact of other product-specific features indicates their lesser influence on the PICP metric. The general trend shows that the most influential features maintain their relative importance across varying α levels, underscoring their robustness.

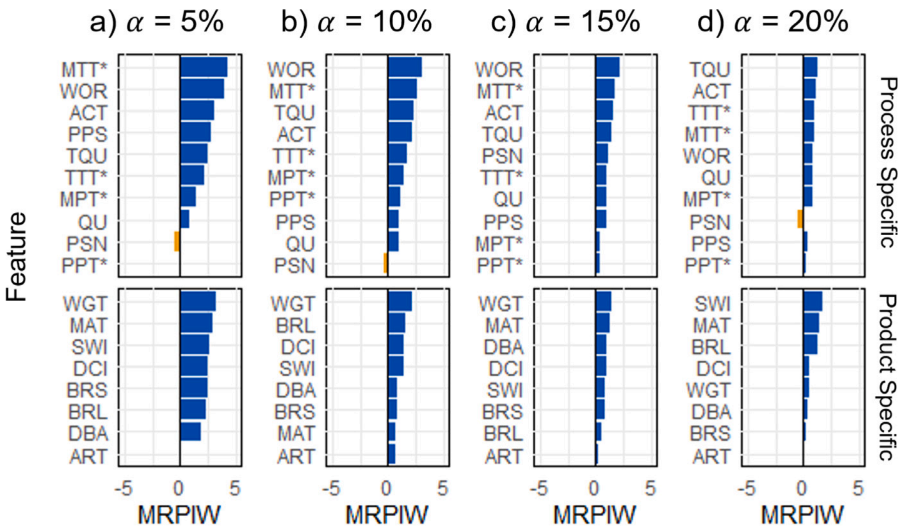


Fig. 7. MRPIW-related PFI for the GBM model for permuted test data.

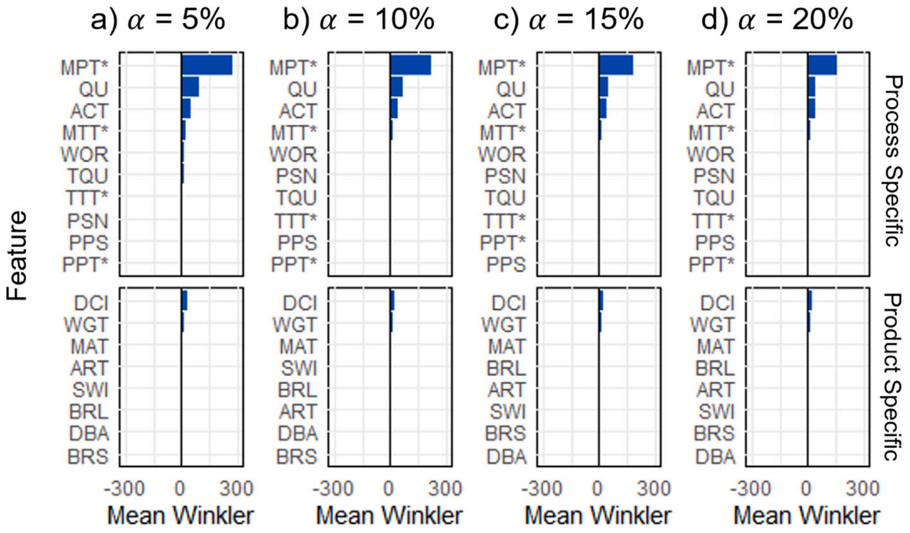


Fig. 8. Mean Winkler score-related PFI for the GBM model for permuted test data.

Considering that the prediction intervals are allowed to change, increases in PICP necessitate further examination of the prediction interval width to assess the interplay between model uncertainty and coverage of ground truths.

Figs. 10 (a), (b), (c) and (d) depict the deviations in MPIW, offering insight into the uncertainty of the ML model's predictions, with wider intervals indicating higher uncertainty. Similar to Figs. 9 (a), (b), (c) and (d), the consistently high deviations for *MPT** suggest the variable's critical role in model uncertainty, with other features generally following the same trend. The relationship between the expected coverage of ground truths based on the significance level and the width of prediction intervals is highlighted by larger deviations in MPIW being observed with decreasing α levels. Regarding the interpretation pertaining to deviations in PICP depicted in Figs. 9 (a), (b), (c) and (d), the increase in coverage implies that the calibration on the permuted data resulted in prediction intervals being generated more conservatively, leading to an increased performance on intact unseen data.

The observations for MPIW are overall congruent with the results for the MRPIW, depicted in Figs. 11 (a), (b), (c) and (d): Importance ranking remain unchanged, with the magnitude and orientation of deviations from the baseline mimicking the feature importance regarding the MPIW metric.

For the analysis of model uncertainty regarding the mean Winkler score, a comparison between Figs. 12 (a), (b), (c) and (d) and Figs. 8 (a), (b), (c) and (d) highlights the overall stability of the mean Winkler score for most features, emphasizing the role of relevant variables in the estimation of prediction intervals. Most notably, the impact on the mean Winkler score is only pronounced for *MPT**, following the trends observed for MPIW and MRPIW of decreasing its absolute impact with increasing significance levels. The remaining variables follow the previously identified importance trends, although their absolute effect on the mean Winkler score is relatively negligible.

The presented observations highlight that the model's inability to leverage the shuffled feature values of the permuted variable during the calibration step results in the prediction intervals being widened accordingly. However, the intact test dataset allows for the model to gain insight from the feature in question, consequently maintaining its point prediction accuracy during testing. This results in increased coverage of ground truths due to overestimation of prediction interval widths. Although this yields increased returns in PICP (see Figs. 9) for certain features, the mean Winkler score seems to accentuate only the most relevant ones.

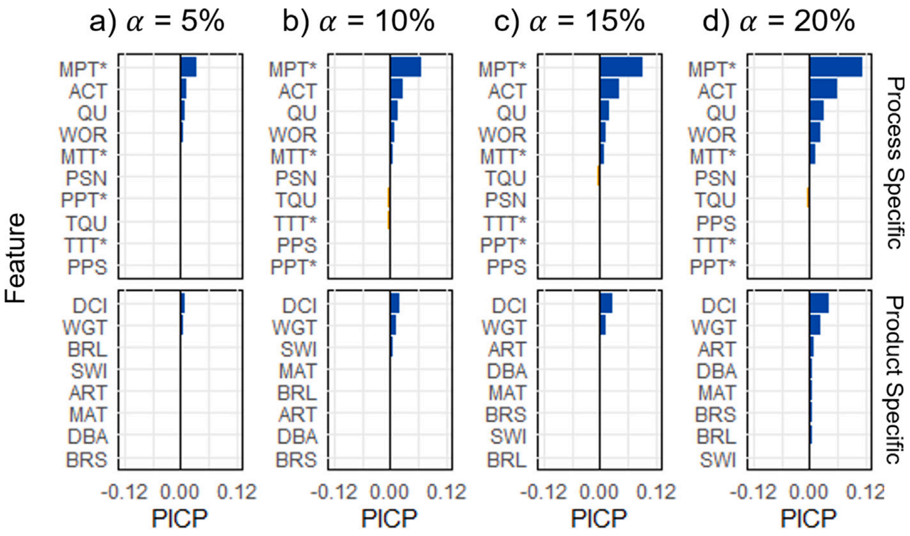


Fig. 9. PICP-related PFI for the GBM model for permuted calibration data.

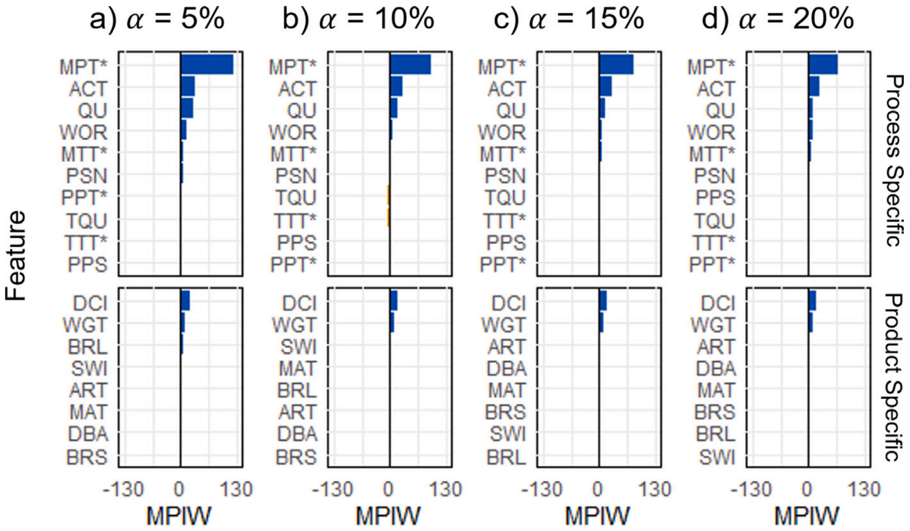


Fig. 10. MPIW-related PFI for the GBM model for permuted calibration data.

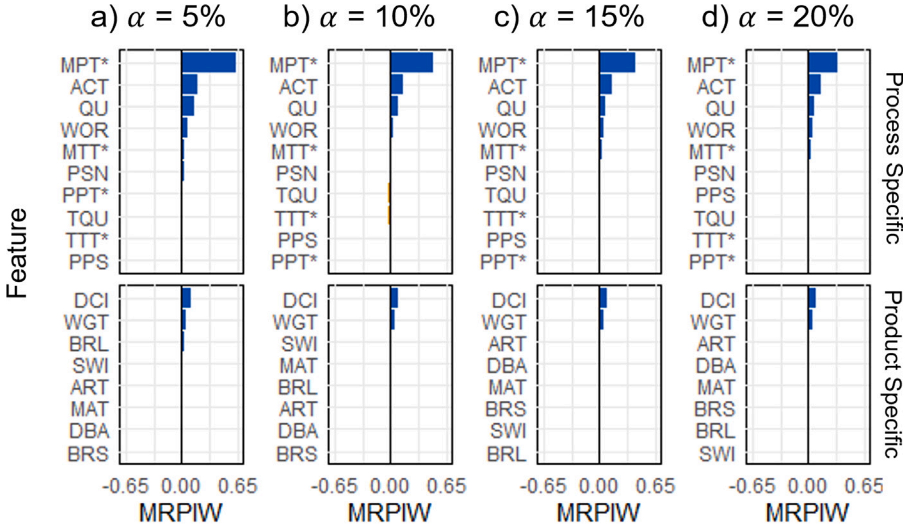


Fig. 11. MRPIW-related PFI for the GBM model for permuted calibration data.

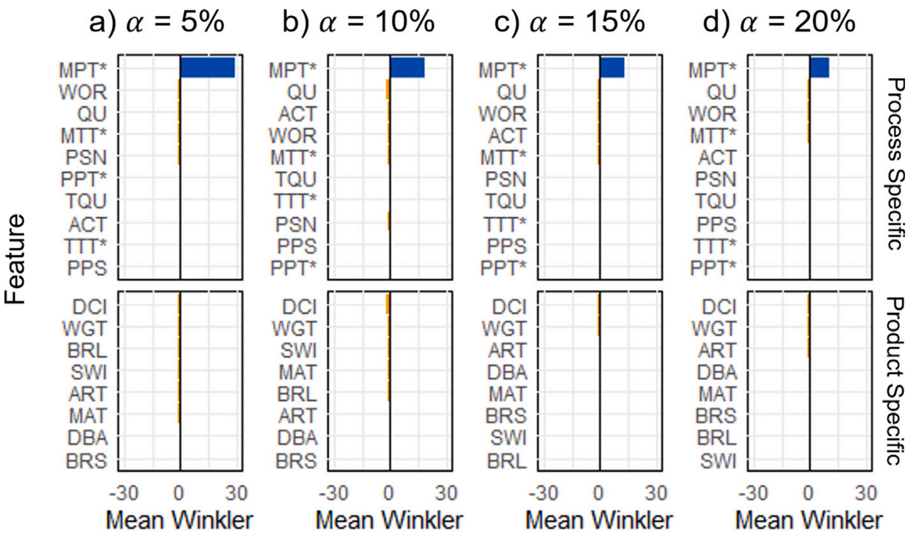


Fig. 12. Mean Winkler score-related PFI for the GBM model for permuted calibration data.

5.3.3. Permutation feature importance on calibration and test data

While the previous analysis allowed the model to leverage the predictive qualities of all variables from the test dataset for its point predictions, the following analysis extends the permutation to corresponding feature values within the test set as well. The resulting deviations from the baseline (see Table 8) regarding PICP, MPIW, MRPIW and mean Winkler score were examined and compared to the previous results.

Figs. 13 (a), (b), (c) and (d) present the results of PFI regarding PICP, showing a distribution of positive and negative deviations from the baseline, with the magnitude of deviations increasing with rising α values. The importance rankings across significance levels conform mostly with the corresponding rankings observed in Figs. 6 (a), (b), (c) and (d) and Figs. 9 (a), (b), (c) and (d). Differences can be observed regarding the bi-directionality of deviations as well as the increase in the number of variables with a noticeable impact on the PICP. As the significance levels decrease, the importance of certain features, particularly within product specifications at $\alpha = 5\%$, becomes less pronounced with minimal impact on the coverage of ground truths. The absolute effect on the PICP metric is marginal across significance levels, with the highest impact being recorded for *QU* at $\alpha = 20\%$, deviating by slightly more than 1%. Thus, the model's prediction interval coverage is relatively unaffected by the permutations, reflecting effective calibration of the model's prediction intervals with minimal fluctuations.

Since the calibration process remains unchanged, relevant results pertaining to the MPIW can be retrieved from Figs. 10 (a), (b), (c) and (d) of Section 5.3.2. However, as Figs. 14 (a), (b), (c) and (d) depict, the results regarding the MRPIW metric rather show similarities to Figs. 7 (a), (b), (c) and (d). Most notably, the magnitude of deviations from the baseline as well as slight differences in the feature importance rankings indicate the impact of inaccuracies in the point prediction on the relative prediction interval width: Considering that the calculation of the MRPIW is tied to the point prediction (see Section 4.3) and comparing the results to Fig. 11, the rise in MRPIW can be mainly attributed to the model generally reducing the predicted processing time as a reaction to the erroneous values introduced through feature permutation. Although the MRPIW metric conveys relevant insight about the quality of prediction intervals, its convoluted relationship with the model's point predictions complicates the analysis of model uncertainty and emphasizes the analyses from Section 5.3.2.

Mean Winkler score-related PFI is depicted in Figs. 15 (a), (b), (c) and (d) and shows notable similarities across alpha levels and feature categories towards the results depicted in Figs. 8 (a), (b), (c) and (d). These similarities pertain to the overall feature importance ranking, indicating the previously established relevance of variables like *MPT**, *QU*, *ACT* and *DCI*, although the magnitude in deviations decreased notably compared to the results of the former analysis. Similar to the results for MRPIW, a comparison with previously presented results for the mean Winkler score indicates that these similarities stem from inaccuracies in the point prediction due to the permuted test data, with a subsequent effect on the performance regarding UQ. In particular, Figs. 12 (a), (b), (c) and (d) indicated that calibration on the permuted feature generally mitigates the impact of erroneous feature values, with *MPT** increasing the mean Winkler score by less than 30 points and negligible changes for other variables. Consequently, any increases in the deviation of the mean Winkler score in Figs. 15 (a), (b), (c) and (d), although partially mitigated through the calibration, are predominantly affected by the changes in point predictions.

6. Discussion

Our proposed uncertainty-aware PFI enables the general assessment of a feature's impact on an ML model's predictive performance, specifically focusing on how perturbations in feature values influence the prediction intervals around predictions. In our work, we leverage SCP to construct prediction intervals and then examine three distinct scenarios of feature-value permutation to isolate how each stage – calibration vs. testing – affects uncertainty. This section discusses how these scenarios tie into practical settings of predictive process monitoring, compares them with conventional approaches, and highlights the principal insights derived from our findings.

Rationale behind the three permutation scenarios: A key contribution of our study is the design of three specific scenarios for permuting feature values – (1) in the test data only, (2) in the calibration data only, and (3) in both calibration and test data – and evaluating their impact on CP-based uncertainty quantification. These scenarios collectively capture a spectrum of real-world conditions in predictive process monitoring, especially when data come from event logs that are prone to changes over time and across different stages of model usage.

In the first scenario, only the test data are permuted. This setup simulates on-the-fly perturbations in new process instances, reflecting how, in many production systems, real-time data can be incomplete, noisy, or suffer from sensor or user-entry errors. Because the calibration data remain untouched, the intervals retain their original width, allowing us to isolate how deployment-phase noise influences the reliability of point predictions and final coverage. In the second scenario, only

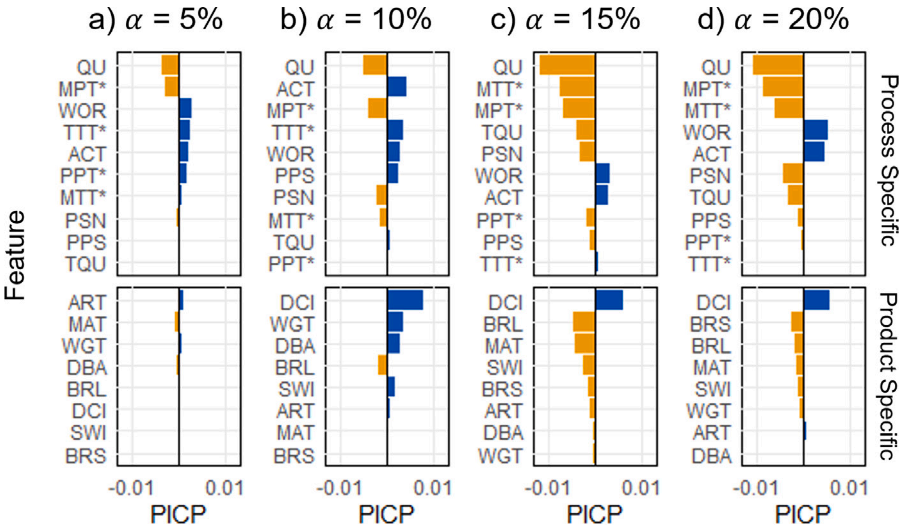


Fig. 13. PICP-related PFI for the GBM model for permuted calibration and test data.

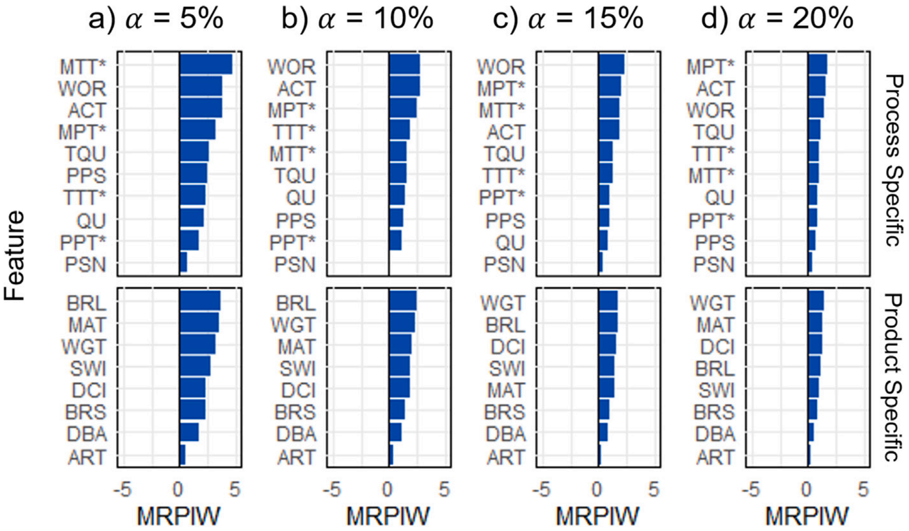


Fig. 14. MRPIW-related PFI for the GBM model for permuted calibration and test data.

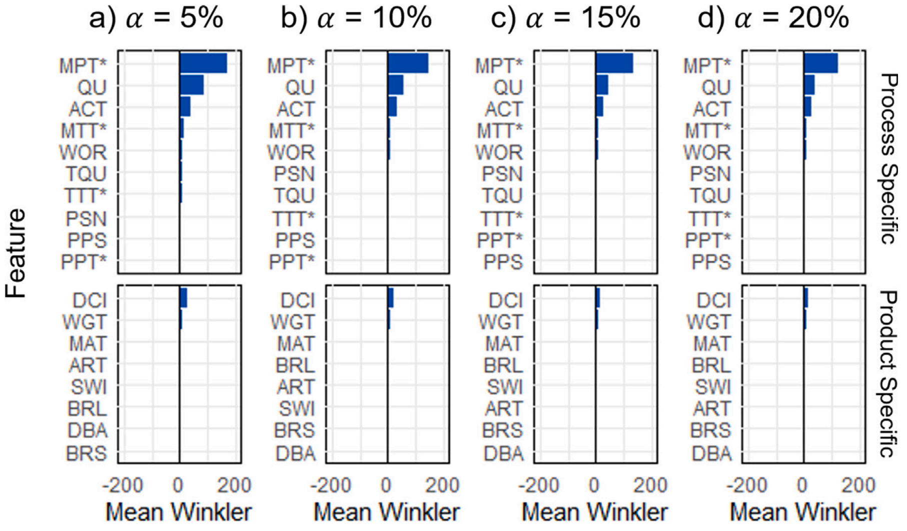


Fig. 15. Mean Winkler score-related PFI for the GBM model for permuted calibration and test data.

the calibration data are permuted, which reveals how flawed interval construction arises if the historical or offline logs used to size these intervals are inconsistent with present conditions. Although the test data stay accurate, shifting even one critical feature in calibration can lead to systematically biased intervals that either over- or under-estimate the true uncertainty. In the third scenario, both calibration and test data are simultaneously permuted, representing a worst-case environment where logs are incomplete or outdated and newly arriving data are similarly unreliable. This dual disturbance provides insight into whether the model can still maintain acceptable coverage despite major deviations in both the data used to form interval widths and the data used for ongoing forecasts.

Because process mining involves the exploration of event logs that may span months or years, the three scenarios address several core challenges. Continual drifts and heterogeneity in the event logs mean that the training and calibration stages might be disconnected from current process realities—machines could be replaced, operators could introduce novel practices, and new product variants might emerge, all of which can appear at the test stage (scenario 1) or already distort historical data (scenario 2). Partial traces further complicate matters because process monitoring is often near-real-time, meaning that any missing or noisy features at either calibration or runtime can shift coverage and interval width. Complex feature spaces, in which domain-specific attributes such as weight or quantity intersect with hierarchical structures (for instance, case-level vs. event-level data), highlight why conventional point-focused explanations (e.g., SHAP) might not suffice to understand how intervals change under feature corruption. Meanwhile, industrial constraints frequently limit data-quality improvement efforts, so managers must know whether legacy logs or real-time data capture is more critical to fix. These three scenarios clarify whether uncertain intervals stem primarily from flawed calibration, noisy inputs at runtime, or the combined effect of both.

Overall, systematically permuting features in test data, calibration data, or both offers a realistic view of how incomplete, drifting, or noisy event logs in process-aware information systems affect conformal intervals. The findings allow practitioners to identify which types of data corruption most strongly undermine coverage, thereby guiding whether to prioritize cleaning historical logs or enhancing the reliability of incoming data streams.

Comparison and trade-offs with other UQ methods: Many conventional UQ approaches, such as Bayesian methods or ensembles, focus on capturing parameter uncertainties or averaging over multiple models to derive prediction intervals. These techniques are powerful for accounting for both epistemic and aleatoric uncertainty but typically do not split off a separate calibration set to guarantee finite-sample coverage. By contrast, our CP–based approach inherently uses a dedicated calibration set to produce coverage guarantees under mild assumptions (e.g., exchangeability). This structural separation enables the three-scenario design, allowing us to examine, in isolation, how errors in historical logs (calibration data) versus real-time data (test data) affect interval width and coverage. Recalibrating intervals with permuted calibration data offers direct insights into how historical inaccuracies can inflate or deflate intervals while permuting the test data evaluates the model's sensitivity to real-time noise. In practical settings where process mining often encounters evolving data distributions, this scenario-based approach can highlight whether data-quality investments are more urgent in the logs used for calibration or ongoing data-capture processes. From a computational perspective, repeatedly recalibrating CP intervals to assess feature permutations can be more demanding than a single-pass Bayesian or ensemble-based approach. However, the additional overhead can yield more granular insights into which aspects of the data pipeline – historical calibration or real-time inputs – pose the greatest risk to coverage guarantees. For organizations that rely on accurate intervals to manage production resources or detect anomalies, pinpointing the exact source of coverage issues can be worth the extra computation.

Isolating the feature importance regarding prediction intervals from feature importance regarding point forecast: When feature shuffling is applied at the testing stage, as in the first and third experimental setting (Section 5.3.1 and Section 5.3.3 respectively), the integrity of the test dataset is compromised as a result. This reduction in dataset quality degrades the quality of point forecasts, subsequently affecting the utility of the prediction intervals. In particular for these scenarios, noticeable similarities for the MRPIW as well as mean Winkler metrics regarding importance rankings and the relative magnitude of deviations were observable. Comparing the results for these metrics to those of the second experiment scenario (see Section 5.3.2) reveals distinct differences regarding the expression of variable impact on these metrics. Due to the test data being kept intact and the model's point prediction being able to retain its accuracy, these configurations ensure that the measured effect is not diluted by changes in the point forecast. Thus, the results for this scenario highlight the impact of feature permutation solely on the calibration result and the quality of prediction intervals. For the underlying dataset and use case, the findings from the second experiment setting underscore the crucial role of process-specific features and the negligible role of product-specific features in calibrating prediction intervals. In the remaining scenarios, the model's point forecast mitigated the effect of the permuted feature value by modifying its prediction accordingly, subsequently altering the position of the prediction intervals and diluting the measured effect.

Informativeness of UQ measures for PFI and their practical implications: An important aspect of our study lies in interpreting how the employed uncertainty metrics – particularly PICP, MPIW, MRPIW, and the Winkler score – translate into both experimental outcomes and real-world decision-making in manufacturing. In many industrial settings, tighter (narrower) prediction intervals can aid in scheduling and resource allocation by reducing idle times and process bottlenecks. However, intervals that become too narrow risk overlooking genuine variability, potentially causing unexpected overruns and reactive reallocations. By contrast, intervals that are too wide may safeguard against missing the true duration of a task but can lead to inefficient resource usage.

Findings from our experiments further emphasize that these metrics often diverge in how they respond to feature permutations. Apart from PICP, no other metric exhibited relevant negative deviations (which would signify improvement) when a feature was permuted. Increases in MPIW generally indicate that the model requires broader intervals, which can reduce the quality and usefulness of its predictions for practical planning. Yet in scenarios that alter test data, changes in MRPIW and the mean Winkler score partly stem from the model's tendency to lower predicted values when input features appear corrupted, thereby inflating the relative interval width. Because MRPIW and the Winkler score are compound metrics, it is vital to interpret their fluctuations with caution: a higher MRPIW may reflect a more uncertain model reacting to irregular inputs rather than merely indicating poor calibration.

In concrete manufacturing contexts, this trade-off between coverage and interval tightness poses critical questions. PICP addresses reliability by measuring the proportion of true values captured within the intervals, which can be particularly valuable when missing a true value incurs steep penalties (e.g., unplanned downtime or delays). In one of our case studies, ensuring sufficiently high coverage for welding activities helped mitigate the costly ripple effects of underestimating task durations. However, excessively increasing the coverage often necessitated substantially wider intervals, as tracked by MPIW, limiting the granularity of resource scheduling. The Winkler score, in turn, penalizes intervals that fail to encompass the true value yet also discourages overbroad predictions, capturing the delicate balance between safety margins and operational efficiency.

Ultimately, whether narrower intervals or higher coverage is preferable depends on the cost structure of schedule deviations, rework, and machine downtime in a specific production environment. A plant manager might prefer intervals that are moderately wide to ensure reliable forecasts, while a less failure-sensitive assembly line may aim for tighter intervals in pursuit of higher throughput. The insights gained from PICP, MPIW, MRPIW, and the Winkler score collectively illuminate these trade-offs by revealing intricate details about how each permutation experiment – or real-world data anomaly – affects model confidence and calibration quality. By examining them in conjunction, practitioners can better identify scenarios where additional data-cleaning efforts, sensor upgrades, or model refinements offer the greatest payoff in balancing efficiency and reliability.

Practical limitations and domain adaptation: We tested our approach in a regression setting tailored to continuous outcome prediction in a manufacturing environment for a PPM problem. However, in classification tasks with severe class imbalance, the validity of SPC and PFI can become more challenging to maintain, particularly for minority classes. While CP remains formally valid under exchangeability assumptions, highly skewed label distributions may limit its ability to generate tight, reliable intervals for underrepresented outcomes. One way to mitigate this limitation is to adopt more elaborate calibration strategies. Another option involves specialized sampling techniques before applying CP.

Beyond manufacturing, domains like healthcare or finance often confront rapidly evolving data distributions and domain-specific constraints. For instance, healthcare data can be inherently temporal, requiring additional causal modeling or domain knowledge to handle risk factors effectively. Financial data similarly necessitates adaptive conformal methods that can update prediction intervals as market indicators shift. By adjusting calibration procedures or leveraging domain insights, our uncertainty-aware PFI framework can be adapted successfully to a wide range of high-stakes applications, maintaining robust interval estimation and feature-importance insights despite domain-specific challenges. In addition, our current implementation assumes that the exchangeability assumption remains reasonably intact and that both calibration and test data share similar distributions. Situations involving severe covariate shifts, non-stationarity in event logs, or real-time streaming data may demand more sophisticated CP variants (such as online or adaptive conformal approaches).

Moreover, organizations operating under stringent data privacy or regulatory guidelines may face limitations in applying both feature permutations and conformal predictions. In such cases, anonymization or federated learning solutions might be required to preserve compliance, potentially reducing the granularity or availability of calibration data. Methods for adaptively incorporating these constraints, while still maintaining valid intervals, represent a promising direction for extending the current framework.

Future Work: While our study presents a novel integration of PFI with UQ within the CP framework, several avenues remain open for further exploration and enhancement. Our current application focuses on process mining for manufacturing, which presents unique challenges due to its event-driven and temporal data characteristics. Extending our methodology to other domains within process mining, such as healthcare, finance, or autonomous systems, would demonstrate its versatility and robustness across different data types and application contexts.

In addition to evaluating traditional UQ measures, exploring alternative metrics could enhance the comprehensiveness of predictive uncertainty assessment. Various uncertainty measures, such as Bayesian credible intervals, entropy-based measures, or adversarial uncertainty measures, could be combined to look at various aspects of uncertainty and make it easier to understand which features are most important. Our method recognizes the challenges associated with permuting features in the calibration data, which can disrupt the exchangeability assumption central to conformal prediction. Future research could investigate techniques to mitigate this issue, such as conditional permutation methods that preserve the joint distribution of features or incorporate robustness checks to assess the impact of exchangeability violations. Additionally, developing adaptive conformal methods that relax the exchangeability assumption while maintaining valid uncertainty estimates would be a significant advancement.

Future work could explore integrating our PFI and UQ framework with knowledge graph-based reasoning to enhance interpretability (Li et al., 2024a). Leveraging structured semantic information and explicit rules can provide more transparent and understandable explanations of model predictions, especially in complex, event-driven domains (Li et al., 2024b). This integration would allow for the incorporation of domain-specific knowledge, facilitating the identification of meaningful feature relationships and enhancing the overall explainability of the model. Understanding the causal relationships between features and predictions is also crucial for developing robust and reliable models. Future research could also integrate causal inference techniques with our PFI and UQ framework to disentangle correlation from causation. This integration would identify truly influential features and enhance the model's ability to generalize to unseen scenarios.

Finally, conducting longitudinal studies and deploying our methodology in real-world settings would provide valuable insights into its practical utility and impact. Such studies could assess how the importance of uncertainty-aware features influences decision-making processes, model trust, and operational efficiency in dynamic environments. Additionally, collaborating with industry partners to apply our framework to live data streams would validate its effectiveness and adaptability in real-time applications.

7. Conclusion

This study presents a novel methodology for integrating PFI into the domain of UQ. Our approach, structured into four primary stages – data preprocessing, model training, UQ, and XAI – provides a comprehensive framework for predictive process monitoring. The initial data preprocessing stage ensures high-quality inputs. This refined event log serves as a strong foundation for subsequent model training, ensuring appropriate model performance. In the UQ stage, SCP are employed to generate reliable prediction intervals, and evaluated using PICP, MRPIW, and mean Winkler score. The final XAI stage leverages PFI to elucidate the model's inner workings. By evaluating feature importance at various stages, permuting the calibration or test data or both, we were able to examine the results under various aspects. This stage contributes to the correct interpretation of the measured effects and provides a guideline for evaluating feature importance in the context of UQ.

Our methodology successfully integrates advanced ML techniques with rigorous UQ and XAI practices. By systematically combining these elements, we create a robust framework that can not only be leveraged to enhance predictive performance but also contribute to model interpretability. This dual focus on UQ and explainability addresses critical challenges in the predictive process monitoring domain associated with black-box algorithms, fostering trust and accountability in automated decision-making systems. Through the evaluation of three distinct PFI for UQ approaches, our study highlights the necessity of isolating the impact of feature permutation on UQ metrics from those related to point forecasts. The findings emphasize the importance of maintaining test data integrity and leveraging a variety of UQ metrics for a holistic understanding of model behavior. By doing so, we provide a nuanced understanding of how features contribute to both predictions and the confidence in these predictions, ultimately enhancing the robustness and reliability of predictive models.

In conclusion, our proposed methodology represents a significant advancement in predictive process monitoring by integrating XAI with UQ. This comprehensive approach ensures that predictive models are not only accurate but also transparent and reliable, thereby promoting their adoption and trust across various domains.

CRediT authorship contribution statement

Nijat Mehdiyev: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Maxim Majlatow:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Peter Fettke:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT 4 to enhance language clarity, coherence, and rephrasing. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research was funded in part by the German Federal Ministry of Education and Research, Germany under grant number 01IS24048C (project EINHORN).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Production analysis event log

Dataset and use case overview

The dataset utilized for the alternative use case is the Production Analysis event log (Levy, 2014) and stems from a manufacturing context, which exhibits similarities with the primary use case description detailed in Section 4: various articles are being produced and each activity of the production process is documented as an event in the log, accompanied by the event- and trace-level specifications regarding the produced item. Similarly, the product specifications dictate the manufacturing procedures, including rework to fulfill qualitative product requirements.

The original event log consists of 4543 events grouped into 225 cases with 55 unique activities performed using 31 unique resources. For preprocessing and feature engineering the primary use case and the steps performed in Mehdiyev et al. (2024a) served as guidelines. Notably, the activity (*ACT*) and resource (*RSC*) columns were subject to rigorous changes due to the redundancy of shared information, commonly stemming from resource information being reiterated in the activity column and leading to a false uniqueness. Therefore, activities were split from the resources, such as machines or tools, they were executed with, leading to a corrected number of 25 unique activities. Furthermore, we identified tasks that were interrupted, for example, due to changes in worker shifts, which partitions planned processing times. Such events were aggregated into a single event, correcting relevant feature values, such as those regarding produced quantities, accordingly. Additionally, previous activities, their processing times as well as the next planned activities were added via feature engineering.

Table 9
Evaluation of the final RF model regarding uncertainty quantification on test data of the Production Analysis event log.

α-Value	$PICP$	$MPIW$	$MRPIW$	W
5%	0.98	1588.2	11.6	2016.9
10%	0.92	855.4	6.2	1395.9
15%	0.88	652.9	4.8	1149.7
20%	0.81	476.8	3.48	982.0

Model evaluation

The dataset was then partitioned into training, calibration, and test sets in a 6:2:2 ratio, yielding 1,536 events for training, 513 events for calibration, and 520 events for testing. With the only product-specific feature being the part description (*ART*), describing the specific article or product being manufactured, the remainder of variables contains process-specific information. n accordance with the primary use case, hyperparameter optimization was performed as described in Section 4.2, with the RF model (min_n= 3, mtry= 6, 527 trees) achieving the best performance across MAE and RMSE with 180 min and 356 min respectively, followed by the GBM and DT models and, lastly, the GLM model. An additional evaluation of test data yielded an MAE 158 min and an RMSE of 281 min, with the next best model being GBM with 159 and 312 min respectively. Both evaluations highlight the superior performance of the RF model, especially with regards to the RMSE metric, scoring a 10% lower RMSE then the next best model when evaluated on test data. Thus, further analysis regarding UQ is conducted by employing the RF model.

The evaluation of model uncertainty for the Production Analysis event log entails the application of SCP to the final RF model and, as in the primary use case (see Section 5.2), utilizing the calibration data for fine-tuning prediction intervals. Afterward, the model was evaluated on the test data, with the results summarized in Table 9, presenting MPIW, MRPIW, PICP, and mean Winkler score across various significance levels α.

The achieved coverage probabilities for the final RF model on the Production Analysis event log consistently exceed the nominal coverage levels corresponding to the α values. Specifically, at $\alpha = 5\%$ (nominal coverage of 95%), the PICP is 98%, indicating over-coverage due to overly conservative prediction intervals. As α increases from 5% to 20%, the PICP decreases from 98% to 81%, yet remains above the nominal coverage levels (95% to 80%), suggesting that the intervals are consistently wider than necessary. Concurrently, the MPIW and MRPIW decrease with increasing α, reflecting narrower intervals but still indicating excessive width at lower α levels.

Comparing these results with the German Manufacturer dataset in Table 8, we observe that the PICP values are closer to the nominal coverage levels and occasionally slightly below them (e.g., PICP of 94% at $\alpha = 5\%$). The notably wider prediction intervals for the Production Analysis dataset can be attributed to several key differences in its characteristics compared to the German Manufacturer dataset. Firstly, the Production Analysis dataset exhibits a significantly higher standard deviation in processing times (550.9 min) relative to its mean (325.2 min), indicating substantial variability and dispersion in the target variable. In contrast, the German Manufacturer dataset has a lower standard deviation (117.08 min) with respect to its mean processing time (100.74 min), suggesting more consistent processing times across cases. Secondly, the Production Analysis dataset is considerably smaller, containing only 225 cases and 2,569 events, whereas the German Manufacturer dataset includes 32,831 cases and 165,433 events. The limited size of the Production Analysis dataset reduces the model's capacity to learn complex patterns effectively, often resulting in less precise predictions and necessitating wider prediction intervals to maintain the desired coverage levels. Additionally, W values are substantially higher for the Production Analysis dataset across all α

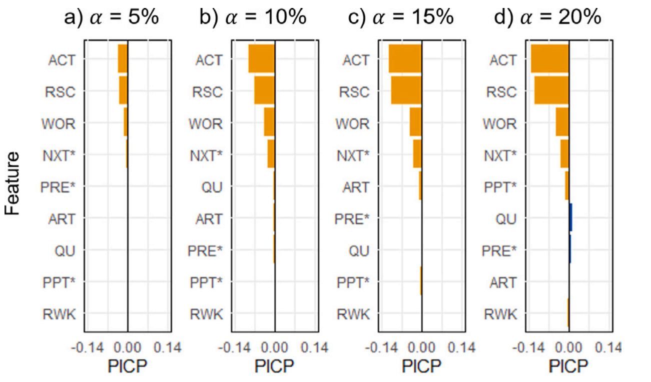


Fig. 16. PICP-related PFI for the RF model for permuted test data.

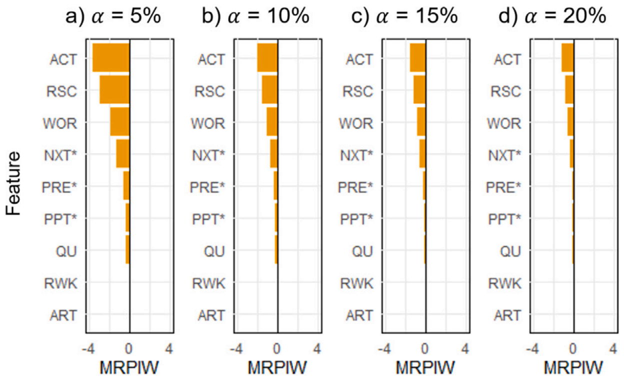


Fig. 17. MRPIW-related PFI for the RF model for permuted test data.

levels (e.g., $W = 2016.9$ at $\alpha = 5\%$) compared to the German Manufacturer dataset (e.g., $W = 409.3$ at $\alpha = 5\%$). The excessive W scores in the Production Analysis dataset reflect the combination of wider intervals (as indicated by higher MPIW and MRPIW) and over-coverage. This inefficiency can be attributed to the dataset's higher variability and smaller size, which compel the model to produce conservative estimates to achieve the desired coverage, albeit at the expense of interval sharpness.

Permutation feature importance on test data

The PFI analysis for the uncertainty metrics at varying significance levels α reveals consistent patterns in feature influence for the calibrated model. Across all metrics, the features *ACT*, *RSC*, and *WOR* consistently demonstrate high importance, underscoring their critical impact on model performance. The engineered feature *NXT** (next activity) also shows moderate influence but ranks slightly lower, with its counterpart *PRE** (previous activity) showing almost no relevance at all. Comparing the results for PICP (see Fig. 16) against MRPIW (see Fig. 17), higher deviations in the PICP are documented for increasing α levels, whereas for MRPIW the inverse trend is exhibited, suggesting reduced sensitivity to individual features at higher significance levels at the cost of ground truth coverage. Lower-ranked features like *QU* and *PPT** show minimal impact, indicating limited influence on interval coverage. For the mean Winkler score (see Fig. 18), which balances interval width and accuracy, *ACT* and *RSC* again exhibit strong importance as well as robust scores for all features across different α values. These identified trends show strong similarities to the results from the German Manufacturer dataset (see Figs. 6–8).

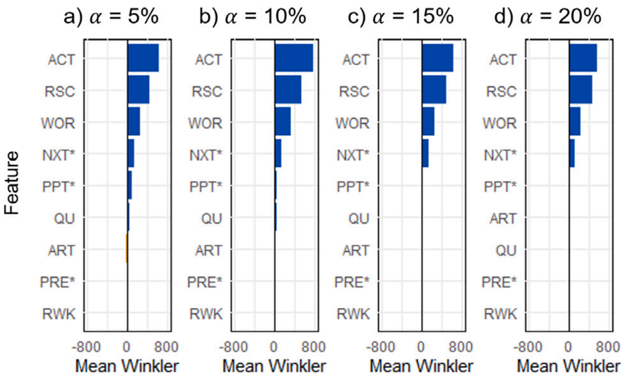


Fig. 18. Mean Winkler score-related PFI for the RF model for permuted test data.

Permutation feature importance on calibration data

The analysis of PFI across UQ metrics, namely PICP (see Fig. 19), MPIW (see Fig. 20), MRPIW (see Fig. 21), and mean Winkler score (see Fig. 22), reveals distinct patterns in feature relevance with respect to the model calibrated through split conformal inference. For PICP, the most impactful features consistently include *RSC*, *ACT*, and *WOR* across various levels of α, suggesting their impact on the model's predictive coverage. Notably, the impact of *PPT** appears slightly more prominent at the lowest α values (5%), indicating that the role of certain features in maintaining prediction interval coverage may vary with the calibration confidence level. Since the calibration data is permuted, which directly influences the interval width, differences in feature impacts on the MPIW can be observed (see Fig. 20). Although the prediction intervals primarily show increases, indicating heightened model uncertainty, only marginal variations are observable in feature rankings. This trend extends to the MRPIW and mean Winkler score as well, exhibiting similar trends regarding the relative impact of features on the examined uncertainty metric. In summary, *RSC* and *ACT* consistently emerge as primary contributors across all metrics, highlighting their importance in maintaining effective uncertainty estimates. Secondary features, such as *PPT** and *WOR*, exhibit variable influence depending on the calibration level (α), indicating that their role in UQ metrics depends on the tolerance level for prediction error. Furthermore, notable differences in feature rankings compared to only permuting test data can be identified, primarily *RSC* overtaking *ACT* as the most impactful variable.

Permutation feature importance on calibration and test data

The permutation of calibration data leads to more conservative prediction intervals, which is observed through generally higher values in MPIW, as noted in previous figures (see Fig. 20). However, permuting both calibration and test data, introduces an opportunity to analyze the interaction between the adapted prediction intervals and the modified model predictions, as seen across PICP (see Fig. 23), MRPIW (see Fig. 24), and mean Winkler score metrics (see Fig. 25). For PICP, significant features are consistently observed across α values, with *ACT* playing a prominent role, particularly at medium and high α levels (10%–20%). The inclusion of test data permutation highlights the robustness of *ACT* in sustaining coverage requirements despite the disturbances introduced in both the calibration and test datasets. MRPIW demonstrates the interaction between the adaptation of the CP towards the permuted calibration data and the adaptation of model predictions to the permuted test data, highlighting the fine-grained role of features like *WOR*, *PPT** and *NXT** in managing prediction intervals at the lowest α level (5%). For higher significance levels, however, the relevance of these features diminishes. For the mean Winkler score, similar trends and feature rankings as in Fig. 18 can be observed, although slight decreases in the absolute values are exhibited.

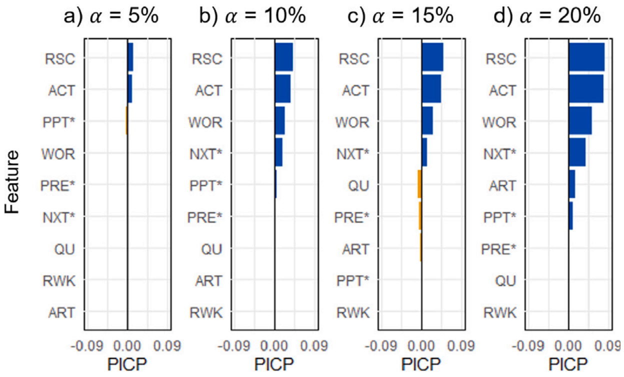


Fig. 19. PICP-related PFI for the RF model for permuted calibration data.

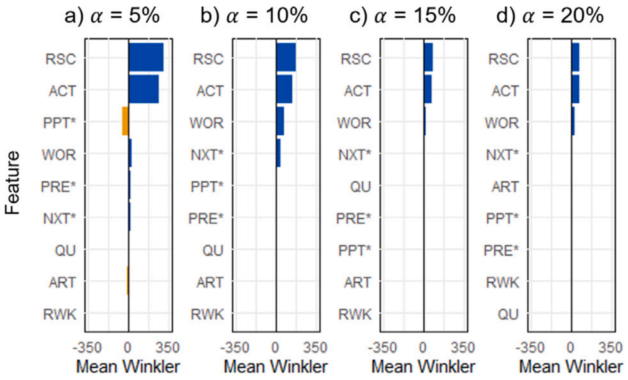


Fig. 22. Mean Winkler score-related PFI for the RF model for permuted calibration data.

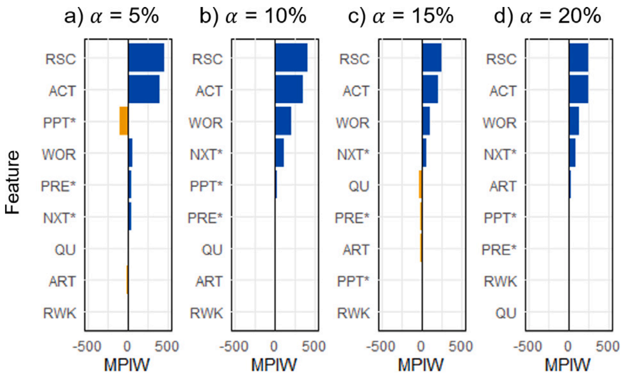


Fig. 20. MPIW-related PFI for the RF model for permuted calibration data.

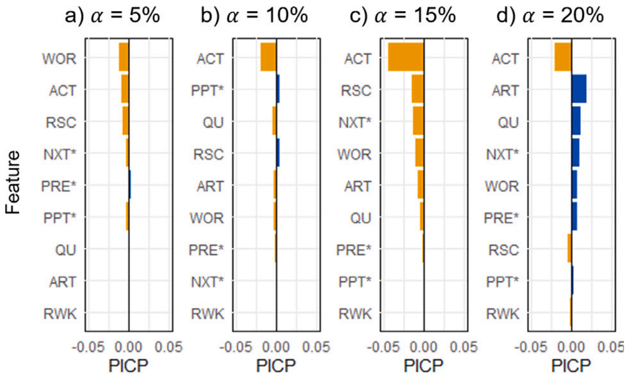


Fig. 23. PICP-related PFI for the RF model for permuted calibration and test data.

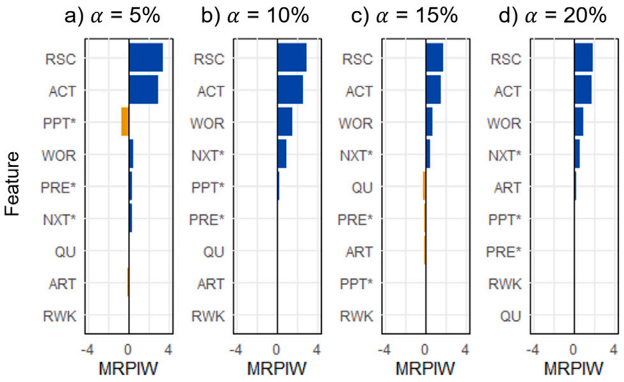


Fig. 21. MRPIW-related PFI for the RF model for permuted calibration data.

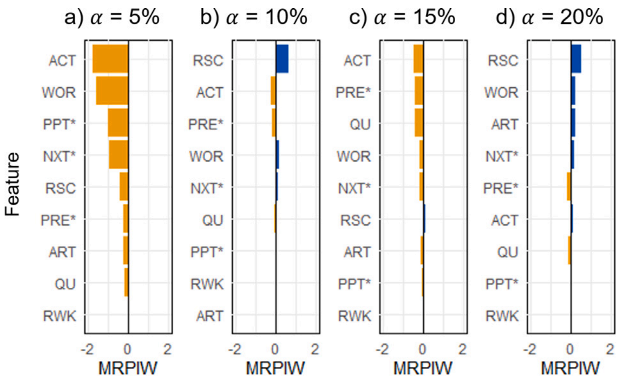


Fig. 24. MRPIW-related PFI for the RF model for permuted calibration and test data.

In summary, the analysis underscores the centrality of *ACT* and *RSC* across all UQ metrics in adapting to the permutations in calibration and test data. These features are pivotal in the model's recalibration process, while *PPT** and *WOR* offer conditional importance that varies with α, particularly in stricter predictive requirements.

To conduct a dedicated analysis of feature impacts on model uncertainty, it is essential to methodologically separate the influence of features on UQ metrics from their influence on model point predictions. When both calibration and test data are permuted, the observed effects on UQ metrics result from a combination of feature impact on both interval adjustments (UQ) and model output accuracy (point predictions). This dual impact complicates the interpretation of each feature's contribution to uncertainty. Isolating these aspects by analyzing UQ feature impacts independently of point prediction effects showed substantial differences and is crucial for obtaining clear insights into each feature's role in managing prediction intervals and maintaining predictive coverage.

Appendix B. Conformal prediction calibration sensitivity analysis

Regarding the sensitivity analysis conducted in Section 5.2, Figs. 26–29, 30–33, 34–37, and 38–41 depict the effects of varying calibration set sizes across different levels of $alpha$ for the PICP, MPIW, MRPIW and mean Winkler score metrics for each of the employed ML models (GBM, GLM, DT and RF) respectively. These findings support the findings and identified trends from Section 5.2, highlighting the superior performance of the GBM model across varying $alpha$ values.

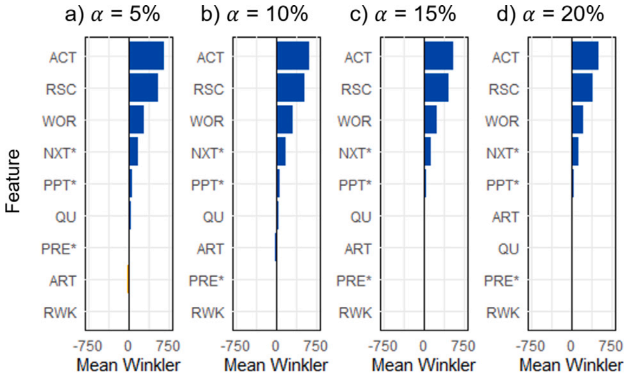


Fig. 25. Mean Winkler score-related PFI for the RF model for permuted calibration and test data.

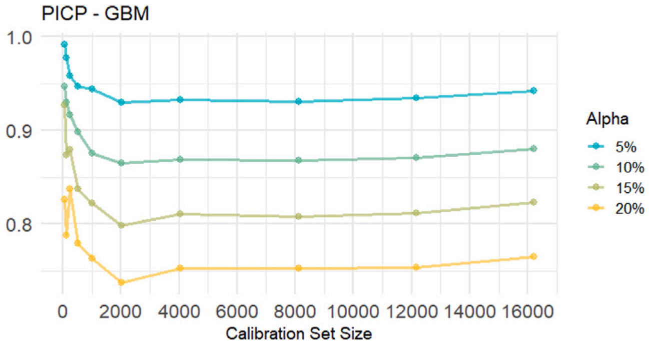


Fig. 26. Evaluation of the fitted GBM model regarding sensitivity of PICP on test data across varying sizes of calibration data.

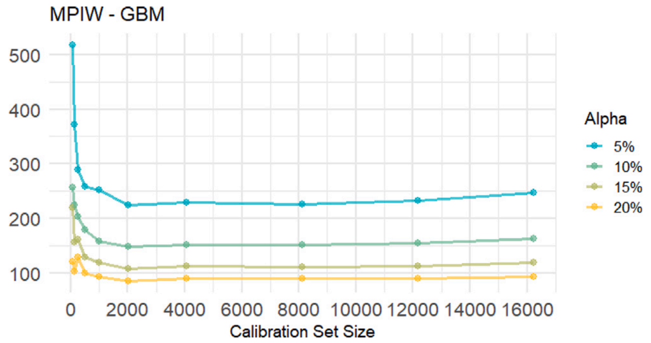


Fig. 27. Evaluation of the fitted GBM model regarding sensitivity of MPIW on test data across varying sizes of calibration data.

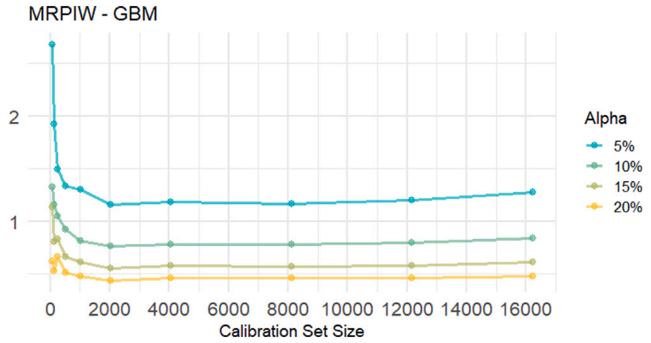


Fig. 28. Evaluation of the fitted GBM model regarding sensitivity of MRPIW on test data across varying sizes of calibration data.

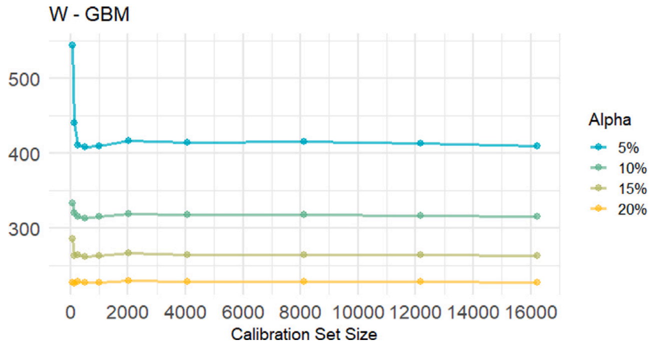


Fig. 29. Evaluation of the fitted GBM model regarding sensitivity of W on test data across varying sizes of calibration data.

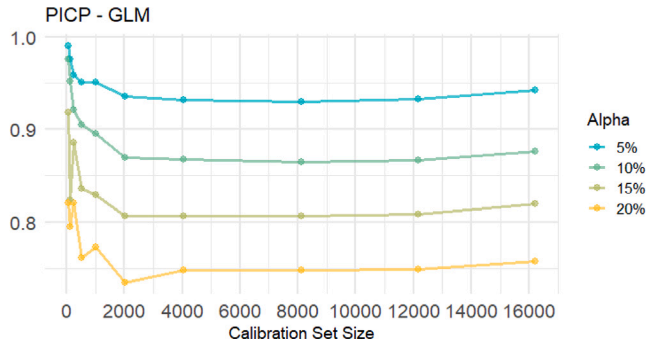


Fig. 30. Evaluation of the fitted GBM model regarding sensitivity of PICP on test data across varying sizes of calibration data.

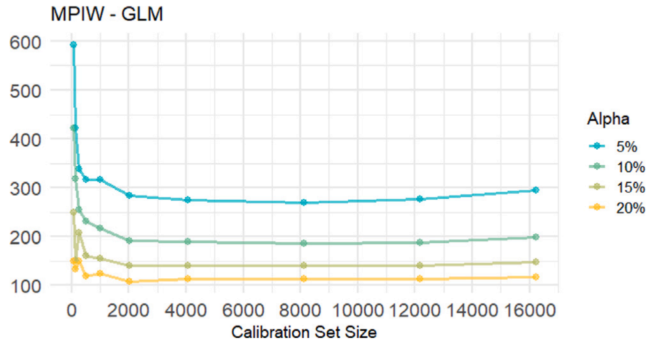


Fig. 31. Evaluation of the fitted GBM model regarding sensitivity of MPIW on test data across varying sizes of calibration data.

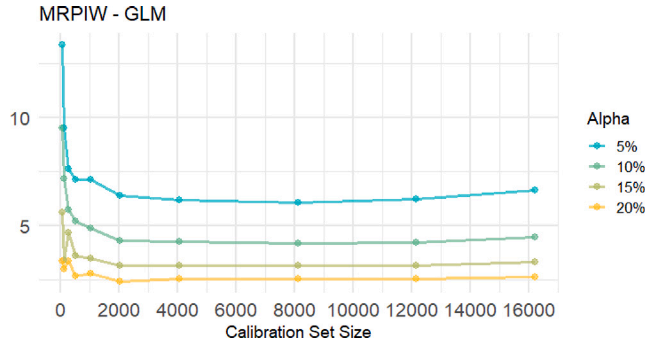


Fig. 32. Evaluation of the fitted GBM model regarding sensitivity of MRPIW on test data across varying sizes of calibration data.

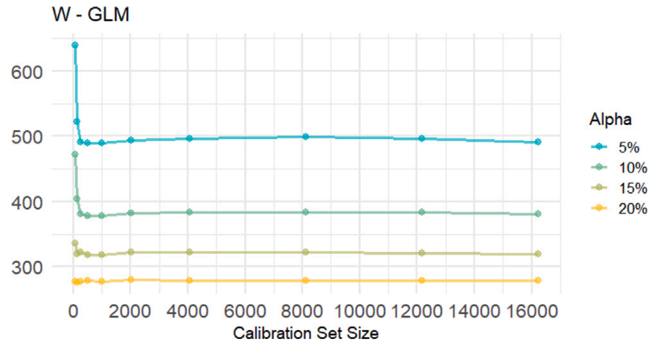


Fig. 33. Evaluation of the fitted GBM model regarding sensitivity of W on test data across varying sizes of calibration data.

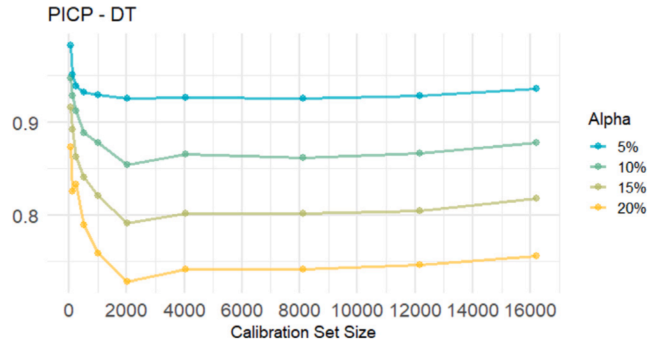


Fig. 34. Evaluation of the fitted GBM model regarding sensitivity of PICP on test data across varying sizes of calibration data.

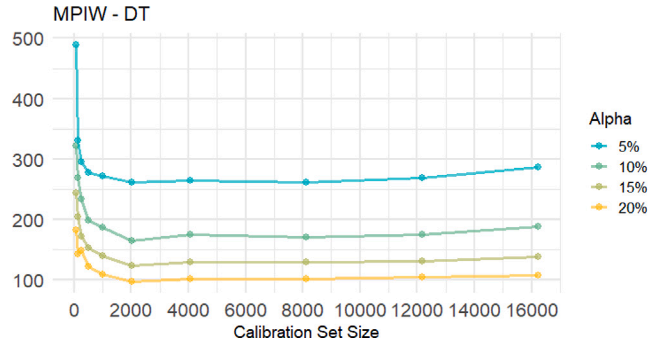


Fig. 35. Evaluation of the fitted GBM model regarding sensitivity of MPIW on test data across varying sizes of calibration data.

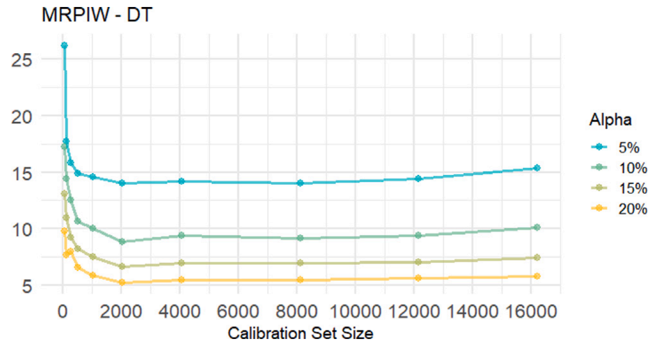


Fig. 36. Evaluation of the fitted GBM model regarding sensitivity of MRPIW on test data across varying sizes of calibration data.

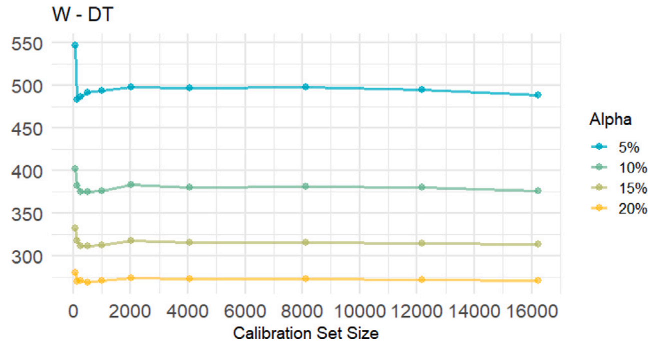


Fig. 37. Evaluation of the fitted GBM model regarding sensitivity of W on test data across varying sizes of calibration data.

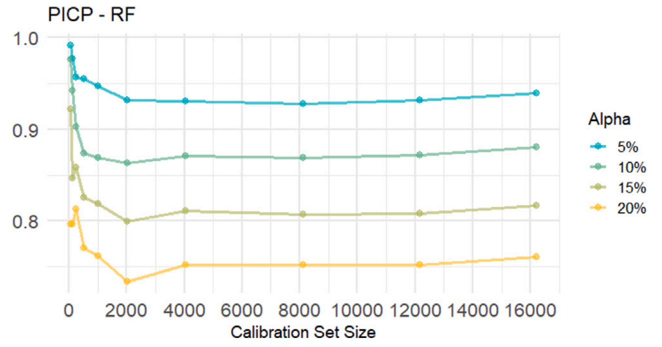


Fig. 38. Evaluation of the fitted GBM model regarding sensitivity of PICP on test data across varying sizes of calibration data.

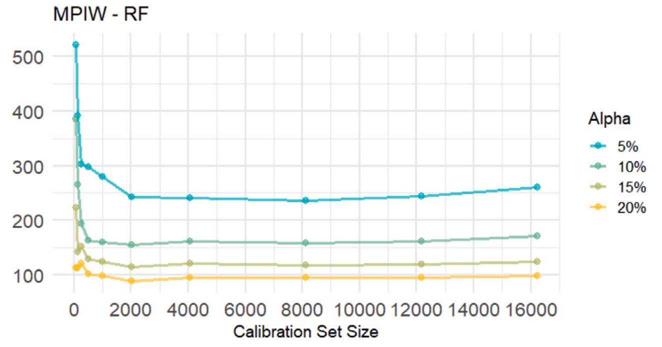


Fig. 39. Evaluation of the fitted GBM model regarding sensitivity of MPIW on test data across varying sizes of calibration data.

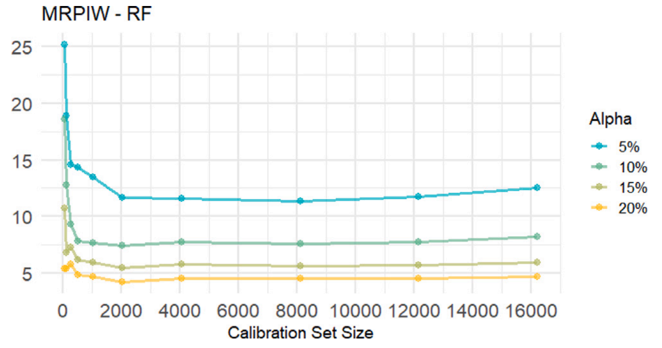


Fig. 40. Evaluation of the fitted GBM model regarding sensitivity of MRPIW on test data across varying sizes of calibration data.

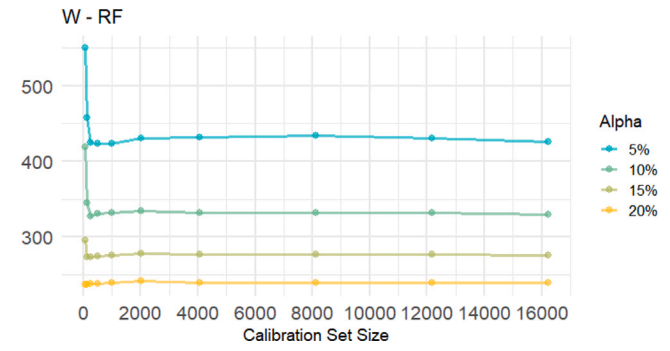


Fig. 41. Evaluation of the fitted GBM model regarding sensitivity of W on test data across varying sizes of calibration data.

Data availability

The data that has been used is confidential.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Inf. Fusion 99, 101805.

Angelopoulos, A.N., Bates, S., et al., 2023. Conformal prediction: A gentle introduction. Found. Trends® Mach. Learn. 16 (4), 494–591.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Baniecki, H., Biecek, P., 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. Inf. Fusion 102303.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al., 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 401–413.

Bousdekis, A., Kerasiotis, A., Kotsias, S., Theodoropoulou, G., Miaoulis, G., Ghazanfarpour, D., 2023. Modelling and predictive monitoring of business processes under uncertainty with reinforcement learning. Sensors 23 (15), 6931.

Breuker, D., Matzner, M., Delfmann, P., Becker, J., 2016. Comprehensible predictive models for business processes. Mis Q. 40 (4), 1009–1034.

Bukhsh, Z.A., Saeed, A., Stipanovic, I., Doree, A.G., 2019. Predictive maintenance using tree-based classification techniques: A case of railway switches. Transp. Res. Part C: Emerg. Technol. 101, 35–54.

Chander, B., John, C., Warrier, L., Gopalakrishnan, K., 2024. Toward Trustworthy Artificial Intelligence (TAI) in the context of explainability and robustness. ACM Comput. Surv..

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Chiaburu, T., Haußer, F., Bießmann, F., 2024. Uncertainty in XAI: Human perception and modeling approaches. Mach. Learn. Knowl. Extr. 6 (2), 1170–1192.

Coma-Puig, B., Carmona, J., 2022. Non-technical losses detection in energy consumption focusing on energy recovery and explainability. Mach. Learn. 111 (2), 487–517.

De Koninck, P., De Weerdt, J., vanden Broucke, S.K., 2017. Explaining clusterings of process instances. Data Min. Knowl. Discov. 31 (3), 774–808.

Di Francescomarino, C., Ghidini, C., Maggi, F.M., Milani, F., 2018. Predictive process monitoring methods: Which one suits me best? In: International Conference on Business Process Management. Springer, pp. 462–479.

Dong, J., Chen, S., Miralinaghi, M., Chen, T., Li, P., Labi, S., 2023. Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. Transp. Res. Part C: Emerg. Technol. 156, 104358.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Comput. Surv. 55 (9), 1–33.

Emmert-Streib, F., Yli-Harja, O., Dehmer, M., 2020. Explainable artificial intelligence and machine learning: A reality rooted perspective. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 10 (6), e1368.

Evermann, J., Rehse, J.-R., Fettke, P., 2017. Predicting process behaviour using deep learning. Decis. Support Syst. 100, 129–140.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20 (177), 1–81.

Foygel Barber, R., Candes, E.J., Ramdas, A., Tibshirani, R.J., 2021. The limits of distribution-free conditional predictive inference. Inf. Inference: A J. IMA 10 (2), 455–482.

Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in recurrent neural networks. Adv. Neural Inf. Process. Syst. 29.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56 (Suppl 1), 1513–1589.

Ghanem, R., Higdon, D., Owhadi, H., et al., 2017. Handbook of Uncertainty Quantification, Vol. 6. Springer.

Gibbs, I., Candes, E., 2021. Adaptive conformal inference under distribution shift. Adv. Neural Inf. Process. Syst. 34, 1660–1672.

Graves, A., 2011. Practical variational inference for neural networks. Adv. Neural Inf. Process. Syst. 24.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI—Explainable artificial intelligence. Sci. Robot. 4 (37), eaay7120.

Hastie, T., Qian, J., Tay, K., 2021. An introduction to glmnet. CRAN R Repositary 5, 1–35.

Hill, D., Masoomi, A., Torop, M., Ghimire, S., Dy, J., 2024. Boundary-aware uncertainty for feature attribution explainers. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 55–63.

Hsieh, C., Moreira, C., Ouyang, C., 2021. Dice4el: interpreting process predictions using a milestone-aware counterfactual approach. In: 2021 3rd International Conference on Process Mining. ICPM, IEEE, pp. 88–95.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. 110 (3), 457–506.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Lakshmanan, G.T., Duan, S., Keyser, P.T., Curbera, F., Khalaf, R., 2011. Predictive analytics for semi-structured case oriented business processes. In: Business Process Management Workshops: BPM 2010 International Workshops and Education Track, Hoboken, NJ, USA, September 13-15, 2010, Revised Selected Papers 8. Springer, pp. 640–651.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K., 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296, 103473.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-free predictive inference for regression. J. Amer. Statist. Assoc. 113 (523), 1094–1111.

Lei, J., Rinaldo, A., Wasserman, L., 2015. A conformal prediction approach to explore functional data. Ann. Math. Artif. Intell. 74, 29–43.

Levy, D., 2014. Production analysis with process mining technology. http://dx.doi.org/10.4121/UUID:68726926-5AC5-4FAB-B873-EE76EA412399, URL https://data.4tu.nl/articles/_/12697997/1.

Li, G., Huang, Q., Liu, C., Wang, G., Guo, L., Liu, R., Liu, L., 2024a. Fully automated diagnosis of thyroid nodule ultrasound using brain-inspired inference. Neurocomputing 582, 127497.

Li, H., Wang, C., Huang, Q., 2024b. Employing iterative feature selection in fuzzy rule-based binary classification. IEEE Trans. Fuzzy Syst..

Liao, Q.V., Zhang, Y., Luss, R., Doshi-Velez, F., Dhurandhar, A., 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 10. pp. 147–159.

Löfström, H., Löfström, T., Johansson, U., Sönströd, C., 2024. Calibrated explanations: With uncertainty information and counterfactuals. Expert Syst. Appl. 246, 123154.

Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya, U.R., 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Comput. Methods Programs Biomed. 226, 107161.

MacKay, D.J., 1995. Bayesian neural networks and density networks. Nucl. Instrum. Methods Phys. Res. Sect. A: Accel. Spectrometers, Detect. Assoc. Equip. 354 (1), 73–80.

Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C., 2014. Predictive monitoring of business processes. In: Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings 26. Springer, pp. 457–472.

Malinin, A., Gales, M., 2018. Predictive uncertainty estimation via prior networks. Adv. Neural Inf. Process. Syst. 31.

Márquez-Chamorro, A.E., Resinas, M., Ruiz-Cortés, A., 2017. Predictive monitoring of business processes: a survey. IEEE Trans. Serv. Comput. 11 (6), 962–977.

Marx, C., Park, Y., Hasson, H., Wang, Y., Ermon, S., Huan, L., 2023. But are you sure? an uncertainty-aware perspective on explainable ai. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 7375–7391.

Mehdiyev, N., Fettke, P., 2020. Local post-hoc explanations for predictive process monitoring in manufacturing. arXiv preprint arXiv:2009.10513.

Mehdiyev, N., Fettke, P., 2021. Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring. Interpret. Artif. Intell.: A Perspect. Granul. Comput. 1–28.

Mehdiyev, N., Majlatow, M., Fettke, P., 2023. Explainable artificial intelligence meets uncertainty quantification for predictive process monitoring.

Mehdiyev, N., Majlatow, M., Fettke, P., 2024a. Counterfactual explanations in the big picture: An approach for process prediction-driven job-shop scheduling optimization. Cogn. Comput. 1–27.

Mehdiyev, N., Majlatow, M., Fettke, P., 2024b. Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective. Ann. Oper. Res. 1–40.

Mohseni, S., Zarei, N., Ragan, E.D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4), 1–45.

Nakao, Y., Stumpf, S., Ahmed, S., Naseer, A., Strappelli, L., 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. ACM Trans. Interact. Intell. Syst. (TiiS) 12 (3), 1–30.

Osband, I., Blundell, C., Pritzel, A., Van Roy, B., 2016. Deep exploration via bootstrapped DQN. Adv. Neural Inf. Process. Syst. 29.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Adv. Neural Inf. Process. Syst. 32.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., et al., 2023. The role of explainable AI in the context of the AI act. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1139–1150.

Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., 2002. Inductive confidence machines for regression. In: Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13. Springer, pp. 345–356.

Portolani, P., Brusaferri, A., Ballarino, A., Matteucci, M., 2022. Uncertainty in predictive process monitoring. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Springer, pp. 547–559.

Rizzi, W., Di Francescomarino, C., Maggi, F.M., 2020. Explainability in predictive process monitoring: when understanding helps improving. In: International Conference on Business Process Management. Springer, pp. 141–158.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

Saeed, W., Omlin, C., 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl.-Based Syst. 263, 110273.

Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. J. Mach. Learn. Res. 9 (3).

Shoush, M., Dumas, M., 2022. When to intervene? prescriptive process monitoring under uncertainty and resource constraints. In: International Conference on Business Process Management. Springer, pp. 207–223.

Slack, D., Hilgard, A., Singh, S., Lakkaraju, H., 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. Adv. Neural Inf. Process. Syst. 34, 9391–9404.

Teinemaa, I., Dumas, M., Rosa, M.L., Maggi, F.M., 2019. Outcome-oriented predictive process monitoring: Review and benchmark. ACM Trans. Knowl. Discov. Data (TKDD) 13 (2), 1–57.

Tibshirani, R.J., Foygel Barber, R., Candes, E., Ramdas, A., 2019. Conformal prediction under covariate shift. Adv. Neural Inf. Process. Syst. 32.

Toccaceli, P., Gammerman, A., 2019. Combination of inductive mondrian conformal predictors. Mach. Learn. 108, 489–510.

Vale, D., El-Sharif, A., Ali, M., 2022. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. AI Ethics 2 (4), 815–826.

Van Der Aalst, W., 2012. Process mining. Commun. ACM 55 (8), 76–83.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic learning in a random world, Vol. 29. vol. 29, Springer.

Vovk, V., Petej, I., Nouretdinov, I., Manokhin, V., Gammerman, A., 2020. Computationally efficient versions of conformal predictive distributions. Neurocomputing 397, 292–308.

Watson, D., O'Hara, J., Tax, N., Mudd, R., Guy, I., 2024. Explaining predictive uncertainty with information theoretic shapley values. Adv. Neural Inf. Process. Syst. 36.

Weber, P., Carl, K.V., Hinz, O., 2024. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Manag. Rev. Q. 74 (2), 867–907.

Weytjens, H., De Weerdt, J., 2022. Learning uncertainty with artificial neural networks for predictive process monitoring. Appl. Soft Comput. 125, 109134.

Xu, C., Xie, Y., 2021. Conformal prediction interval for dynamic time-series. In: International Conference on Machine Learning. PMLR, pp. 11559–11569.

Yang, S., Yee, K., 2024. Towards reliable uncertainty quantification via deep ensemble in multi-output regression task. Eng. Appl. Artif. Intell. 132, 107871.