

Mehdiyev, Nijat; Majlatow, Maxim; Fettke, Peter

Article — Published Version

Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective

Annals of Operations Research

Provided in Cooperation with:

Springer Nature

Suggested Citation: Mehdiyev, Nijat; Majlatow, Maxim; Fettke, Peter (2024) : Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective, Annals of Operations Research, ISSN 1572-9338, Springer US, New York, NY, Vol. 347, Iss. 2, pp. 991-1030,
<https://doi.org/10.1007/s10479-024-05943-4>

This Version is available at:

<https://hdl.handle.net/10419/323301>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<http://creativecommons.org/licenses/by/4.0/>



Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective

Nijat Mehdiyev^{1,2} · Maxim Majlatow^{1,2} · Peter Fettke^{1,2}

Received: 3 April 2023 / Accepted: 11 March 2024 / Published online: 3 April 2024
© The Author(s) 2024

Abstract

In the rapidly evolving landscape of manufacturing, the ability to make accurate predictions is crucial for optimizing processes. This study introduces a novel framework that combines predictive uncertainty with explanatory mechanisms to enhance decision-making in complex systems. The approach leverages Quantile Regression Forests for reliable predictive process monitoring and incorporates Shapley Additive Explanations (SHAP) to identify the drivers of predictive uncertainty. This dual-faceted strategy serves as a valuable tool for domain experts engaged in process planning activities. Supported by a real-world case study involving a medium-sized German manufacturing firm, the article validates the model's effectiveness through rigorous evaluations, including sensitivity analyses and tests for statistical significance. By seamlessly integrating uncertainty quantification with explainable artificial intelligence, this research makes a novel contribution to the evolving discourse on intelligent decision-making in complex systems.

Keywords Explainable artificial intelligence (XAI) · Uncertainty quantification (UQ) · Predictive process monitoring · Information systems (IS)

1 Introduction

In today's highly competitive and complex business environment, organizations are under constant pressure to optimize their performance and decision-making processes. According to Herbert Simon, enhancing organizational performance relies on effectively channeling

✉ Nijat Mehdiyev
nijat.mehdiyev@dfki.de
Maxim Majlatow
maxim.majlatow@dfki.de
Peter Fettke
peter.fettke@dfki.de

¹ German Research Center for Artificial Intelligence (DFKI), Campus D 3.2, 66123 Saarbrücken, Saarland, Germany

² Saarland University, Campus D 3.2, Saarbrücken 66123, Saarland, Germany

finite human attention towards critical data for decision-making, necessitating the integration of information systems (IS), artificial intelligence (AI) and operations research (OR) insights (Simon, 1997). Recent OR research provides evidence in support of this proposition, as the discipline has witnessed a transformation due to the abundant availability of rich and voluminous data from various sources coupled with advances in machine learning (ML) (Frazzetto et al., 2019). As of late, heightened academic attention has been devoted to prescriptive analytics, a discipline that suggests combining the results of predictive analytics with optimization techniques in a probabilistic framework to generate responsive, automated, restricted, time-sensitive, and ideal decisions (Lepenioti et al., 2020).

The increasing prominence of data-driven solutions in operational decision-making is evident, particularly in the domain of manufacturing intelligence (Mehdizadeh and Fettke, 2021). One of the pivotal applications of this trend is the use of predictive analytics for production planning and scheduling. However, to harness its full potential to support production planners, certain limitations must be addressed. One of the significant gaps is that the majority of studies primarily concentrate on the prediction of non-technical parameters, such as demand and supply fluctuations. These parameters often serve as constraints or are integrated into the objective function of selected optimization frameworks. On the other hand, technical aspects inherent to the production process—like operational yield, production lead time, quality concerns, and potential system failures—remain underexplored (Chaari et al., 2014). This discrepancy can be attributed to the scarcity of pertinent data from information systems at the shop floor level. As a result, many current estimations regarding technical parameters are grounded more in intuition or assumptions rather than concrete data, leading to results that often fall short of optimal.

Another considerable gap in the field pertains to the output produced during the predictive analytics stage. Upon closer examination of studies that integrate data-driven parameter estimation before optimization, it becomes apparent that they overwhelmingly produce point forecasts (Mitrentsis and Lens, 2022). This approach, however, leads to the application of deterministic optimization methods, which may not fully capture the complexities and uncertainties of real-world scenarios. Despite the existence of numerous optimization methods that consider uncertainty, as proposed in Mula et al. (2006), their integration with the preceding predictive analytics stage has yet to be established. Consequently, a rising demand exists for the production of ML outputs that can precisely and comprehensively capture and quantify the predictive uncertainty within the specific operational research context being examined.

Lastly, even if uncertainty associated with the optimization parameter of interest can be estimated, a more actionable and advantageous approach would involve explaining its underlying source. This can be accomplished through an explainable artificial intelligence (XAI) approach that identifies input patterns that lead to uncertain predictions. By pinpointing the specific input features contributing to predictive uncertainty, practitioners can gain insights into regions where training data is sparse or where specific features exhibit anomalous behavior (Antorán et al., 2020). These insights would alert necessary adjustments to the model's decision-making process or outcomes prior to its subsequent operationalization.

To address the three identified gaps, we propose a multi-stage ML approach that incorporates uncertainty awareness and explainability. We demonstrate the effectiveness of this approach through its application to a real-world production planning scenario. The contribution of this study is multifaceted. To address the first gap, we use a supervised learning approach to probabilistically estimate a production-related parameter, specifically the processing time of production events. To achieve this objective, we employ process event data sourced from manufacturing execution systems (MES). These systems are process-aware information systems (PAIS) that facilitate the coordination of underlying operational pro-

cesses and capture the digital footprints of process events during execution. The resulting event log consists of sequentially recorded events associated with a particular case, along with various attributes such as timestamps, resources (human or machine) responsible for process execution, and other case-specific details. To be more precise, the problem at hand is formulated as a predictive process monitoring problem. This necessitates the utilization of specific pre-processing, encoding, and feature engineering techniques to account for the inherent business and operational process data requirements.

To tackle the second gap, we utilize Quantile Regression Forests (QRF), an ML method developed specifically to estimate conditional quantiles for high-dimensional predictor variables (Meinshausen, 2006). QRF is an extension of the traditional random forests technique, offering a non-parametric and precise approach for estimating prediction intervals. These prediction intervals provide valuable information on the uncertainty of model outcomes, allowing for a better understanding of the predictive power of the model and its limitations.

To bridge the third gap, we offer an explanation of the main drivers of uncertainty by examining the impact of feature values on prediction intervals. To accomplish this, we utilize local and global post-hoc explanations using SHapley Additive Explanations (SHAP) (Lundberg and Lee, 2017). Our approach differs from the state-of-the-art use of this technique in that we use the prediction interval width as the output, which provides a direct explanation of feature attributions to uncertainty. Furthermore, we refine our explanations on a granular level, such as for different uncertainty profiles or individual production activities, resulting in a more nuanced understanding of the underlying drivers of uncertainty.

The remainder of this paper is structured as follows: Sect. 2 outlines the real-world scenario that serves as the backdrop for our proposed method. Section 3 details the core methodology, while Sect. 4 describes the experimental design and evaluation metrics. Section 5 provides an exhaustive analysis of our method's performance. This is followed by Sect. 6, which delves into the practical and scientific implications of these results. Section 7 reviews pertinent literature, and Sect. 8 offers the final remarks.

2 Motivating usage scenario

In this section, we outline the production processes relevant to the collected data and clarify how the recommended methodology is applied in practice. This sets the stage for understanding the context. Importantly, the suggested approach for process prediction, which incorporates both uncertainty and explainability, is adaptable across various planning scenarios. This case study is part of a joint research project with a medium-sized German manufacturer specializing in custom and standardized vessel components. The production process involves multiple stages and utilizes materials such as stainless steel, aluminum, and carbon steel, requiring specialized equipment and expertise.

At the start of the manufacturing process, customer orders are sourced from the partner's product catalog. Once an order is received, the manufacturing firm assesses its priority and determines the required sequence of production activities, which may be either preset or slightly adjusted based on the customer's specifications. These specifications encompass a variety of attributes, such as *article group identifier*, *material group identifier*, *weight*, *bend radius*, *base diameter*, *sheet width*, *quantity*, and *welding specifications*. These factors significantly impact the time needed for each production activity. Despite possessing a sequence of activities for each customer order, process experts presently depend on intuition

Table 1 Production process event log

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

or experience-based estimations to ascertain their duration. This inability to quantify this vital time-specific production parameter results in suboptimal planning outcomes.

To address this issue, the partner has implemented an MES solution to capture the process execution details of production activities for each customer order. In our use case, the process data adheres to a particular structure. Each customer order is represented by a case with a unique case identifier, and a process case comprises the causal and temporal sequence of several events related to the production of the corresponding customer order. A process event encompasses the activity describing the production step executed, the start and end timestamps of execution, case attributes such as customer order specifications detailed above, and event-specific attributes like machine or human resources responsible. The examined use case involves 30 distinct activities, including *forming of material on dishing presses, manual welding, plasma welding, surface grinding, manual sanding, deburring of edges, etc.* All process execution data of historical customer orders are exported and stored in an event log. Table 1 presents an excerpt from the event log for illustrative purposes.

Using historical event data, experts can now accurately calculate the duration of each production activity, also known as event processing time. This is done by measuring the time difference between the start and end timestamps for each activity. In a highly competitive market, precise estimation of these processing times is crucial for effective planning. With this information, experts can forecast the total production time needed to complete an entire order. To better understand how event processing time relates to other time metrics commonly discussed in the field, please refer to Fig. 1. In this context, event processing time is a component of the overall cycle time, including waiting or idle periods.

Upon identifying the target parameter of interest, probabilistic machine learning solutions should be employed to generate data-driven estimations and corresponding uncertainty information. This is supplemented with relevant explanation mechanisms, allowing users to

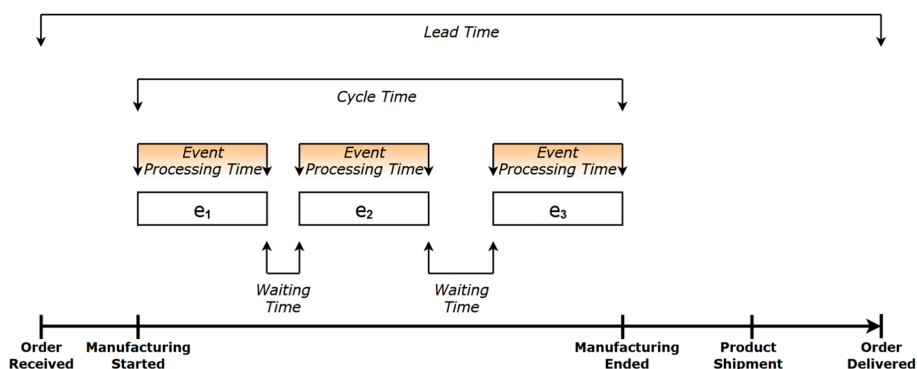


Fig. 1 Relationship of event processing time with other time concepts as described in Dumas et al. (2018)

comprehend the model uncertainty related to individual activity durations. The uncertainty-aware outputs are then utilized as input for decision augmentation scenarios or as input for the adopted optimization approach to generate production plans.

3 Methodology

The study introduces a method that integrates IS and AI to address challenges in OR. More specifically, the primary objective is to develop an ML-based solution using process event data from relevant information systems, namely MES (see Fig. 2). Important considerations in this regard include the need to quantify uncertainty and ensure explainability as part of the OR framework. This section outlines the methodology, which includes defining and preparing process event data for supervised learning, employing QRF for uncertainty quantification, constructing uncertainty profiles, and using the SHAP method for explaining predictive model uncertainty.

3.1 Process data preparation

This section describes the procedure for converting a process event log data from MES into a tabular dataset and formulating the duration prediction for each activity in the running traces as a supervised learning task. To accomplish this, it is crucial to identify the input variables from the examined running traces and align them with the respective target values. For clarity, we initially introduce notations and formal definitions for elements like events, event logs, traces, partial traces, and event duration, drawing from established literature (Polato et al., 2014; van der Aalst, 2016; Teinemaa et al., 2019).

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

- $a \in \mathcal{A}$ is the corresponding process activity;
- $c \in \mathcal{C}$ is the case *id*;
- $t_{start} \in \mathcal{T}_{start}$ is the start timestamp of the event (defined as seconds since 1/1/1970 which is a Unix epoch time representation);
- $t_{complete} \in \mathcal{T}_{complete}$ is the completion timestamp of the event;
- $v_1, \dots, v_n$ represents the list of event specific attributes, where $\forall 1 \leq i \leq n : v_i \in \mathcal{V}_i$, $\mathcal{V}_i$ denoting the domain of the i^{th} attribute.

1) Process Data Preparation

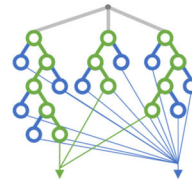
Original Event Log

Case	Activity	Start	End	Quantity	Weight
100001	Processing	2023-01-01 08:00:00	2023-01-01 08:05:00	1	100
100001	Waiting	2023-01-01 08:05:00	2023-01-01 08:10:00	1	100
100001	Processing	2023-01-01 08:10:00	2023-01-01 08:15:00	1	100
100001	Waiting	2023-01-01 08:15:00	2023-01-01 08:20:00	1	100
100001	Processing	2023-01-01 08:20:00	2023-01-01 08:25:00	1	100

Processed Event Log

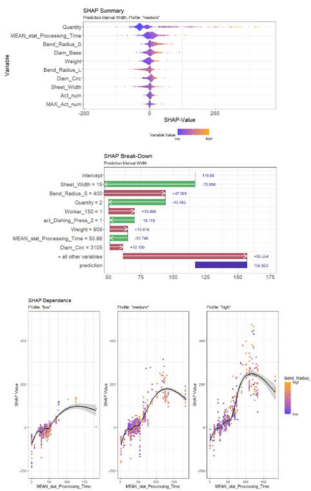
Case	Activity	Start	End	Quantity	Weight	Assessment	Assessment	Assessment	Assessment
100001	Processing	2023-01-01 08:00:00	2023-01-01 08:05:00	1	100	Processing	Processing	Processing	Processing
100001	Waiting	2023-01-01 08:05:00	2023-01-01 08:10:00	1	100	Waiting	Waiting	Waiting	Waiting
100001	Processing	2023-01-01 08:10:00	2023-01-01 08:15:00	1	100	Processing	Processing	Processing	Processing
100001	Waiting	2023-01-01 08:15:00	2023-01-01 08:20:00	1	100	Waiting	Waiting	Waiting	Waiting
100001	Processing	2023-01-01 08:20:00	2023-01-01 08:25:00	1	100	Processing	Processing	Processing	Processing

2) QRF Model Training

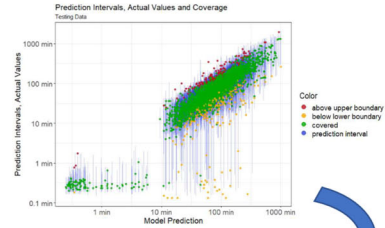


Point Prediction Prediction Interval

5) SHAP-Analysis



3) Prediction Intervals



4) Uncertainty Profiles

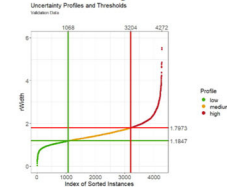


Fig. 2 Overview of the proposed uncertainty explainability approach

Consequently, $\mathcal{E} = \mathcal{A} \times \mathcal{C} \times \mathcal{T}_{start} \times \mathcal{T}_{complete} \times \mathcal{V}_1 \times \dots \times \mathcal{V}_n$ is defined as the universe of events. Moreover, we define the following project functions given the event $e \in \mathcal{E}$:

- $p_a: \mathcal{E} \rightarrow \mathcal{A}, p_a(e) = a,$
- $p_c: \mathcal{E} \rightarrow \mathcal{C}, p_c(e) = c,$
- $p_{t_{start}}: \mathcal{E} \rightarrow \mathcal{T}_{start}, p_{t_{start}}(e) = t_{start},$
- $p_{t_{complete}}: \mathcal{E} \rightarrow \mathcal{T}_{complete}, p_{t_{complete}}(e) = t_{complete},$
- $p_{v_i}: \mathcal{E} \rightarrow \mathcal{V}_i, p_{v_i}(e) = v_i, \forall 1 \leq i \leq n$

Definition 2 (Traces and Event Log) A trace $\sigma \in \mathcal{E}^*$ is a finite sequence of events $\sigma_c = \langle e_1, e_2, \dots, e_{|\sigma_c|} \rangle$, for which each $e_i \in \sigma$ occurs no more than once and $\forall e_i, e_j \in \sigma, p_c(e_i) = p_c(e_j) \wedge p_{T_S}(e_i) \leq p_{T_S}(e_j)$, if $1 \leq i < j < |\sigma_c|$. The event log $\mathcal{E}_C$ is defined as a set of completed traces, $\mathcal{E}_C = \{\sigma_c \mid c \in \mathcal{C}\}$.

Definition 3 (Partial Traces) Two options are given for obtaining partial traces, depending on the predictive process monitoring use case. Defined over σ the following $hd^i(\sigma_c)$ and $tl^i(\sigma_c)$ generate the prefixes and suffixes respectively as follows:

- selection operator ($\cdot$): $\sigma_c(i) = \sigma_i, \forall 1 \leq i \leq n;$
- $hd^i(\sigma_c) = \langle e_1, e_2, \dots, e_{\min(i,n)} \rangle$ for $i \in [1, |\sigma_c|] \subset \mathbb{N}$
- $tl^i(\sigma_c) = \langle e_w, e_{w+1}, \dots, e_n \rangle$ where $w = \max(n - i + 1, 1);$

- $|\sigma| = n$ (i.e. the cardinality or length of the trace).

We denote the set of partial traces generated by the $tl^i(\sigma_c)$ function as γ . These partial traces form the basis for constructing a tabular dataset that predicts the duration of the remaining events in a running trace. To elucidate, our proposed approach in this study is performed before initiating a case. Nonetheless, the model and data structure can be adapted to accommodate updates following each event within the running traces, if needed. Hence, the application of the $tl^i(\sigma_c)$ function proves pertinent for shaping the training data structure.

In the literature of predictive process monitoring, various process performance indicators (PPIs) often serve as targets of interest. These targets can range from cost and quality metrics to time-related indicators. The proactive analysis of such time-based metrics is instrumental in enhancing both the operational and strategic capabilities of organizations. In this study, the focus is on the processing time of an event, which is defined as the duration of the corresponding activity. This duration is computed as the difference between the completion and start timestamps of the event:

Definition 4 (*Event Processing Time/Labeling*) Given a non-empty trace $\sigma \neq \langle \rangle \in \mathcal{E}^*$, a labeling function $resp : \mathcal{E} \rightarrow \mathcal{Y}$, also referred to as annotation function, maps an event $e \in \sigma$ to the corresponding value of its response variable $resp(e) \in \mathcal{Y}$. We define the event processing time as our response variable, calculated as follows:

$$resp(e) = p_{t_{complete}}(e) - p_{t_{start}}(e), \quad (1)$$

with the domain of the defined response variable being $\mathcal{Y} \subset \mathbb{R}^+$.

Definition 5 (*Feature Extraction*) The feature extraction function in this study is defined as a function $feat : \mathcal{E}^* \rightarrow \mathcal{X}^*$ which extracts the feature values from a given non-empty trace $\sigma \neq \langle \rangle \in \mathcal{E}^*$, with $\mathcal{X} \in \mathbb{R}^{dim}$ denoting the domain of the features and dim being the input dimension. For a given trace $\sigma_c = \langle e_1, e_2, \dots, e_{|\sigma_c|} \rangle$, the feature extraction function $feat$ generates a set of features $(x_{i,1}, \dots, x_{i,dim})$ for each event e_i . In addition to case-specific and event-specific feature values, the feature extraction function enables the retrieval of intra-case-specific features, such as n-grams.

For a set of predictor variables $\mathbf{x} = (x_1, \dots, x_p)$ and a response variable y , we define $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ as the dataset of associated $(\mathbf{x}, y)$ values, with N denoting total amount of observations. The dataset is split into three parts: $D = D_{train} \cup D_{val} \cup D_{test}$, in order to use D_{train} for training, D_{val} for hyperparameter optimization and as a calibration set to derive uncertainty profiles (see Sect. 5.2), and D_{test} for evaluation, with N_{train} , N_{val} , N_{test} being the respective amount of instances in each subset.

3.2 Interval prediction with quantile regression forests

Random Forests (RF) is a robust machine learning method that captures complex non-linear relationships in data (Breiman, 2001). The primary objective of RF is to produce accurate predictions by forming an ensemble of decision trees during the training phase. The model then outputs either the mode of the classes for classification tasks or the mean prediction for regression tasks based on these individual trees. However, the standard RF approach does not delve into the complete conditional distribution of the response variable.

For a comprehensive understanding of such distributions, QRF was introduced as an extension of RF (Meinshausen, 2006). This approach aims to estimate conditional quantiles,

offering deeper insights into both the distribution of the data and the uncertainty associated with predictions. The capability to produce prediction intervals renders QRF valuable across multiple OR applications. In these contexts, accurately portraying predictive uncertainties is vital for informed decision-making and strategic planning.

In accordance with (Meinshausen, 2006), let θ be a random parameter vector guiding tree growth within the RF, $T(\theta)$ the associated tree, $\mathcal{B}$ the space of a data point X with dimensionality p and $R_\ell \subseteq \mathcal{B}$ be a rectangular subspace for any leaf ℓ of every tree within the RF. Any $x \in \mathcal{B}$ can be allocated to exactly one leaf of any tree of the RF such that $x \in R_\ell$, thus denoting the specific leaf of the corresponding tree via $\ell(x, \theta)$. First, the weight function $w_i(x, \theta)$ is defined for each observation i and tree $T(\theta)$, given by

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{\ell(x, \theta)}\}}}{\#\{j : X_j \in R_{\ell(x, \theta)}\}}, \quad (2)$$

where $1_{\{X_i \in R_{\ell(x, \theta)}\}}$ is an indicator function that equals to 1 if the observation i falls in the leaf node corresponding to $\ell(x, \theta)$, and $\#\{j : X_j \in R_{\ell(x, \theta)}\}$ is the number of observations that fall in that same leaf node.

Next, the weights are averaged over multiple trees to obtain the final weight function $w_i(x)$:

$$w_i(x) = k^{-1} \sum_{t=1}^k w_i(x, \theta_t), \quad (3)$$

where k is the number of trees and θ_t represents the t -th tree.

Finally, the estimated conditional quantile function $\widehat{F}(y \mid X = x)$ is calculated as a weighted sum of the indicator function $1_{\{Y_i \leq y\}}$ for each observation i , where Y_i is the response variable:

$$\widehat{F}(y \mid X = x) = \sum_{i=1}^n w_i(x) 1_{\{Y_i \leq y\}}. \quad (4)$$

Using this estimated conditional quantile function, the quantile function $Q_\alpha(x)$ for a given quantile level α can be obtained as

$$Q_\alpha(x) = \inf\{y : \widehat{F}(y \mid X = x) \geq \alpha\}. \quad (5)$$

Based on the quantile function $Q_\alpha(x)$, prediction intervals of specific levels can be derived using

$$I_{1-2\alpha}(x) = [Q_\alpha(x), Q_{1-\alpha}(x)], \quad (6)$$

For the prediction of the conditional mean, as yielded by regular RF, QRF allow the utilization of

$$\widehat{F}(y \mid X = x) = \sum_{i=1}^n w_i(x) Y_i. \quad (7)$$

In the field of OR, prediction intervals serve as a robust tool for quantifying uncertainty in diverse decision-making scenarios. Unlike standard prediction models, which provide only a single-point estimate, prediction intervals offer a range of possible outcomes. This range enhances the reliability of decisions by considering the inherent variability in the underlying model. For instance, prediction intervals may serve multiple purposes in OR: they can evaluate the probability of exceeding specific thresholds, function as parameters in

robust optimization models, or assess risks across various decision options. Incorporating these intervals into the decision-making framework enhances predictive accuracy and, as a result, optimizes operational efficiency.

3.3 Explanation of predictive uncertainty with SHAP

Post-hoc XAI techniques serve as powerful instruments for illuminating the decisions made by machine learning algorithms. According to Lipovetsky and Conklin (2001), local methods aim to elucidate a given prediction model by employing a simpler explanation model g . In this framework, the explanation model g utilizes a simplified set of inputs, x' , which are mapped to the original inputs through a specific function $x = h_x(x')$. Local methods endeavor to approximate $f(h_x(z'))$ by $g(z')$ when z' is closely related to x' . In the context of additive feature attribution methods, the explanation model g is represented as a linear combination of M binary variables z'_i . The equation is as follows:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i \quad (8)$$

where M signifies the total number of simplified input features, and the coefficients ϕ_i are real numbers indicating the impact of each feature. Summation of these effects offers an approximation of the examined prediction $f(x)$.

Various methods are available for addressing additive feature attribution, such as LIME, DeepLift, among others (Zhou et al., 2022). Each technique contributes unique insights, enriching the academic dialogue in the field. In this study, we utilize the SHAP methodology, which is underpinned by robust theoretical foundations. Before delving into the specifics of the SHAP method, it is important to articulate the shift in our explanatory objective for this study.

Rather than focusing on the point prediction given by the predictive model $f(x)$, our attention in this study is centered on elucidating the predictive uncertainty as quantified by the width of the prediction interval $W(x)$:

$$W(x) = Q_{1-\alpha}(x) - Q_{\alpha}(x) \quad (9)$$

In this equation, $W(x)$ is the function responsible for calculating the width of the prediction interval for a specific input x . The function $Q_{1-\alpha}(x)$ determines the upper $(1 - \alpha)$ quantile of the prediction distribution corresponding to x , while $Q_{\alpha}(x)$ calculates its lower α quantile. Essentially, $W(x)$ computes the difference between these upper and lower quantiles, serving as an indicator of predictive uncertainty. A wider prediction interval implies greater uncertainty, whereas a narrower one suggests higher model confidence. By investigating the influence of various features on these prediction intervals, we aim to provide insights into the factors affecting predictive uncertainty.

A widely-adopted method for analyzing the role of each feature in machine learning output is through the use of Shapley regression values. In our specific focus on predictive uncertainty, we evaluate the influence of a specific feature i on $W(x)$ by retraining the model for every subset S taken from $F \setminus \{i\}$, where F is the full set of all features. Two distinct models are thus derived: $W_{S \cup \{i\}}$ and W_S , which include and exclude the feature i , respectively. The difference $W_{S \cup \{i\}}(x_{S \cup \{i\}}) - W_S(x_S)$ identifies the influence of feature i on the prediction interval width. Given the potential for varying interactions between i and other features in S , all feasible subsets $S \subseteq F \setminus \{i\}$ are considered. The Shapley values for $W(x)$ are computed according to

the following equation:

$$\phi_i^W = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [W_{S \cup \{i\}}(x_{S \cup \{i\}}) - W_S(x_S)] \quad (10)$$

Shapley regression values provide a mathematically rigorous means for feature attribution. However, their computational intensity can be a limiting factor, especially for complex, high-dimensional models. This bottleneck led to the development of SHAP, a method designed to approximate Shapley values more efficiently (Lundberg and Lee, 2017). SHAP leverages various algorithmic optimizations, such as kernel approximations and tree-based algorithms, to substantially reduce computational cost. In doing so, it bridges the gap between the theoretical rigor of Shapley regression values and the practical necessity for computational efficiency. The SHAP values for $W(x)$ are obtained through a mathematical framework that satisfies local accuracy, missingness, and consistency:

$$\phi_i(W, x) = \sum_{z' \subseteq x'} \frac{|z'|! (M - |z'| - 1)!}{M!} [W_x(z') - W_x(z' \setminus i)], \quad (11)$$

where $W_x(z') = W(h_x(z'))$ signifies the conditional expectation of $W(x)$, conditioned on z' .

In our study, we have used KernelSHAP, a specific variant of the SHAP method. KernelSHAP is non-parametric and model-agnostic, which can be applied to any machine learning model. This approach employs a kernel-based approximation to estimate Shapley values, significantly reducing computational time while maintaining a high level of accuracy:

$$\begin{aligned} \Omega(g) &= 0, \\ \pi_{x'}(z') &= \frac{(M-1)}{(M \text{ choose } |z'|) |z'| (M - |z'|)}, \\ L(W, g, \pi_{x'}) &= \sum_{z' \in Z} [W(h_x^{-1}(z')) - g(z')]^2 \pi_{x'}(z'), \end{aligned} \quad (12)$$

In summary, this study employs a comprehensive approach to feature attribution, utilizing the SHAP methodology to examine predictive uncertainty. Rather than limiting the scope to traditional additive feature attribution methods, we expand our focus to quantify how each variable influences the level of uncertainty in model predictions. The uncertainty arises from the model's inherent limitations in capturing the complex, possibly non-linear, relationships between the features and the target variable. This complexity is compounded by the presence of noise in the data, potential interactions between features, and other unobserved variables that are not part of the model. Therefore, the uncertainty is not a property of the features themselves but is a manifestation of the model's limitations and the inherent variability in the data. By employing SHAP, we aim to decompose this predictive uncertainty into contributions from each feature. This allows us to understand which features are most responsible for the uncertainty in our predictions. Such insights are invaluable for both model refinement and for guiding decision-making processes in operational settings. For instance, if a particular feature is found to contribute significantly to predictive uncertainty, efforts could be made to obtain more accurate or additional data for that feature, thereby potentially reducing the uncertainty. Our goal is not merely to correlate the uncertainty source with the SHAP values but to dissect the predictive uncertainty into understandable components tied to each feature. This nuanced understanding aids in improving the model and provides actionable insights for decision-makers.

4 Experiment settings

This section outlines the dataset used, the software tools employed for implementing our proposed approach, the methodology for hyperparameter optimization in QRF model, and the metrics chosen for evaluating model performance.

4.1 Dataset overview

Building upon the use case outlined in Sect. 2, the manufacturing partner provided an event log. This log included a variety of cases and within those cases, activities, durations, and additional data pertinent to specific tasks. Following a thorough preprocessing phase, a more refined dataset was produced. This refined dataset maintains the fundamental features of the original event log and serves as a strong foundation for both data analysis and the creation of predictive models. The primary goal of this use case is to accurately predict the time duration required for the processing of individual activities, while explicitly excluding any waiting times. To achieve this goal, the preprocessing strategy consisted of several specific steps:

- Removal of outliers, such as cases or events that either lack specifications or have an end timestamp that occurs before the start timestamp.
- Application of one-hot encoding to categorical variables, which include the type of activity performed or the identifiers of the resources involved.
- Aggregation of events that are related to the same activity into a single event. This includes summing up the processing durations and involved resources. In this stage, one-hot encoding is also expanded to include additional categorical variables such as the machinery used.
- Feature engineering to create new variables. These new variables include preceding and following activities, the processing duration of the prior activity, and statistical measures related to the processing time of the current activity.
- Iterative feature selection in conjunction with model training and evaluation to eliminate variables that do not contribute effectively to predictive accuracy.

Table 2 provides summary statistics for the used event log. The initial dataset encompasses 11,943 cases, featuring a collective of 48,577 events related to 30 distinct activities. The dataset exhibited a mean processing time of 96.9 min and a standard deviation of 135.5 min. Additionally, the mean trace length was calculated to be 4.06 events per case. Following the aggregation, cleansing, and feature engineering processes, the finalized dataset comprises 200 predictor variables. These variables are broken down into 42 trace-specific and 158 event-specific variables, accompanied by a target variable that signifies the event processing time.

The dataset was initially partitioned into training and test sets using a chronological separation of traces with a ratio of 92.5% for training and 7.5% for testing. This approach ensures that all process events associated with a single production case remain within the same data partition. Such a method aligns well with the use-case requirements, particularly in the context of model deployment. By adhering to this partitioning strategy, the integrity of production cases is preserved, thereby facilitating more reliable and relevant model training and evaluation. Details concerning the training and test set splits are also available in Table 2.

Table 2 Summary of dataset characteristics

	Train	Test	Complete Data Set
Number of events	45,188	3,389	48,577
Number of traces	11,051	892	11,943
Unique activities	30	27	30
Mean processing time (min)	97.2	93.3	96.9
Std. deviation of processing time (min)	137.4	106.5	135.5
Mean trace length	4.09	3.8	4.06

Table 3 Hyperparameter optimization settings for BART, DT, LR, QRF and XGBoost models

Model	Parameter	Range	Model	Parameter	Range
BART	prior_outcome_range	[0, 5]	QRF	min_n	[2, 32]
	prior_terminal_node_expo	[1, 3]		mtry	[40, 90]
	trees	[1, 2000]		trees	[50, 100]
DT			XGBoost	learn_rate	(0, 0.1)
	cost_complexity	[0, 0.1]		loss_reduction	[0, 30]
	min_n	[2, 32]		min_n	[2, 32]
	tree_depth	[5, 30]		mtry	[2, 200]
				sample_size	(0, 1)
LR				trees	[1, 2000]
	penalty	$[10^{-5}, 10^5]$		tree_depth	[5, 30]

4.2 Model training and hyperparameter optimization

To assess the performance of the predictive models, a comparative analysis was conducted involving QRF, Bayesian Additive Regression Trees (BART), Decision Trees (DT), Linear Regression (LR), and eXtreme Gradient Boosting (XGBoost). DT and LR were chosen for their inherent interpretability, while BART, QRF, and XGBoost were included as black-box models due to their superior performance with tabular datasets. Hyperparameter optimization was employed to identify the optimal model settings for the regression task at hand.

Table 3 presents the range of hyperparameters explored for each model type. A parameter grid was formulated for each model by randomly sampling from the hyperparameter search space, restricted to a total of 216 permutations of parameter values per model. This number of permutations allowed for a reasonable approximation of optimal settings, balancing accuracy with computational efficiency.

The hyperparameter tuning process leveraged 10-fold cross-validation, in line with best practices for credible model comparison (Molinaro et al., 2005). Moreover, an overlapping sliding window approach was adopted to align with the project's use-case requirements. Specifically, the training dataset was chronologically divided into eleven equal segments, based on case identifiers. Each cross-validation fold consisted of adjacent segments serving as the training and validation sets, with each set containing roughly 1,000 unique traces. In this manner, the validation set from one fold was used as the training set for the subsequent

fold. Consequently, each of the 216 hyperparameter combinations underwent optimization once for each cross-validation fold. This rigorous approach to hyperparameter optimization ensured both robustness and relevance in the model comparison, satisfying both analytical and computational constraints.

4.3 Evaluation metrics

To gauge the efficacy of the selected machine learning techniques, two key performance indicators were examined: point prediction accuracy and model uncertainty. Furthermore, statistical significance tests were employed to facilitate a rigorous comparison between the various models. Point prediction focuses on the model's ability to accurately forecast specific outcomes, while model uncertainty assesses the reliability of these predictions. Statistical significance tests offer an additional layer of validation, determining whether observed performance differences between models are indeed meaningful.

4.3.1 Point prediction

Regarding point predictions, multiple metrics exist for evaluating the quality of regression models. These metrics aim to quantify the prediction error relative to the actual ground truth values. In this study, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) serve as the selected evaluation criteria.

The RMSE is calculated by taking the square root of the average squared differences between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (13)$$

However, as argued by (Willmott and Matsuura, 2005), RMSE can be influenced significantly by outliers or the distribution of error magnitudes, potentially making it less reliable for assessing model quality in certain scenarios. To address this, MAE serves as an alternative metric. The MAE is obtained by averaging the absolute differences between the predicted and actual values:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (14)$$

Building on the insights from (Chai and Draxler, 2014), which asserts that RMSE is a more dependable metric when errors follow a Gaussian distribution, this study will employ both RMSE and MAE. Utilizing these two metrics provides a more comprehensive evaluation of model performance, each compensating for the other's limitations.

4.3.2 Uncertainty quantification

The second dimension of evaluation focuses on the soundness of uncertainty quantification. Here, we employ metrics that measure how well the generated prediction intervals capture the actual values. Specifically, we use Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) as suggested by Shrestha and Solomatine (2006). Additionally, we incorporate Mean Relative Prediction Interval Width (MRPIW) based on the work by Jørgensen et al. (2004), Klas et al. (2011).

PICP quantifies the proportion of actual target values encompassed by the prediction intervals. It is calculated as the ratio of instances where the actual target value falls within the prediction interval to the total number of instances:

$$PICP = \frac{1}{N} \sum_{i=1}^N c_i, \quad c_i = \begin{cases} 1, & y_i \in [L_i, U_i] \\ 0, & y_i \notin [L_i, U_i] \end{cases} \quad (15)$$

MPIW is another metric that helps assess the quality of prediction intervals. It is obtained by averaging the widths of these intervals, specifically the distances between their upper and lower limits:

$$MPIW = \frac{1}{N} \sum_{i=1}^N (U_i - L_i) \quad (16)$$

However, since MPIW provides an absolute measure of interval width, Jørgensen et al. (2004), Klas et al. (2011) recommend using a metric that considers the relative width of prediction intervals in relation to point estimates. Thus, MRPIW is introduced. It averages the relative widths (rWidth) of intervals, which are calculated as the ratio of the prediction interval width to the corresponding point estimate:

$$MRPIW = \frac{1}{N} \sum_{i=1}^N rWidth_i, \quad rWidth_i = \frac{(U_i - L_i)}{\hat{y}_i} \quad (17)$$

According to Jørgensen et al. (2004), a lower MRPIW value suggests reduced model uncertainty, given a constant PICP. To holistically assess model uncertainty, this study will use all three metrics: PICP, MPIW, and MRPIW.

4.3.3 Statistical significance tests

In machine learning assessment, the Friedman–Nemenyi test functions as a compelling instrument for evaluating the efficacy of multiple algorithms, especially when these algorithms are gauged using cross-validation techniques. Cross-validation provides a dependable measure for the generalizable performance of various machine learning models. Incorporating the cross-validation scores within the Friedman–Nemenyi analytical framework results in a robust and well-rounded evaluation.

The Friedman test ranks the machine learning algorithms based on their average performance metrics across each fold of the cross-validation process (Friedman, 1937; Milton, 1939). This leads to the computation of the Friedman statistic (χ_F^2), which is subsequently used for comparing and evaluating the performance of the algorithms under study:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_{j=1}^k R_j^2 - \frac{k(k+1)^2}{4} \right] \quad (18)$$

where N is the number of cross-validation folds, k is the number of machine learning approaches being assessed, and R_j is the sum of the ranks for the j^{th} approach across all folds. A higher χ_F^2 value indicates a statistically significant difference in the performance of the evaluated methods.

Should the Friedman test indicate statistical significance, the subsequent step is to apply the Nemenyi post-hoc test to compute the critical difference (CD) (Nemenyi, 1963; Garcia

and Herrera, 2008). The (CD) is calculated using the formula:

$$CD = q \sqrt{\frac{k(k+1)}{6N}} \quad (19)$$

In this context, CD is the benchmark for discerning whether the performance disparities between any two methodologies are statistically significant, based on their average rankings.

Utilizing the Friedman–Nemenyi test suite offers multiple benefits for model evaluation. Firstly, its non-parametric characteristic ensures that the assessment remains robust even when the data distribution is non-normal. Secondly, the method’s inherent capability to simultaneously compare multiple machine learning algorithms provides a comprehensive evaluation landscape. Finally, incorporating a post-hoc test reduces the likelihood of committing Type I errors, a crucial aspect when conducting multiple comparisons. Collectively, these features make the Friedman–Nemenyi testing framework a statistically sound and computationally efficient tool for identifying the most effective machine-learning approach for a specific task.

4.4 Software tools

For the tasks of data processing and feature engineering, the “*tidyverse*” collection of libraries was employed, with a particular emphasis on the “*dplyr*” library. The QRF model was implemented in R, primarily using the “*tidymodels*” and “*ranger*” libraries. These libraries also facilitated the comparative assessment of other models, such as XGBoost. Additionally, the libraries employed included “*glmnet*” for linear regression, “*rpart*” for decision trees, and “*dbarts*” for Bayesian Additive Regression Trees (BART). In the realm of model evaluation, “*PMCMR*” and “*PMCMRplus*” were used for conducting significance tests. The calculation of SHAP values was executed through the “*kernelshap*” library, in collaboration with “*doParallel*” for parallel computing. For visualization purposes, “*ggplot2*,” “*ggstatsplot*,” and “*ggbeeswarm*” were the primary libraries employed. This array of specialized libraries enabled a comprehensive approach to data preparation, model implementation, evaluation, and visualization, aligning well with the project’s analytical and predictive objectives.

5 Results

This section analyzes the results of the proposed approach. First, the results of the model evaluation are presented, entailing the construction of a baseline for comparative model analysis, the analysis of the sensitivity of hyperparameters for the QRF model as well as the evaluation of soundness with regards to UQ. Next, the construction of uncertainty profiles is presented, followed by a description of key findings. Lastly, an analysis of model uncertainty utilizing SHAP is conducted, comprising examinations on a local and global level as well as on the basis of uncertainty profiles.

5.1 Model evaluation

This subsection presents a detailed examination of several facets pertaining to model performance. It commences by introducing a baseline model. This baseline acts as a comparative measure and is derived from the manufacturer’s current methodology for estimating processing times, as introduced in Sect. 2. Following establishing the baseline, the subsection

transitions into a comparative analysis focused on point predictions. This evaluation leverages the outcomes from the hyperparameter optimization phase to assess the efficacy of the top-performing models. Subsequently, the section explores hyperparameter sensitivity for the QRF model. This part aims to ascertain the extent to which exhaustive hyperparameter tuning is requisite for achieving optimal performance. Finally, the subsection wraps up with an analysis targeting model reliability in terms of uncertainty. This is conducted by juxtaposing the performances of QRF and BART, specifically in their ability to quantify uncertainty.

5.1.1 Comparative analysis for point predictions

As outlined in Sect. 2, the process planners employ domain expertise and historical data to estimate processing times for individual production steps. Despite incorporating various factors such as event-specific details and historical cases, these estimates have been found to be unreliable in multiple instances. Documented in the manufacturer's planning data, these estimates serve as a baseline for our point prediction evaluation. For comparative analysis, we focus on models optimized through hyperparameter tuning, as discussed in Sect. 4.2. The settings for each optimized model are summarized in Table 4.

Table 5 offers a side-by-side comparison of model performance for point prediction, presenting validation and test datasets. Average results and standard deviations are presented for the validation data. The baseline performance, represented by MAE values of 55.6 and 51.9 and RMSE values of 117.8 and 99.5 for the validation and test datasets respectively,

Table 4 Hyperparameter optimization results for BART, DT, LR, QRF and XGBoost models

Model	Parameter	Value	Model	Parameter	Value
BART	prior_outcome_range	5	QRF	min_n	20
	prior_terminal_node_expo	1		mtry	70
	trees	1000		trees	100
DT			XGBoost	learn_rate	0.0163
	cost_complexity	2.51e−05		loss_reduction	0.164
	min_n	32		min_n	6
	tree_depth	15		mtry	99
LR				sample_size	0.923
	penalty	1.17		trees	1832
				tree_depth	12

Table 5 Comparative analysis for model performance (point prediction)

Dataset and metric	Baseline	BART	DT	LR	QRF	XGBoost
Validation data						
MAE	55.6 ± 6.81 ^a	38.5 ± 4.58	44.5 ± 6.06	43 ± 5.38	37.4 ± 4.91	36.5 ± 4.73
RMSE	117.8 ± 11.1	76.1 ± 10	95 ± 13.3	80.1 ± 12.4	81.4 ± 11.6	78.3 ± 11.1
Test data						
MAE	51.9	36.3	41.5	43.6	35.51	33.4
RMSE	99.5	64.7	80.5	75.9	66.88	64

^aStandard Deviation across Cross-Validation Folds

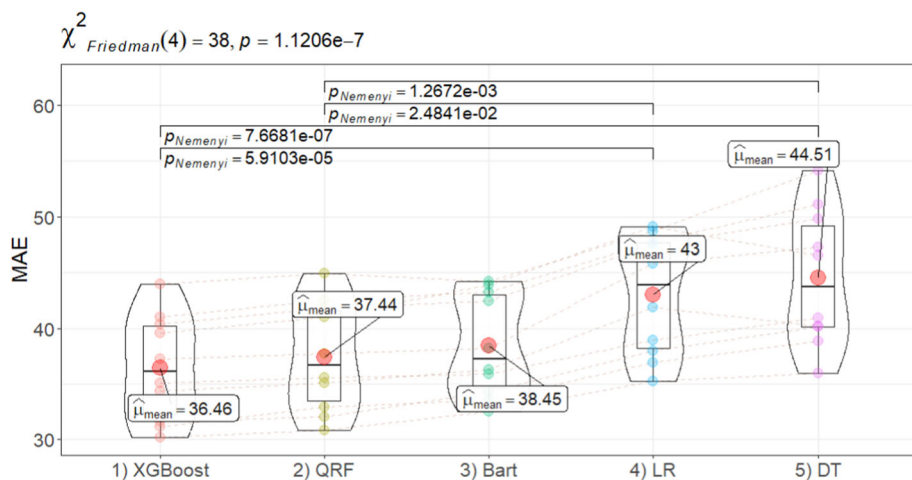


Fig. 3 Friedman–Nemenyi test results for ML model comparison

serves as our reference point. Across all evaluation metrics, every ML model tested surpassed this baseline. Noteworthy among these are the QRF and XGBoost models, which stand out as the most effective. XGBoost achieves the lowest MAE and RMSE values across both datasets, recording MAE scores of 35.5 and 33.4, and RMSE scores of 78.3 and 64. These figures are closely followed by the QRF model, which posts MAE and RMSE values of 37.4 and 35.51, and 81.4 and 66.88, respectively.

Figure 3 presents the results of the Friedman–Nemenyi test for predictive models BART, DT, LR, QRF, and XGBoost, utilizing a 10-fold overlapping sliding window cross-validation. The Friedman test produces a chi-squared value of 38 and a markedly small p-value of $1.1206e-7$, signifying significant performance disparities among the models under consideration. To deepen our understanding, we applied the Nemenyi test for pairwise comparisons. Figure 3 illustrates these comparisons, where the presence of connecting lines between models denotes statistically significant differences in performance. The test confirms that the performance advantages of QRF and XGBoost over inherently interpretable models are statistically significant, not merely coincidental.

Notably, the absence of a line between XGBoost and QRF implies their performance difference is statistically negligible. Although XGBoost records a slightly lower MAE than QRF, this discrepancy is not statistically meaningful. Consequently, both models are considered statistically equivalent for this specific analysis. In practical terms, the choice between QRF and XGBoost should hinge on other factors, such as QRF's ability to quantify prediction uncertainty or XGBoost's scalability and diverse performance advantages. In our context, QRF's inherent capability to quantify model uncertainty makes it the preferable option over XGBoost.

5.1.2 Sensitivity analysis of hyperparameter optimization for QRF

The sensitivity of the QRF model to various hyperparameter settings is examined in Fig. 4. This figure presents mean MAE values for each hyperparameter configuration of the QRF model across all validation folds, as outlined in Sect. 4.2. Vertical bars in the figure indicate the corresponding standard deviations. The model demonstrates a robust performance with

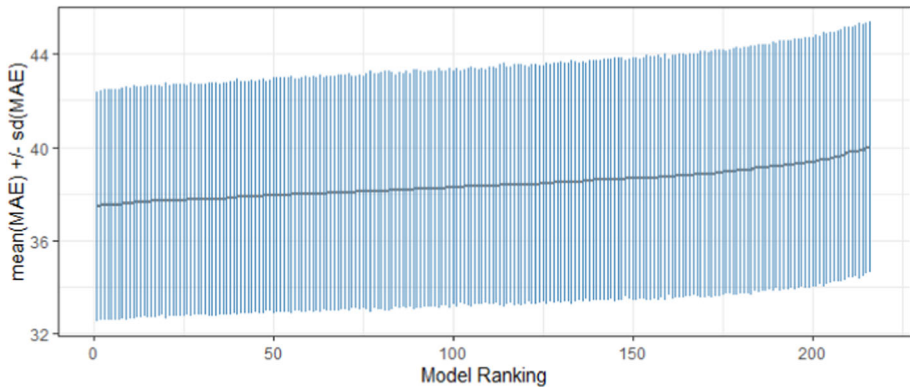


Fig. 4 Sensitivity analysis of hyperparameter setting for QRF based on 10-fold overlapping sliding window cross validation

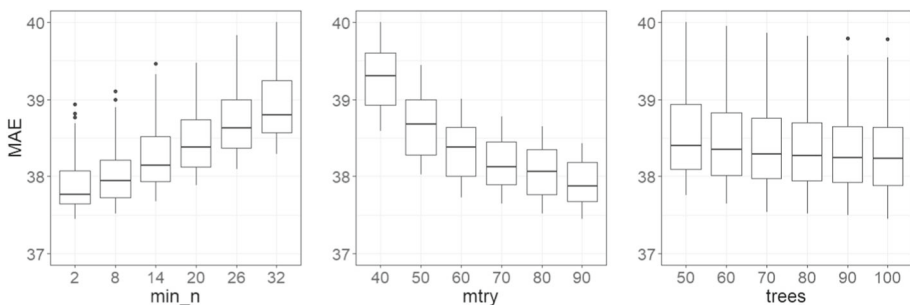


Fig. 5 Sensitivity analysis of hyperparameter setting for QRF based on 10-fold overlapping sliding window cross validation for each of the metrics *min_n*, *mtry*, *trees*

a relatively narrow gap of approximately 2.5 min between the best and worst outcomes. Additionally, the standard deviation fluctuates within a modest range of 4.9 to 5.5 min across the validation folds. Overall, the QRF model shows commendable stability with respect to hyperparameter variations.

Figure 5 displays a boxplot evaluating the performance of the QRF model based on MAE values across different hyperparameter settings. The hyperparameters under consideration include the number of variables randomly sampled at each split (*mtry*), the total number of trees in the forest (*trees*), and the minimum number of observations in terminal nodes (*min_n*). The MAE values predominantly range between 37.4 and 40.0, highlighting the model's general robustness to hyperparameter variation. This is particularly noteworthy given the extensive range of settings evaluated, encompassing *min_n* values from 2 to 32, *mtry* from 40 to 90, and *trees* from 50 to 100 (see Sect. 4.2).

It is crucial to acknowledge that optimizing individual hyperparameters in isolation may yield suboptimal results, given their complex interdependencies. For instance, the best-performing QRF model in our study had a *min_n* of 20, a *mtry* of 70, and *trees* of 100 (see Table 4). This complexity often renders simplistic, univariate optimization approaches ineffective. While the stability of MAE across various settings is advantageous, it should not overshadow the intricate relationships among hyperparameters and their collective impact on the model's performance.

5.1.3 Evaluation of uncertainty soundness

To evaluate the quality of uncertainty estimates, we conducted a qualitative comparison between QRF and BART. The models were configured to produce prediction intervals, as described in Sect. 3.2 for QRF and analogously for BART. Following the approach in He et al. (2017), Ehsan et al. (2019), the models were compared using PICP, MPIW, and MRPIW metrics. The comparative results are summarized in Table 6. While QRF and BART show similar performance in terms of PICP and MPIW, they differ substantially in their MRPIW values. Specifically, QRF's MRPIW on the validation data is 3.01, significantly lower than BART's 11.86. This pattern is consistent on the test data, where QRF scores an MRPIW of 1.74 compared to BART's 4.83. Further, the standard deviation for QRF's MRPIW (0.352) is much lower than that of BART (4.29). This suggests that QRF offers more stable performance and better adapts its prediction intervals across validation folds.

The notable reduction in MRPIW for QRF underscores its effectiveness in generating more reliable prediction intervals. This is supported by lower standard deviations, which testify to QRF's robustness in providing reliable uncertainty estimates. Although both models offer similar coverage (PICP) and comparable MPIW results ($p_{Nemenyi} = 0.53$), QRF outperforms BART significantly in MRPIW ($p_{Nemenyi} = 0.0016$) on the validation data, as visualized in Fig. 6.

Regarding the evaluation of individual predictions, two key figures, namely Figs. 7 and 8, offer salient observations. These figures not only provide point predictions and associated residuals but also present the coverage ensured by each model's prediction intervals on the test data. It is noteworthy that the prediction intervals provided by the QRF model adapt according to the magnitude of the corresponding point predictions, a feature conspicuously absent in the BART model. In more quantifiable terms, the minimal width of the prediction

Table 6 Comparative analysis of model uncertainty for BART and QRF

Dataset and metric	BART	QRF	Dataset and Metric	BART	QRF
Validation data			Test data		
PICP	0.949 ± 0.01^a	0.949 ± 0.006	PICP	0.919	0.912
MPIW	250.93 ± 44.5	243.44 ± 28.3	MPIW	179.2	171.8
MRPIW	11.87 ± 4.29	3.01 ± 0.352	MRPIW	4.83	1.74

^aStandard Deviation across Cross-Validation Folds

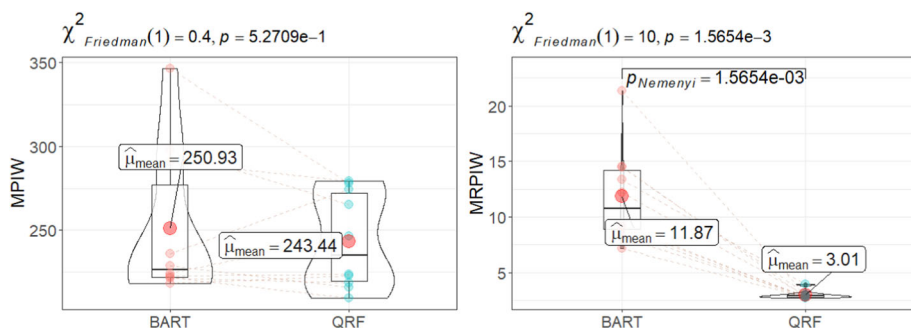


Fig. 6 Friedman–Nemenyi test results for BART and QRF for each of the uncertainty metrics *MPIW* and *MRPIW*

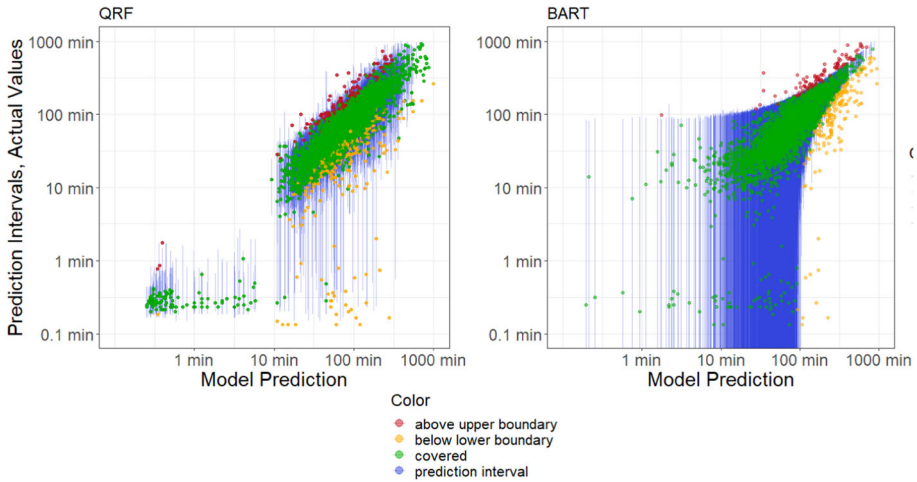


Fig. 7 Visualization of model predictions, actual values, prediction intervals and coverage for BART and QRF for test data. The points represent the relationship between the model predictions on the x-axis and the actual values on the y-axis. Corresponding prediction intervals are depicted as blue vertical lines, spanning from the value of the upper boundary to the value of the lower boundary. The color of the points indicate if the actual value was captured by the prediction interval (green) or not (orange for actual values below lower boundary, red for actual values above upper boundary). A logarithmic scale was used to allow a clearer depiction of values below the lower boundary

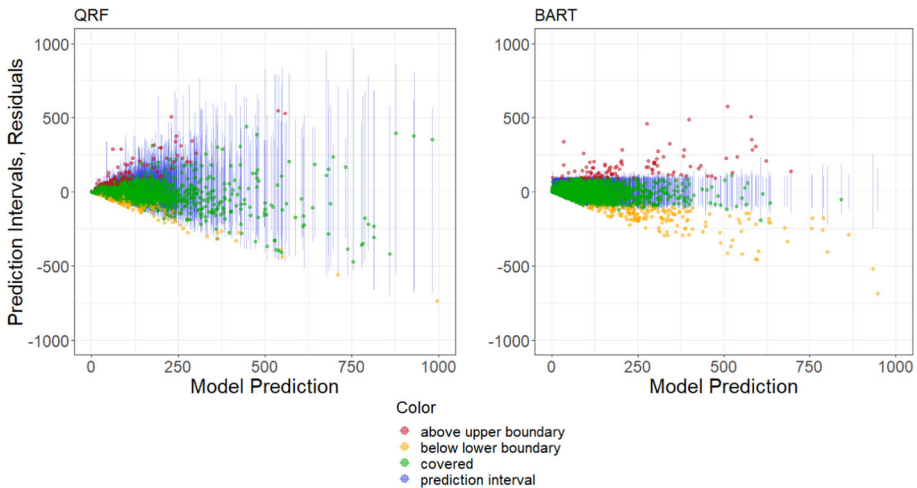


Fig. 8 Visualization of residuals, prediction intervals and coverage for BART and QRF for test data. The points represent the relationship between the model predictions on the x-axis and residual values on the y-axis. Corresponding prediction intervals are depicted as blue vertical lines, spanning from the value of the upper boundary to the value of the lower boundary reduced by the model prediction. The color of the points indicate if the actual value was captured by the prediction interval (green) or not (orange for actual values below lower boundary, red for actual values above upper boundary)

interval for BART stands at 80.3 min with a standard deviation of 18.2 min. This statistical observation critically hampers BART's utility in the context of UQ, particularly for lower target values. In stark contrast, the QRF model manifests a much more adaptive behavior, with a minimal prediction interval width of 0.22 min and a notably wider standard deviation of 194.6 min. Such adaptability becomes especially pertinent for model predictions falling below the 100-minute mark. Further, Fig. 7 elucidates that BART sets a minimal value for the lower boundary of its prediction intervals, stemming from its intrinsic restriction to non-negative values. To summarize, the QRF model exhibits a more nuanced capability in tailoring its prediction intervals based on specific point predictions, thereby making it substantially more fitting for applications that necessitate reliable uncertainty estimates.

To further validate the relevance and soundness of the UQ results delivered by the QRF model, an interview was conducted with a process expert from the manufacturing partner. The endorsement from the process expert provides valuable qualitative validation for the QRF model's approach to estimating model uncertainties. This is significant because the expert has a deep understanding of the complexities and variabilities in the manufacturing process, offering a real-world perspective that complements statistical evaluations.

The use of a dedicated dashboard in our evaluation for model evaluation was a key factor in making the complex QRF model accessible to the expert. The dashboard not only showcased the model's predictions and related uncertainties but also allowed the expert to interact with data on both trace and event levels. This interactive component provides a dynamic way to test the model's capabilities and limitations, adding another layer to its validation process. Moreover, the dashboard is a prototype for how the model could be integrated into existing management systems, illustrating its operational feasibility. The expert's affirmation speaks to the model's practical utility. By examining the dashboard, the expert was able to relate the model's UQ outputs to everyday operational decisions. The best- and worst-case scenarios generated by the model, as manifested in the prediction intervals, can serve as actionable guidance for production planning, helping to mitigate risks and optimize resource allocation.

Additionally, the expert's assessment of the model as "intuitively comprehensible" indicates that the QRF model could be integrated into existing workflows with minimal disruption. Its "adaptive prediction intervals" were also deemed "sufficiently sound and satisfactory", underlining the model's ability to adapt to the unique characteristics of individual production steps, thus enhancing its real-world applicability. In summary, the expert's validation is more than just an endorsement; it provides a multi-faceted evaluation that underscores the QRF model's robustness, practical utility, and adaptability. This aligns well with the model's statistical validation, thereby reinforcing its viability as a reliable tool for uncertainty quantification in complex manufacturing processes.

5.2 Uncertainty profile construction and evaluation

In response to the operational needs expressed by process experts, we have extended our model to include a system of uncertainty profiles—specifically categorized as low, medium, and high. The aim is to offer a nuanced lens through which model predictions can be evaluated, providing practitioners with an effective tool for risk mitigation and decision-making. While prediction intervals are useful as raw uncertainty measures, they may not be immediately interpretable in a practical setting.

An initial attempt was made to categorize uncertainty via percentile-based profiling, focusing on the widths of prediction intervals. However, this approach proved to be suboptimal; the width of the prediction interval was observed to correlate strongly with the actual output

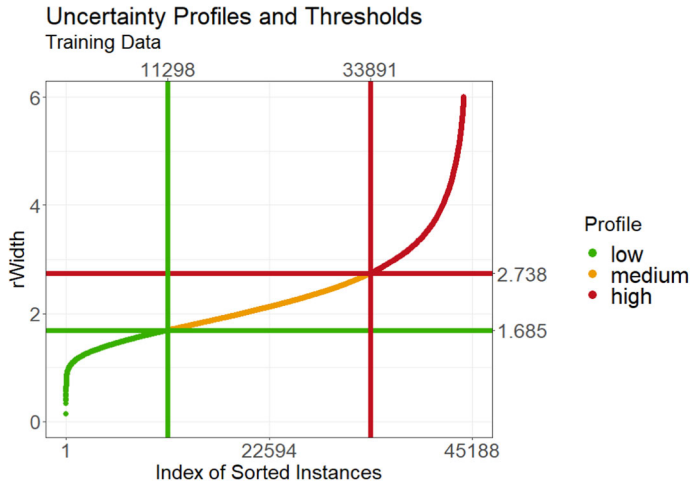


Fig. 9 Visualization of uncertainty profile thresholds for validation data. First, data set was sorted in ascending order of $rWidth$ values and a numerical index was introduced to depict the order. Each point in the plot depicts its index on the x-axis and the corresponding $rWidth$ value on the y-axis. The green and red vertical line divide points respectively at the 25th and 75th percentile. The green and red horizontal lines depict the corresponding $rWidth$ values, which respectively separate the “low” from the “medium” ($rWidth = 1.685$) and the “medium” from the “high” profile ($rWidth = 2.738$). The points are colored to represent their affiliation with the corresponding profile: green for “low”, orange for “medium” and red for “high”

values. Consequently, activities with longer durations exhibited inflated prediction intervals, and the reverse was true for activities with shorter durations. To address this issue, we turned to an alternative metric: relative width intervals, as delineated in Sect. 4.3. This normalized approach accounts for the inherent variability in activity durations, thereby providing a more accurate and reliable representation of associated uncertainties. By leveraging relative width intervals to categorize uncertainties, we offer process experts an enhanced understanding of the confidence levels for each prediction. This, in turn, allows for informed decision-making concerning potential adjustments in operational processes (see Fig. 9).

The construction of uncertainty profiles leverages the training dataset for calibration, which involves scoring the dataset using the fitted QRF model and calculating the $rWidth$ for each model prediction. Instances were then sorted in ascending order of $rWidth$ values, and the 25th and 75th percentile thresholds were used to define the uncertainty profiles. Values below the 25th percentile threshold were classified as “low” profile, with $rWidth$ values below 1.685, while values above the 75th percentile threshold were classified as “high” profile, with $rWidth$ values above 2.738. The remaining values were assigned to the “medium” profile. Figure 9 visualizes the results, with the vertical lines indicating the 25th and 75th percentile thresholds and the horizontal lines representing the $rWidth$ values corresponding to these thresholds.

The evaluation metrics for each uncertainty profile from training data are presented in Table 7. The PICP value is notably high, registering at 0.997 across all profiles. This can be attributed to the model’s high level of familiarity with the training data, ensuring almost maximum prediction interval coverage. When examining the MPIW metric, the “high” profile shows the widest prediction intervals, with a value of 260.8. This is followed by the “medium” profile at 203.7, and the “low” profile at 174.4. This result aligns with the expectation that higher uncertainty profiles will naturally have larger prediction intervals. As for MRPIW, the

Table 7 Evaluation metrics for the QRF model for uncertainty profiles

Dataset	PICP	MPIW	MRPIW	MAE	Baseline	Number of events
Training Data	0.997	210.6	2.69	21.3	55.6	45,188
Profile: “low”	0.997	174.4	1.42	23.4	64.1	11,297 (25%)
Profile: “medium”	0.997	203.7	2.15	20.5	55.6	22,594 (50%)
Profile: “high”	0.997	260.8	5.06	20.8	50.4	11,297 (25%)
Test Data	0.912	171.8	1.74	35.5	51.9	3,389
Profile: “low”	0.893	131.9	1.30	31.8	54.8	1,948 (58%)
Profile: “medium”	0.938	220.2	2.10	39.8	76.0	1,201 (35%)
Profile: “high”	0.944	269.8	3.80	41.1	77.3	240 (7%)

trend indicates the “high” profile leading with a value of 5.06. It is followed by the “medium” profile at 2.15 and the “low” profile at 1.42. This distribution is consistent with the initial construction methodology of the uncertainty profiles, which relies on MRPIW values.

To evaluate the model’s uncertainty on the test dataset, we follow a methodology analogous to the one applied to the training dataset. First, the test dataset is scored to calculate the rWidth, PICP, MPIW, and MRPIW metrics. Subsequently, using the uncertainty profile thresholds defined during the calibration step, individual test data predictions are categorized into corresponding uncertainty profiles. Table 7 presents these metrics for each profile, enabling a comparative assessment with the training data.

In the context of PICP, the “low” profile exhibits reduced coverage with a value of 0.893, whereas the “medium” and “high” profiles register higher coverage values of 0.938 and 0.944 respectively. This phenomenon underscores that the QRF model offers improved coverage for predictions with elevated uncertainty levels, albeit at the expense of larger prediction intervals. Consequently, this highlights a trade-off between prediction coverage and uncertainty. For MPIW, the “low” profile has been optimized with a value of 131.9, while the “medium” and “high” profiles manifest extended prediction interval ranges, registering values of 220.2 and 269.8, respectively. It is crucial to note that these variances are partly attributable to the imbalanced structure of the test dataset (see Sect. 4.1). The MRPIW trends for the test data align closely with those observed for the validation data. Additionally, the test dataset’s profile allocation distribution indicates that 58% of events are categorized under the “low” profile, 35% under the “medium,” and 7% under the “high” profile. This distribution suggests that the test dataset, deliberately extracted to represent a chronological sample from the complete dataset, is imbalanced.

Regarding the Mean Absolute Error (MAE), the test data reveals a nuanced pattern: the “low” profile records the smallest MAE of 31.8, followed by the “medium” and “high” profiles with MAEs of 39.8 and 41.1, respectively. This suggests that the MAE increases with the model’s uncertainty levels. Furthermore, the QRF model demonstrates superior performance in generating point estimates across all uncertainty profiles when compared to the baseline predictions. Remarkably, the Mean Absolute Error (MAE) for instances categorized under the “high” uncertainty profile outperforms even the baseline results for instances within the “low” profile. This outcome accentuates the robustness of the QRF model, not just in terms of uncertainty quantification, but also in the accuracy of its point estimates.

In scenarios involving urgent orders and time-sensitive deadlines, process experts emphasized the importance of proactively identifying critical stages in production. They also called for enhanced methods for estimating best-case and worst-case scenarios. As outlined in

Sect. 5.1.3, interviews with these experts confirmed that the existing results met their requirements. The incorporation of uncertainty profiles into the production planning process offers two substantial advantages. First, it streamlines the communication of model uncertainty among stakeholders. By categorizing uncertainties into distinct profiles—low, medium, and high—these profiles provide a user-friendly mechanism to quickly assess the level of risk or reliability associated with each prediction. This categorization allows for a more targeted discussion and enables decision-makers to quickly identify areas requiring further scrutiny or alternative planning. Second, the uncertainty profiles contribute to the optimization of the production schedule. They not only provide point estimates but also prediction intervals for each production step. This additional layer of information facilitates more robust planning by considering not just the most likely outcomes but also possible variations. Planners can therefore sequence production steps more effectively, taking into account both the estimated processing times and their associated uncertainties. For example, tasks falling under the “high” uncertainty profile may be scheduled with added buffer times or could trigger additional verification steps to manage risk.

5.3 SHAP analysis for model uncertainty

This subsection delves into the analysis of SHAP values to examine the influence of individual features on the resulting prediction intervals. We perform this analysis on two distinct levels: a local level, concentrating on a single data instance, and a global level, which assesses the model’s general performance. Initially, we scrutinize SHAP values at the local level, targeting the prediction interval of a designated instance. To deepen our understanding, we extend this local analysis to encompass SHAP values for lower and upper prediction boundaries and point estimation. This multifaceted examination helps us understand the variation in feature contributions under different uncertainty levels.

Transitioning to the global level, we direct our focus to the model’s overall behavior. This stage of analysis involves evaluating both the SHAP feature importance rankings and the corresponding SHAP summary plots for the prediction intervals. Subsequently, we draw comparisons between SHAP summary plots across different uncertainty profiles, categorized as “low,” “medium,” and “high.” This comparative study enables us to identify disparities in feature contributions under fluctuating uncertainty conditions. To complete the global analysis, we present the SHAP Dependence plots for selected variables within each uncertainty profile. This final step allows us to detect emerging trends or patterns in how features interact across diverse uncertainty scenarios.

5.3.1 Local SHAP analysis

Figure 10 elucidates the influence of various factors on the prediction interval width for a particular test instance. This instance, classified under the “low” uncertainty profile, is related to the “Dishing Press” activity carried out on the “Dishing Press 5” machine. In the plot, variables are ranked along the y-axis based on their absolute impact on the prediction interval width, with the most influential variables appearing at the top. For this specific instance, the number of items produced (“Quantity = 4”) emerges as the most significant factor in widening the prediction interval, extending it by 82.1 min. Conversely, the historical mean processing time for an item in the relevant article group (“Mean Processing Time = 25.27”) is the principal factor in contracting the prediction interval, reducing its width by 33.3 min. Given the many variables, the plot focuses on the top ten factors based on their

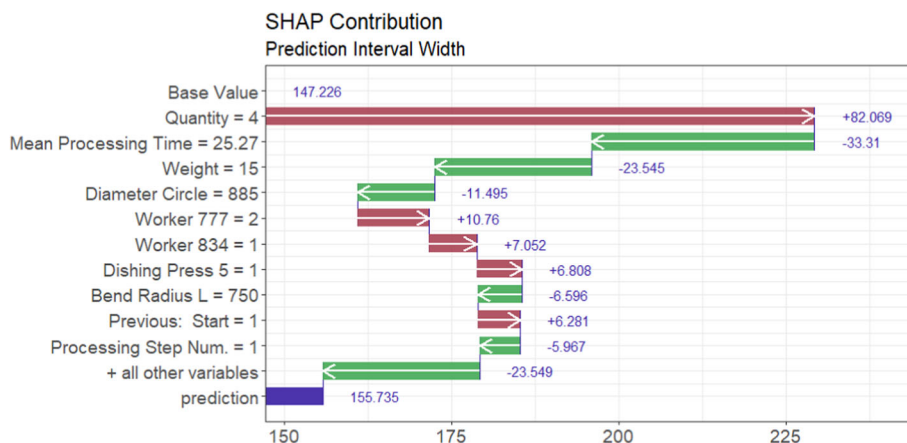


Fig. 10 SHAP contribution plot, depicting the impact of feature values of a specific instance to the corresponding predicted interval width. The horizontal bars depict the relationship between variable values, as seen on the y-axis, and the corresponding impact on the final prediction, as seen on the x-axis. The y-axis starts with the base SHAP value, followed by the ten most impactful variables based on their SHAP value for the instance, concluding with an entry representing the collective contribution of the remaining variables as well as the final prediction. Bars colored in red indicate an increase in interval width, while green-colored bars indicate a decrease in interval width and are supported by arrows as a visual aid. The precise SHAP-contributions, the base value and the final prediction are incorporated as well

absolute contribution to the interval width. The residual impact of the remaining variables is consolidated under the label “+ all other variables.” These collectively contribute to a further narrowing of the prediction interval by 23.5 min, resulting in a final prediction interval width of 155.7 min.

This comprehensive insight into variable influence is a critical tool for process experts. By dissecting the constituent elements that contribute to prediction interval widths, the experts gain a nuanced understanding of how different variables either amplify or attenuate uncertainty. This empowers them not only to anticipate fluctuations in production cycles but also to strategize mitigations for undesirable variances. For example, knowing that the quantity of produced items significantly widens the prediction interval could lead to a reassessment of batch sizes to optimize workflow.

Figures 11, 12, and 13 offer an in-depth examination of the SHAP contribution plots for point predictions, as well as the lower and upper boundaries of the prediction interval. These figures collectively elucidate the distinct roles played by specific variables across different facets of the prediction framework. For point predictions, the weight value (“Weight = 15”) is especially noteworthy, decreasing the predicted processing time by 10.2 min. One of the bend radius variables (“Bend Radius S = 75”) follows as the second most influential factor, increasing the predicted time by 5.5 min. While some variables display parallel trends in their SHAP values across different prediction aspects, it is essential to highlight the differences in the contributing factors between point predictions and prediction intervals.

Turning attention to the lower boundary, a SHAP base value of 22.0 min is registered. The diameter variable (“Diameter Circle = 885”) follows item quantity as the second most influential factor, decreasing the prediction by 4.2 min. When evaluating the upper boundary, however, the dynamics shift dramatically. The SHAP base value here is 90.7 min, and the average statistical processing time for an item (“Mean Processing Time = 25.27”) emerges as the most significant contributor, reducing the upper boundary by 34.0 min. This variance in

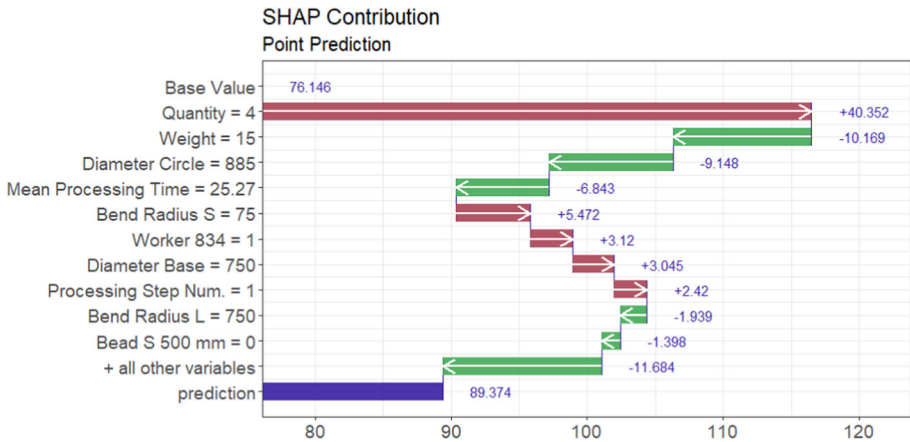


Fig. 11 SHAP contribution plot, in the same fashion as Fig. 10, depicting the impact of feature values of a specific instance on the point prediction

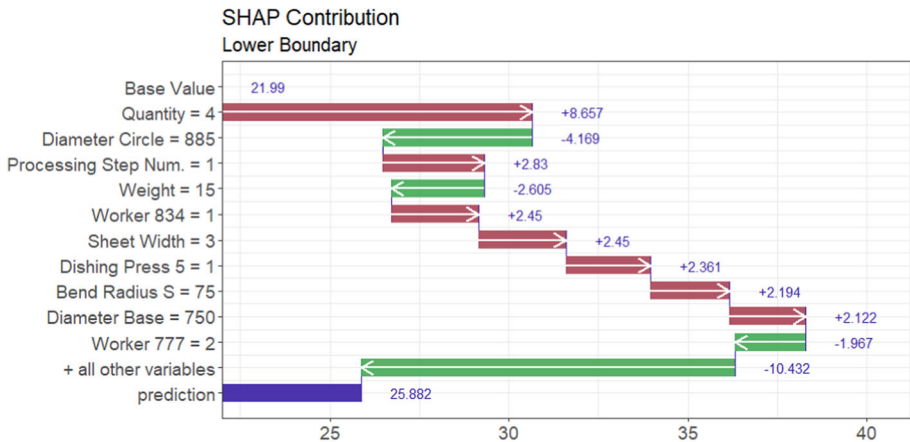


Fig. 12 SHAP contribution plot, in the same fashion as Fig. 10, depicting the impact of feature values of a specific instance on the lower boundary of the prediction interval

the impact of variables on the lower and upper boundaries reveals the complex interactions that define prediction intervals. Specifically, “Mean Processing Time” exerts a considerable influence on shaping the upper boundary, while its impact on the lower boundary is marginal. This discrepancy is instrumental in illustrating the role of the upper boundary in widening the prediction interval, confirming its dominant influence in defining interval widths.

For the lower boundary (see Fig. 12), a SHAP base value of 22.0 min is documented, with the diameter variable (“Diameter Circle = 885”) value having the most significant impact after item quantity (“Quantity = 4”) on the prediction, decreasing the predicted value by 4.2 min. In contrast, for the upper boundary (see Fig. 13), the SHAP base value of 169.2 min is impacted most by the average statistical processing time for an item of the underlying article group (“Mean Processing Time = 25.27”), decreasing the predicted value by 34.0 min. Furthermore, observed discrepancies in the ranking of variables further illustrate differences in the influence of variable values on the predicted subject. For instance, the “Mean Processing

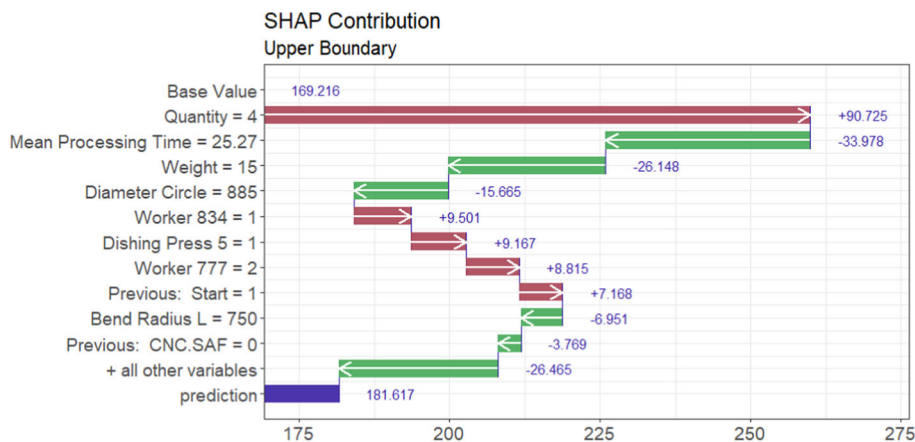


Fig. 13 SHAP contribution plot, in the same fashion as Fig. 10, depicting the impact of feature values of a specific instance on the upper boundary of the prediction interval

Time” shows a remarkable impact on the calculation of the upper boundary, while being relatively insignificant with regards to the calculation of the lower boundary. Additionally, similar trends regarding the impact of variable values and their ranking are registered when comparing the SHAP contribution of the upper boundary with the prediction interval width (see Fig. 10), indicating the dominant nature of the upper boundary towards increasing the prediction interval.

The preliminary assessment, conducted in tandem with process experts, confirms that SHAP analysis considerably amplifies experts’ comprehensive understanding of the factors influencing uncertainty. Specifically, it empowers experts to single out and validate key features that affect the range of specific prediction intervals. One of the standout features of SHAP analysis is its ability to dissect the influence exerted by each variable on the width of the prediction interval width. By doing so, it provides experts with actionable insights that guide them toward identifying variables that either widen or narrow the prediction range. In practical terms, this clarity helps in pinpointing variables that are the primary contributors to increased uncertainty or greater prediction reliability.

Further enriching the analysis, SHAP values across different facets—such as the lower and upper boundaries of the prediction interval, as well as the interval’s width—is compared. This side-by-side analysis provides a unique lens to examine the relationships and dependencies among variables. Such comparative scrutiny reveals which variables have a more pronounced impact at the extremes of the prediction interval. As a result, critical factors that become particularly influential under extreme conditions can be identified. This multi-dimensional understanding allows for more targeted decision-making, whether for mitigating risk or optimizing performance.

5.3.2 Global SHAP analysis

Figure 14 provides a global perspective on model behavior by showcasing the SHAP feature importance derived from the test dataset. The focus is on the top ten variables exerting the highest absolute impact on the prediction of interval widths. Key variables such as the quantity of produced items (“Quantity”) and the average statistical processing time for items in the respective article group (“Mean Processing Time”) emerge as the most significant.

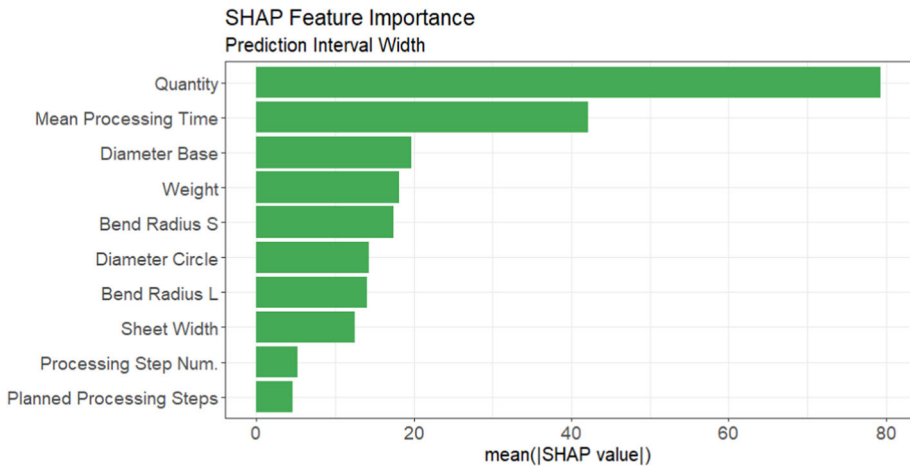


Fig. 14 SHAP feature importance plot, depicting the impact of feature values on the final model prediction on a global level. This plot visualizes the ten most important variables in descending order, as seen on the y-axis, and their corresponding impact via the length of horizontal bars, as seen on the x-axis. The importance of a variable is calculated by averaging the absolute SHAP values documented for the corresponding variable for the test data set

These are followed by variables tailored to specific cases, like the diameter of the base of the manufactured item (“Diameter Base”) and its weight (“Weight”). Additional variables of note include feature-engineered factors such as the event’s position within the planned production schedule (“Processing Step Num.”) and the total count of planned steps within the corresponding production trace (“Planned Processing Steps”). This visualization offers a synthesized yet comprehensive understanding of the variables most instrumental in influencing model uncertainty.

Figure 15 presents a SHAP summary plot for a nuanced understanding of how feature values correlate with their respective impact on predicted interval widths. The plot is constrained to the top 10 most influential variables, listed in descending order of importance, with “Quantity” being the most impactful. Each variable displayed on the y-axis has points signifying feature values—represented by color—and corresponding SHAP values—indicated by their x-axis position. For example, the “Mean Processing Time” variable largely associates lower values with negative SHAP values, effectively narrowing the prediction interval. In contrast, higher values result in positive SHAP values, thereby widening the interval. Another variable, “Bend Radius S,” demonstrates low impact at low values and a more substantial impact at higher values, although it doesn’t show a clear trend in either increasing or decreasing the prediction interval width. Interestingly, while “Quantity” stands out as a top contributor to interval width and, by extension, model uncertainty, its values don’t reveal a straightforward correlation with the level of model uncertainty. This analysis reveals a complex interaction among variables; the impact of one variable on prediction intervals often depends on the values of other variables. Such insights underscore the necessity of considering multiple variables in conjunction to arrive at a more accurate and holistic understanding of model behavior and uncertainty.

The employment of global SHAP explanations, including both SHAP summary plots and SHAP feature importance, yields multiple benefits for process experts aiming to understand complex machine learning models. These tools elucidate the nuanced interplay between input

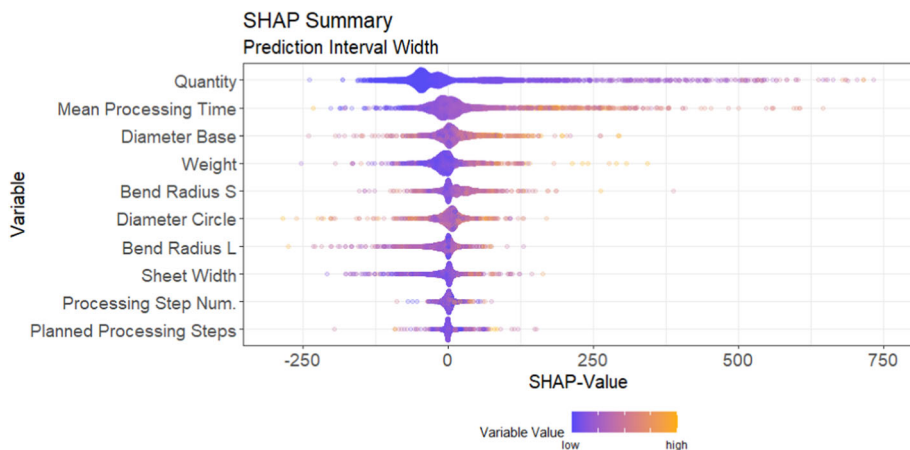


Fig. 15 SHAP summary plot for the ten most impactful variables, depicting the relationship between variable values and corresponding SHAP values. The variables are represented on the y-axis, SHAP values on the x-axis. For the visualization of the distribution of SHAP values, a mixed approach of beeswarm and violin plot was chosen: Each point represents an instance, with the color of the point depicting the relative variable value, its position on the x-axis representing the SHAP value and its position on the y-axis within the bounds of the variable complying with the density of the area

variables and model predictions, affording a comprehensive view of the model’s behavior. Such insights are invaluable for fine-tuning the model to achieve optimal performance. SHAP summary plots serve a dual role: they offer an aggregated perspective of feature importance across all instances and help identify overarching trends. This facilitates the recognition of features that have a consistent impact on model predictions. On the other hand, SHAP feature importance rankings provide a detailed, instance-specific breakdown of feature contributions. This granularity enables experts to pinpoint the key variables that influence specific outcomes. In addition to these tools, SHAP Dependence plots illustrate how features interact with each other, revealing potential synergies or redundancies. These plots are particularly useful for understanding the nuances of feature interaction in the context of uncertainty profiles, an aspect further discussed in Sect. 5.3.3. Together, these global SHAP explanations not only enhance interpretability but also build a framework of transparency, trust, and comprehensibility. This is particularly important for process experts who rely on data-driven decision-making processes.

5.3.3 SHAP analysis of uncertainty profiles

This section focuses on refining the model’s uncertainty explainability by examining variations in uncertainty profiles. Figures 16, 17, and 18 present the SHAP summary plots for “low,” “medium,” and “high” uncertainty profiles, respectively. A side-by-side analysis of these figures reveals nuanced differences in model behavior across the individual profiles. Although “Quantity” and “Mean Processing Time” consistently emerge as the most influential variables across all profiles, the ranking of other significant features varies. For example, the weight of the processed material (“Weight”) holds varying levels of importance depending on the profile. In the “low” and “medium” profiles, “Weight” ranks as the fourth most impactful variable, while in the “high” profile, it falls to the sixth position. Furthermore, the distribution of SHAP values for specific variables varies across profiles. Notably, the

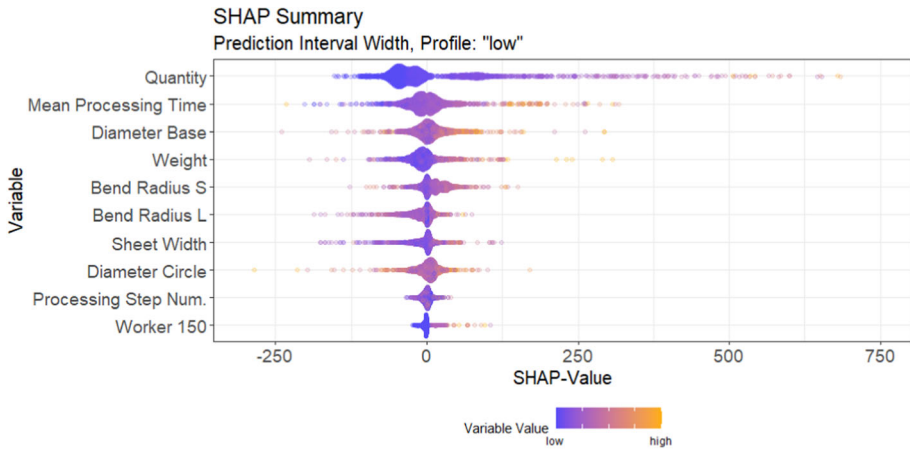


Fig. 16 SHAP summary plot for the profile “low”. The same approach as in Fig. 15 was used, but the data was restricted to instances pertaining to the “low” uncertainty group

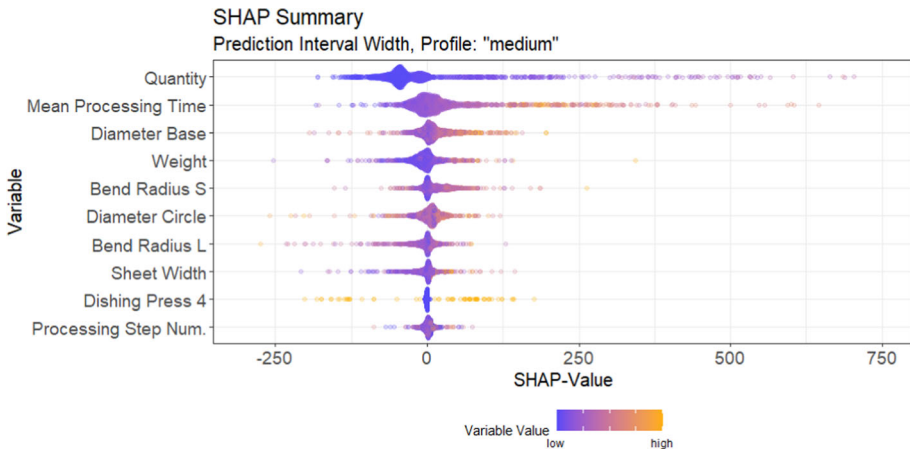


Fig. 17 SHAP summary plot for the profile “medium”. The same approach as in Fig. 15 was used, but the data was restricted to instances pertaining to the “medium” uncertainty group

distribution for the variable “Bend Radius S” shows a rightward skew in both the “low” and “high” profiles, while maintaining a more balanced distribution in the “medium” profile. These observations highlight the complexity of the model and the necessity for individualized interpretations based on the specific profile under examination. This detailed understanding allows for targeted model tuning and more effective decision-making processes.

The SHAP dependence plots serve as an advanced tool for diving deeper into the behavior of specific variables and their impact on prediction intervals. Figure 19 illustrates SHAP dependence plots for two key variables: “Diameter Base” and “Weight.” Each plot is tailored to an individual uncertainty profile—low, medium, and high. The plots reveal a general upward trend in SHAP values with increasing “Diameter Base” values across all profiles. However, subtle variations in SHAP value distributions are discernible between profiles. A color gradient representing the “Weight” variable adds an extra layer of insight, suggesting

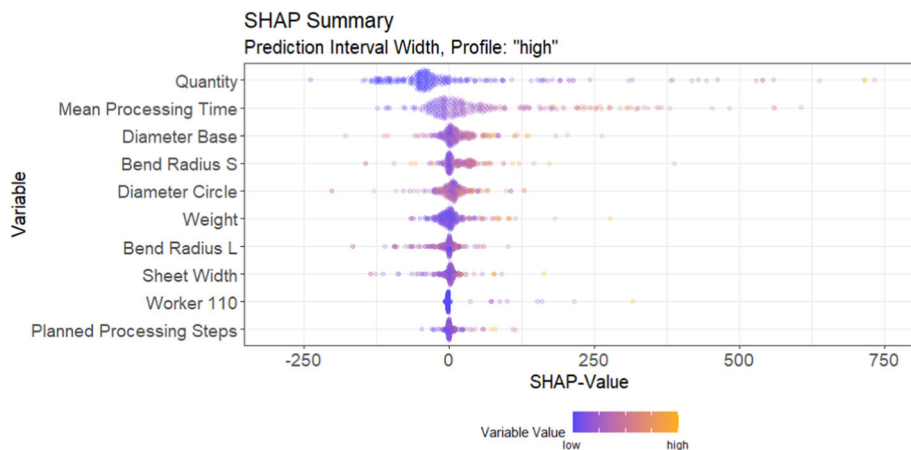


Fig. 18 SHAP summary plot for the profile “high”. The same approach as in Fig. 15 was used, but the data was restricted to instances pertaining to the “high” uncertainty group

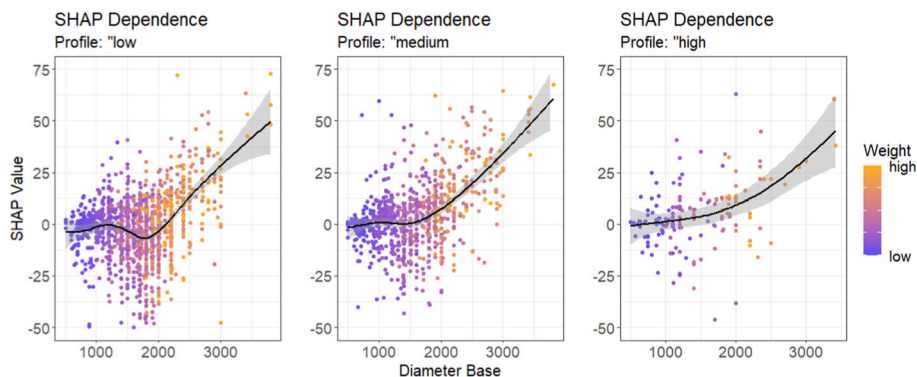


Fig. 19 SHAP dependence plot for the variable “Diameter Base”, with the secondary variable “Weight” for the “low”, “medium” and “high” uncertainty profile. Each point represents the relationship between the “Diameter Base” value of an instance, as seen on the x-axis, the corresponding SHAP value, as seen on the y-axis, and the corresponding relative “Weight” value, provided via color coding. A black smoothing curve, calculated via a general additive model, provides a visual aid for each plot

a positive correlation between “Diameter Base” and “Weight” in terms of their impact on SHAP values. Particularly notable is a trend in the “low” profile, where a region of negative SHAP values appears for “Diameter Base” values between 1500 and 2000. This specific behavior is attenuated in the “medium” profile and completely absent in the “high” profile.

The use of global SHAP analysis on data subsets distinguished by varying uncertainty profiles offers a multi-faceted view into the model’s behavior. This approach enhances both the model’s robustness and its applicability across diverse operational conditions. By focusing on these distinct subsets, process experts can pinpoint variables that have a pronounced influence in specific contexts of uncertainty. This fine-grained understanding enables targeted model refinement, illuminating paths for performance improvement. Furthermore, examining these data subsets can reveal inconsistencies in the model’s sensitivity to certain features across different uncertainty profiles. Such insights are valuable for understanding the model’s

limitations and its resilience to diverse input conditions, thereby aiding in the calibration of its predictive capabilities. The end result of this comprehensive approach is a machine learning model better equipped for real-world applications. It allows for the development of customized strategies to mitigate risks and uncertainties in various scenarios, thus enhancing the model's utility and trustworthiness in decision-making processes.

6 Discussion

6.1 Relevance for operations research

The relevance of our methodology for OR is multifaceted, with each component—be it uncertainty estimation or explanation—having distinct implications. The proposed approach aligns with the “predict-then-optimize” model commonly found in OR. In this approach, ML is used to predict essential parameters of an optimization model before or simultaneously as the optimization models are solved (Mišić and Perakis, 2020). The core of our work lies in the “predict” phase, utilizing the QRF technique to produce forecasts that come with quantified uncertainty. These forecasts then may serve as inputs for optimization models. Unlike traditional OR models that often rely on deterministic or overly simplistic stochastic parameters, our approach captures the intricate, possibly non-linear, nature of uncertainty. This refined understanding of uncertainty is then integrated into various OR models, thereby improving the robustness and reliability of the optimization solutions. Essentially, our work lays a sophisticated foundation for the subsequent optimization phase.

More specifically, focusing on predictive process monitoring, our primary concern is comprehending dynamic system behavior. Within the OR framework, our problem is formulated to improve system responses to fluctuating inputs. Our approach refines decision-making by narrowing the range of options, thus enabling the integration of data analytics into operational optimization. The use case, centered on quantifying uncertainty in process predictions, extends traditional OR problems to evaluate system performance under diverse conditions. Moreover, our model provides multi-level insights into predictive uncertainty, valuable for both tactical and operational planning. These insights are particularly useful for resource allocation and production scheduling in our partner manufacturing firm. For example, the use of SHAP analysis to identify the most influential features contributing to overall uncertainty parallels resource allocation challenges in OR. This understanding enables targeted interventions and resource reallocations to minimize uncertainty in real-world processes.

The overarching objective of our research, akin to many OR initiatives, is to offer robust decision support. By quantifying uncertainty, we provide decision-makers with a comprehensive understanding of potential outcomes, including best-case, worst-case, and most-likely scenarios. This aligns with OR's emphasis on risk mitigation, where decisions account for the variability and uncertainty of outcomes. Given the specific context, it is clear that our approach is deeply rooted in OR methodologies. OR's extensive toolkit has been pivotal in shaping our research, making it rigorous and applicable to real-world challenges. This mutually beneficial relationship ensures that our contributions are both grounded in established methods and innovative in the field of predictive process monitoring.

6.2 Implications for domain experts

The utilization of a multi-stage machine learning approach that integrates uncertainty awareness and explainability holds significant implications for decision-makers across diverse business operations. Through the use of machine learning models that account for uncertainty, experts in a given field can gain a deeper understanding of potential outcomes and their associated variability. This enhanced comprehension can facilitate more efficient decision-making and ultimately lead to improved risk management. The heightened awareness of uncertainty leads to improved operational processes in organizations, promoting a culture of decision-making based on data, which ultimately results in increased efficiency and effectiveness. The integration of ML explanation components in the decision-making process offers the fundamental benefit of being able to discern the fundamental factors that contribute to predictive uncertainty. This factor provides decision-makers with the ability to concentrate their endeavors on mitigating the pertinent origins of unpredictability, thereby enhancing the resilience and trustworthiness of the decision-making mechanism.

Moreover, the knowledge acquired from our uncertainty-aware explainable approach can be utilized to enhance resource allocation, facilitating organizations to give priority to resources in domains with significant volatility and alleviate related risks. This solution can also provide benefits for domain experts in the areas of strategic planning and organizational adaptability. By comprehending the magnitude of uncertainties, individuals can formulate more reliable and adaptable strategic plans that correspond with the objectives of the organization and ensure sustained prosperity. In addition, the capacity to measure and clarify uncertainties provides professionals with the necessary assets to adapt to evolving circumstances and address possible interruptions, augmenting the competitiveness of the enterprise in a dynamic commercial setting.

6.3 Theoretical/scientific implications

The use of our proposed approach that integrates uncertainty awareness and explainability has noteworthy theoretical and scientific implications for the domains of prescriptive analytics, OR, and AI. This study contributes to the advancement of scientific knowledge by addressing gaps in the existing literature. Specifically, it emphasizes the significance of integrating technical production parameters, producing machine learning outputs that account for uncertainties, and elucidating the origins of such uncertainties. The incorporation of these components within the decision-making framework has the potential to enhance the efficacy of the model, produce more resilient optimization results, and foster a deeper comprehension of intricacies inherent in practical scenarios.

From a methodological perspective, the proposed approach expands the use of ML methods, such as QRF and SHAP, to generate prediction intervals and attribute uncertainty to particular input features. The progress made in this field not only facilitates a more thorough comprehension of the inherent uncertainties in problems related to OR but also lays the groundwork for the creation of novel methodologies and techniques that further augment the combination of uncertainty and explanation in models used for optimization. Consequently, forthcoming studies may utilize these methodological advancements to develop novel approaches that tackle a diverse range of intricate commercial challenges.

Finally, by illustrating its applicability to real production planning scenarios, the proposed method makes a significant addition to the scientific community. This use case serves as a proof-of-concept, demonstrating how the multi-stage ML strategy is effective at man-

aging uncertainty and delivering useful insights. The successful application of the suggested strategy in a practical setting may inspire additional investigation and study in related areas, fostering interdisciplinary cooperation and encouraging the creation of new theories, methodologies, and applications that advance scientific understanding generally.

6.4 Threats to validity

While the proposed methodology shows promising results in the field of predictive process monitoring, it's crucial to acknowledge potential threats to the study's validity. Recognizing these limitations not only provides a more complete understanding of the study's constraints but also encourages further research aimed at addressing these issues.

The validity of the findings can be significantly influenced by the quality and representativeness of the data utilized in this study. The case study's findings might not be generalizable if the data used don't accurately reflect the real-world scenario or if they have biases, inconsistencies, or errors. Moreover, it is imperative to have an adequately large sample size to mitigate the impact of random variations or anomalies on the results. The validity of the study may be impacted by the assumptions made during the development of the ML models. It is important to note that the assumptions regarding the underlying distribution of the data and the interactions between variables may not be applicable in all scenarios. Thus, the efficacy of the suggested methodology may exhibit variability contingent upon these aforementioned factors.

While SHAP provides insights into the sources of uncertainty, the scope for interpreting and clarifying these explanations may be limited. The understandability of contributing factors can be hindered by complex interactions among variables or high-dimensional data, posing challenges for domain experts. Future research should focus on creating more accessible and understandable explanations, thereby improving communication with stakeholders. By addressing these potential limitations, subsequent studies can refine the proposed methodology and deepen the overall understanding of UQ and XAI within the context of predictive process monitoring.

7 Related work

The principal objective of OR is to harmonize methods for effective managerial decision-making. Integral to this aim is integrating information systems and decision-support tools, as articulated by Simon (1997). The relationship between OR and AI is mutually beneficial; AI approaches often require the solution of optimization problems—a core component of OR methodologies. Conversely, AI techniques find application in OR for predicting crucial parameters and formulating heuristics for complex optimization tasks (Bennett and Parrado-Hernández, 2006). In the specialized area of conditional-stochastic optimization, the work by Bertsimas and Kallus (2020) illustrates the promise of using predictive analytics to estimate conditionally expected costs for various inputs. This addresses the challenging task of minimizing uncertain costs in the presence of incomplete information. The study sets the stage for innovative applications in prescriptive analytics by showing how predictive techniques can solve complex optimization issues. Further exploring this synergy, the paper by Bengio et al. (2021) delves into the combinative potential of ML and combinatorial optimization. The authors advocate for a novel approach that views optimization problems as data points.

This enables the identification of problem distributions and enhances decision-making capabilities, moving beyond the limitations of traditional heuristics.

Traditionally, the OR field has been constrained by limited data availability and computational resources, necessitating reliance on models derived from microeconomic theory, game theory, optimization, and stochastic models (Mišić and Perakis, 2020). However, the modern landscape has evolved significantly due to advancements in computational power and algorithms, as well as increased data availability. This evolution has elevated AI to a central role in OR. Specifically, AI has been instrumental in enhancing our understanding of underlying processes in areas such as scheduling, thereby enabling more efficient operations (Isaksson et al., 2018). Data-driven analytics have demonstrated effectiveness across a wide spectrum of OR challenges. These range from capacity planning (Youn et al., 2022) and production planning (Usuga Cadavid et al., 2020) to distribution planning (Kumar et al., 2020) and inventory management (van Jaarsveld and Scheller-Wolf, 2015). Further applications include transportation (Chung et al., 2017), sales and operations planning (Thomé et al., 2012), as well as dynamic pricing and revenue management (Xue et al., 2016). The application of data-driven decision-making within these sectors yields substantial benefits. Among these are an enhanced return on investment, optimized asset utilization, and an increase in market value (Mehdiyev and Fettke, 2021).

In this study, we focus on a specific ML problem, namely predictive process monitoring, which is a technique within the broader field of process mining that includes process discovery, conformance checking, and process enhancement (Van Der Aalst et al., 2012). Predictive process monitoring leverages historical execution data to provide users with predictions about a target of interest for a given process execution (Maggi et al., 2014). Process mining encompasses a set of techniques aimed at extracting valuable insights from data generated by process-aware information systems during process execution. It serves as an intermediary between process science (including OR) and data science (encompassing fields such as predictive and prescriptive analytics), offering methods for data-driven process analysis (van der Aalst, 2022). As illustrated in Fig. 20 and presented in Rehse et al. (2019), there are three central prediction tasks based on the target of interest and its characteristics: process outcome prediction (Teinemaa et al., 2019), next event prediction (Tax et al., 2017; Evermann et al., 2017), and remaining time prediction (Verenich et al., 2019; Teinemaa et al., 2018).

Numerous review articles have been published on the subject of predictive process monitoring. For example, Di Francescomarino et al. (2018) classified 51 process prediction methods based on their prediction targets using a value-driven framework. These methods exhibited different prediction architectures and were categorized into various categories,

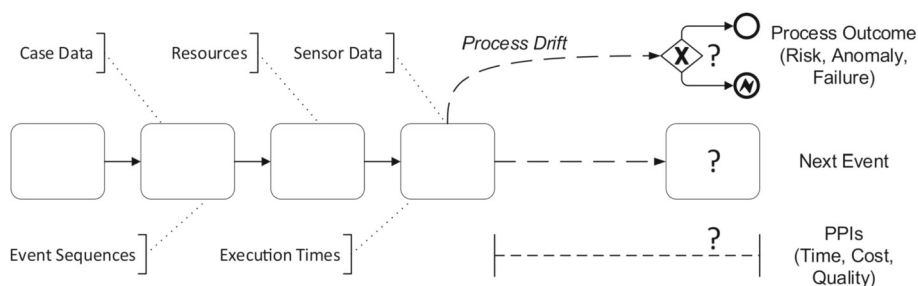


Fig. 20 Overview of predictive process analytics (Rehse et al., 2019)

including categorical outcome, costs, inter-case metrics, risk, sequence of values, and time. Teinemaa et al. (2019) conducted a systematic review and proposed a taxonomy for outcome-oriented predictive process monitoring. The authors identified and compared 14 relevant papers based on several criteria, including classification algorithm, filtering, prefix extraction, sequence encoding, and trace bucketing. Additionally, an experimental evaluation capturing the impact of different qualitative criteria was conducted using the authors' own implementation. Verenich et al. (2019) conducted a survey on methods for predicting remaining time in business processes, examining and comparing 25 relevant papers published between 2008 and 2017 based on criteria such as application domain, input data, prediction algorithm, and process awareness. A quantitative comparison was performed via a benchmark of 16 remaining time prediction methods on various publicly available datasets.

While black-box ML algorithms excel in predictive accuracy for process monitoring, their inherent opacity often leaves users reliant on less effective but transparent models (Neu et al., 2022; Arrieta et al., 2020). In this landscape, XAI has arisen as a vital area of research, aimed at bridging the gap between AI performance and human interpretability. By doing so, XAI seeks to improve user trust and facilitate effective collaboration between intelligent systems and human operators (Mehdiyev and Fettke, 2021; Guidotti et al., 2018). Several comprehensive reviews have contributed to the understanding of various aspects of explanatory techniques within AI and ML (Emmert-Streib et al., 2020).

One critical focus has been the exploration of local versus global explanation methods. These methods differ in their approaches and implications, offering customized strategies for explanation based on the context of their application (Adadi and Berrada, 2018). Extending this, research has gone into understanding the interaction between specific explanation techniques and the models they seek to make transparent. The categorization of these techniques as either model-centric or model-agnostic has been instrumental in guiding their application (Angelov et al., 2021). Furthermore, an inclusive approach has been adopted to involve diverse stakeholders in the design and deployment stages of explanation techniques. This inclusivity allows for a more nuanced implementation that meets varied requirements and perspectives (Arrieta et al., 2020). Simultaneously, there has been a concerted effort to elucidate the overarching goals of explanatory mechanisms. These studies reveal the motivations and intended outcomes driving their development and deployment (Mehdiyev and Fettke, 2021). The multi-disciplinary nature of this research has enabled the incorporation of insights from cognitive and social sciences. This approach enriches our understanding of the human factors that affect the comprehension and acceptance of machine-generated explanations (Miller, 2019). Additionally, the contextual elements have been considered in assessing explanatory techniques, underlining their importance in a real-world, decision-making environment. Evaluation criteria and benchmarks have also been established for a rigorous and objective assessment of explanatory methods' efficacy and utility (Vilone and Longo, 2021; van der Waa et al., 2021).

In the context of predictive process monitoring, the focus has frequently been on employing XAI techniques, primarily through post-hoc explanation mechanisms (Harl et al., 2020; Stevens et al., 2022; Velmurugan et al., 2021; Mehdiyev and Fettke, 2021, 2020). Another emerging trend in ML research and predictive process monitoring is UQ, which aims to capture and effectively communicate the inherent uncertainties in model predictions. UQ offers an additional layer of transparency, augmenting the comprehensibility of decisions derived from machine intelligence (Bhatt et al., 2021). The utility of UQ extends to aiding stakeholders in ascertaining when to trust model predictions, thus enhancing the functionality of automated decision systems. By quantifying and incorporating uncertainty systematically, UQ creates more robust and reliable decision-making frameworks, particularly when infor-

mation is ambiguous (Ghanem et al., 2017). The benefits of employing UQ methodologies are multidisciplinary, proving useful in sectors ranging from engineering and finance to environmental management (Smith, 2013). Several recent efforts have explored uncertainty within predictive process monitoring (Weytjens and De Weerd, 2022; Shoush and Dumas, 2022).

Despite the individual advancements in UQ and XAI, the intersection of these two fields remains largely unexplored in the academic literature. Limited studies have considered the bidirectional integration of UQ and XAI, focusing on clarifying the origins of uncertainties and investigating the uncertainties embedded within explanations themselves (Slack et al., 2021; Antorán et al., 2020; Moosbauer et al., 2021). This article endeavors to address this research gap. It aims to make the uncertainties associated with ML models more accessible to domain experts, particularly within the context of predictive process monitoring. To the best of our knowledge, this study is unique in its approach to merge UQ and XAI methodologies specifically for predictive process monitoring problems. It is a pioneering effort in this nascent field, targeting the formulation of more transparent, robust, and reliable decision support

8 Conclusion

In summary, this study presents a comprehensive approach to addressing the complex problem of predicting the time to completion for various manufacturing processes, while also quantifying and explaining the associated uncertainties. By leveraging advanced machine learning techniques such as QRF and SHAP analysis, we have been able to generate explainable, uncertainty-aware predictions that are crucial for real-world applications. These predictions serve as a sophisticated preparatory layer for subsequent optimization steps, aligning closely with the “predict-then-optimize” paradigm prevalent in OR. Our comparative analysis provides a robust framework for evaluating the efficacy of our model against both industry practices and a range of alternative predictive models. The inclusion of statistical tests and real-world feedback from process owners adds further credibility to our findings. By offering both theoretical and empirical validation, we believe this work makes a significant contribution to the fields of ML and OR, particularly in the context of manufacturing. The methodology developed here is not only applicable to the specific use case presented but also holds promise for broader applications, thereby opening avenues for future research and practical implementations.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IS21006B (project ExPro).

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138–52160.
- Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5), 1424.
- Antorán, J., Bhatt, U., Adel, T., Weller, A., & Hernández-Lobato, J. M. (2020). Getting a clue: A method for explaining uncertainty estimates. arXiv preprint [arXiv:2006.06848](https://arxiv.org/abs/2006.06848).
- 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. *Information Fusion*, 58, 82–115.
- Bengio, Y., Lodi, A., & Prouvost, A. (2021). Machine learning for combinatorial optimization: A methodological tour d’horizon. *European Journal of Operational Research*, 290(2), 405–421.
- Bennett, K. P., & Parrado-Hernández, E. (2006). The interplay of optimization and machine learning research. *The Journal of Machine Learning Research*, 7, 1265–1281.
- Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044.
- Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., & Tickoo, O. (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).
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chari, T., Chaabane, S., Aissani, N., & Trentesaux, D. (2014). Scheduling under uncertainty: Survey and research directions. In: *2014 International Conference on Advanced Logistics and Transport (ICALT)* (pp. 229–234). IEEE.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250.
- Chung, S. H., Ma, H. L., & Chan, H. K. (2017). Cascading delay risk of airline workforce deployments with crew pairing and schedule optimization. *Risk Analysis*, 37(8), 1443–1458.
- Di Francescomarino, C., Ghidini, C., Maggi, F.M., & Milani, F. (2018). Predictive process monitoring methods: Which one suits me best? In *Business Process Management: 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9–14, 2018, Proceedings 16* (pp. 462–479). Springer.
- Dumas, M., La Rosa, M., Mendling, J., Reijers, H. A., Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). Introduction to business process management. *Fundamentals of Business Process Management*, 1, 1–33.
- Ehsan, B. M. A., Begum, F., Ilham, S. J., & Khan, R. S. (2019). Advanced wind speed prediction using convective weather variables through machine learning application. *Applied Computing and Geosciences*, 1, 100002.
- Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6), 1368.
- Evermann, J., Rehse, J.-R., & Fettke, P. (2017). Predicting process behaviour using deep learning. *Decision Support Systems*, 100, 129–140.
- Frazzetto, D., Nielsen, T. D., Pedersen, T. B., & Šikšnys, L. (2019). Prescriptive analytics: A survey of emerging trends and technologies. *The VLDB Journal*, 28, 575–595.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
- Garcia, S., & Herrera, F. (2008). An extension on “statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons. *Journal of Machine Learning Research*, 9(12), 1.
- Ghanem, R., Higdon, D., Owhadi, H., et al. (2017). *Handbook of Uncertainty Quantification* (Vol. 6). Springer.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1–42.
- Harl, M., Weinzierl, S., Stierle, M., & Matzner, M. (2020). Explainable predictive business process monitoring using gated graph neural networks. *Journal of Decision Systems*, 29(sup1), 312–327.
- He, J., Wanik, D. W., Hartman, B. M., Anagnostou, E. N., Astitha, M., & Frediani, M. E. (2017). Nonparametric tree-based predictive modeling of storm outages on an electric distribution network. *Risk Analysis*, 37(3), 441–458.
- Isaksson, A. J., Harjunoski, I., & Sand, G. (2018). The impact of digitalization on the future of control and operations. *Computers & Chemical Engineering*, 114, 122–129.

- Jørgensen, M., Teigen, K. H., & Moløkken, K. (2004). Better sure than safe? over-confidence in judgement based software development effort prediction intervals. *Journal of Systems and Software*, 70(1), 79–93.
- Klas, M., Trendowicz, A., Ishigai, Y., & Nakao, H. (2011). Handling estimation uncertainty with bootstrapping: Empirical evaluation in the context of hybrid prediction methods. In *2011 International Symposium on Empirical Software Engineering and Measurement* (pp. 245–254).
- Kumar, R., Ganapathy, L., Gokhale, R., & Tiwari, M. K. (2020). Quantitative approaches for the integration of production and distribution planning in the supply chain: A systematic literature review. *International Journal of Production Research*, 58(11), 3527–3553.
- Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50, 57–70.
- Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4), 319–330.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems. NIPS'17* (pp. 4768–4777).
- 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* (pp. 457–472). Springer.
- Mehdiyev, N., & Fettke, P. (2020). Prescriptive process analytics with deep learning and explainable artificial intelligence. In *28th European Conference on Information Systems (ECIS). An Online AIS Conference*.
- 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. In *Interpretable Artificial Intelligence: A Perspective of Granular Computing* (pp. 1–28).
- Mehdiyev, N., & Fettke, P. (2021). Local post-hoc explanations for predictive process monitoring in manufacturing. In *29th European Conference on Information Systems (ECIS). An Online AIS Conference*.
- Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research*, 7, 983–999.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Milton, F. (1939). A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 34(205), 109.
- Mišić, V. V., & Perakis, G. (2020). Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1), 158–169.
- Mitrentsis, G., & Lens, H. (2022). An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. *Applied Energy*, 309, 118473.
- Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15), 3301–3307.
- Moosbauer, J., Herbinger, J., Casalichio, G., Lindauer, M., & Bischl, B. (2021). Explaining hyperparameter optimization via partial dependence plots. *Advances in Neural Information Processing Systems*, 34, 2280–2291.
- Mula, J., Poler, R., García-Sabater, J. P., & Lario, F. C. (2006). Models for production planning under uncertainty: A review. *International Journal of Production Economics*, 103(1), 271–285.
- Nemenyi, P. B. (1963). *Distribution-free Multiple Comparisons*. Department of Mathematics, Princeton University.
- Neu, D. A., Lahann, J., & Fettke, P. (2022). A systematic literature review on state-of-the-art deep learning methods for process prediction. *Artificial Intelligence Review*, 1, 1–27.
- Polato, M., Sperduti, A., Burattin, A., & de Leoni, M. (2014). Data-aware remaining time prediction of business process instances. In: *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 816–823). IEEE.
- Rehse, J.-R., Mehdiyev, N., & Fettke, P. (2019). Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *KI-Künstliche Intelligenz*, 33, 181–187.
- Shoush, M., & Dumas, M. (2022). When to intervene? prescriptive process monitoring under uncertainty and resource constraints. In *Business Process Management Forum: BPM 2022 Forum, Münster, Germany, September 11–16, 2022, Proceedings* (pp. 207–223). Springer.
- Shrestha, D. L., & Solomatine, D. P. (2006). Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2), 225–235.
- Simon, H. A. (1997). The future of information systems. *Annals of Operations Research*, 71, 3–14.
- Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34, 9391–9404.
- Smith, R. C. (2013). *Uncertainty Quantification: Theory, Implementation, and Applications* (Vol. 12). SIAM.

- Stevens, A., De Smedt, J., & Peepkorn, J. (2022). Quantifying explainability in outcome-oriented predictive process monitoring. In *Process Mining Workshops: ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31–November 4, 2021, Revised Selected Papers* (pp. 194–206). Springer.
- Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive business process monitoring with LSTM neural networks. In *Advanced Information Systems Engineering: 29th International Conference, CAiSE 2017, Essen, Germany, June 12–16, 2017, Proceedings 29* (pp. 477–492). Springer.
- Teinemaa, I., Dumas, M., Leontjeva, A., & Maggi, F. M. (2018). Temporal stability in predictive process monitoring. *Data Mining and Knowledge Discovery*, 32, 1306–1338.
- Teinemaa, I., Dumas, M., Rosa, M. L., & Maggi, F. M. (2019). Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2), 1–57.
- Thomé, A. M. T., Scavarda, L. F., Fernandez, N. S., & Scavarda, A. J. (2012). Sales and operations planning: A research synthesis. *International Journal of Production Economics*, 138(1), 1–13.
- Usuga Cadavid, J. P., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020). Machine learning applied in production planning and control: A state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing*, 31, 1531–1558.
- van der Aalst, W. M. P. (2022). Process mining: A 360 degree overview. Lecture Notes in Business Information Processing. In W. M. P. van der Aalst & J. Carmona (Eds.), *Process Mining Handbook* (Vol. 448, pp. 3–34). Springer.
- van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 103404.
- Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blicke, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., & Buijs, J. (2012). Process mining manifesto. In *Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9* (pp. 169–194). Springer.
- van Jaarsveld, W., & Scheller-Wolf, A. (2015). Optimization of industrial-scale assemble-to-order systems. *INFORMS Journal on Computing*, 27(3), 544–560.
- van der Aalst, W. M. P. (2016). *Process mining: Data science in action*, 2nd edn.
- Velmurugan, M., Ouyang, C., Moreira, C., & Sindhgatta, R. (2021). Evaluating fidelity of explainable methods for predictive process analytics. In *Intelligent Information Systems: CAiSE Forum 2021, Melbourne, VIC, Australia, June 28–July 2, 2021, Proceedings* (pp. 64–72). Springer.
- Verenich, I., Dumas, M., Rosa, M. L., Maggi, F. M., & Teinemaa, I. (2019). Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4), 1–34.
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106.
- Weytjens, H., & De Weerd, J. (2022). Learning uncertainty with artificial neural networks for predictive process monitoring. *Applied Soft Computing*, 125, 109134.
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82.
- Xue, Z., Wang, Z., & Ettl, M. (2016). Pricing personalized bundles: A new approach and an empirical study. *Manufacturing & Service Operations Management*, 18(1), 51–68.
- Youn, S., Geismar, H. N., & Pinedo, M. (2022). Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. *Production and Operations Management*, 31(12), 4407–4423.
- Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022). Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 9623–9633).