Artificial Intelligence Review (2025) 58:378
https://doi.org/10.1007/s10462-025-11399-0

Interpretable and explainable machine learning methods for predictive process monitoring: a systematic literature review

Nijat Mehdiyev[1,2] · **Maxim Majlatow**[1,2] · **Peter Fettke**[1,2]

Received: 4 July 2024 / Accepted: 16 September 2025

Abstract
This study presents a systematic literature review on the explainability and interpretability of machine learning models within the context of predictive process monitoring. Given the rapid advancement and increasing opacity of artificial intelligence systems, understanding the "black-box" nature of these technologies has become critical, particularly for models trained on complex operational and business process data. Using the PRISMA framework, this review systematically analyzes and synthesizes the literature of the past decade, including recent and forthcoming works from 2025, to provide a timely and comprehensive overview of the field. We differentiate between intrinsically interpretable models and more complex systems that require post-hoc explanation techniques, offering a structured panorama of current methodologies and their real-world applications. Through this rigorous bibliographic analysis, our research provides a detailed synthesis of the state of explainability in predictive process mining, identifying key trends, persistent challenges and a clear agenda for future research. Ultimately, our findings aim to equip researchers and practitioners with a deeper understanding of how to develop and implement more trustworthy, transparent and effective intelligent systems for predictive process analytics.

Keywords Explainable artificial ingelligence (XAI) · Interpretable machine learning · Predictive process monitoring · Process mining

✉ 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, 66123 Saarbrücken, Saarland, Germany

Published online: 14 October 2025

1 Introduction

Business process management (BPM) has long served as a foundational discipline for the systematic analysis, monitoring and optimization of organizational processes to enhance operational efficiency and strategic alignment. At the intersection of BPM and data science, process mining has emerged as a powerful paradigm, transforming voluminous, low-level event data captured by process-aware information systems (PAIS) into actionable evidence-based knowledge (Van Der Aalst 2012). By applying analytical techniques such as automated process discovery and conformance checking to event logs, which meticulously record activity sequences, timestamps and resource allocations, organizations can move beyond standard process models. This methodological rigor allows for the data-driven visualization of "as-is" processes, the quantification of deviations and the identification of bottlenecks, thereby providing a robust foundation for intelligent systems designed to diagnose inefficiencies and recommend improvements.

Building upon these diagnostic capabilities, predictive process monitoring (PPM) has become a rapidly evolving branch of process mining that leverages machine learning (ML) to predict the future states and outcomes of running process instances (Di Francescomarino et al. 2018). While early PPM approaches often relied on traditional classifiers and regressors trained on manually engineered features from event logs, recent advancements have been dominated by deep learning architectures. These sophisticated models can directly process complex sequential dynamics to predict next activities, remaining times and compliance risks with ever-increasing accuracy (Evermann et al. 2017; Mehdiyev and Fettke 2021). This progression has yielded significant operational value by enabling proactive, data-driven decision support.

However, the very complexity that drives the high performance of these advanced models simultaneously renders them opaque, creating a significant "black-box" problem (Guidotti et al. 2018). This lack of transparency is a critical barrier to adoption, particularly in high-stakes domains where understanding the rationale behind a prediction is as important as the prediction itself. For these powerful tools to be trusted and effectively integrated into operational decision-making, it is essential for stakeholders to comprehend the reasoning that underpins their outputs (Márquez-Chamorro et al. 2017). Consequently, a growing body of research has begun to focus on enhancing the interpretability and explainability of these models within the PPM context.

Despite the surge in academic interest, the literature on explainable and interpretable PPM remains fragmented. The rapid proliferation of studies has resulted in a scattered landscape of knowledge, making it difficult for researchers to identify critical gaps and for practitioners to make informed decisions about the most suitable methodologies for their needs. This paper addresses this gap by presenting a systematic literature review (SLR), conducted according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) framework (Page et al. 2021), to provide a comprehensive and structured synthesis of the field. We survey the literature of the last decade, distinguishing between intrinsically interpretable models and black-box approaches that require post-hoc explanation techniques, to map the current state of research and outline future directions.

Building on this rigorous systematic survey, this paper offers the following key contributions to the study of explainable and interpretable predictive process monitoring:

- *Comprehensive, structured panorama of the field* This review consolidates scattered work into a single, well-organized synthesis that links application domains, benchmark datasets and predictive tasks to the explainable artificial intelligence (XAI) techniques employed.
- *Unified taxonomy and critical appraisal of methods* All intrinsically interpretable models and post-hoc explanation approaches are classified and compared in terms of their transparency mechanisms, data requirements and present limitations when applied to event logs.
- *Systematic evaluation audit* A detailed examination of experimental designs, quantitative metrics, qualitative user studies, and functional, application-grounded and human-grounded tests yields harmonized guidelines for judging explanation quality and reproducibility.
- *Evidence-based research agenda* Cross-study comparison reveals under-explored domains, dataset biases, untested method combinations and missing evaluation evidence, thereby outlining concrete gaps and priorities for future work.
- *Actionable guidance for practitioners and researchers* By matching domain characteristics, data availability and explanation needs to suitable XAI approaches, the review provides a decision-support framework that promotes transparency, reliability and user trust in real-world predictive-process-monitoring deployments.

The remainder of this paper is organized as follows. Section 2 details the methodology of our systematic literature review, outlining the formal foundations of PPM and XAI, the research questions guiding our analysis and the PRISMA-based protocol for literature search, selection and synthesis. Section 3 presents a comprehensive discussion of our findings, systematically addressing our research questions by analyzing the application contexts, the landscape of interpretable and explainable AI methods and the evaluation paradigms employed in the reviewed literature. Section 4 provides a broader discussion, situating our contributions relative to existing surveys and exploring the key challenges, open issues, and the practical, scientific and theoretical implications of our findings. Finally, Sect. 5 outlines promising directions for future work, and Sect. 6 concludes the paper with a summary of its contributions.

2 Methodology

Our systematic literature review employs a rigorously structured methodology aligned with the PRISMA guidelines to ensure transparency and reproducibility (Page et al. 2021). It unfolds in six tightly linked subsections. We first ground the study in Formal Foundations (Section 2.1), defining core process-mining concepts, the predictive process-monitoring pipeline and the distinction between interpretable and explainable machine-learning models. Building on this base, we articulate the Rationale and Objectives (Section 2.2), converting the field's open issues into precise research questions. The next three subsections detail how these goals are operationalized. Information Sources, Search Strategy and Selection Process (Section 2.3) specifies where and how the literature was retrieved, while Eligibility Criteria (Section 2.4) formalizes the inclusion and exclusion rules that guard the review's scope and rigor. Data Collection and Synthesis Methods (Section 2.5) explains how evidence is

extracted and thematically integrated through template analysis. Finally, Study Selection and Descriptive Analysis (Section 2.6) reports the descriptive analysis of PRISMA-based screening results, presenting the corpus on which all later analyses rest.

2.1 Formal foundations

The section explores core aspects of explainable and interpretable predictive process monitoring. It begins with primary ideas and formal definitions crucial to process mining, followed by a focus on predictive process monitoring, including the essential components of the data pipeline and various problem areas. In addition, it differentiates between interpretable and explainable ML, providing foundational understanding through formal definitions and relevant methods. This structured approach ensures a clear presentation of essential background, preparing for further study in the intersection of ML, interpretability/explainability and predictive process monitoring.

2.1.1 Predictive process monitoring

To enable a precise understanding of predictive process monitoring, we first introduce the formal definitions of its key data constructs such as events, traces, event logs, and their transformation into features and labels for supervised learning, based on established literature in the field (Polato et al. 2014; Teinemaa et al. 2019; De Leoni et al. 2015).

Definition 1 (**Event**) An *event* is denoted by the tuple $e = (a, c, t_{start}, t_{complete},$

$v_1, \ldots, v_n)$, where $a \in \mathcal{A}$ is a categorical variable denoting the process activity, $c \in \mathcal{C}$ is a categorical variable signifying the unique identifier for the trace, also called *case ID*, $t_{start} \in \mathcal{T}_{start}$ and $t_{complete} \in \mathcal{T}_{complete}$ represent the event's commencement and completion timestamp (utilizing an epoch time representation like Unix) respectively, and $v_1, \ldots, v_n$ denoting the event-specific attributes, where $\forall 1 \leq i \leq n : v_i \in \mathcal{V}_i$ denote the domain of the i^{th} attribute. Consequently, these variables create a multi-dimensional space for the universe of events $\mathcal{E}$.
In essence, an event in the context of predictive process monitoring is a multi-faceted entity characterized by its activity type, its association with a specific process trace, its start and completion times and any additional attributes that may be relevant. These elements collectively define a multi-dimensional space $\mathcal{E}$ which can be thought of as the set of all possible events that could occur in the system under study. The exemplary Table 1, derived from a manufacturing scenario, depicts an event in each row, with the first event being characterized by its *Activity* "Plasma Welding", its *Start Time* "2019-04-18 06:26:47", its *End Time* "2019-04-18 09:51:25", the resource (*Worker ID*), the *Processing Time* "03:24:38" as well as other variables. Based on Definition 1 we now define traces and partial traces:

Definition 2 (**Trace, Partial Trace, Prefix and Suffix**) A *trace* $\sigma \in \mathcal{E}^*$ is a finite sequence of unique events $\sigma = \langle e_1, e_2, \ldots, e_{|\sigma|} \rangle$, with $|\sigma|$ denoting the amount of events in the trace, also called *trace length*, ordered chronologically and pertaining to a shared trace identifier $c \in \mathcal{C}$, also called *case ID*. We denote the set of all possible traces by $\mathcal{S} \subseteq \mathcal{E}^*$, with each trace $\sigma \in \mathcal{S}$ belonging to this universe. A *partial trace* is a subsequence $\sigma' = \langle e_{i_1}, e_{i_2}, \ldots, e_{i_k} \rangle$

Table 1 Process event log sample

Case ID	Activity	Start time	End time	Worker ID	...	Processing time
162384	Plasma	2019-04-18	2019-04-18	409	..	03:24:38
	Welding	06:26:47	09:51:25			
162384	Grinding	2019-04-18	2019-04-18	108	..	06:55:44
	Weld. Seam	12:11:30	19:07:14			
162384	Dishing	2019-04-23	2019-04-23	150		07:43:40
	Press (2)	10:50:31	18:34:11			
162384	Beading	2019-04-24	2019-04-24	726		09:37:32
		10:20:13	19:57:45			
162384	X-Ray	2019-04-25	2019-04-25	703		00:00:09
	Examination	10:26:23	10:26:32			
162384	Edge	2019-04-26	2019-04-26	742		03:41:49
	Deburring	09:08:38	17:50:27			
..	..	..	..	..	..	
177566	3D Micro-	2021-06-21	2021-06-21	139		03:21:59
	step	07:04:38	10:26:37		..	
..	..	..	..	..	..	..

Exemplary event log, depicting the trace identifier (*Case ID*), timestamps for *Start time* and *End time*, the executed *Activity*, the executing resource (*Worker ID*), as well as a label (*Processing time*)

of a given trace σ, where $1 \leq i_1 < i_2 < \ldots < i_k \leq |\sigma|$ and $1 \leq k < |\sigma|$. A partial trace also shares the same unique identifier $c \in \mathcal{C}$ as its parent trace σ. The set of all possible partial traces derived from σ is denoted by $\mathcal{S}_{\sigma'}$.

The *prefix* and *suffix* denote specific types of partial traces, yielded by employing the $hd^i(\sigma_c)$ and $tl^i(\sigma_c)$ functions, respectively. This is realized by employing a selection operator (.): $\sigma(i) = \sigma_i, \forall i \in [1, |\sigma|] \subset \mathbb{N}$, such that $hd^i(\sigma) = \langle e_1, e_2, \ldots, e_{\min(i,|\sigma|)} \rangle$ and $tl^i(\sigma) = \langle e_w, e_{w+1}, \ldots, e_{|\sigma|} \rangle$, where $w = \max(1, |\sigma| - i + 1)$.
In Table 1, two traces are depicted with the the *Case ID*s "162374" and "177566". The first trace starts with "Plasma Welding" and concludes with "Edge Deburring", while the second trace is initiated with "3D Microstep" and terminated after "Surface Polishing", with the events pertaining to a trace following a chronological order.

Definition 3 (Event Log) An *event log* is denoted by the set Log, where $Log = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ and $\sigma_i \in \mathcal{S}$ for $1 \leq i \leq n$, $n \in \mathbb{N}^+$. Each σ_i is a trace as previously defined. The event log Log is a collection of traces that may or may not share the same unique identifiers $c \in \mathcal{C}$.

Based on Definition 3, Table 1 represents an excerpt from an event log. Such event logs can be utilized to extract features and labels, which can then be leveraged for the construction of predictive models:

Definition 4 (Feature Extraction) Feature extraction is a mapping function denoted by $\phi : \mathcal{E} \cup \mathcal{S} \rightarrow \mathcal{X}$, where $\mathcal{E}$ is the set of all possible events, $\mathcal{S}$ is the set of all possible traces, and $\mathcal{X}$ is the feature space. Given an event $e \in \mathcal{E}$ or a trace $\sigma \in \mathcal{S}$, the function ϕ transforms it into a feature vector $x \in \mathcal{X}$.

For event-level feature extraction, $\phi_{\text{event}} : \mathcal{E} \rightarrow \mathcal{X}_{\text{event}}$ maps each event e to a feature vector x_{event} in the event-level feature space $\mathcal{X}_{\text{event}}$, while for trace-level feature extraction, $\phi_{\text{trace}} : \mathcal{S} \rightarrow \mathcal{X}_{\text{trace}}$ maps each trace σ to a feature vector x_{trace} in the trace-level feature space $\mathcal{X}_{\text{trace}}$.

Definition 5 (**Labeling**) Let $\mathcal{Y}$ be the set of all possible response variable values. For a non-empty trace $\sigma \neq \langle\rangle$ such that $\sigma \in \mathcal{S}$ and $\mathcal{S} \subseteq \mathcal{E}^*$, the labeling function $\text{resp}_{event} : \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{Y}$, $\text{resp}(e, \sigma) = y$ maps an event e within the trace σ to its respective response variable value $y \in \mathcal{Y}$ and is defined for all $e \in \sigma$ and $\sigma \in \mathcal{S}$. The labeling function $\text{resp}_{trace} : \mathcal{S} \rightarrow \mathcal{Y}$, $\text{resp}(\sigma) = y$ maps a trace σ to its respective response variable value $y \in \mathcal{Y}$ and is defined for all $\sigma \in \mathcal{S}$.

The concepts of feature extraction and labeling serve as a mechanisms to associate specific attributes or outcomes with individual events within a trace. By mapping each event or trace to a response variable, the labeling function facilitates the transformation of raw event data into a format amenable to analytical or ML methods. This enables researchers and practitioners to derive insights, make predictions or evaluate hypotheses based on the labeled data. The feature extraction and labeling functions thus acts as bridges between the raw, multi-dimensional event space and the target outcomes or attributes, thereby enriching a dataset for more advanced analyses. On the basis of previous definitions, we are now able to formalize the concept of supervised learning in the context of predictive process monitoring:

Definition 6 (**Supervised Learning**) Supervised learning is a paradigm in ML where a predictive model is constructed based on a labeled dataset. The dataset $\mathcal{D}$ is generated from an event log Log, feature extraction function $\phi : \mathcal{E} \cup \mathcal{S} \rightarrow \mathcal{X}$, and a use-case-dependent labeling function $\text{resp} : \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{Y}$ or $\text{resp} : \mathcal{S} \rightarrow \mathcal{Y}$. Each entry in $\mathcal{D}$ is a tuple (x, y), where $x \in \mathcal{X}$ is a feature vector and $y \in \mathcal{Y}$ is the corresponding response variable.

The dataset $\mathcal{D}$ is partitioned into training $\mathcal{D}_{\text{train}}$, validation $\mathcal{D}_{\text{val}}$ and testing $\mathcal{D}_{\text{test}}$ subsets. A predictive model $f : \mathcal{X} \rightarrow \mathcal{Y}$ is trained on $\mathcal{D}_{\text{train}}$ by minimizing a loss function $\mathcal{L}(f(x), y)$.

The validation set $\mathcal{D}_{\text{val}}$ is utilized for hyperparameter tuning and to mitigate the risk of overfitting. The testing set $\mathcal{D}_{\text{test}}$ is employed to evaluate the generalization performance of the model, providing an unbiased assessment of its predictive capabilities.

It should be noted that supervised learning on the event level can be considered a special case of trace-level supervised learning, in that partial traces of length one are being employed. With a variety of predictive process monitoring application scenarios (see Fig. 1), we provide definitions for predominant prediction tasks:

Definition 7 (**Process Outcome Prediction**) Given a labeling function $\text{resp}_{\text{outcome}} : \mathcal{S} \rightarrow \mathcal{Y}_{\text{outcome}}$ mapping each (partial) trace σ to its final outcome y_{outcome}, the predictive model $f_{outcome} : \mathcal{X} \rightarrow \mathcal{Y}_{\text{outcome}}$ is constructed via supervised learning to approximate this function.

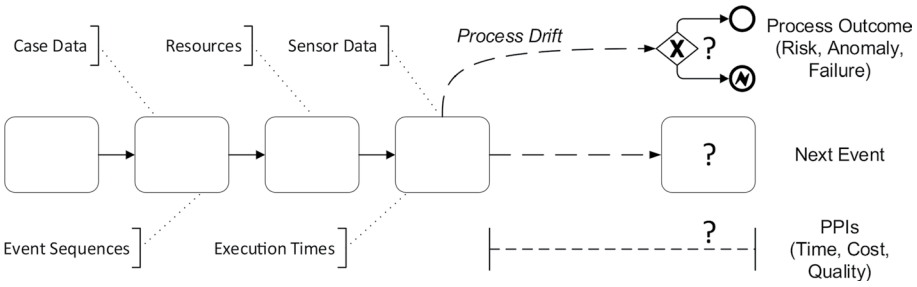

Fig. 1 Sources of input data accumulated in an event log and predictands of supervised learning (Rehse et al. 2018)

Definition 8 (**Next Event Prediction**) Given a labeling function $\text{resp}_{\text{next}} : \mathcal{E} \times \mathcal{S} \rightarrow \mathcal{E}_{\text{next}}$ mapping each event e within a trace σ to its subsequent event e_{next}, the predictive model $f_{next} : \mathcal{X} \rightarrow \mathcal{E}_{\text{next}}$ is constructed via supervised learning to approximate this function.

Definition 9 (**Process Performance Indicator (PPI) Prediction**) Given a labeling function $\text{resp}_{\text{PPI}} : \mathcal{S} \rightarrow \mathcal{Y}_{\text{PPI}}$ mapping each (partial) trace σ to a performance metric y_{PPI}, the predictive model $f_{PPI} : \mathcal{X} \rightarrow \mathcal{Y}_{\text{PPI}}$ is constructed via supervised learning to approximate this function.

Process data facilitates the development of predictive models that serve various objectives. These include the identification of the next likely activity (Evermann et al. 2017; Sindhgatta et al. 2020), the process outcome prediction (Mehdiyev and Fettke 2021; Rizzi et al. 2020), anomaly detection (Böhmer and Rinderle-Ma 2020; Pauwels and Calders 2019a) and remaining time prediction (Polato et al. 2014, 2018).

When it comes to developing accurate, reliable and suitable models for the specific application context, the complexity and variability inherent in modern business processes may pose significant challenges. Additionally, the complexity of the models required to make such predictions is rising in tandem with the demand for more sophisticated estimations. Specifically, opaque models frequently achieve high predictive accuracy, which makes them appealing choices. Having said that, the complexity of these models presents a significant disadvantage, as they can be extremely difficult to grasp. For practical applications, where it is essential to comprehend the reasoning behind predictions to establish trust and make decisions, this is a significant limitation that must be considered (Márquez-Chamorro et al. 2017; Di Francescomarino et al. 2018). As a result, the development of models that strike a balance between accuracy and interpretability continues to be a significant challenge in the field of predictive process monitoring despite the fact that this area has tremendous potential.

2.1.2 Interpretable and explainable AI

To clearly distinguish between interpretable and explainable machine learning models in the context of predictive process monitoring, we now present formal definitions and classifications of the main model types and explanation techniques, grounded in established research (Guidotti et al. 2018; Barredo Arrieta et al. 2020).

Definition 10 (**Intrinsically Interpretable Model**) Let $\mathcal{M}$ be the class of predictive models. A model $f \in \mathcal{M}$ is termed an *intrinsically interpretable model* if it possesses a humanly interpretable internal structure, denoted by $\mathcal{I}(f)$, such that $\mathcal{I}(f) : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z}$ is the space of humanly interpretable representations.

Considering a production process scenario where the objective is to predict the remaining time until case completion, an intrinsically interpretable approach might involve using a DT that makes its predictions based on a small set of easily interpretable features, such as the type of activity and the duration of the previous event. Because DTs are inherently interpretable, the model satisfies the interpretability constraints $\mathcal{I}(f)$ intrinsically. Among approaches that are commonly considered intrinsically interpretable, Stierle et al. (2021) differentiate between rule-based (for example (evolutionary) decision rules (Malioutov et al. 2017; Márquez-Chamorro et al. 2017)), regression-based (for example logistic regression (Teinemaa et al. 2016)), tree-based (for example decision trees (DTs) (Allah Bukhsh et al. 2019)) and probabilistic models (for example Bayesian networks (Dey and Stori 2005)). Additionally, algorithmically transparent approaches like k-Nearest Neighbors (k-NN) (Kumar et al. 2005) as well as generalized additive models (GAMs) (Coussement et al. 2010) are generally considered transparent as well (Barredo Arrieta et al. 2020). Nonetheless, it is worth noting that these white-box models are often outperformed by more complex, opaque models in terms of predictive accuracy (Guidotti et al. 2018).

Definition 11 (**Black-Box Model**) Let $\mathcal{M}$ be the class of predictive models. A model $f \in \mathcal{M}$ is termed a *black-box model* if its internal structure is not readily humanly interpretable, denoted by $\mathcal{I}(f) = \emptyset$.

The characteristics of black-box models encompass a complexity in their behavior and decision making processes which necessitate post-hoc explanations for understanding, with deep learning (DL) methods (like convolutional neural networks (CNN), deep feedforward neural networks (DNN) or recurrent neural networks (RNN)) (Mehdiyev and Fettke 2021; Sindhgatta et al. 2020), gradient boosting models (GBM) (Petsis et al. 2022) and RFs (RF) (Verenich et al. 2016) being among the most prominent.

Definition 12 (**Local Post-hoc Explanations**) Let $\mathcal{M}$ be the class of predictive models and $f \in \mathcal{M}$ be a specific model with predictive mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$. A local explanation is denoted by $f_{\text{local}} : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}_{\text{local}}$, where $\mathcal{Z}_{\text{local}}$ is the space of interpretable local representations. For a given instance $(f, x, y) \in \mathcal{M} \times \mathcal{X} \times \mathcal{Y}$, $f_{\text{local}}(f, x, y)$ explains the model's decision $f(x) = y$ in the vicinity of x. Model-agnostic local explanations can take any $f \in \mathcal{M}$ as input, whereas model-specific local explanations are restricted to a subset $\mathcal{M}_{\text{local},f} \subset \mathcal{M}$.

Prominent examples of local post-hoc explanations are individual conditional expectation (ICE) plots (Goldstein et al. 2013), Shapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) or local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), which are model-agnostic approaches. Model-specific approaches finding use in deep neural networks are layer-wise relevance propagation (Montavon et al. 2019) or DeepLIFT (Shrikumar et al. 2017). For tree-based models exhibiting a high complexity, tree Shapley

Additive exPlanations (TreeSHAP) (Lundberg et al. 2020) realizes a model-specific explanation technique.

Definition 13 (**Global Post-hoc Explanations**) Let $\mathcal{M}$ be the class of predictive models and $f \in \mathcal{M}$ be a specific model with predictive mapping $f : \mathcal{X} \to \mathcal{Y}$. A global explanation is denoted by $f_{\text{global}} : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}_{\text{global}}$, where $\mathcal{Z}_{\text{global}}$ is the space of interpretable global representations. The function $f_{\text{global}}(f, \mathcal{X}, \mathcal{Y})$ elucidates the model's overall decision-making mechanism across the entire domain $\mathcal{X}$. Model-agnostic global explanations can take any $f \in \mathcal{M}$ as input, whereas model-specific global explanations are restricted to a subset $\mathcal{M}_{\text{global},f} \subset \mathcal{M}$.

Prominent examples of global, model-agnostic post-hoc explanations are accumulated local effects (ALE) (Apley and Zhu 2020), decision rules (Frank and Witten 1998; Malioutov et al. 2017), feature importance (Fisher et al. 2019), partial dependence plots (PDP) (Friedman 2001) (also in conjunction with ICE plots (Goldstein et al. 2013)) and global surrogate models like CART decision trees (Rutkowski et al. 2014).

2.2 Rationale and objectives

The rationale for carrying out this SLR is firmly grounded in the ever-evolving and fast-paced domain of interpretable and explainable AI. In recent years, there has been also a significant increase in the number of academic studies that concentrate on the implementation of pertinent methodologies and concepts for the purpose of predictive process monitoring. Nevertheless, the rapid proliferation of academic investigation, combined with a lack of comprehensive meta-analytical studies, has resulted in a fragmented landscape of knowledge. The absence of a systematic framework and cohesive integration of knowledge presents notable challenges for researchers and practitioners alike, rendering the synthesis and practical application of existing information a formidable task. The primary objectives of this SLR are focused on providing nuanced understanding of the PPM domain. Through a comprehensive analysis of the existing research landscape, rigorous evaluation of the methodologies used, awareness of gaps, and the provision of unambiguous evidence-based recommendations, our aim is to enhance the quality and reliability of research conducted within this field. This study adheres to answering the following research questions:

RQ1—Application domain Which real-world domains (e.g., finance, healthcare, manufacturing) are most addressed by explainable PPM studies and how do domain characteristics guide model selection and explanation requirements?

RQ2—Benchmark datasets Which public event logs (BPIC series, Sepsis, Helpdesk, etc.) are used most often in explainable PPM, what key features do they contain and how do those features affect benchmarking and generalizability?

RQ3—Application tasks Which predictive tasks, such as process outcome, next event, time-related, or other PPIs, dominate explainable PPM research and how do task demands influence the pairing of models with explanation techniques?

RQ4—Interpretable AI Which families of intrinsically interpretable models (rules, trees, GAMs, Bayesian, k-NN) are favored for PPM and what design aspects ensure their transparency on event-log data?

RQ5—Explainable AI Which post hoc methods (e.g., SHAP, LIME, TreeSHAP, LRP) explain black-box PPM models and how are they distributed across local vs. global and model-agnostic vs. model-specific categories?

RQ6—Study evaluation How do individual explainable PPM studies structure and report their evaluations—covering dataset choice, baselines, predictive metrics and explanation quality measures - to ensure rigor and reproducibility?

RQ7—Quantitative versus qualitative measures What relative strengths and weaknesses emerge when quantitative metrics (fidelity, stability, sparsity) are compared with qualitative user studies in assessing explanation usefulness in PPM?

RQ8—Evaluation paradigms How are functional, application-grounded and human-grounded evaluation paradigms applied in PPM research, and what insights do they yield about explanation quality and decision support?

2.3 Information sources, search strategy, selection process

We have explored various online databases including ACM Digital Library, AIS eLibrary, IEEE Xplore, Science Direct and SpringerLink to gather relevant publications. These databases, which include but are not limited to topic-specific literature, were searched via queries. The search queries are specified as follows: Each query includes one of the terms "business process prediction", "predictive process monitoring", "prescriptive process analytics", or "process mining" and are combined with either of the terms "expla*", "interpretab*" or "XAI" via the AND-operator, in order to narrow the results to domain-specific subjects. Where it was possible, the following query was used to yield any potentially relevant literature from a database: Q_{comp}= (expla* OR interpret* OR XAI) AND ("process mining" OR "business process prediction" OR "predictive process monitoring" OR "prescriptive process analytics"). The Symbol "*", as in "expla*", is being used as a wildcard if a database allowed the usage of wildcards. In databases that did not allow using wildcards, the terms "explanation", "explainable" and "explainability" were used instead of "expla*", as well as "interpretable" and "interpretability" instead of "interpret*". Table 2 presents a concise summary of the composition and usage of queries in case Q_{comp} could not be processed by a database.

The inconsistencies between the search tools of each of the aforementioned databases make it challenging to conduct a systematic literature search using only the specified queries. In order to conduct an exhaustive search, the queries were applied to the title, keywords and complete text where it was possible:

- For the ACM digital library, the "Search items from"-option was set to "The ACM full-text collection", the queries were searched within "Anywhere" (see "search within"-option) and the filter "research article" was applied to reduce the number of irrelevant results.
- For the AIS eLibrary, the queries were searched within "all fields" and restricted to "peer-reviewed only" articles.
- For the IEEE Xplore, the queries were searched using the "command search"-tool
- For the Springer Nature Link, due to the large volume of irrelevant results, the search was restricted to the content type "article" as well as to the subdiscipline "artificial intelligence".

Table 2 Summary of employed search queries for retrieval of relevant literature

Representation	Search query	Used for querying databases
Q_1	"Business process prediction"	False
Q_2	"Predictive process monitoring"	False
Q_3	"Prescriptive process analytics"	False
Q_4	"Process mining"	False
Q_5	"Expla*"	False
Q_6	"Interpretab*"	False
Q_7	"XAI"	False
$Q_{1,5}$	Q_1 AND Q_5	True
$Q_{1,6}$	Q_1 AND Q_6	True
$Q_{1,7}$	Q_1 AND Q_7	True
$Q_{2,5}$	Q_2 AND Q_5	True
$Q_{2,6}$	Q_2 AND Q_6	True
$Q_{2,7}$	Q_2 AND Q_7	True
$Q_{3,5}$	Q_3 AND Q_5	True
$Q_{3,6}$	Q_3 AND Q_6	True
$Q_{3,7}$	Q_3 AND Q_7	True
$Q_{4,5}$	Q_4 AND Q_5	True
$Q_{4,6}$	Q_4 AND Q_6	True
$Q_{4,7}$	Q_4 AND Q_7	True
Q_{comp}	(Q_1 OR Q_2 OR Q_3 OR Q_4) AND (Q_5 OR Q_6 OR Q_7)	True

Following the database querying, the resulting literature was filtered using pre-defined criteria (for details, see Sect. 2.4). Subsequently, a forward and backward search was conducted on the results to capture additional topic-relevant publications that could not be discovered by searching the databases directly, including relevant articles from the arXiv outlet as well.

2.4 Eligibility criteria

Studies retrieved through a systematic search may nevertheless exhibit characteristics that are not topic-specific for this systematic review, necessitating additional screening to meet research rigor. Therefore, inclusion and exclusion criteria for the literature are defined. The identified literature must satisfy all of the predefined inclusion criteria while also not meeting any of the exclusion criteria in order to be considered for inclusion. A comprehensive list of all inclusion and exclusion criteria can be found in Table 3.

These criteria were applied in the following manner: After querying a database, the title and abstract of each of the resulting publications were analyzed respectively with regard to the inclusion and exclusion. This represents first filtering step after the retrieval of literature. The next filtering step takes place by expanding the analysis from title and abstract to the full text of each publication that passed the first filtering step. Based on the results of the second filtering step, a forward and backward search was conducted, which immediately applied filtering with the previously described inclusion and exclusion criteria. No temporal limits were imposed on the database searches, although it must be acknowledged that relevant studies may have been omitted if they were not indexed in the selected sources.

Table 3 Inclusion and exclusion criteria

Representation	Criteria for	Description
IN_1	Inclusion	Publication outlet is a peer-reviewed source e.g. journal, conference proceedings, etc
IN_2	Inclusion	Publication addresses PPM tasks
IN_3	Inclusion	Publication incorporates XAI methodology
IN_4	Inclusion	Publication is written in English
EX_1	Exclusion	Publication outlet is not a peer-reviewed source and not identified by forward-/backward search
EX_2	Exclusion	Publication does not address PPM tasks
EX_3	Exclusion	Publication neither incorporates XAI methodology nor uses any interpretable methods
EX_4	Exclusion	Publication does not use an event log
EX_5	Exclusion	Publication is not written in English

2.5 Data collection process and synthesis methods

The primary phase of our data collection procedure entails the methodical extraction of pertinent information from every chosen study. This encompasses, though is not exclusively confined to, the study's aims, predictive process monitoring and explainability approaches, results, and issues or contextual factors that are essential for comprehending its impact on the discipline. In order to uphold uniformity and precision, a standardized data extraction form is employed, encompassing all essential particulars that will subsequently prove pivotal in the synthesis and analysis stages.

After the completion of data collection, the research proceeds to the subsequent phase, known as a qualitative synthesis of studies. In this phase, the primary methodology employed is template analysis proposed by King (King 2012), which offers a flexible yet methodical framework for the thematic arrangement and understanding of textual data. The process of template analysis encompassed a series of fundamental stages, beginning with formulating an initial template. To ensure that our qualitative synthesis remains tightly aligned with the eight research questions (**RQ1**–**RQ8**), we re-engineered the template so that each top-level theme directly corresponds to one RQ. A dedicated branch now captures the evidence required for every question: *application domain* (**RQ1**) aggregates references to the sector or context addressed; *benchmark datasets* (**RQ2**) records which event logs are employed, their salient attributes and notes on accessibility; *Application Tasks* (**RQ3**) distinguishes outcome, next-event, time-related and other PPI predictions; *intrinsic models* (**RQ4**) collects details on rule-, tree-, GAM-, Bayesian- and k-NN-based approaches together with the features that make them transparent; *post-hoc methods* (**RQ5**) stores information on

SHAP, LIME, PDP, and similar techniques, tagging whether they provide local or global, model-agnostic or model-specific explanations; *evaluation design* (**RQ6**) registers dataset choices, baselines, predictive metrics and explanation-quality measures; *evaluation type* (**RQ7**) contrasts quantitative metrics (e. g., fidelity, stability, sparsity) with qualitative user studies; and *evaluation paradigms* (**RQ8**) logs whether assessments are functional, application-grounded, or human-grounded.

During coding, any newly encountered concept was inserted beneath its corresponding RQ branch, while redundant or overly granular codes were merged. Iteration ceased once no further themes emerged and the structure provided full coverage of the data. This RQ-driven template guarantees that every extracted datum feeds directly into answering a specific research question, thereby streamlining later aggregation and ensuring a transparent, auditable chain of evidence from primary study to final synthesis.

2.6 Study selection and descriptive analysis

The selection process commenced with the identification of records through an extensive search across multiple databases and registers, including ACM, AIS, IEEE, Science Direct, Springer Link as well as additional backward and forward searches. This initial step identified a total of 1,415 records as potentially eligible for inclusion. Each record was subjected to a careful screening process. Titles and abstracts were reviewed to determine their relevance to the study's inclusion criteria, which led to the exclusion of 1,279 records for not meeting the specified research scope and objectives as per inclusion criteria defined in Table 3. Consequently, 136 publications were selected for retrieval and further evaluation. In the eligibility assessment phase, the full texts of these studies were meticulously examined to ascertain their suitability for inclusion in the review. During this phase, articles were excluded based on predefined exclusion criteria (see Table 3), predominantly for not using event logs or not addressing PPM tasks. This resulted in the exclusion of an additional 29 articles. The culmination of this rigorous selection process was the inclusion of 107 studies in the final review. These studies were deemed to align closely with the research objectives and met all the criteria set forth for the systematic review. No additional reports of included studies were identified, affirming the thoroughness of the search and selection strategy. The transparent and systematic approach to study selection, as evidenced by the PRISMA flow diagram (see Fig. 2), aims to ensure a high level of confidence in the comprehensiveness and relevance of the studies included in this review. This process underscores the robustness and reliability of the findings and discussions that will be presented, providing a solid foundation for the synthesis and analysis that follow.

For metadata analysis, the publication venue, year and associated keywords of each article were examined: Of the 107 studies reviewed, 53 appeared in peer-reviewed journals, 51 in conference proceedings and three as arXiv preprints (see Fig. 3). Except for the arXiv entries, all venues comply with the peer-review standards mandated by systematic literature review protocols. Nonetheless, to ensure comprehensiveness, arXiv submissions identified via backward-search were retained.

Regarding the publishing date of identified literature, Fig. 4 illustrates the publications per year and publication medium in the form of a stacked bar chart. On closer examination, a spike in the amount of publications around the year 2020 can be observed. The majority of retrieved literature was published in 2020 and onward (76 out of 107 articles), with

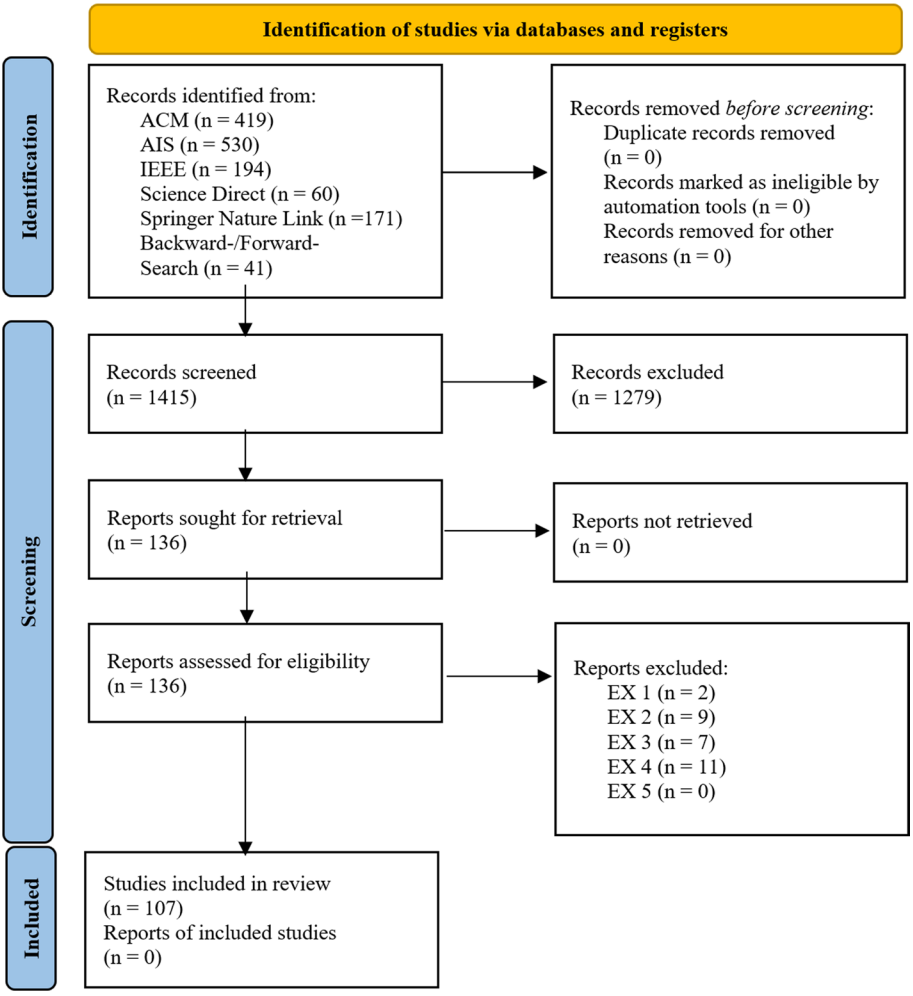

Fig. 2 Flowchart depicting the retrieval and selection of retrieved publications, following the PRISMA approach

2020, 2024 and 2022 being the years with the most publications (48 out of 107 articles), suggesting an increased relevance regarding the adoption of interpretable ML approaches for predictive process monitoring.

For the analysis of keywords, either chosen by the authors or proposed by the publication outlet, the identified articles were visualized via a circle packing chart depicted in Fig. 5, illustrating the employed keywords and corresponding frequency of occurrence. Visually, larger circles depict a more frequent use of the keyword (or phrase) within the circle compared to smaller circles, with "Process Mining" emerging as the most prevalent phrase, occurring in 41 publications. It is noteworthy that different representations of the same concepts were used, such as "explainable artificial intelligence" and "Explainable AI" being used as a key-phrase to depict the domain of an article. For the visualization, keywords describing the same concepts were grouped together under a single keyword. The analysis

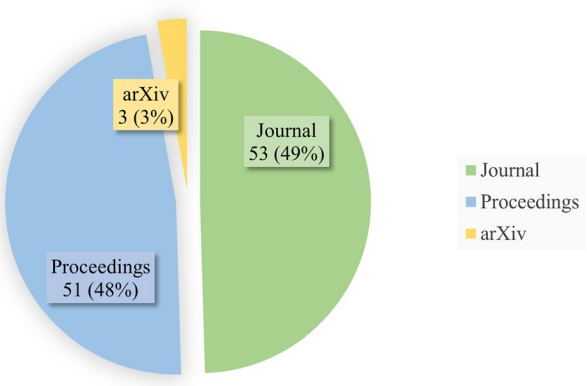

Fig. 3 Number of identified publications per publication outlet

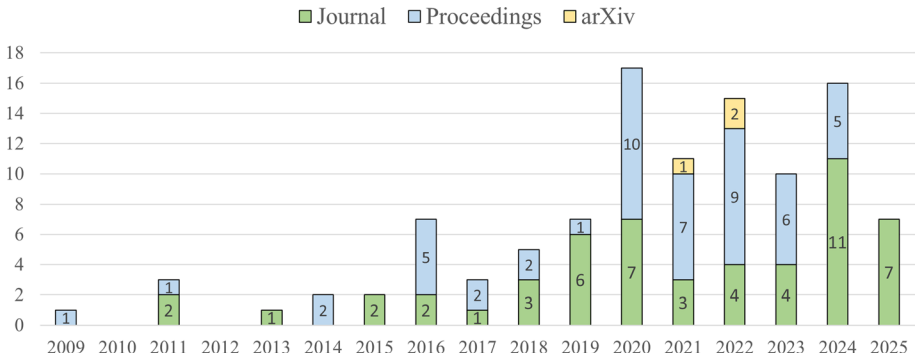

Fig. 4 Number of identified publications per publication outlet grouped by year of publication

of keywords shows, that approximately a third of the articles (36 out of 107) aimed to contribute directly to the XAI domain. Considering the search process for relevant literature, the variety in employed keywords and their formulation outlines the challenges in the adequate formulation of search queries in order to cover various iterations of the terminology specific to the XAI-domain.

3 Discussion of findings

This section presents the findings of the literature review and is systematically divided into four key subsections, each addressing a specific aspect of our research. Section 3.1 explores the application domains of the approaches described in the found articles. This part provides an in-depth look at the implications of our results in different domains and highlights prevalent application fields. Section 3.2 analyzes the employed approaches and ML models as well as the utilized explanation methods. Lastly, Sect. 3.3 examines the evaluation of employed explanation techniques. Each of these subsections collectively contributes to a comprehensive understanding of our research findings, offering a multi-faceted view of our study's impact and significance.

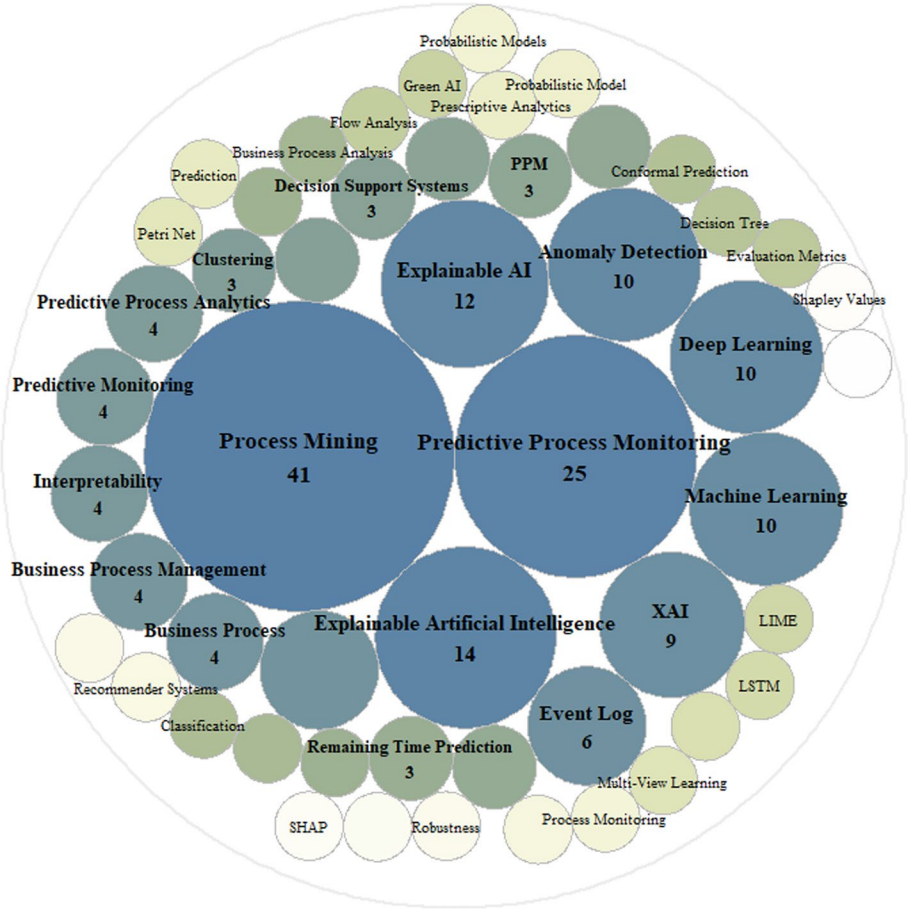

Fig. 5 Circle packing diagram of usage and frequency of article keywords

3.1 Application context

This subsection presents the examination of the retrieved publications, encompassing identified application domains **(RQ1)**, used benchmark datasets **(RQ2)** and central application tasks **(RQ3)**. For the remainder of this section, we refer to Tables 5, 6, 7, and 8 for a detailed documentation of application domains and tasks, as well as utilized datasets identified in the retrieved literature.

3.1.1 Application domain

For the identification of the application domain, the properties of the used data sets, as well as explicit statements of the authors, were analyzed and aggregated. These characteristics allow for the distinction between domain-agnostic and domain-specific applications of the presented approaches and give insight into the work areas covered in the literature (see Tables 5, 6, 7, and 8). As the most prevalent application domains finance (represented in 55

out of 107 articles), healthcare (31 out of 107 articles), customer support related services (25 out of 107 articles) and manufacturing (16 out of 107 articles) were identified. Approximately two fifths of the publications (42 out of 107) were identified as domain-agnostic, due to their independence towards the field of application, thus, demonstrating the transferability of the underlying methodology. For the rest of the publications, the transferability of findings from these studies to other domains is potentially challenging due to the unique structures of event logs, domain-specific methodologies, and tailored analytical approaches inherent in their respective fields. Considering the close relationship between the application domain and the data sets utilized for model training and evaluation, the following section provides a deeper analysis of the benchmark data sets used in the retrieved articles.

3.1.2 Benchmark datasets

Since the employed datasets dictate the possible application domains, examining the utilized event logs not only provides information about the presented application domains, but also about the degree of transferability and adaptiveness of the approaches presented in the analyzed articles. Figure 6 is a treemap diagram depicting the usage of various event logs, arranged by the frequency in ascending order, with the size of each area correlating to the amount of publications that used the corresponding dataset.

For the systematic literature review, we distinguished the business process intelligence challenge (BPIC) logs from four additional high-frequency datasets due to their prominence in predictive process monitoring. Other logs were either inaccessible or appeared too infrequently to merit individual discussion. The BPIC event logs span multiple real-world domains: BPIC 2011 comprises anonymized diagnoses and treatment records from the gynecology department of a Dutch academic hospital (healthcare). BPIC 2012 and BPIC 2017 both pertain to a Dutch financial institute, with the former event log covering personal-loan applications and the latter an upgraded loan process (finance). BPIC 2016 captures customer interactions (web, messaging, call centre) at the Dutch Employee Insurance Agency during unemployment-benefit requests (insurance). BPIC 2018 covers EU direct-payment applications by German farmers under the European Agricultural Guarantee Fund (finance). BPIC 2019 documents purchase-order handling and invoice-matching workflows at a multinational paints and coatings company (finance). BPIC 2013 originates from Volvo IT's

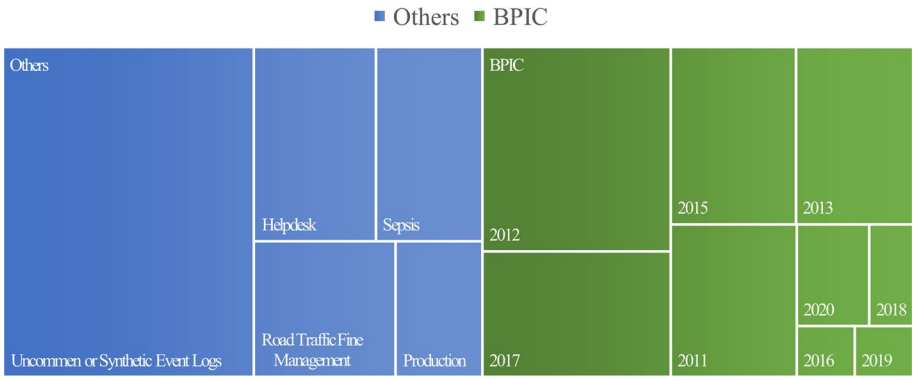

Fig. 6 Treemap diagram representing the usage of various event logs

VINST incident-management system and is therefore allocated to customer-support services. BPIC 2015 records municipal construction-permit applications in five Dutch cities and, despite its public-administration context, lacked sufficient representation in the literature to warrant a distinct domain label. BPIC 2020 introduces two years of business-travel and expense-management events for Eindhoven University of Technology employees, including travel permits, domestic and international expense declarations, prepaid costs and payment requests, and is classified under the finance domain as it reflects administrative travel-expense management.

Beyond BPIC, the four most frequently used additional datasets are Helpdesk, which pertains to customer-support services; Production, which involves manufacturing processes; Sepsis, which covers clinical healthcare pathways; and Road Traffic Fine Management, which relates to law-enforcement procedures. All other datasets were either synthetic, inaccessible, or employed too infrequently to be listed explicitly here. Table 4 illustrates the most frequently used event logs and provides a brief description as well as application domain.

In the found literature, the BPIC dataset catalogue is predominantly employed, with 57% (61 out of 107 articles) using at least one of the provided datasets. The usage of the same data over various publications facilitates the benchmarking of results, which is one of the main reasons for the utilization of the BPIC event logs stated within the articles.

Table 4 Predominantly used event logs within analyzed literature

Event log	Description	Application domain
BPIC 2011 (van Dongen 2011)	Academic hospital's care process	Healthcare
BPIC 2012 (van Dongen 2012)	Loan application process	Finance
BPIC 2013 (Steeman 2013)	Incident management processes	Customer support
BPIC 2014 (van Dongen 2014)	Incident and change management processes	Customer support
BPIC 2015 (van Dongen 2015)	Building permit application processes	Government
BPIC 2016 (Dees and van Dongen 2016)	Unemployment-benefit request process	Finance
BPIC 2017 (van Dongen 2017)	Loan application process	Finance
BPIC 2018 (van Dongen and Borchert 2018)	Subsidy application and payment process	Finance
BPIC 2019 (van Dongen 2019)	Purchase order handling process	Finance
BPIC 2020 (van Dongen 2020)	University's travel permit	Finance
Helpdesk (Polato 2017)	Ticketing process of a help desk	Customer support
Production (Levy 2014)	Manufacturing process event log	Manufacturing
Road Traffic Fine Management (de Leoni and Mannhardt 2015)	Fine management process	Customer support
Sepsis (Mannhardt 2016)	Hospital treatment process	Healthcare

Another reason is the open-source nature of these datasets, making them easily accessible to the public and therefore contributing to the transparency and replicability of the presented approaches. Lastly, all of the BPIC datasets are real-life event logs, facilitating approaches that aim to be grounded in reality. Regarding the frequency of utilization, the BPIC 2012 event log was employed the most (utilized in 26 out of 107 articles), thus contributing to the finance domain being the prevalent application domain. With 39% of articles (42 out of 107) implementing their approach on at least two event logs from differing application domains, 43% (46 out of 107 articles) evaluated their approaches on two or more datasets, examining the robustness of the proposed methodology across data from different sources.

3.1.3 Application tasks

The adoption of certain ML models depends heavily on the prediction tasks at hand. Especially in process prediction, there are prevalent prediction tasks that entail certain types of explanations as well as corresponding explanation objects and subjects. Since the prediction task is integral for the selection of the employed ML model and accompanying explanation methods, this section presents the application tasks of the retrieved articles and categorizes them into the following four groups: The first group deals with the prediction of process outcomes. These predictions often involve classifying events, traces or trace segments into predefined categories, such as identifying anomalies within a process at runtime. The second group focuses on the prediction of the next event in an unfinished process trace. In scenarios involving non-deterministic processes, various features, context factors and preceding events within the trace play a pivotal role in influencing the subsequent activity. The third and fourth group deals with the prediction of process performance indicators, with the third group particularly encompassing predictions of time-related PPI, such as the remaining time until completion for an event or an unfinished process trace. The fourth group is comprised of PPI prediction tasks unrelated to time, such as the prediction of context variables, costs and others. First, publications that aimed for the prediction of the next event are being presented, followed by those that predicted process outcomes. Afterwards, articles that predicted time-related process performance indicators are being presented, and lastly, literature with other process performance indicators prediction tasks.

3.1.3.1 Process outcome prediction Process outcome prediction emerges as a central theme within the reviewed body of literature, illustrating its prevalence and significance in diverse application contexts. Approximately 60% of the retrieved literature (65 out of 107 articles) explicitly focus on tasks related to the prediction of the outcomes of various processes. The correct prediction of process outcomes harbors considerable relevance across various domains, with finance and healthcare at the forefront among the analyzed literature (see Fig. 7). In each of these domains, the ability to foresee and act upon future outcomes provides a strategic advantage, making process outcome prediction an invaluable tool in operational decision-making and strategic planning.

The diversity of prediction tasks addressed within these articles underscores the adaptability of PPM techniques. These include trace classification or clustering, as seen in the works of De Koninck et al. (2017), De Oliveira et al. (2020a), De Oliveira et al. (2020b), Di Francescomarino et al. (2016), Francescomarino et al. (2019), Folino et al. (2017) and

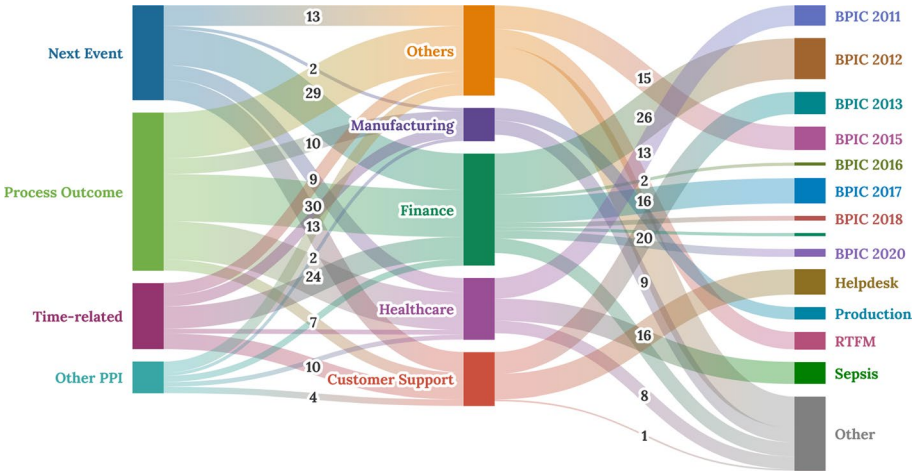

Fig. 7 Sankey-diagram representing the application task, the application domains and the corresponding application datasets. The line width represents the amount of scenarios found in the analyzed literature

Verenich et al. (2016). Anomaly detection, another prevalent focus, is extensively explored by Böhmer and Rinderle-Ma (2020), García-Bañuelos et al. (2017), Irarrázaval et al. (2021) and Pauwels and Calders (2019a, 2019b). Additionally, specific application-driven predictions such as maintenance (Allah Bukhsh et al. 2019), risk detection (Conforti et al. 2016) and insurance reclamation (De Leoni et al. 2015) further demonstrate the contextual specificity of the methodologies applied. Other articles examining process outcome prediction include Agarwal et al. (2022), Bezerra et al. (2009), Bezerra and Wainer (2011, 2013), Diamantini et al. (2024), Elkhawaga et al. (2023, 2024), Folino et al. (2011, 2024, 2025), Galanti et al. (2020, 2023a), Gupta et al. (2015), Harl et al. (2020), Horita et al. (2016), Huang et al. (2022), Khemiri et al. (2018), Kim et al. (2024), Lakshmanan et al. (2011), Maggi et al. (2014), Maita et al. (2025), Malashin et al. (2025), Mehdiyev and Fettke (2020a, 2020b, 2021), Mehdiyev et al. (2021), Montoya et al. (2023), Myers et al. (2018), Ouyang et al. (2021), Pasquadibisceglie et al. (2021, 2024), Porouhan (2024), Prasidis et al. (2021), Rauch et al. (2024), Rehse et al. (2019), Rizzi et al. (2020, 2024), Saini et al. (2020), Sarno et al. (2020), Savickas et al. (2014), Sindhgatta et al. (2020), Stevens and De Smedt (2022a), Stevens et al. (2022b, 2022c), Teinemaa et al. (2016), Tripathi et al. (2024), van Zelst et al. (2020), Velmurugan et al. (2021a, 2021b) and Völzer et al. (2023).

A crucial aspect of preparing data for process outcome prediction involves a pivotal decision point where all available data must be aggregated into a format that is amenable to the employed machine learning model. This transformation process is indispensable for aligning event log data with the specific requirements of various process outcomes. Such transformations may involve the normalization of numerical and categorical data formats, the aggregation of event attributes as well as the engineering and selection of new features to enhance the predictive capability of process monitoring systems. These adaptations are crucial not only for achieving accurate predictions but also for ensuring the robustness and transferability of predictive models across various domains. The documented literature thus highlights the intricate interplay between data characteristics and predictive model performance in process outcome prediction.

3.1.3.2 Next event prediction The prediction of the next event in an unfinished process trace is the second most prevalent application task within the retrieved literature, accounting for 30 out of 107 articles. Among the analyzed publications, next event prediction is predominantly employed in domains such as finance, healthcare and customer support related fields, where it facilitates real-time decisions distinct from the broader scope of process outcome prediction. This task is predominantly aimed at enhancing production processes through forward planning capabilities through its operational immediacy. The prediction of the next event often leads to immediate adjustments in the process execution, differing from the more strategic or overarching implications of process outcome predictions. While next event prediction shares the predictive process monitoring goal with process outcome prediction, the former uniquely focuses on the short-term sequence of activities within a process trace. For example, studies like those by Lakshmanan et al. (2011) and Unuvar et al. (2016) not only predict the next event but also extend to forecast subsequent activities up to the completion of a process trace. Moreover, next event prediction sometimes serves as a secondary outcome of broader research aims, as noted in the works of Maggi et al. (2014), where the main focus isn't solely on predicting the next event but encompasses a wider scope of process analysis. Similarly, Verenich et al. (2017, 2019) implement this prediction as an implicit function within their models, assigning probabilities to possible future states of a process trace. Other articles examining next event prediction include Agarwal et al. (2022), Aversano et al. (2023), Böhmer and Rinderle-Ma (2018, 2020); Brunk et al. (2021), De Leoni et al. (2015), Gerlach et al. (2022), Hanga et al. (2020), Hsieh et al. (2021); Kim et al. (2024), Majumdar et al. (2023), Mayer et al. (2021), Pasquadibisceglie et al. (2023, 2024a, 2024b), Rauch et al. (2024), Rehse et al. (2019), Rizzi et al. (2024), Savickas and Vasilecas (2018), Sindhgatta et al. (2020), Tama et al. (2020), Weinzierl et al. (2020), Wickramanayake et al. (2022a, 2022b) and Zilker et al. (2023).

3.1.3.3 Time related prediction Time-related prediction tasks within PPM are fundamentally geared towards forecasting temporal parameters that directly impact process efficiency and outcome. These tasks typically employ regression models to estimate variables such as the duration of tasks, intervals between events, or the completion time of ongoing processes. The complexity of these predictions stems from the need to precisely model the time-dependent aspects of process flows, which requires a deep understanding of process dynamics and the factors that influence time variations. Exemplary time-related prediction problems encompass the prediction of the timestamp of the next event (Böhmer and Rinderle-Ma 2018, 2020), the prediction of execution times of activities for a given trace (Rehse et al. 2019; Verenich et al. 2017, 2019) and the prediction of remaining time until completion for a given unfinished trace (De Leoni et al. 2015; Ouyang et al. 2021; Sindhgatta et al. 2020; Galanti et al. (2020, 2023a, 2023b). Other articles examining time related prediction tasks include Cao et al. (2023a, 2023b), Guo et al. (2024), Hermann et al. (2024), Mayer et al. (2021), Mehdiyev et al. (2024, 2025a, 2025b), Padella et al. (2022), Polato et al. (2018), Saha et al. (2024) and Toh et al. (2022).

Among the analyzed literature, these types of predictions were especially relevant in the finance sector and for customer support related tasks (see Fig. 7), predicting the time until a case is closed or a resolution is reached. As processing times can directly influence cus-

tomer satisfaction, enabling to manage expectations and allocate resources more efficiently is a central motivation for these prediction tasks. The intricacies involved in time-related prediction include the preprocessing of event logs where significant temporal features must be identified and extracted. These transformations involve handling large datasets with time-stamped events, dealing with missing time entries, as well as correcting or filtering anomalies in time data. Additionally, the selection of the right regression techniques and their calibration to align with the specific characteristics of the process at hand is vital with regard to ensuring the accuracy of the model's predictions.

3.1.3.4 Other process performance indicator predictions Beyond the time-related PPIs, PPM also encompasses a broad spectrum of other PPI-related prediction tasks, which are crucial for enhancing operational efficiency and strategic decision-making across various domains. The analyzed publications demonstrate that the quantification and estimation of these PPIs is specifically tailored to meet the unique needs of each application context and is implemented predominantly within the domains of finance and customer support (refer to Fig. 7). The goals of predicting these PPIs are multifaceted, with applications that typically aim to provide actionable insights that can lead to improved process outcomes. Examples include Bayomie et al. (2022), who develop a numeric indicator for event-case correlation, essential for understanding process interdependencies, as well as Coma-Puig and Carmona (2022), who focus on quantifying and predicting non-technical energy loss, which is a critical factor in operational efficiency. Similarly, Fu et al. (2021) work on quantifying and predicting customer experience scores, pivotal in customer support services. Galanti et al. (2020, 2023a, 2023b) predict costs associated with process traces to determine their relevance, an approach also adopted by Mayer et al. (2021) for comparable cost estimations. Additionally, Petsis et al. (2022) predict the number of patient visits, which is vital for resource planning in healthcare settings. The remaining articles examining prediction tasks related to other process performance indicators include Aguilar Magalhães et al. (2025), Hermann et al. (2024), Montoya and Astudillo (2023), Park et al. (2024), Rizzi et al. (2024), Saha et al. (2024) and Trescato et al. (2024).

3.1.3.5 Further insights The analysis of the surveyed literature reveals a significant emphasis on classification tasks in 87 unique studies, with 30 articles focusing on next event prediction and 65 on process outcome prediction. Regression tasks, featured in 32 out of 107 articles, primarily targeted time-related PPIs in 23 articles, while 15 articles examined other process-related PPIs. For a detailed understanding, the Sankey diagram in Fig. 7 consolidates the information from Tables 5, 6, 7, and 8 and visualizes the connections between the application tasks, domains and datasets utilized in these studies. This illustration highlights the predominance of process outcome predictions, followed by next event and time-related predictions. The finance domain is most frequently addressed, largely due to its prominent representation in the BPIC datasets, with the BPIC 2012 event log used in about a quarter of the articles (26 out of 107).

Regarding the interplay between utilized datasets, the application task and domain, our analysis suggests that the limited availability of publicly accessible process logs may substantially influence the scope and diversity of application domains and tasks within predic-

tive process monitoring, effectively restricting the range of research topics and curtails the generalizability and applicability of models and techniques across various industries. The dominance of certain datasets, like the BPIC 2012 and BPIC 2017 in finance or BPIC 2011 and Sepsis in healthcare, illustrates how the availability of domain-specific datasets has a potential to skew research focus toward particular industries and problem types.

3.2 Interpretable and explainable AI for PPM

This subsection turns to the methodological backbone of explainable and interpretable PPM. Guided by **RQ4** and **RQ5**, we first review the classes of *intrinsically interpretable models* reported in the literature and discuss the structural features, such as rule transparency, additive decompositions and proximity-based reasoning, that make these models understandable when applied to event-log data. We then survey the range of *post-hoc explanation techniques* used to illuminate black-box predictors, grouping them by the scope of their insights (local versus global) and their dependency on a specific model architecture (model-agnostic versus model-specific). A consolidated overview of the evidence extracted from individual studies is provided in Tables 9, 10, 11.

3.2.1 Interpretable AI for PPM

Intrinsically interpretable models such as DTs, linear regression and rule-based systems are favored for their transparency and ease of understanding, making them particularly suitable for domains where interpretability is critical for compliance and operational transparency. These models allow stakeholders to comprehend how predictions are made, which is crucial in sectors like healthcare and finance where decisions based on model predictions can have significant consequences.

Within the surveyed publications, DTs emerged as the most prevalent interpretable AI models, featured in 22 out of 64 articles employing white-box prediction models. The following articles provide a conceptual overview of the versatile utilization of these interpretable models: While Lakshmanan et al. (2011) implemented a binary DT using C4.5 on a synthetic event log to simulate an insurance claim scenario, focusing on predicting process outcomes, Maggi et al. (2014) developed a more sophisticated framework that classifies traces of an event log based on specific scenarios and use cases, building C4.5 decision trees to predict not only process outcomes, but the next events as well. The latter framework was operationalized within the ProM framework (van Dongen et al. 2005) and validated on the BPIC 2011 event log, with performance metrics including accuracy, AUROC, F-1 scores and ROC characteristics. Allah Bukhsh et al. (2019) applied Classification and Regression Trees (CART) method and evaluated the model performance alongside a RF and gradient boosting trees to predict maintenance requirements for railway switches. The models were assessed based on accuracy, F-1 scores, kappa and misclassification rates. De Leoni et al. (2015) implemented the proposed framework as a plug-in for the ProM framework and, given an event log as input, mines a process model yielding either a corresponding DT (C4.5, see Quinlan (1993) and Mitchell (1997)) or Regression Tree (RepTree, see Witten et al. (2011)). As application tasks, the presented framework allows for predicting upcoming events, process outcomes or the remaining time until process completion and was evaluated on the BPIC 2016 event log. Di Francescomarino et al. (2016) introduced a PPM framework

Table 5 Categorization of application task, application domain and utilized event log in the found literature

	Application Task				Application Domain						Event log														
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fine man.	Sepsis	Other
Agarwal et al. (2022)	■	■				■	■		■	■		■				■						■	■	■
Aguilar Magalhães et al. (2025)				■	■																			■
Aversano et al. (2023)	■					■			■	■		■				■								■
Bayomie et al. (2022)				■	■	■			■	■			■	■		■								
Bezerra et al. (2009)		■					■		■	■														■
Bezerra and Wainer (2011)		■				■																		■
Bezerra and Wainer (2013)		■					■		■	■														■
Böhmer and Rinderle-Ma (2018)	■		■		■	■				■		■								■				

Table 5 (continued)

	Application				Application						Event log													
	Task				Domain																			
	Class.		Regr.								BPIC										Misc.			
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fine man.	Sepsis	Other
Böhmer and Rinderle-Ma (2020)		■							■					■										■
Böhmer and Rinderle-Ma (2020)	■		■		■	■				■		■								■				
Brunk et al. (2021)	■				■	■				■		■	■											
Allah Bukhsh et al. (2019)		■							■															■
Cao et al. (2023a)			■			■										■		■						
Cao et al. (2023b)			■		■	■			3	1						■				■				
Coma-Puig and Carmona (2022)				■					■															■
Conforti et al. (2016)		■				■											■							

Table 5 (continued)

	Application				Application						Event log													
	Task				Domain																			
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Help-desk	Production	Road traffic fine man.	Sepsis	Other
De Koninck et al. (2017)		■			■	■				■			■											■
De Leoni et al. (2015)	■	■	■			■									■									
De Oliveira et al. (2020aa)		■							■	■														■
De Oliveira et al. (2020b)		■					■																	
Diamantini et al. (2024)		■							■															■
Di Francescomarino et al. (2016)		■					■	1	■	■	■			■										
Francescomarino et al. (2019)		■					■				■													
Elkhawaga et al. (2023)		■				■	■	■	■	■						■					■	■	■	

Table 5 (continued)

	Application				Application						Event log													
	Task				Domain																			
	Class.		Regr.								BPIC										Misc.			
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fine man.	Sepsis	Other
Elkhawaga et al. (2024)		■				■	■		■	■						■						■	■	
Folino et al. (2011)		■							■															■
Folino et al. (2017)		■			■								■											
Folino et al. (2024)		■					■				■												■	
Folino et al. (2025)		■					■				■													
Fu et al. (2021)				■					■															■
Galanti et al. (2020)		■	■	■	■	■				■		■	■							■				■
Galanti et al. (2023a)		■	■	■	■	■			■	■		■	■							■		■		■

Table 6 Categorization of application task, application domain and utilized event log in the found literature

	Application Task				Application Domain						Event log														
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Oth-ers	Do-main ag-nos-tic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Help-desk	Pro-duc-tion	Road traf-fic fine man.	Sep-sis	Other
Galanti et al. (2023b)			■	■					■															■
García-Bañuelos et al. (2017)		■							■													■		
Gerlach et al. (2022)	■					■				■									■					■
Guo et al. (2024)			■		■	■		■	■	■		■		■						■	■			
Gupta et al. (2015)		■				■																		■
Hanga et al. (2020)	■				■	■			■	■		■	■							■		■		
Harl et al. (2020)		■				■										■								
Hermann et al. (2024)			■	■				■																■
Horita et al. (2016)		■							■													■		
Hsieh et al. (2021)	■					■						■												
Huang et al. (2022)		■				■										■								
Irarrázaval et al. (2021)		■							■															■
Khemiri et al. (2018)		■						■																■

Table 6 (continued)

	Application				Application						Event log														
	Task				Domain																				
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Oth-ers	Do-main ag-nos-tic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Help-desk	Pro-duc-tion	Road traf-fic fine man.	Sep-sis	Other
Kim et al. (2024)	■	■				■	■		■	■	■	■		■									■	
Lakshmanan et al. (2011)	■	■				■																		■
Maggi et al. (2014)	■	■					■				■													
Maita et al. (2025)		■				■						■												■
Majumdar et al. (2023)	■								■															■
Malashin et al. (2025)		■				■																		■
Mayer et al. (2021)	■		■	■				■																■
Mehdiyev and Fettke (2020a)		■			■								■											
Mehdiyev and Fettke (2020b)		■				■									■									
Mehdiyev and Fettke (2021)		■						■																■
Mehdiyev et al. (2021)		■				■																		■
Mehdiyev et al. (2024)			■					■																■
Mehdiyev et al. (2025a)			■					■																■

Table 6 (continued)

	Application				Application						Event log													
	Task				Domain																			
	Class.		Regr.								BPIC										Misc.			
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fine man.	Sepsis	Other
Mehdiyev et al. (2025b)			■					■													■			■
Montoya and Astudillo (2023)				■		■										■								
Montoya et al. (2023)		■							■															■
Myers et al. (2018)		■						■																■
Ouyang et al. (2021)		■	■			■	■		■	■		■												
Padella et al. (2022)			■		■	■				■			■											■
Park et al. (2024)				■					■															■
Pasquadibisceglie et al. (2021)		■				■	■	■		■	■	■									■		■	

Table 7 Categorization of application task, application domain and utilized event log in the found literature

	Application Task				Application Domain						Event log														
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain Agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fineman	Sepsis	Other
Pasquadibisceglie et al. (2023)	■				■	■				■		■	■			■			■					■
Pasquadibisceglie et al. (2024a)	■				■	■	■		■	■				■		■			■	■			■	■
Pasquadibisceglie et al. (2024b)	■				■	■	■		■	■			■			■			■	■				■
Pasquadibisceglie et al. (2024)		■							■															■
Pauwels and Calders (2019a)		■				■											■							
Pauwels and Calders (2019b)		■				■			■	■				■			■							■
Petsis et al. (2022)				■			■																	■
Polato et al. (2018)			■		■				■											■		■		
Porouhan (2024)		■					■																	■
Prasidis et al. (2021)		■				■										■								
Rauch et al. (2024)	■	■			■	■			■	■		■							■	■		■		

Table 7 (continued)

	Application				Application						Event log														
	Task				Domain																				
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Others	Domain Agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Help-desk	Production	Road traffic fine man	Sepsis	Other
Rehse et al. (2019)	■	■	■				■																	■
Rizzi et al. (2020)		■					■				■													■
Rizzi et al. (2024)	■	■		■			■				■													
Saha et al. (2024)			■	■					■															■
Saini et al. (2020)		■							■															■
Sarno et al. (2020)		■				■																		■
Savickas et al. (2014)		■							■	■														■
Savickas and Vasilecas (2018)	■					■			■	■		■		■										■
Sindhgatta et al. (2020)		■	■			■	■		■	■	■	■		■										
Sindhgatta et al. (2020)	■				■	■			■	■		■	■	■						■				
Stevens and De Smedt (2022a)		■					■	■	■	■	■			■							■		■	
Stevens et al. (2022b)		■					■	■	■	■				■							■		■	
Stevens et al. (2022c)		■				■			■	■				■		■						■		

Table 7 (continued)

	Application				Application							Event log													
	Task				Domain																				
	Class.		Regr.									BPIC										Misc.			
Publication	Next event	Process outcome	Time-related	Other PPI	Customer support	Finance	Healthcare	Manufacturing	Oth-ers	Do-main Ag-nos-tic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Help-desk	Pro-duc-tion	Road traf-fic fine man	Sep-sis	Other
Tama et al. (2020)	■				■	■	■		■	■		■	■							■		■	■	
Teinemaa et al. (2016)		■				■																		■
Toh et al. (2022)			■			■																		■
Trescato et al. (2024)				■			■																	■
Tripathi et al. (2024)		■			■				■				■	■										
Unuvar et al. (2016)	■								■															■
van Zelst et al. (2020)		■					■			■													■	
Velmurugan et al. (2021a)		■				■	■	■		■		■									■		■	
Velmurugan et al. (2021b)		■				■	■	■		■		■									■		■	

Table 8 Categorization of application task, application domain and utilized event log in the found literature

	Application				Application						Event log													
	Task				Domain																			
	Class.		Regr.								BPIC										Misc.				
Publication	Next event	Process outcome	Time-related	Other PPI	Customer Support	Finance	Healthcare	Manufacturing	Others	Domain Agnostic	BPIC 2011	BPIC 2012	BPIC 2013	BPIC 2015	BPIC 2016	BPIC 2017	BPIC 2018	BPIC 2019	BPIC 2020	Helpdesk	Production	Road traffic fineman	Sepsis	Other
Verenich et al. (2016)		■					■				■													
Verenich et al. (2017)	■		■		■	■				■		■								■				
Verenich et al. (2019)	■		■		■	■	■		■	■		■								■		■		■
Völzer et al. (2023)		■							■													■		
Weinzierl et al. (2020)	■				■	■				■								■		■				
Wickramanayake et al. (2022a)	■					■				■		■				■								
Wickramanayake et al. (2022b)	■					■						■												
Zilker et al. (2023)	■						■																■	

that was integrated into the ProM framework to allow operation during runtime. This framework utilizes either frequency-based or sequence-based encoding for event logs, which are then processed using either agglomerative clustering, DBSCAN or K-Means Clustering. Following clustering, the framework allows for DTs and RFs to be employed as classification models alongside manual optimization of certain hyperparameters. Building on this foundation, Francescomarino et al. (2019) developed a subsequent version of the previous framework that also operates within ProM and introduces enhancements in the clustering stage by incorporating two distinct methods: model-based clustering, as outlined by Fraley and Raftery (2003), for frequency-based encoding and DBSCAN for sequence-based encoding.

Bayesian networks (8 articles) and linear or logistic regression models (12 articles), constitute for the further most prevalent approaches in the reviewed literature. Bayesian networks have been utilized for transparent analysis of event logs, tackling tasks such as next event prediction, process outcome forecasting and anomaly detection. As exemplary work, Brunk et al. (2021) employed a Dynamic Bayesian Network with a manually defined structure in order to predict the next event within a given trace of an event log. This approach aimed at differentiating attributes of the event log that are the cause or the effect of a given process and was evaluated on the BPIC 2012 and BPIC 2013 data sets. For benchmarking, implementations of probabilistic finite automata and n-grams were utilized to compare accuracy and various approaches presented in other publications for the given event logs. Similarly, linear or logistic regression models have been applied in diverse contexts to enhance decision-making processes. Agarwal et al. (2022) proposed a decision support system employing logistic regression for process outcome and next event prediction, while Stevens and De Smedt (2022a) and Stevens et al. (2022b) presented a methodology for process outcome prediction with a strong focus on the evaluation of model explanations. Teinemaa et al. (2016) presented an approach of predicting the process outcome for two real-life event logs by employing techniques from text-mining in order to encode process traces. A logistic regression model has been utilized as a classifier for said task and was benchmarked for computation time, F-1 scores and earliness, though it was noted that it was outperformed by RF models across all evaluation metrics.

Other white-box approaches, such as k-means clustering, heuristic rule-based clustering, and methods integrating multiple interpretable AI models, were explored across 39 of the 64 articles employing white-box models. These articles leveraged a variety of mixed approaches for diverse PPM tasks. For instance, Böhmer and Rinderle-Ma (2020) introduced sequential prediction rules in the context of next event prediction and evaluated their approach ("LoGo") on the BPIC 2012 and Helpdesk data sets based on the mean absolute error and accuracy, comparing their approach to LSTM and RNN models. These rules predict the next activity at a general level for specific event log traces, using probability-based heuristics as classifiers when no general rules apply. Conforti et al. (2016) introduced "PRISM," a risk detection model that operates in real-time during process execution, utilizing dedicated sensors developed from a risk-annotated process model. This model triggers alerts when predefined risk conditions are met and employs a similarity measure to proactively identify and manage risks in similar instances. Folino et al. (2017) present a rule-based clustering approach employing propositional patterns.

These studies showcase the adaptability and efficacy of white-box approaches in addressing specific predictive needs in process monitoring, enhancing both the interpretability and

Table 9 Categorization of employed ML and explanation methods in the found literature, segmented into model interpretability, explanation scope, explanation relation and explanation format

	Interpretable AI				Explainable AI																							
									Scope														Relation		Format			
	White-Box				Black-Box				Local							Global												
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Agarwal et al. (2022)			■											■				■		■				
Aguilar Magalhães et al. (2025)			■		■		■	■							■				■	■	■		■	
Aversano et al. (2023)					■									■						■	■			■
Bayomie et al. (2022)				■										■				■		■		■		
Bezerra et al. (2009)				■										■				■		■		■		
Bezerra and Wainer (2011)				■										■				■		■		■		
Bezerra and Wainer (2013)				■										■				■		■		■		
Böhmer and Rinderle-Ma (2018)				■										■						■	■			

Table 9 (continued)

	Interpretable AI				Explainable AI																					
									Scope												Relation		Format			
	White-Box				Black-Box				Local						Global											
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Böhmer and Rinderle-Ma (2020)				■										■						■	■			
Böhmer and Rinderle-Ma (2020)				■										■				■		■		■		
Brunk et al. (2021)	■													■				■		■	■			
Allah Bukhsh et al. (2019)		■	■			■	■					■			■			■	■	■	■	■		■
Cao et al. (2023a)					■									■					■			■		■
Cao et al. (2023b)					■									■				■	■		■	■		■
Coma-Puig and Carmona (2022)						■							■				■		■		■			■
Conforti et al. (2016)				■										■						■	■	■		

Table 9 (continued)

	Interpretable AI				Explainable AI																				
									Scope												Relation		Format		
	White-Box				Black-Box				Local							Global									
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
De Koninck et al. (2017)								■	■										■			■		■
De Leoni et al. (2015)		■	■											■				■		■	■	■		■
De Oliveira et al. (2020a)				■										■						■		■		■
De Oliveira et al. (2020b)				■										■					■					■
Diamantini et al. (2024)		■		■	■	■	■	■										■	■			■	■	
Di Francescomarino et al. (2016)		■												■				■		■		■		
Francescomarino et al. (2019)		■												■				■		■		■		
Elkhawaga et al. (2023)			■			■								■					■			■		

Table 9 (continued)

	Interpretable AI				Explainable AI																				
									Scope												Relation		Format		
	White-Box				Black-Box				Local						Global										
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Elkhawaga et al. (2024)			■			■									■		■	■	■		■			
Folino et al. (2011)		■		■										■				■		■		■		
Folino et al. (2017)				■										■				■		■		■		
Folino et al. (2024)					■			■				■							■		■			■
Folino et al. (2025)					■							■							■		■			■
Fu et al. (2021)				■										■				■						
Galanti et al. (2020)					■								■						■		■			■
Galanti et al. (2023a)					■	■							■				■		■		■			■

Table 10 Categorization of employed ML and explanation methods in the found literature, segmented into model interpretability, explanation scope, explanation relation and explanation format

	Interpretable AI				Explainable AI																					
									Scope												Relation		Format			
	White-Box				Black-Box				Local							Global										
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Galanti et al. (2023b)						■											■		■					■
García-Bañuelos et al. (2017)				■										■						■			■	
Gerlach et al. (2022)					■									■					■		■			■
Guo et al. (2024)					■	■								■	■					■	■			■
Gupta et al. (2015)		■												■				■		■		■		
Hanga et al. (2020)					■									■				■	■		■			
Harl et al. (2020)					■									■						■	■			■
Hermann et al. (2024)				■										■				■		■		■		
Horita et al. (2016)		■		■										■				■		■	■	■		■
Hsieh et al. (2021)					■				■										■					■
Huang et al. (2022)					■				■					■					■			■		■
Irarrázaval et al. (2021)		■		■										■				■		■		■		
Khemiri et al. (2018)		■												■				■		■		■		

Table 10 (continued)

	Interpretable AI				Explainable AI																				
									Scope												Relation		Format		
	White-Box				Black-Box				Local							Global									
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Kim et al. (2024)		■			■	■	■										■	■	■	■	■			■
Lakshmanan et al. (2011)		■												■				■		■		■		
Maggi et al. (2014)		■		■										■				■				■		
Maita et al. (2025)				■			■					■						■	■		■			■
Majumdar et al. (2023)	■													■						■		■	■	
Malashin et al. (2025)		■		■	■	■	■					■					■		■					■
Mayer et al. (2021) &				■	■	■			■		■	■				■			■	■	■			■
Mehdiyev and Fettke (2020a)					■									■					■			■		■
Mehdiyev and Fettke (2020b)					■										■	■			■		■			■
Mehdiyev and Fettke (2021)					■						■		■						■		■			■
Mehdiyev et al. (2021)					■	■					■	■	■		■	■		■	■		■	■		■
Mehdiyev et al. (2024)					■	■	■		■										■		■		■	■
Mehdiyev et al. (2025a)					■			■			■					■				■	■			■

Table 10 (continued)

	Interpretable AI				Explainable AI																				
									Scope											Relation		Format			
	White-Box				Black-Box				Local							Global									
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Mehdiyev et al. (2025b)		■	■			■		■							■				■		■			■
Montoya and Astudillo (2023)				■										■				■		■		■		■
Montoya et al. (2023)				■					■					■				■		■		■		■
Myers et al. (2018)				■										■				■		■		■		
Ouyang et al. (2021)						■						■			■				■		■			■
Padella et al. (2022)						■			■				■						■		■			■
Park et al. (2024)				■										■				■		■	■	■		
Pasquadibisceglie et al. (2021)					■									■					■			■		■

Table 11 Categorization of employed ML and explanation methods in the found literature, segmented into model interpretability, explanation scope, explanation relation and explanation format

	Interpretable AI				Explainable AI																							
									Scope												Relation		Format					
	White-Box				Black-Box				Local						Global													
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Pasquadibisceglie et al. (2023)								■		■					■					■	■			■
Pasquadibisceglie et al. (2024a)					■					■										■	■			■
Pasquadibisceglie et al. (2024b)					■					■					■					■	■			■
Pasquadibisceglie et al. (2024)			■		■	■	■										■		■					■
Pauwels and Calders (2019a)	■													■				■		■		■		■
Pauwels and Calders (2019b)	■													■				■		■		■		■
Petsis et al. (2022)						■							■				■		■		■			■
Polato et al. (2018)				■										■						■	■			■
Porouhan (2024)		■												■				■		■		■		■
Prasidis et al. (2021)	■													■				■		■		■		■
Rauch et al. (2024)	■			■										■				■		■	■			■

Table 11 (continued)

	Interpretable AI				Explainable AI																				
									Scope											Relation		Format			
	White-Box				Black-Box				Local						Global										
Publication	Bayesian network	Decision tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Feature importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Rehse et al. (2019)					■									■	■				■		■	■	■	■
Rizzi et al. (2020)							■					■	■						■		■			■
Rizzi et al. (2024)		■	■		■	■	■	■			■	■	■						■		■			■
Saha et al. (2024)				■										■				■		■	■			
Saini et al. (2020)				■										■				■		■		■		
Sarno et al. (2020)				■										■				■		■	■			
Savickas et al. (2014)	■													■				■		■		■		■
Savickas and Vasilecas (2018)	■													■				■		■		■		■
Sindhgatta et al. (2020)						■						■			■				■		■			■
Sindhgatta et al. (2020)					■									■					■		■			■
Stevens and De Smedt (2022a)		■	■		■	■	■						■		■				■	■	■	■		■
Stevens et al. (2022b)			■		■		■						■	■					■	■	■			■
Stevens et al. (2022c)		■		■	■	■							■	■	■				■	■	■	■		

Table 11 (continued)

	Interpretable AI				Explainable AI																					
									Scope												Relation		Format			
	White-Box				Black-Box				Local							Global										
Publication	Bayesian network	De-ci-sion tree	Lin./log. regression	Other	Deep learning	Gradient boosting	Random forest	Other	Counterfactual	Feature importance	ICE	LIME	Shapley-based	Other	Fea-ture im-por-tance	PDP	Shap-ley-based	Other	Mod-el-ag-nostic	Mod-el-spe-cific	Nu-mer-ic	Rule-based	Tex-tual	Vi-su-al
Tama et al. (2020)		■		■										■				■		■	■	■		■
Teinemaa et al. (2016)			■			■								■				■	■			■		
Toh et al. (2022)						■							■						■		■			
Trescato et al. (2024)				■										■				■		■	■			■
Tripathi et al. (2024)				■									■				■		■					
Unuvar et al. (2016)		■												■				■		■		■		■
van Zelst et al. (2020)				■										■				■		■	■			■
Velmurugan et al. (2021a)						■						■	■						■		■			
Velmurugan et al. (2021b)						■						■	■						■		■			

applicability of predictive models in real-world scenarios. However, these models often encounter limitations in handling complex datasets or sophisticated predictive tasks where higher-dimensional interactions are present. This underscores a common scenario in predictive modeling where the simplicity and transparency of white-box models can lead to diminished predictive performance compared to black-box models.

3.2.2 Explainable AI for PPM

Black-box approaches such as DL, GBMs and RFs are chosen for their sophisticated modeling capabilities and superior performance on complex datasets. These models excel in environments where the primary focus is on predictive accuracy and handling high-dimensional data with complex patterns. However, these gains in performance come at the cost of reduced interpretability. The internal workings of these models are often opaque or overly complex, making it challenging to discern which features are influencing the predictions, thereby complicating efforts to validate and trust the model's decisions. This trade-off necessitates a balanced approach, especially in industries where the stakes are high. Regarding the retrieved literature, the number of articles relying on opaque models (59 out of 107 articles) is slightly below those utilizing interpretable models (64 out of 107). However, considering the literature retrieval process, specifically the exclusion of articles which omit to provide explainability to employed black-box models, a large amount of publication relying on opaque models in PPM was filtered out during the identification, screening and detailed assessment of articles.

3.2.2.1 Black-box models DL methods, such as DNN, RNN, especially LSTM, stand out for their ability to detect and learn complex patterns in extensive datasets. However, the multi-layered architecture that contributes to their strength also obscures the reasoning behind their decisions, making them less interpretable than simpler models. Among the surveyed publications, 38 out of 59 articles utilizing black-box models employed **DL**. Exemplary applications include Mehdiyev and Fettke (2020a, 2020b, 2021) utilized DNNs across their studies, focusing on high-performing models and post-hoc explainability. Galanti et al. (2020) utilized a conventional LSTM, while Hanga et al. (2020) performed a comparative analysis between a conventional and bidirectional LSTM, comparing both against the results of similar studies. Similarly, Rehse et al. (2019) utilized an LSTM, exploring potentials of explainability within process prediction in the context of Industry 4.0 (see Rehse et al. 2018). While Huang et al. (2022) focused solely on using an LSTM in their "LORELEY" approach, tailored for event log analysis, Hsieh et al. (2021) introduced an innovative approach that combines a DNN and an LSTM into an ensemble, implementing "DiCE4EL" - a variant of "DiCE" (Mothilal et al. 2020). The former framework uniquely merges the strengths of both neural network architectures to enhance predictive accuracy while providing explainability adapted from established methodologies. In their study, Sindhgatta et al. (2020) tailored their approach by using a bidirectional LSTM in one case, while opting for an ensemble of two bidirectional LSTMs in two additional cases, based on the application task. Weinzierl et al. (2020) presented "XNAP", a model-specific approach that employs a bidirectional LSTM RNN that is able to propagate feature relevance scores from one layer to another, thus providing insight into the model's decision process. Building on Sindhgatta et al. (2020) and Wickramanayake et al. (2022a) introduced two architectures using

ensembles of bidirectional LSTM models. They further developed an explanation framework in Wickramanayake et al. (2022b) using this architecture. In a similar vein, Stevens et al. (2022c) as well as Stevens and De Smedt (2022a) employed LSTM models, with the former integrating an XGBoost model for benchmarking and the latter using a CNN and RF for comprehensive model evaluation.

GBM approach allows for optimizing complex loss functions and handling various types of data, including unbalanced datasets. Unlike deep learning, which uses a holistic approach through layers, gradient boosting focuses on improving predictions incrementally, which can lead to better performance on structured data. However, the sequential nature of boosting can make these models computationally intensive and still difficult to interpret due to the aggregation of numerous small models, each contributing to the final outcome. While they share the high performance of deep learning in complex tasks, their operational intricacy often prevents a clear understanding of how specific features influence predictions. GBMs were utilized in a variety of the surveyed studies to assess predictive methodologies, with 26 out of 59 articles employing black-box approaches opting for these models: In the study by Stevens and De Smedt (2022a), GBMs like XGBoost were part of a broader ensemble that included various predictive models such as GLRM, logistic regression and logit leaf model, along with CNN, LSTM and RF. These models were evaluated across multiple event logs including BPIC 2011, BPIC 2015, Production and Sepsis, with a focus on process outcome predictions. The performance was assessed based on AUROC scores, with this diversified model application being guided by the "X-MOP" framework, which assists in choosing the appropriate model for specific tasks. Stevens et al. (2022c) further explored these models, comparing white-box and black-box approaches in terms of functional complexity, monotonicity and parsimony. Velmurugan et al. (2021b) examined the stability of the LIME and SHAP explanation methods for process outcome predictions. They employed logistic regression as a white-box model and compared it with an XGBoost black-box model, evaluating their performance on the BPIC 2012, Production and Sepsis event logs while considering different data encoding techniques. Additionally, Ouyang et al. (2021), Petsis et al. (2022), Sindhgatta et al. (2020), Velmurugan et al. (2021a) and Verenich et al. (2019) all employed XGBoost models to assess post-hoc explainability techniques, further highlighting the adaptability and utility of gradient boosting in predictive process monitoring.

RFs are a robust and versatile machine learning approach that combines multiple DTs to enhance predictive accuracy and prevent overfitting. While RFs are more interpretable than deep learning models due to their reliance on DTs, the ensemble nature still limits transparency compared to single-tree models. In the reviewed literature, RF models were frequently used for process outcome prediction tasks, either alone or in comparison with other machine learning methods, with 16 out of 59 articles employing black-box approaches using the RF model. Allah Bukhsh et al. (2019) employed RF alongside DT and GBMs and compared their predictive performance. Similarly, Teinemaa et al. (2016) contrasted RF and logistic regression models in their methodology. Rizzi et al. (2020) adopted RF and enhanced its performance through iterative retraining based on prior explanations. Verenich et al. (2016) presented an approach that builds a RF on top of an event log after the corresponding traces have been clustered using one of two proposed clustering algorithms. In similar fashion, Verenich et al. (2017) used a RF model as a classifier for activities within traces after matching them to a discovered process model from the event log.

While the majority of research focuses on the aforementioned types, a subset of studies explores alternative black-box models that do not fit into these conventional categories. These models, often designed for specific use-cases, integrate unique methodologies to enhance predictive accuracy while addressing the interpretability challenges inherent in black-box approaches. Out of the 59 articles using black-box models 10 publications utilized approaches which are considered opaque, with the following publications exemplifying models that do not fit into the previously discussed categories: De Koninck et al. (2017) introduced an approach for trace clustering, utilizing a modified "search for explanations for clusters of process instances" (SECPI) architecture (De Weerdt and vanden Broucke 2014). This method employs SVMs for each identified cluster to identify the minimal set of features that keep an instance within its designated cluster. Meanwhile, Verenich et al. (2016, 2017, 2019) expanded their methodologies by incorporating clustering and two process model discovery components, adding an interpretable layer to their black-box approaches.

3.2.2.2 Post-hoc explanation methods Post-hoc explanation methods exhibit a variety of differences, depending on the model that is explained, as well as the application context and PPM task that is being tackled. In particular, the following characteristics are differentiated: regarding explanation scope, local and global explanations are distinguished, with the former focusing on explanations pertaining to individual model predictions and the latter referring to the general workings of the examined model. The model relation differentiates between model-specific explanation methods, which leverage the intricacies of the model methodology, and model-agnostic explanation methods, which can be applied regardless of the utilized model. Lastly, the output format of the explanation can be in numeric, textual, rule-based, or visual form as well as a mixture thereof.

3.2.2.3 Local post-hoc explanation methods Local XAI methods focus on revealing the relevance of variables for predictions on a single data point. These explanations do not necessarily uncover general model behavior but provide valuable insight into specific prediction instances.

3.2.2.4 Counterfactual explanations Counterfactual explanations is a contrastive method of providing insight by presenting conditions, specifically certain variable values, under which the prediction score would exceed or fall below a certain threshold compared to its original score. These explanations aim to identify the least amount of intervention in order to flip a prediction label for classification tasks or bring the prediction score across a certain threshold for regression tasks. Counterfactual explanations have informative characteristics and provide actionable advice for attaining specific prediction scores. However, the fact that an exhaustive search for counterfactual explanations is likely to suffer from a combinatorial explosion for categorical variables and that it can be expected to find various such explanations necessitates an implementation that is suitable for its corresponding application context. Figure 8 is an example of a visual counterfactual explanation from Hsieh et al. (2021), illustrating the original instance as well as counterfactual instances with modified feature values that result in the prediction score exceeding a given threshold. In a similar fashion, Hsieh et al. (2021) implemented counterfactual explanations using a tabular visualization

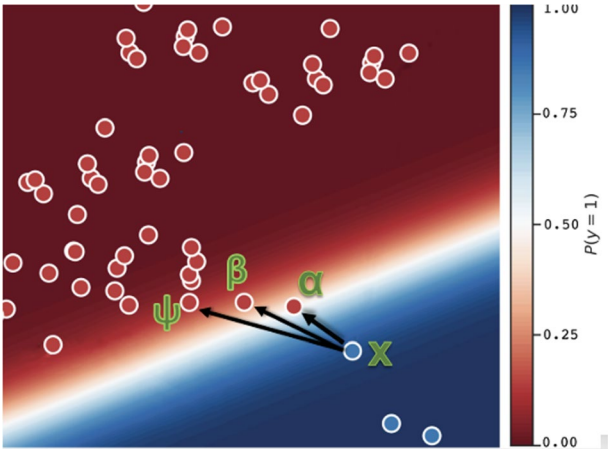

Fig. 8 Counterfactual explanation as it is implemented by Hsieh et al. (2021), demonstrating the original instance and the found counterfactual instances

(A_SUBMITTED, 112, $15,500),
(A_PARTLYSUBMITTED, 112, $15,500),
A_PREACCEPTED, 112, $15,500),
(A_ACCEPTED, 10939, $15,500),
Prediction: O_SELECTED
Milestone: A_FINALISED
Counterfactual: *What would I have had to change for the loan to be A_FINALISED?*

(a)

Counterfactual 1		Counterfactual 2		Counterfactual 3	
Activity	**Resource**	**Activity**	**Resource**	**Activity**	**Resource**
A_SUBMITTED	112	A_SUBMITTED	112	A_SUBMITTED	112
A_PARTLYSUBMITTED	112	A_PARTLYSUBMITTED	112	A_PARTLYSUBMITTED	112
A_PREACCEPTED	112	A_PREACCEPTED	10910	A_PREACCEPTED	10939
A_ACCEPTED	10931	W_Complete request	10912	W_Handling leads	10939
A_FINALISED	10931	A_ACCEPTED	10932	A_ACCEPTED	11189
—	—	A_FINALISED	10932	O_SELECTED	11189
—	—	—	—	A_FINALISED	11189

(b)

Fig. 9 Counterfactual explanation as implemented by Hsieh et al. (2021). **a** demonstrates the original instance, whereas **b** demonstrates the counterfactual explanations and the features that have been altered to achieve the desired prediction—in this case, the acceptance of a loan of $15,500

for the altered features of the counterfactual explanations, as seen in Fig. 9. Further counterfactual explanations for PPM can be found in De Koninck et al. (2017), Huang et al. (2022), Mayer et al. (2021) and Padella et al. (2022).

3.2.2.5 Individual conditional expectation (ICE) ICE plots are a model-agnostic approach that illustrate the impact of an iterated feature for a single data point. Algorithmically, the value of a given variable of an instance is iterated over its observed values for categorical variables or over certain ranges for numerical variables, and the resulting change in the prediction score is captured. In practice, ICE plots can be visualized for an individual instance or for a group of instances in a single plot, depending on the use case, although the latter

approach qualifies as a global explanation. Figure 10 is an example of an ICE plot from Mehdiyev and Fettke (2021), illustrating the changes of prediction scores for each single instance within a group (visualized as one line per instance) across value changes of the "Overall Equipment Effectiveness" variable. A visualization such as Fig. 10 facilitates the identification of and differentiation between global and local model behavior. Other publications employing ICE are Mayer et al. (2021) and Mehdiyev et al. (2021).

3.2.2.6 Local interpretable model-agnostic explanations (LIME) LIME ((Ribeiro et al. 2016)) explains an individual prediction by training a simple, interpretable surrogate model in the neighborhood of the instance. It generates perturbed samples near that point, queries the black-box model for their predictions, weights those samples by proximity (using a locality kernel), and fits an interpretable model on an interpretable representation of the data. When the surrogate achieves good local fidelity, its parameters provide a locally faithful account of which features drove the prediction. Across the analyzed literature, LIME was used as an explanation technique in the following works: Allah Bukhsh et al. (2019) (see Fig. 11a), Mayer et al. (2021), Mehdiyev et al. (2021), Ouyang et al. (2021) (Fig. 11b), Rizzi et al. (2020), Sindhgatta et al. (2020), Velmurugan et al. (2021a), and Velmurugan et al. (2021b). Notably, Velmurugan et al. (2021b) adopted the style of Visani et al. (2021), estimating feature contributions with LIME across ten surrogate models to assess the stability of the explanations. Although LIME benefits from interpretable surrogate models, identifying and clustering instances that belong to a specific locality is a substantial challenge for non-image data and depends heavily on the use case. To address this, Mehdiyev and Fettke (2020a) proposed a modified, model-specific approach conceptually based on LIME and K-LIME (Hall et al. 2017). They used neural codes from the last hidden layer of a DNN as vectors for distance calculation between instances, thereby defining localities from

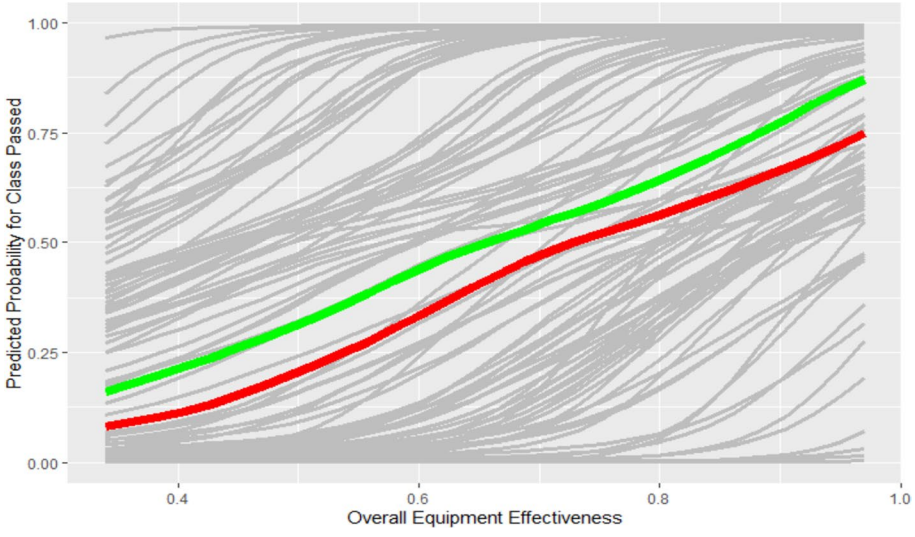

Fig. 10 Example of an ICE plot (Mehdiyev and Fettke 2021) with the green line depicting a true positive instance and the red line depicting a true negative instance

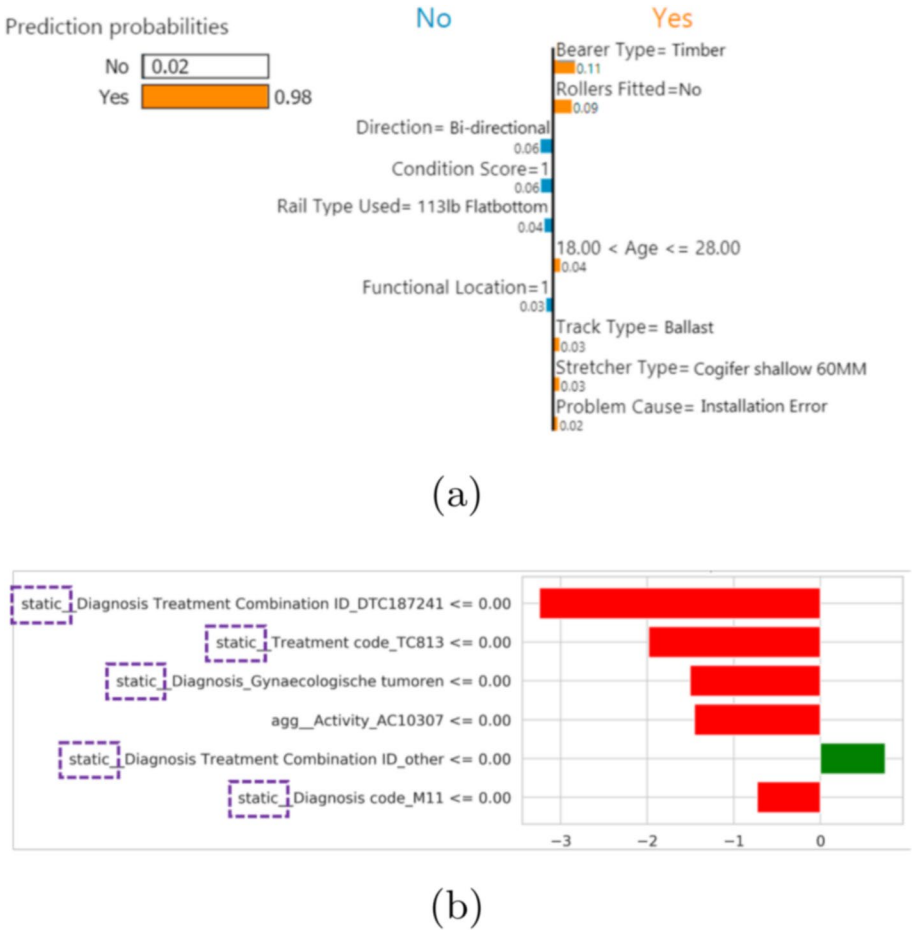

Fig. 11 Example for LIME as it is implemented by **a** Allah Bukhsh et al. (2019) and **b**Ouyang et al. (2021) for PPM

the model's learned representations. Rehse et al. (2019) reported a similar idea, also using neural codes from the last hidden layer to identify localities for specific instances.

3.2.2.7 Shapley-based local explanations Shapley (1953), from cooperative game theory, allocate a final payoff among players by averaging each player's marginal contribution across all possible orderings. In the context of local explanations for ML models, features act as the players and the model's prediction for a specific instance is treated as the payoff (often relative to a baseline), so each feature receives a contribution that reflects its influence on that prediction. Because exact computation requires evaluating all coalitions and grows exponentially with the number of features, practical methods use approximations. SHAP (Lundberg and Lee 2017) provides a unified framework with model-agnostic and model-specific estimators, including Kernel SHAP, Linear SHAP, and Deep SHAP, to produce

local attributions for individual instances. An example of such a local Shapley-based explanation is shown in Fig. 12, illustrating the implementation by Mehdiyev and Fettke (2021).

Other local explanation methods For LSTMs, layerwise relevance propagation (LRP) (Lapuschkin et al. 2015; Arras et al. 2017) is a local, model-specific attribution method that reveals the impact of each feature on the prediction for a given instance, as demonstrated by Harl et al. (2020), Sindhgatta et al. (2020), Stevens et al. (2022c), Weinzierl et al. (2020), Wickramanayake et al. (2022a), and Wickramanayake et al. (2022b). Although presented here as a local XAI method, Sindhgatta et al. (2020) and Stevens et al. (2022c) report only global explanations derived from LRP attributions. Related to LRP, Hanga et al. (2020) propose a model-specific approach for LSTMs in next-event prediction that allocates probability scores to candidate events. For an unfinished trace, the model encodes the trace as a graph and displays the estimated probability for each predicted activity. While this provides users with a confidence measure, the interpretability of these probabilities is highly use-case dependent and the approach does not explain how the probabilities were formed. De Koninck et al. (2017) employ SECPI, which trains an SVM, an inherently non-interpretable model, to determine the minimum set of characteristics a trace must retain to stay in its assigned cluster. This primarily explains the clustering method. The authors define "explainable" instances as "instances for which such an explanation can be extracted from the underlying SVM," an interpretation that may warrant further discussion. Huang et al. (2022) present LORELEY, an approach based on LORE (Guidotti et al. 2019), which, similar to LIME, creates local explanations by training a decision tree within the instance's

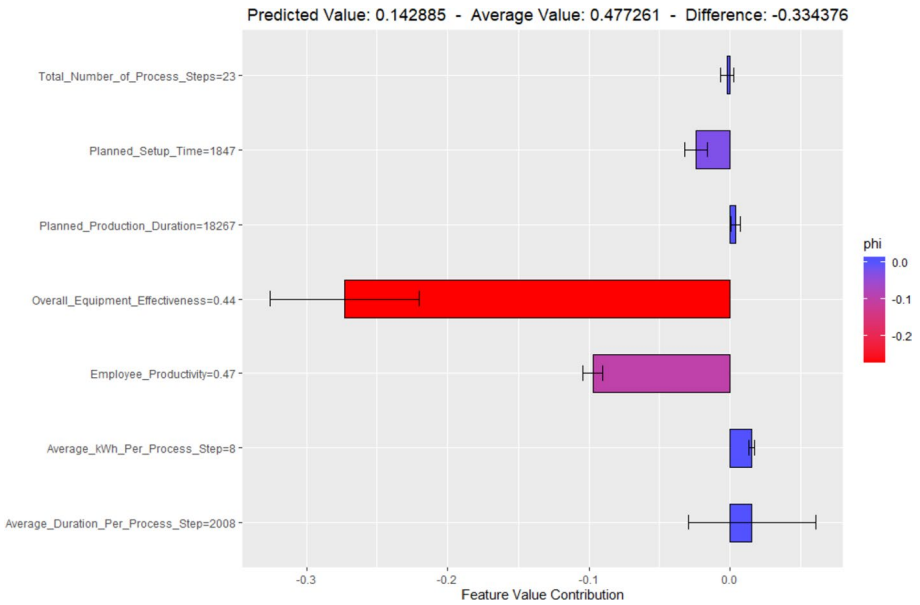

Fig. 12 Example of a Shapley-based local explanation as it is implemented by Mehdiyev and Fettke (2021), illustrating feature impact on the predictions score using bars, with their length and color representing the contribution of the corresponding feature. The specific feature values as well as the numerical value of their contribution are visible on the axes, the prediction score, the average prediction score and the difference due to feature impact are displayed at the top of the plot

neighborhood to capture local model behavior. LORELEY adapts algorithms for trace similarity, distance, and clustering to predictive process monitoring. Because the surrogate is a decision tree, these explanations can also serve as counterfactuals.

3.2.2.8 Global post-hoc explanation methods While local explanations zoom in on individual predictions, global explanations aim at describing interdependence and relationships between variable expressions and model predictions on a general level, giving insight about the underlying data as well as the model that was trained on said data. Global explanations enable the assessment of the general model behavior by domain experts and allow for uncovering discrepancies between model behavior and domain knowledge.

3.2.2.9 Shapley-based global explanations Local SHAP attributions can be aggregated to reveal global model behavior. This has been demonstrated by Galanti et al. (2020) (Fig. 13) and by Petsis et al. (2022) (Fig. 14). In practice, common global visualizations include SHAP summary plots, which display the distribution of SHAP values for every feature across the entire scored dataset, and SHAP dependence plots, which are similar in spirit to PDP and use Shapley values to show how variation in a feature relates to its contribution to the prediction. Beyond ranking features by mean absolute contribution, summary plots convey directionality and heterogeneity across instances. Dependence plots can be enhanced by coloring points by a second feature to reveal potential interactions. Analysts often compute global importance by averaging absolute SHAP values, create class-specific summaries for classification tasks, and stratify results by cohorts to compare populations. Shapley-based approaches are attractive because they rest on a clear cooperative game theoretic foundation and yield additive, instance-level attributions that aggregate naturally to the global level. The reference point for these explanations can be set to specific subsets of the dataset, which increases applicability across use cases. As with any attribution method, results depend on the background data, on correlations between features, and on the coverage of the feature space, so it is good practice to report the chosen background set and to validate patterns across subgroups.

3.2.2.10 Feature importance Feature importance (Gevrey et al. 2003; McDermid et al. 2021) is an umbrella term for methods that quantify how much each feature contributes to a model's predictions. These techniques are often used to summarize global behavior, while some implementations can also be adapted to provide local views for individual instances. A widely used approach is permutation feature importance (Fisher et al. 2019). For each feature in turn, its values are shuffled across the dataset, the model is re-scored, and the change in error is recorded. Repeating this across features yields a ranking of influential variables; however, this procedure does not explicitly capture interaction effects. The permutation approach is employed by Ouyang et al. (2021), Sindhgatta et al. (2020), Stevens and De Smedt (2022a) and Stevens et al. (2022c). For LSTMs, feature importance can be derived from layerwise relevance propagation by averaging relevance scores per variable across the scored dataset, as shown by Harl et al. (2020), Sindhgatta et al. (2020), Stevens et al. (2022c), Weinzierl et al. (2020), Wickramanayake et al. (2022a) and Wickramanayake et al. (2022b). Another option is leave-one-feature-out retraining in the style of Feng et al.

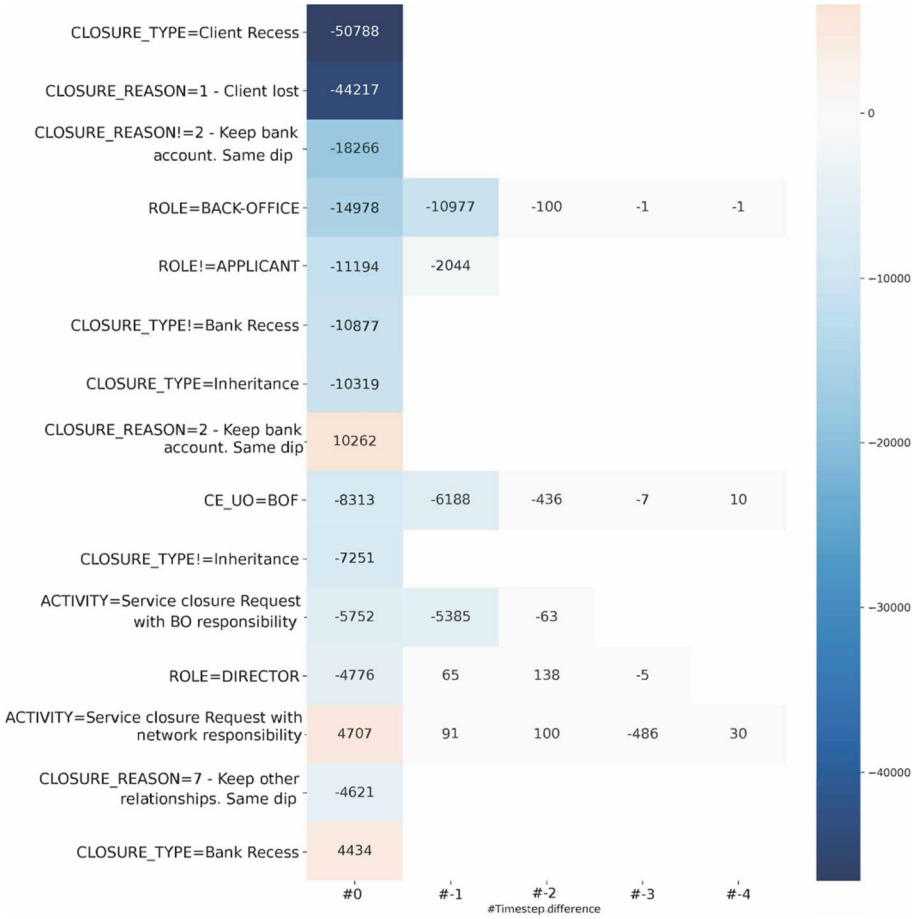

Fig. 13 Shapley-based global explanation as it is implemented by Galanti et al. (2020), illustrating frequency of features and corresponding values when they were significantly relevant for the prediction by using a heatmap

(2013), where a model is retrained without a given feature and the change in performance is measured; Allah Bukhsh et al. (2019) adopt this strategy. Galanti et al. (2023a) and Stevens et al. (2022c) apply SHAP feature importance and SHAP summary plots, which aggregate local SHAP values for each variable over the dataset to show overall impact and variation. For DNN, connection weight-based importance following Gedeon (1997) has been used by Mehdiyev and Fettke (2020b) and Rehse et al. (2019) to characterize global behavior. For tree-based models, such as XGBoost in Stevens et al. (2022c), feature importance can be computed from the average contribution to impurity reduction (for example, Gini-based purity).Fig. 15 illustrates an example from Mehdiyev and Fettke (2020b), showing

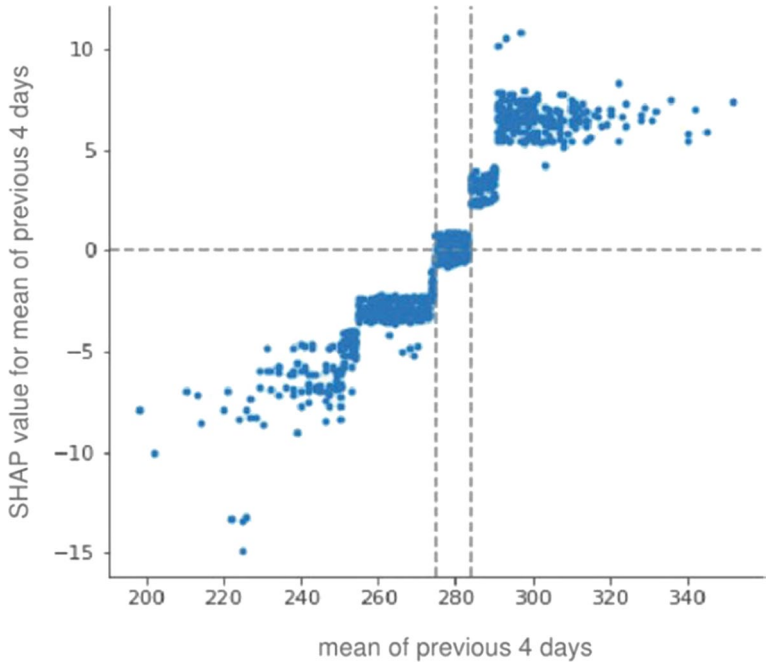

Fig. 14 SHAP dependence plot as it is implemented by Petsis et al. (2022)

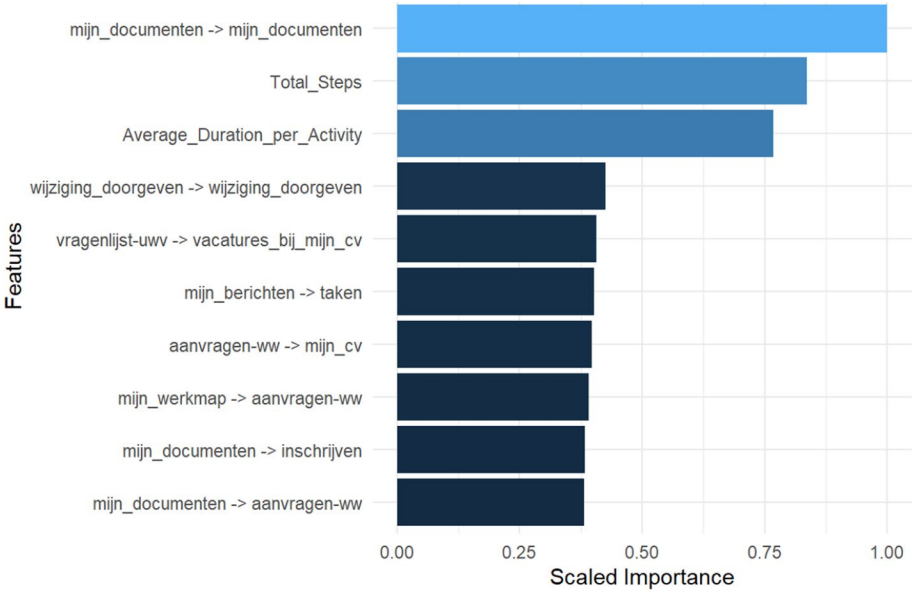

Fig. 15 Feature importance as described by Mehdiyev and Fettke (2020b)

the scaled importance of the ten most influential features in a bar plot, where bar length and color convey each feature's impact on the prediction score.

3.2.2.11 Partial dependence plots (PDP) PDP Friedman (2001) provides a model-agnostic, global view of how the expression of a single feature influences a model's prediction while averaging out the effects of all other features. The core idea is straightforward. Select a grid of values for the feature of interest. For each grid value, replace that feature in every instance of the dataset with the grid value, score the modified data with the trained model, and compute the average prediction. Repeating this across the grid traces the marginal effect of the feature on the prediction score. For categorical variables, the procedure is performed per category; for numeric variables, it is performed over a set of representative values such as quantiles or evenly spaced points. PDPs are popular because they are easy to read and can reveal global trends such as monotonic relationships, thresholds, and regions of diminishing or increasing returns. They support model validation by domain experts who can compare the learned relationship against domain expectations. There are important limitations. Because PDPs average over the joint distribution of the remaining features, they do not directly reveal feature interactions and they can be misleading when the feature of interest is strongly correlated with others. The averaging also masks heterogeneous effects that may differ across subgroups or individual instances. From a computational perspective, the cost scales with the number of instances and the number of grid points chosen for the feature, which can be substantial for high-cardinality categorical variables or finely gridded continuous variables. Figure 16 shows an example from Mehdiyev and Fettke (2020b). The PDP depicts the mean prediction score as a function of the variable "Average Duration per Process Step," with separate colored lines for different age groups. The plot indicates that higher average duration per step is associated with lower predicted scores, and the separation between the age-group curves suggests that age contributes meaningfully to the prediction score as well.

3.3 Evaluation of explainability and interpretability for PPM

Evaluating explainability and interpretability in ML is a multifaceted task that requires a careful comparison of methodological choices, each with distinct strengths and caveats. This section contrasts quantitative and qualitative evaluation strategies and situates them within the complementary paradigms of functional, application, and human-grounded evaluations (Doshi-Velez and Kim 2017). Assessing the value of an explanation is inherently multi-dimensional. Guided by **RQ6**–**RQ8**, this subsection examines how the reviewed work designs, executes, and reports evaluations of explainable and interpretable PPM methods. First, we summarize the *study-level protocols* reported in the literature−such as data splits, baselines, predictive metrics, and explanation-quality measures−addressing the concerns of **RQ6**. Next, we contrast the evidence produced by *quantitative* metrics (fidelity, stability, robustness to sampling, and computational cost) with insights from *qualitative* user studies (expert judgment of usefulness, clarity, and trust), thereby tackling **RQ7**. Finally, we map individual evaluations to the three paradigms "functional," "application-grounded," and "human-grounded," and discuss the decision-support insights each yields for PPM practice,

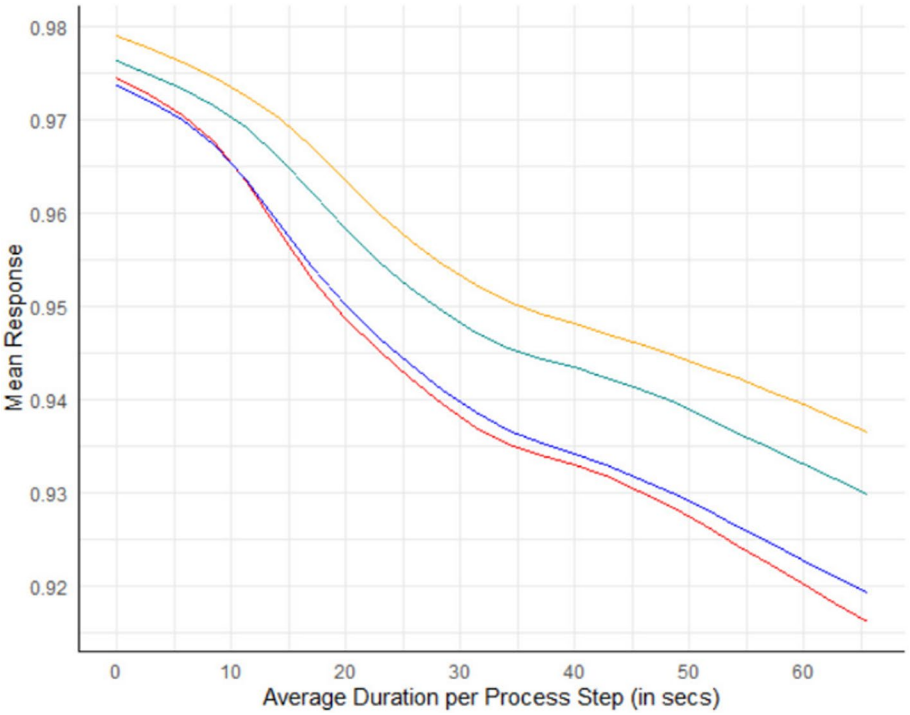

Fig. 16 Partial dependence plots as applied by Mehdiyev and Fettke (2020b)

as required by **RQ8**. Throughout, we highlight recurring methodological choices—such as the definition of background data, the handling of correlated features, and the selection of user tasks—and identify gaps that still impede rigorous and practice-relevant assessment of explanation quality in predictive-process-monitoring contexts.

3.3.1 Evaluation design and reporting

In the analyzed literature, the evaluation of proposed XAI methods varied with characteristics of the underlying method, its users and goals, the model in need of explanations as well as the application context. This section presents exemplary evaluation approaches of the analyzed articles (see Tables 13, 14, 15 and 16), illustrating an excerpt from a broad range of evaluation aspects regarding explainability.

De Koninck et al. (2017) evaluate their implementation of SECPI by comparing the runtime in seconds, the length of explanations, i.e. the number of created rules that explain why an instance belongs to a specific cluster, as well as the relative amount of "explainable" instances, i.e. the relative amount of instances for which the employed SVM was able to find minimal sets of rules that allow the instance to stay in its allocated cluster.

Folino et al. (2017) evaluate their approach for extracting explanations for trace clustering by providing clustering rules on "explanation complexity", i.e. the number of rules needed to justify a trace's allocation to a specific cluster, as well as interestingness and compared the results to an explainable M5Rules (Holmes et al. 1999) implementation.

Galanti et al. (2023a) employ a two-part approach to evaluating their utilized explanation approach: First, explanations are evaluated on their soundness based on statistical analysis and domain knowledge. Second, a user evaluation with 20 participants was conducted, with the participants solving 18 tasks and reporting their personal estimation of the difficulty of said tasks. Afterwards, usability and user experience have been captured using questionnaires.

Hsieh et al. (2021) evaluate the quality of their counterfactual explanations with regard to diversity, plausibility, proximity, sparsity and whether the explanations can incorporate categorical features. In this context, diversity refers to the amount of different counterfactual explanations created, plausibility refers to the soundness of the counterfactual explanations based on domain knowledge, proximity refers to the proximity of the counterfactual explanations and the instance given as input based on the distance measurement, sparsity refers to the mean amount of modified features that constitute a counterfactual explanation for the instance given as an input. The evaluation incorporates a statistical approach as well as the evaluation of explanations for specific traces.

Mehdiyev and Fettke (2020a) used the coefficient of determination (R^2-value) for the surrogate model for each locality in order to reveal the quality of the surrogate capturing the behavior of the underlying model. Due to the surrogate models being inherently interpretable Decision Trees, the provided explanations were not evaluated individually.

Stevens and De Smedt (2022a) evaluate their employed XAI-methods with regard to functional complexity, level of disagreement and parsimony: For the authors, in this context, functional complexity refers to a metric, similar to the measurement of permutation feature importance, that captures how easily a prediction can be manipulated when altering certain feature values, level of disagreement (Lakkaraju et al. 2017) refers to discrepancies with regard to the prediction score between the underlying model and corresponding surrogate models, and parsimony refers to the trade-off between the simplicity of provided explanations and the performance, i.e. accuracy, of the underlying model.

Velmurugan et al. (2021a) differentiate internal and external fidelity, referring to the definition of fidelity from Messalas et al. (2019): External fidelity measures the similarity between the predictions of the underlying model and corresponding surrogate model, whereas internal fidelity focuses on the decision-making process of the models, specifically on the amount of similarities between these models. The authors focused on the internal fidelity of LIME and SHAP and for its measurement, instances were perturbed ten times and the mean absolute percentage error between the task model and surrogate model was documented.

Velmurugan et al. (2021b) evaluated the stability, referring to Visani et al. (2021), aiming at measuring the consistency of explanations for the same or similar instances. In particular, the stability of the identified most important features (a subgroup of features residing in the top quartile with regard to the weight distribution) as well as the stability of corresponding weights was examined. The authors used this approach to evaluate the employed LIME and SHAP methods.

3.3.2 Evaluation type: quantitative versus qualitative evaluation

The evaluation of explainability methodologies is a multifaceted task, encompassing the adoption of both qualitative and quantitative methodologies. The significance of quantita-

tive metrics in the evaluation of XAI is emphasized by both Li et al. (2021) and Rosenfeld (2021). Li's research reveals that no single method exhibits superiority across all metrics, underscoring the need for a comprehensive evaluation framework. On the other hand, Rosenfeld proposes four distinct metrics that can be employed to quantify the explanatory nature of XAI systems. Nauta et al. (2023) underscore the imperative of conducting a thorough and all-encompassing evaluation, wherein the authors present twelve distinct properties that warrant careful assessment. Nevertheless, it is worth noting that anecdotal evidence and user studies are commonly employed in the evaluation of XAI. This observation implies that a comprehensive approach that integrates both qualitative and quantitative methodologies is required (Mohseni et al. 2021). Of the 107 papers reviewed for XAI in predictive process monitoring, a majority did not engage in any formal evaluation, while only a sixth (18 articles) employed quantitative or qualitative methods, and only two integrated both. This indicates a gap in the current research practices, where the nuances and user-centric aspects crucial for the adoption and trustworthiness of XAI systems might be overlooked. The hypothesis here is that integrating both quantitative and qualitative methods can provide a more holistic understanding of an AI system's explainability, balancing the objectivity of numerical data with the depth of descriptive analysis.

3.3.3 Evaluation paradigm: application, human and functional-grounded methods

Transitioning from the dichotomy of quantitative and qualitative evaluations, the framework proposed by Doshi-Velez offers a more granular understanding of XAI evaluation through functional, application and human-grounded methodologies (Doshi-Velez and Kim 2017). Functional-grounded evaluation delves into the theoretical and technical soundness of explanations. It's a critical approach for ensuring that the XAI methods align with established cognitive and computational frameworks, as highlighted by Mehdiyev et al. (2021). This approach is vital for the foundational integrity of XAI systems, ensuring that they are not only effective but also theoretically sound.

Application-grounded evaluation shifts the focus to the practical impact of XAI, examining how explainers influence specific decision-making tasks. This methodology is crucial for assessing the real-world utility of XAI, ensuring that the explanations provided are not only understandable but also actionable and beneficial in practical scenarios. Meanwhile, human-grounded evaluation, as discussed by Mohseni et al. (2021), centers on the user's perspective, measuring how effectively an XAI system's explanations foster trust and understanding among its human users. This approach is paramount for the user-centric development of XAI systems, ensuring that they meet the actual needs and expectations of the people they are designed to assist.

Within the retrieved literature, predominance of functional and human-grounded approaches was observed, yet the overall engagement in comprehensive evaluation was limited. This indicates a recognition of the importance of diverse evaluative lenses but also hints at the challenges and complexities inherent in implementing such multifaceted methodologies. While the field acknowledges the need for a broad spectrum of evaluation strategies, the practical implementation is still catching up, requiring more robust frameworks and tools to facilitate these comprehensive assessments.

In conclusion, the evaluation of XAI systems is an intricate task, necessitating a balanced and thorough approach that encompasses both quantitative and qualitative methods,

as well as functional, application and human-grounded evaluations. The current research landscape shows a tendency towards quantitative methods and reveals a significant gap in formal evaluation practices. To advance the field of XAI and ensure the development of effective, reliable and user-centered systems, a more rigorous and holistic approach to evaluation is imperative. As the field continues to evolve, embracing this multifaceted evaluation paradigm will be crucial for the maturation and widespread adoption of explainable and trustworthy AI systems (Table 12).

4 Challenges and implications

4.1 Related surveys and contributions

The PPM field has been the subject of numerous studies and SLRs, each contributing valuable insights into different aspects of this rapidly evolving domain. This section contrasts the focus and contributions of prominent related studies, particularly review articles with the distinctive elements of our study, particularly emphasizing our exploration of interpretable and explainable AI within predictive process monitoring (see Table 17)

Di Francescomarino et al. (2018); Maggi et al. (2014); Márquez-Chamorro et al. (2017) and Teinemaa et al. (2019) have provided comprehensive overviews of predictive process monitoring tasks, computational methods and evaluation approaches. They discuss various computational predictive methods, from statistical techniques to ML approaches, and provide valuable insights into the applications and performance of various models. While these studies offer a substantial understanding of predictive process monitoring, they do not focus explicitly on interpretability and explainability. At most, these studies include a discussion of some interpretable AI methods, but XAI approaches, particularly those going beyond inherent model transparency, are not addressed at all. Kubrak et al. (2022) delve into prescriptive process monitoring, incorporating elements of XAI and interpretable AI. However, their focus is predominantly on prescriptive analytics, and while they mention relevant XAI papers, they do not provide an extensive overview of studies in this area, leaving a gap for a more focused and detailed exploration.

Stierle et al. (2021) stand out as one of the few studies aiming to provide a systematic review of XAI approaches specifically for predictive process monitoring. They categorize literature according to purpose, evaluation method and model complexity, differentiating between intrinsically interpretable models and opaque models requiring post-hoc explanations. However, being a research-in-progress paper and considering the rapid advancements and proliferation of research in this field, the scope of their review is somewhat limited. Our study addresses this by providing a more comprehensive and up-to-date review of XAI in predictive process monitoring. Furthermore, while Mehdiyev and Fettke (2021) and Elkhawaga et al. (2022) discuss the necessity of XAI for predictive process monitoring and propose frameworks for building relevant solutions, they do not provide an SLR. Similarly, the article by Mathew et al. (2025) presents an in-depth narrative survey of emerging XAI methods. While they explicitly evaluate the efficacy of various methods, their narrative review does not follow a formal systematic protocol and omits applications in PPM. Chou et al. (2022) present a systematic review of counterfactual and causability methods in explainable AI. Their work focuses squarely on explainability theory, algorithms and

Table 12 Categorization of employed ML and explanation methods in the found literature, segmented into model interpretability, explanation scope, explanation relation and explanation format

	Interpretable AI				Explainable AI																							
									Scope												Relation		Format					
	White-Box				Black-Box				Local						Global													
Publication	Bayesian Network	Decision Tree	Lin./Log. Regression	Other	Deep Learning	Gradient Boosting	Random Forest	Other	Counterfactual	Feature Importance	ICE	LIME	Shapley-based	Other	Feature Importance	PDP	Shapley-based	Other	Model-agnostic	Model-specific	Numeric	Rule-based	Textual	Visual
Verenich et al. (2016)							■	■						■					■					■
Verenich et al. (2017)							■	■						■					■		■			■
Verenich et al. (2019)						■		■						■					■		■			■
Völzer et al. (2023)				■										■				■		■	■	■		■
Weinzierl et al. (2020)					■									■						■	■			■
Wickramanayake et al. (2022a)					■									■						■	■			■
Wickramanayake et al. (2022b)					■									■						■	■			■
Zilker et al. (2023)					■								■	■					■	■	■			■

applications, but does not include interpretable or explainable techniques applied to predictive process monitoring, nor does it propose evaluation metrics for explanation quality. Rivera-Lazo et al. (2023) conduct a PRISMA-based systematic literature review on attention mechanisms in process mining. They highlight attention as a post-hoc explainability method for sequence prediction, but they do not assess interpretability or explainability metrics. Hoogendoorn et al. (2023) perform a semi-systematic survey of explainability in process mining using the BPI Challenge 2020 as a case study. Their analysis covers discovery and compliance models with post-hoc explanations, yet it neither addresses predictive monitoring nor includes evaluation frameworks for explanation quality. These contributions are valuable in demonstrating applied examples and discussing frameworks, but they do not offer a broad overview of the field.

In contrast, our contribution lies in the systematic and focused exploration of interpretable and explainable AI in predictive process monitoring. We build on the foundation laid by previous surveys but go further by explicitly focusing on XAI approaches. Our study systematically collects and synthesizes the latest research, providing a nuanced understanding of the characteristics, capabilities and limitations of various XAI methods. We aim to fill the gaps left by previous studies, offering a comprehensive review that not only maps the current landscape but also critically assesses methodologies, identifies research gaps and provides clear, evidence-based recommendations for researchers and practitioners. Our SLR thus contributes to a more organized, centralized understanding of XAI in predictive process monitoring, supporting informed decision-making and guiding future research in this vital area.

4.2 Challenges and open issues

The critical exploration of explainable and interpretable AI surfaces a multitude of challenges and open issues, pivotal among which is the frequent omission of proper evaluation. A significant proportion of studies in the field prioritize the accuracy of ML algorithms, often relegating the evaluation of explainability and interpretability to a secondary concern. This singular focus not only undermines XAI's core tenet of making complex algorithms understandable to humans, but also jeopardizes the utility of these systems in practical scenarios where understanding the "why" behind decisions is as important as the decisions themselves.

For those studies that do venture into the evaluation of their XAI approaches, many anchor themselves firmly in either qualitative or quantitative domains. The resultant analyses are thereby one-dimensional, offering a sliver of insight into either the measurable effectiveness or the subjective user experience of the explanations generated. What this dichotomy fails to capture is the nuanced interplay between these two facets in real-world applications. A more comprehensive, multifaceted approach is called for: one that synthesizes both quantitative precision and qualitative depth to yield a richer, more rounded assessment of XAI methods.

The preference for using benchmark datasets, such as the BPIC datasets, tends to amplify this issue. These datasets allow for rigorous quantitative analysis, yet they simultaneously constrain the possibility of qualitative assessment due to the lack of access to domain experts. These experts are crucial for interpreting the results within a meaningful context, ensuring that the explanations provided by XAI systems align with domain-specific knowledge and practical realities. Further complicating the landscape is the issue of transferabil-

Table 13 Categorization of employed explanation evaluation methods for PPM approaches

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Agarwal et al. (2022)	■										
Aguilar Magalhães et al. (2025)	■										
Aversano et al. (2023)		■		■		■			■		
Bayomie et al. (2022)	■										
Bezerra et al. (2009)	■										
Bezerra and Wainer (2011)	■										
Bezerra and Wainer (2013)	■										
Böhmer and Rinderle-Ma (2018)	■										
Böhmer and Rinderle-Ma (2020)		■	■				■				
Böhmer and Rinderle-Ma (2020)	■										
Brunk et al. (2021)	■										

Table 13 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Allah Bukhsh et al. (2019)	■										
Cao et al. (2023a)	■										
Cao et al. (2023b)	■										
Coma-Puig and Carmona (2022)	■										
Conforti et al. (2016)	■										
De Koninck et al. (2017)		■		■	■					■	
De Leoni et al. (2015)	■										
De Oliveira et al. (2020a)	■										
De Oliveira et al. (2020b)	■										
Diamantini et al. (2024)	■										
Di Francescomarino et al. (2016)	■										

Table 13 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Francesco-marino et al. (2019)	■										
Elkhawaga et al. (2023)		■		■		■		■	■	■	
Elkhawaga et al. (2024)		■		■		■					
Folino et al. (2011)	■										
Folino et al. (2017)		■	■		■					■	
Folino et al. (2024)	■										
Folino et al. (2025)	■										
Fu et al. (2021)	■										
Galanti et al. (2020)	■										
Galanti et al. (2023a)		■	■		■					■	

Table 14 Categorization of employed explanation evaluation methods for PPM approaches

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Galanti et al. (2023b)	■										
García-Bañuelos et al. (2017)		■	■	■	■		■			■	
Gerlach et al. (2022)	■										
Guo et al. (2024)	■										
Gupta et al. (2015)	■										
Hanga et al. (2020)	■										
Harl et al. (2020)	■										
Hermann et al. (2024)	■										
Horita et al. (2016)	■										
Hsieh et al. (2021)	■										
Huang et al. (2022)	■										
Irarrázaval et al. (2021)	■										

Table 14 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Khemiri et al. (2018)	■										
Kim et al. (2024)	■										
Lakshmanan et al. (2011)	■										
Maggi et al. (2014)	■										
Maita et al. (2025)		■	■				■				
Majumdar et al. (2023)		■	■				■				
Malashin et al. (2025)	■										
Mayer et al. (2021)	■										
Mehdiyev and Fettke (2020a)		■		■		■		■			
Mehdiyev and Fettke (2020b)	■										
Mehdiyev and Fettke (2021)	■										

Table 14 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Mehdiyev et al. (2021)	■										
Mehdiyev et al. (2024)		■		■		■					
Mehdiyev et al. (2025a)		■	■				■				
Mehdiyev et al. (2025b)		■		■		■					■
Montoya and Astudillo (2023)	■										
Montoya et al. (2023)	■										
Myers et al. (2018)	■										
Ouyang et al. (2021)	■										
Padella et al. (2022)	■										
Park et al. (2024)	■										
Pasquadibisceglie et al. (2021)	■										

Table 15 Categorization of employed explanation evaluation methods for PPM approaches

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Pasquadibisceglie et al. (2023)	■										
Pasquadibisceglie et al. (2024a)	■										
Pasquadibisceglie et al. (2024b)	■										
Pasquadibisceglie et al. (2024)	■										
Pauwels and Calders (2019a)	■										
Pauwels and Calders (2019b)	■										
Petsis et al. (2022)	■										
Polato et al. (2018)	■										
Porouhan (2024)	■										
Prasidis et al. (2021)	■										

Table 15 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Par-si-mo-ny	Sta-bil-ity
Rauch et al. (2024)	■										
Rehse et al. (2019)	■										
Rizzi et al. (2020)	■										
Rizzi et al. (2024)	■										
Saha et al. (2024)	■										
Saini et al. (2020)	■										
Sarno et al. (2020)	■										
Savickas et al. (2014)	■										
Savickas and Vasilecas (2018)	■										
Sindhgatta et al. (2020)	■										
Sindhgatta et al. (2020)	■										
Stevens and De Smedt (2022a)		■		■		■		■	■	■	

Table 15 (continued)

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Stevens et al. (2022b)	■										
Stevens et al. (2022c)	■										
Tama et al. (2020)	■										
Teinemaa et al. (2016)	■										
Toh et al. (2022)	■										
Trescato et al. (2024)	■										
Tripathi et al. (2024)	■										
Unuvar et al. (2016)	■										
van Zelst et al. (2020)	■										
Velmurugan et al. (2021a)		■		■		■		■			
Velmurugan et al. (2021b)		■		■		■					■

Table 16 Categorization of employed explanation evaluation methods for PPM approaches

	Evaluation										
	Performed		Type		Method			Metrics			
Publication	No	Yes	Qualitative	Quantitative	Application-grounded	Functional-grounded	Human-grounded	Fidelity	Functional Complexity	Parsimony	Stability
Verenich et al. (2016)	■										
Verenich et al. (2017)	■										
Verenich et al. (2019)	■										
Völzer et al. (2023)	■										
Weinzierl et al. (2020)	■										
Wickramanayake et al. (2022a)	■										
Wickramanayake et al. (2022b)	■										
Zilker et al. (2023)		■	■				■				

ity. The tendency of studies to narrow their focus to specific domains, such as healthcare or finance, begs the question of how well these solutions can be applied across different fields. This siloed approach to research overlooks the importance of generalization properties, leaving unaddressed the potential for XAI solutions to adapt to and function within a variety of domains.

Moreover, the scarcity of real-world studies beyond those involving BPIC data presents a considerable gap in the literature. The evaluations that do exist often occur in controlled "laboratory" environments, devoid of the economic and organizational contexts that heavily influence the feasibility, scalability and economic viability of XAI solutions for predictive process monitoring. Without the consideration of these broader factors, the evaluations remain theoretical exercises rather than practical analyses.

In this respect, the discussion points to the necessity for XAI research to transcend its current confines. To advance, it must embrace evaluations that not only traverse the spectrum from quantitative to qualitative but also consider the systemic implications of deploying XAI in diverse, real-world settings. By integrating economic and organizational considerations, future research can aspire to develop XAI solutions that are not only technically robust and understandable but also practically implementable and economically sustainable. Such holistic evaluations will provide a crucial bridge between the theoretical promise of XAI and its real-world applicability, ultimately driving the field towards mature, responsible and widespread use of interpretable and explainable systems.

4.3 Practical implications

The practical implications of explainability and interpretability in the realm of predictive process monitoring are profound and multifaceted. As organizations increasingly deploy ML algorithms to predict future process behaviors, the need for these systems to be transparent and comprehensible becomes paramount. XAI bridges the gap between the complexity of ML models and the operational necessity for clarity and accountability in decision-making processes. In industries where process outcomes are critical, such as healthcare, the ability of stakeholders to understand and trust AI-based predictions is not a luxury but a requirement. The practical deployment of XAI in these settings implies that operators and decision-makers can glean insights into the reasoning behind predictions, facilitating informed interventions and strategic planning. For instance, in a manufacturing plant, an interpretable model can illuminate the factors leading to potential equipment failure, enabling preemptive maintenance and reducing downtime (Mehdiyev et al. 2022).

Furthermore, the practicality of explainability extends to the adaptability and scalability of interpretability methods. In the ever-changing landscape of process data, AI systems must provide timely and contextually relevant explanations. The need for explanations to be customizable and aligned with users' varying levels of expertise and objectives. This adaptability ensures that AI serves its intended purpose effectively across different contexts and user groups, a critical consideration in BPM's diverse and dynamic environments. Moreover, XAI can play a pivotal role in regulatory compliance and risk management. In sectors like finance or law, where predictive models are used to make significant decisions, regulators increasingly demand transparency. XAI methods that can elucidate the logic behind loan application processes or patient pathway assessments are beneficial and may soon be mandated as standard practice.

Table 17 Summary and categorisation of related work

Characteristics	Related work												
	Chou et al. (2022)	Di Francescomarino et al. (2018)	Elkhawaga et al. (2022)	Hoogendoorn et al. (2023)	Kubrak et al. (2022)	Maggi et al. (2014)	Márquez-Chamorro et al. (2017)	Mathew et al. (2025)	Mehdiyev and Fettke (2021)	Rivera-Lazo et al. (2023)	Stierle et al. (2021)	Teinemaa et al. (2019)	This article
Is the primary emphasis of the article on interpretability or explainability?	■		■	■				■	■	■	■		■
Does the article include interpretable AI methods for predictive process monitoring?		■			■	■	■				■	■	■
Does the article include explainable AI methods for predictive process monitoring?			■						■	■	■		■
Does the article discuss the evaluation of interpretability or explainability?			■					■	■		■		■
Is the article a completed systematic review of literature?	■	■			■	■	■			■		■	■

However, translating XAI from theory to practice also may entail several complexities. One of the primary concerns is the integration of XAI systems within existing IT infrastructures. Many organizations operate on legacy systems and introducing sophisticated XAI solutions requires careful planning and execution to ensure compatibility and minimal disruption to ongoing operations. Another practical implication is the need for user training and adaptation. The effectiveness of an XAI system is contingent on the end-user's ability to interpret and act upon the explanations provided. This necessitates training programs to enhance the AI literacy of the workforce, ensuring that users can leverage the full potential of XAI in their day-to-day responsibilities. Furthermore, the economic impact of implementing XAI systems must be considered. Organizations need to evaluate the cost-benefit ratio of adopting such technologies, weighing the potential savings from improved process efficiencies against the investment in technology and training. The practical implications of XAI also extend to the continuous monitoring and updating of these systems. As processes evolve and new data becomes available, XAI models must be maintained and retrained to ensure their explanations remain accurate and relevant. This ongoing maintenance requires a commitment to resource allocation and a strategy for long-term management.

In conclusion, the practical implications present a complex array of challenges and opportunities. For XAI to be successfully integrated into predictive process monitoring, organizations must navigate the technical, operational and economic landscapes, balancing the promise of AI-driven insights with the realities of their application in the real world. As the field of XAI matures, this pragmatic approach will likely dictate the success and proliferation of explainable systems in industry.

4.4 Scientific and theoretical implications

The integration of XAI within predictive process monitoring is not just a practical enhancement; it represents a paradigm shift in how scientific inquiry and theoretical development are approached in the context of complex systems. This transformation encompasses fundamental methodological considerations ranging from data representation strategies to evaluation frameworks, requiring systematic reconsideration of established scientific practices.

From a scientific perspective, the incorporation of XAI opens new avenues for research in algorithmic transparency and interpretability. It challenges the conventional black-box approach to ML, calling for novel algorithms and models that are inherently interpretable or can be paired with explanation mechanisms. This need accelerates advancements in areas like feature importance analysis, counterfactual explanations and causal inference models, all of which contribute to a deeper understanding of the underlying mechanics of complex predictive models. Critical methodological foundations require systematic evaluation rather than arbitrary selection, as studies reveal that encoding selection represents a pervasive methodological flaw, with most studies deploying default configurations without rigorous justification, potentially compromising scientific validity (Tavares et al. 2023). The importance of rigorous methodological choices extends beyond individual techniques to encompass the entire analytical pipeline, findings demonstrate that sophisticated symbolic sequence encodings significantly outperform naive approaches, emphasizing how foundational data transformation decisions influence system effectiveness (Leontjeva et al. 2015).

In the theoretical context, XAI stimulates a re-evaluation of existing theories related to decision-making, cognition and information processing. It brings to light questions about

the nature of understanding and trust in automated systems. For instance, what constitutes a "good" explanation in a predictive process monitoring context and how do these explanations impact human decision-making and trust? These fundamental questions require systematic investigation of how various design decisions influence explanation effectiveness, including how encoding methodologies shape the relationship between domain knowledge and computational representations. Studies illustrate how the transition from knowledge-driven to data-driven encoding approaches introduces fundamental tensions between performance optimization and interpretability requirements (Senderovich et al. 2017), while research also establishes that encoding complexity directly impacts model interpretability, with sophisticated transformations potentially obscuring relationships between process characteristics and predictions (Stevens and De Smedt 2022a). The pursuit of answers encourages interdisciplinary collaboration, drawing from fields such as psychology, cognitive science and philosophy to enrich the theoretical underpinnings of XAI.

Furthermore, XAI's focus on interpretability and explainability mandates a rigorous theoretical understanding of the processes being monitored. This requirement not only reinforces the need for domain expertise in model development but also promotes a more symbiotic relationship between domain experts and data scientists. The scientific implications extend to fundamental methodological foundations, as encoding decisions shape assumptions about what constitutes meaningful process knowledge and how it can be communicated through explainable AI systems, Adams et al. (2022) demonstrate that object-centric encoding approaches require specialized consideration of multi-dimensional relationships that traditional explanation methods cannot adequately address. In this context, predictive process monitoring becomes a collaborative scientific endeavor, blending empirical data analysis with domain-specific insights to produce models that are both high-performing and understandable.

The scientific implications of XAI also extend to the validation and evaluation of AI models. Traditional performance metrics like accuracy, precision and recall are no longer sufficient. XAI introduces the need for new metrics and methodologies that can assess the quality of explanations in terms of relevance, completeness and comprehensibility. This evolution requires systematic evaluation frameworks that consider how fundamental design decisions, including encoding methodologies that create cascading effects throughout the analytical pipeline, influence both predictive accuracy and explanation fidelity. This evolution reflects a broader shift in the scientific community's approach to evaluating AI, placing equal emphasis on the interpretability and operational effectiveness of the models.

From a theoretical standpoint, XAI challenges and refines our understanding of concepts like causality, uncertainty and prediction (Mehdiyev et al. 2025a). It encourages a more nuanced exploration of how these elements interplay in complex systems and how they can be effectively communicated to users. This exploration has profound implications for theoretical models across various domains, from supply chain management to healthcare, where understanding the causal relationships and uncertainties inherent in predictive models is crucial for effective decision-making.

In summary, the integration of XAI in predictive process monitoring is catalyzing significant scientific and theoretical advancements. It is driving the development of new algorithms and models, fostering interdisciplinary research, redefining evaluation methodologies and deepening our understanding of complex systems. As the field progresses, the continued exploration of these scientific and theoretical implications, grounded in rigorous

methodological considerations including systematic encoding strategies and comprehensive evaluation frameworks, will be instrumental in realizing the full potential of XAI, not only as a tool for enhanced predictive analytics but also as a beacon for responsible and transparent AI development.

5 Future work

5.1 XAI and other trustworthy AI methods combination for predictive process monitoring

The integration of XAI with complementary trustworthy AI methodologies represents a critical frontier for advancing predictive process monitoring systems. Current research highlights significant gaps in holistic approaches that address the multifaceted nature of trust in automated decision-making environments.

Uncertainty quantification and XAI integration emerge as a primary research priority. Research demonstrates substantial gaps in integrating these techniques effectively, with current approaches representing the first attempts to merge uncertainty quantification with explainable AI within predictive process monitoring contexts (Mehdiyev et al. 2023; Weytjens and De Weerdt 2022; Prasidis et al. 2021; Majlatow et al. 2025). The integration must address bidirectional challenges where uncertainty quantification enhances explanation trustworthiness while explainable methods elucidate sources of model uncertainty (Mehdiyev et al. 2025a, 2023). Future work should develop frameworks that communicate confidence bounds for both predictions and their underlying explanations, addressing current limitations where traditional explanation methods fail to convey reliability information to stakeholders.

Privacy-preserving explainable AI represents another critical convergence area, as organizations increasingly require transparency without compromising sensitive process data (Mannhardt et al. 2019). Recent advances demonstrate comprehensive frameworks that combine XAI with privacy-preserving machine learning, achieving significant improvements in both interpretability scores and privacy adherence compared to existing approaches. Future research should explore federated explanation architectures that enable cross-organizational transparency while maintaining regulatory compliance, developing differential privacy techniques that provide meaningful explanations while protecting individual case confidentiality (Fahrenkrog-Petersen et al. 2023).

Fairness-aware explainable process monitoring demands systematic investigation given the potential for algorithmic bias in resource allocation and case prioritization decisions (Qafari et al. 2019). Current research reveals capabilities for identifying procedure-based bias through explanation quality measurement across different demographic groups. The development of fairness metrics for explanations provides pathways for detecting and mitigating bias in process predictions, particularly crucial where socioeconomic factors might inappropriately influence outcomes. Future work should investigate how explanation systems can detect and communicate potential fairness issues while maintaining predictive accuracy.

Human-centered trustworthy AI design requires attention to practical implementation challenges beyond technical integration (Shneiderman 2022). Research emphasizes that

trustworthy systems must incorporate fundamental principles into AI development procedures while ensuring systems abide by moral and legal requirements. The framework for explainable predictive process monitoring demonstrates that stakeholder trust represents a necessary condition for adoption, with systems lacking explanation capabilities facing significant adoption barriers. Future work should investigate adaptive explanation systems that customize communication strategies based on stakeholder expertise and decision-making contexts, requiring interdisciplinary collaboration between AI researchers, process management experts and human-computer interaction specialists.

Real-time trustworthy AI systems must address the temporal dynamics of trust-building in operational environments. Human-centric monitoring approaches reveal the need for systems that enable automatic report generation while preserving human autonomy through collaborative decision-making frameworks. Future research should develop learning explanation systems that evolve based on user feedback, creating dynamic trust-building capabilities aligned with changing organizational needs (Reinkemeyer 2020). Such adaptive systems would represent significant advancement beyond current static explanation approaches, offering temporal trust-building that aligns with evolving business processes.

Multi-modal trustworthy integration requires novel evaluation methodologies that assess explanation quality alongside privacy preservation, fairness, uncertainty communication and human usability (Chvirova et al. 2024). Research demonstrates the complexity of evaluation through multiple taxonomies based on both applications and evaluation metrics. Future work should prioritize developing standardized frameworks that enable systematic comparison of integrated trustworthy AI approaches while addressing scalability and real-time performance requirements essential for operational process monitoring systems. This integration must consider edge computing architectures that deliver trustworthy explanations in distributed environments without compromising system responsiveness or data security.

The synthesis of these trustworthy AI dimensions requires novel evaluation methodologies that assess explanation quality alongside privacy preservation, fairness, uncertainty communication and human usability. Future work should prioritize developing standardized frameworks that enable systematic comparison of integrated trustworthy AI approaches, while also addressing scalability and real-time performance requirements essential for operational process monitoring systems.

5.2 LLM and XAI integration for predictive process monitoring

The emergence of large language models (LLM) presents transformative opportunities for advancing explainable predictive process monitoring through natural language interfaces and enhanced interpretability mechanisms (Sebin et al. 2024; Bilal et al. 2025). Recent advances demonstrate that LLMs can serve not merely as automation tools but as collaborative partners in process management, fundamentally reshaping how stakeholders interact with and understand predictive analytics systems (Pfeiffer et al. 2025; Berti and Qafari 2023).

Conversational explanation systems represent the most promising integration pathway, addressing critical limitations in current static explanation approaches (Zhang et al. 2025). Research demonstrates that LLM-driven frameworks can transform opaque predictions into auditable, interactive workflows by enabling natural language dialogues grounded in process mining explanations (Fahland et al. 2025; Wang et al. 2024). These systems employ

multi-agent architectures that decompose user queries into specialized tasks, mirroring human problem-solving approaches through assumption probing, hypothesis testing and conclusion refinement (He et al. 2025). Future work should investigate how conversational interfaces can adapt explanations not just to data characteristics but to stakeholders' evolving priorities and domain expertise levels.

Process data abstraction and semantic alignment emerges as a fundamental research challenge requiring systematic investigation. Current approaches demonstrate that LLMs exhibit robust understanding of key process mining abstractions with notable proficiency in interpreting both declarative and procedural process models (Rebmann et al. 2025; Berti et al. 2023). However, effectively embedding process data within language model frameworks while preserving semantic integrity remains complex. Future research should develop standardized methodologies for transforming process mining artifacts into textual representations that maintain temporal dependencies, causality relationships and domain-specific constraints essential for accurate predictive monitoring.

Multi-modal explanation generation presents opportunities for enhanced stakeholder comprehension through diverse communication channels. Research reveals that LLM architectures can integrate dashboards, conversational widgets and visual analytics to present predictions and uncertainty intervals in intuitive format (Mehdiyev et al. 2023; Park et al. 2018). Future work should investigate how language models can orchestrate multiple explanation modalities, automatically selecting appropriate visualization and communication strategies based on user context, query complexity and decision-making requirements. This integration should address the challenge of maintaining consistency across different explanation formats while optimizing for stakeholder understanding.

Domain-specific knowledge integration demands attention to specialized vocabularies and regulatory requirements across different industries. Research indicates that LLMs demonstrate capacity for evaluating fairness concepts in process mining, opening pathways for rapid assessment of event log bias and compliance issues (Gallegos et al. 2024; Berti et al. 2024). Future work should investigate how domain knowledge graphs and specialized corpora can enhance LLM understanding of industry-specific process constraints, regulatory requirements and stakeholder priorities (Vogt et al. 2024). This integration should maintain generalizability while providing deep domain expertise for sectors such as healthcare, finance and manufacturing.

Retrieval-augmented explanation systems offer pathways for maintaining consistency and coherence in generated explanations while incorporating evolving process knowledge. Current approaches employ vector-based indexing mechanisms to rank and incorporate relevant explanations based on semantic similarity (Ehsan and Riedl 2024). Future research should develop sophisticated retrieval strategies that consider temporal dynamics, process evolution and stakeholder feedback to continuously improve explanation quality and relevance. These systems should balance between leveraging historical explanation patterns and adapting to novel process scenarios.

5.3 Explainable predictive process monitoring on stream event data

The convergence of XAI with stream event processing represents a transformative frontier for real-time predictive process monitoring, addressing critical gaps in current approaches that primarily focus on static, batch-oriented explanation generation (Burattin 2022). The

increasing ubiquity of streaming data in modern business environments demands explanation systems capable of providing transparent insights into predictive decisions as events unfold in real-time (Mehdiyev et al. 2015).

Real-time explanation generation emerges as the fundamental challenge requiring systematic investigation. Current research demonstrates that predictive process monitoring systems must provide explanations that are not only accurate but also timely enough to support operational decision-making (Mehdiyev and Fettke 2020a). The main goal of predictive process monitoring involves predicting possible outcomes, execution times and costs using historical data, but traditional explanation approaches fail to accommodate the temporal constraints inherent in streaming environments. Future work should develop explanation frameworks that can generate interpretable insights within milliseconds of receiving new event data, enabling stakeholders to understand and act upon predictions before process conditions change (Mozolewski et al. 2024).

Complex event processing integration presents opportunities for enhanced explanation capabilities through sophisticated pattern recognition and abstraction mechanisms (Krumeich et al. 2015). Research indicates that complex event processing technology enables dynamic processing of multiple events simultaneously, allowing for the expression of causal, temporal, spatial and other relations between events (Mehdiyev et al. 2015; Cugola and Margara 2012). These relationships specify patterns that can be leveraged for real-time event monitoring and explanation generation. Future research should investigate how complex event processing architectures can be augmented with explanation capabilities, enabling the identification and communication of meaningful patterns that drive predictive decisions in streaming environments.

Temporal pattern explanation demands novel approaches for communicating how temporal dependencies influence predictive decisions (Cheikhrouhou et al. 2015). Research reveals that traditional analytics tools are generally not well-suited for complex event processing, particularly when computing temporal or spatial patterns from raw streaming data. Future research should investigate explanation methods that can effectively communicate temporal causality, spatio-temporal dependencies and time-based aggregations to stakeholders who may not possess technical expertise in stream processing concepts (Cheng et al. 2021). This includes developing visualization techniques that can represent temporal explanation patterns in intuitive formats suitable for real-time decision support.

Scalability and latency optimization requires careful consideration of computational trade-offs between explanation quality and system performance (Salama et al. 2019). Current implementations demonstrate that explanation systems must retain highly efficient implementations suitable for data stream processing requirements (Bhat and Raychowdhury 2023). Future work should investigate distributed explanation architectures that can maintain low-latency performance while providing comprehensive interpretability across high-volume event streams. This includes exploring edge computing approaches that can provide local explanations for distributed streaming sources without compromising system responsiveness.

Adaptive explanation strategies should address the dynamic nature of streaming environments where process patterns and stakeholder information needs evolve continuously (Su et al. 2024). Future research should develop explanation systems that can automatically adapt their communication strategies based on changing event patterns, stakeholder feedback and evolving process contexts (Turchi et al. 2024). Such adaptive systems would

represent significant advancement beyond current static explanation approaches, offering dynamic interpretability that aligns with the inherently dynamic nature of streaming business processes and operational decision-making requirements.

5.4 XAI and object-centric process mining and predictions

The emergence of object-centric process mining represents a paradigmatic shift that fundamentally challenges traditional explanation approaches in predictive process monitoring (Gianola et al. 2024; Berti et al. 2023). The transition from single-case perspectives to multi-object analytical frameworks introduce unprecedented complexity for explainable artificial intelligence systems, requiring novel approaches that can illuminate prediction rationales across interconnected object relationships and temporal dependencies (van der Aalst 2023).

Multi-object explanation frameworks represent the most critical research frontier, addressing the fundamental challenge of communicating predictions that span multiple interconnected objects within unified process models (Basmer et al. 2024). Current research demonstrates that object-centric process mining allows events to be related to multiple objects simultaneously, creating complex webs of relationships that traditional explanation methods cannot adequately address (Fioretto and Masciari 2025). Future work should develop explanation architectures capable of tracing prediction influences across object boundaries, enabling stakeholders to understand how decisions about one object type influence predictions for related objects. This requires novel visualization and communication strategies that can represent multi-dimensional causal relationships without overwhelming users with excessive complexity.

Cross-object dependency explanation demands systematic investigation of how temporal and causal relationships between different object types influence predictive decisions (Galanti et al. 2023b). Research reveals that object-centric approaches enable visualization and comprehension of interactions across different object types, emphasizing that performance and compliance issues cannot be understood when objects are considered in isolation (Liss et al. 2023). Future research should explore explanation methods that can effectively communicate cross-object dependencies, particularly in scenarios where predictions for one object type depend on historical patterns or current states of related objects. This includes developing techniques for explaining how object lifecycle interactions contribute to prediction confidence and accuracy.

Unified explanation models require attention to the structural complexity inherent in object-centric process representations (Adams et al. 2023). Current approaches demonstrate that object-centric process mining provides more realistic representations of enterprise data by eliminating the need for repeated data extraction whenever perspectives change, but this structural flexibility introduces significant challenges for maintaining explanation consistency (Adams et al. 2022). Future work should investigate explanation frameworks that can adapt to different object-centric perspectives while maintaining interpretability coherence across various analytical viewpoints. This includes developing explanation architectures that can seamlessly transition between object-specific and cross-object analytical perspectives.

Behavioral pattern explanation presents opportunities for enhanced understanding through specialized explanation approaches tailored to object-centric behavioral patterns (Miri and Jalali 2024; Porsil and van der Aalst 2025). Research indicates that object-centric local process models enable analyzing complex processes by focusing on specific behavioral

patterns that span multiple object types (Peeva et al. 2024). Future research should develop explanation methods specifically designed for object-centric behavioral patterns, enabling stakeholders to understand how localized process behaviors contribute to broader predictive insights. This includes investigating techniques for explaining pattern significance, temporal boundaries and cross-pattern interactions that influence predictive accuracy.

Evaluation methodologies must address the unique challenges of assessing explanation quality in multi-object predictive environments (Aliyeva and Mehdiyev 2024). Traditional explanation evaluation approaches may prove inadequate for object-centric contexts where prediction quality and explanation relevance depend on complex inter-object relationships (Adams and van Der Aalst 2021). Future research should develop specialized evaluation frameworks that can assess explanation effectiveness across multiple object types while considering the temporal and causal complexities inherent in object-centric process models.

6 Conclusion

This SLR was motivated by the urgent need to navigate the rapidly expanding yet fragmented landscape of XAI for PPM. By systematically synthesizing and structuring over one hundred studies published up to 2025, we have provided a comprehensive, evidence-based overview of the field's current state, key achievements and most significant gaps. Our analysis reveals a field in dynamic transition. While early efforts focused on intrinsically interpretable models, the pursuit of higher predictive accuracy has driven a decisive shift towards complex black-box models. Consequently, the field is now heavily reliant on post-hoc explanation methods like SHAP and LIME. However, this progress in model complexity has not been matched by progress in evaluation. The vast majority of studies still prioritize the evaluation of predictive performance over a rigorous assessment of the generated explanations themselves, with a notable scarcity of human-grounded studies to validate their real-world utility. This reveals a critical imbalance: the field is succeeding in generating explanations, but it is not yet systematically verifying if they are meaningful, reliable or useful to human stakeholders.

The primary contributions of this review are therefore threefold. We have presented a comprehensive synthesis of the current research landscape, structured along key dimensions including application domains, datasets, predictive tasks and AI methodologies. Furthermore, we have identified and detailed critical research frontiers where current approaches fall short, particularly regarding the foundational impact of data encoding, the paradigmatic shift to OCPM and the unique challenges of streaming data. Finally, this work establishes a forward-looking research agenda designed to guide future work toward addressing these pressing challenges and advancing the maturity of the field. Looking forward, the future of explainable PPM must be defined by a move from mere generation to meaningful interaction. Our findings call for a concerted research effort in several key areas. Investigators must tackle the foundational challenge of data encoding, as these choices fundamentally shape what a model can learn and explain. A new generation of XAI techniques is required to handle the multi-object dependencies inherent in OCPM. Furthermore, as business processes become increasingly real-time, developing low-latency, adaptive explanation systems for streaming data is no longer optional, but essential. For practitioners, our findings serve as a call for critical evaluation. It is not enough to simply adopt a model that produces an expla-

nation; one must question its fidelity, reliability and relevance to the specific operational context. For researchers, this review is a call to action: to shift focus toward a more holistic, human-centric and rigorous evaluation of explainability.

While this review was conducted with methodological rigor, it is subject to the inherent limitations of any SLR, such as the scope of the queried databases and the specific search terms used. Nonetheless, we are confident that our work provides a robust and essential baseline. Ultimately, the goal of XAI in process mining is not just to open the black box, but to build a durable bridge of trust between human decision-makers and the complex AI that supports them. By addressing the gaps identified in this review, the research community can move beyond generating explanations as a technical artifact and toward delivering them as a truly actionable and trustworthy component of intelligent process management.

Author contributions All authors contributed equally.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was partially funded by the German Federal Ministry of Research, Technology and Space under grant number 01IS24048C (project EINHORN).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Adams JN, Park G, van der Aalst WM (2023) Preserving complex object-centric graph structures to improve machine learning tasks in process mining. Eng Appl Artif Intell 125:106764

Adams JN, Park G, Levich S, Schuster D, van der Aalst WM (2022) A framework for extracting and encoding features from object-centric event data. In: International conference on service-oriented computing, 36–53. Springer

Adams JN, van Der Aalst WM (2021) Precision and fitness in object-centric process mining. In: 2021 3rd international conference on process mining (ICPM), pp 128–135 . IEEE

Agarwal P, Gao B, Huo S, Reddy P, Dechu S, Obeidi Y, Muthusamy V, Isahagian V, Carbajales S (2022) A process-aware decision support system for business processes. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, pp 2673–2681

Aguilar Magalhães R, Andolfato Filho A, Portela Santos EA, Ara A (2025) Active learning in process mining: an active sampling using dimensionality reduction and iterative selection. Int J Data Sci Anal, pp 1–16

Aliyeva K, Mehdiyev N (2024) Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: a use case from the healthcare domain. Inf Sci 657:119987

Allah Bukhsh Z, Saeed A, Stipanovic I, Doree AG (2019) Predictive maintenance using tree-based classification techniques: a case of railway switches. Transp Res Part C Emerging Technol 101:35–54. https://doi.org/10.1016/J.TRC.2019.02.001

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 82(4):1059–1086

Arras L, Montavon G, Müller K-R, Samek W (2017) Explaining recurrent neural network predictions in sentiment analysis. ArXiv **abs/1706.07206**

Aversano L, Bernardi ML, Cimitile M, Iammarino M, Verdone C (2023) A data-aware explainable deep learning approach for next activity prediction. Eng Appl Artif Intell 126:106758

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fus 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Basmer M, Kabierski M, Sahling K, Patecka A, Bala S, Mendling J (2024) A classification of data quality issues in object-centric event data. In: International Conference on Process Mining, pp 311–323. Springer

Bayomie D, Revoredo K, Di Ciccio C, Mendling J (2022) Improving accuracy and explainability in event-case correlation via rule mining. In: 2022 4th international conference on process mining (ICPM), pp 24–31. IEEE

Berti A, Qafari MS (2023) Leveraging large language models (LLMS) for process mining (technical report). arXiv preprint arXiv:2307.12701

Berti A, Schuster D, van der Aalst WM (2023)Abstractions, scenarios, and prompt definitions for process mining with LLMS: a case study. In: International conference on business process management, pp 427–439. Springer

Berti A, Kourani H, Häfke H, Li CY, Schuster D(2024) Evaluating large language models in process mining: capabilities, benchmarks, and evaluation strategies. In: International conference on business process modeling, development and support, pp 13–21 . Springer

Berti A, Montali M, van der Aalst WM (2023) Advancements and challenges in object-centric process mining: a systematic literature review. arXiv preprint arXiv:2311.08795

Bezerra F, Wainer J (2011) Fraud detection in process aware systems. Int J Bus Process Integr Manag 5(2):121–129

Bezerra F, Wainer J (2013) Algorithms for anomaly detection of traces in logs of process aware information systems. Inf Syst 38(1):33–44

Bezerra F, Wainer J, van der Aalst WM (2009) Anomaly detection using process mining. In: International workshop on business process modeling, development and support, pp 149–161. Springer

Bhat A, Raychowdhury A (2023) Non-uniform interpolation in integrated gradients for low-latency explainable-AI. In: 2023 IEEE International symposium on circuits and systems (ISCAS), pp 1–5 . IEEE

Bilal A, Ebert D, Lin B (2025) LLMS for explainable AI: a comprehensive survey. arXiv preprint arXiv:2504.00125

Böhmer K, Rinderle-Ma S (2020) Mining association rules for anomaly detection in dynamic process runtime behavior and explaining the root cause to users. Inf Syst 90:101438

Böhmer K, Rinderle-Ma S (2018) Probability based heuristic for predictive business process monitoring. In: On the move to meaningful internet systems. OTM 2018 conferences: confederated international conferences: CoopIS, C &TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I, pp 78–96. Springer

Böhmer K, Rinderle-Ma S (2020) Logo: combining local and global techniques for predictive business process monitoring. In: Advanced information systems engineering: 32nd international conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, proceedings 32, pp 283–298. Springer

Brunk J, Stierle M, Papke L, Revoredo K, Matzner M, Becker J (2021) Cause vs. effect in context-sensitive prediction of business process instances. Inf Syst 95:101635. https://doi.org/10.1016/j.is.2020.101635

Burattin A (2022) Streaming process mining. Process Mining Handbook 349:3–10

Cao R, Zeng Q, Ni W, Duan H, Liu C, Lu F, Zhao Z (2023) Business process remaining time prediction using explainable reachability graph from gated RNNs. Appl Intell 53(11):13178–13191

Cao R, Zeng Q, Ni W, Lu F, Liu C, Duan H (2023) Explainable business process remaining time prediction using reachability graph. Chin J Electron 32(3):625–639

Cheikhrouhou S, Kallel S, Guermouche N, Jmaiel M (2015) The temporal perspective in business process modeling: a survey and research challenges. SOCA 9(1):75–85

Cheng X, Wang J, Li H, Zhang Y, Wu L, Liu Y (2021) A method to evaluate task-specific importance of Spatio-temporal units based on explainable artificial intelligence. Int J Geogr Inf Sci 35(10):2002–2025

Chou Y-L, Moreira C, Bruza P, Ouyang C, Jorge J (2022) Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf Fusion 81:59–83

Chvirova D, Egger A, Fehrer T, Kratsch W, Röglinger M, Wittmann J, Wördehoff N (2024) A multimedia dataset for object-centric business process mining in it asset management. Data Brief 55:110716

Coma-Puig B, Carmona J (2022) Non-technical losses detection in energy consumption focusing on energy recovery and explainability. Mach Learn 111(2):487–517
Conforti R, Fink S, Manderscheid J, Röglinger M (2016) PRISM—a predictive risk monitoring approach for business processes, pp 383–400
Coussement K, Benoit DF, Van den Poel D (2010) Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Syst Appl 37(3):2132–2143
Cugola G, Margara A (2012) Processing flows of information: from data stream to complex event processing. ACM Comput Surv 44(3):1–62
de Leoni MM, Mannhardt F (2015) Road traffic fine management process. Eindhoven University of Technology
De Leoni M, Van Der Aalst WMP, Dees M (2015) A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. https://doi.org/10.1016/j.is.2015.07.003
De Weerdt J, vanden Broucke S (2014) Secpi: searching for explanations for clustered process instances. In: Sadiq S, Soffer P, Völzer H (eds) Business process management. Springer, Cham, pp 408–415
De Koninck P, De Weerdt J, vanden Broucke SKLM (2017) Explaining clusterings of process instances. Data Min Knowl Disc 31(3):774–808. https://doi.org/10.1007/s10618-016-0488-4
Dees M, van Dongen BFB (2016) Bpi challenge 2016. UWV
Dey S, Stori J (2005) A bayesian network approach to root cause diagnosis of process variations. Int J Mach Tools Manuf 45(1):75–91
De Oliveira H, Augusto V, Jouaneton B, Lamarsalle L, Prodel M, Xie X (2020a) An optimization-based process mining approach for explainable classification of timed event logs. In: 2020 IEEE 16th international conference on automation science and engineering (CASE), pp 43–48. IEEE
De Oliveira H, Augusto V, Jouaneton B, Lamarsalle L, Prodel M, Xie X (2020b) Automatic and explainable labeling of medical event logs with autoencoding. IEEE J Biomed Health Inform 24(11):3076–3084
Diamantini C, Genga L, Mircoli A, Potena D (2024) Evidence-driven appraisal of students' careers using process mining: a case study. J Intell Inf Syst, pp 1–20
Di Francescomarino C, Ghidini C, Maggi FM, Milani F (2018) Predictive process monitoring methods: Which one suits me best? In: International Conference on Business Process Management, pp 462–479 . Springer
Dongen BF, Medeiros AKA, Verbeek HMW, Weijters AJMM, Aalst WMP (2005) The prom framework: a new era in process mining tool support. In: Ciardo G, Darondeau P (eds) Applications and theory of petri nets 2005. Springer, Berlin, pp 444–454
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
Ehsan U, Riedl M (2024) Explainable AI reloaded: challenging the XAI status quo in the era of large language models. In: Proceedings of the halfway to the future symposium, pp 1–8
El-khawaga G, Abu-Elkheir M, Reichert M (2022) XAI in the context of predictive process monitoring: an empirical analysis framework. Algorithms 15(6):199
Elkhawaga G, Elzeki OM, Abu-Elkheir M, Reichert M (2024) Why should i trust your explanation? an evaluation approach for xai methods applied to predictive process monitoring results. IEEE Trans Artific Intell 5(4):1458–1472
Elkhawaga G, Abu-Elkheir M, Reichert M (2023) A rule-based evaluation method of local explainers for predictive process monitoring. In: 2023 IEEE international conference on data mining workshops (ICDMW), pp 922–930. IEEE
Evermann J, Rehse J-R, Fettke P (2017) Predicting process behaviour using deep learning. Decis Support Syst 100:129–140
Fahland D, Fournier F, Limonad L, Skarbovsky I, Swevels AJ (2025) How well can a large language model explain business processes as perceived by users? Data Knowl Eng 157:102416
Fahrenkrog-Petersen SA, Aa H, Weidlich M (2023) Optimal event log sanitization for privacy-preserving process mining. Data Knowl Engi 145:102175
Feng D, Chen F, Xu W (2013) Efficient leave-one-out strategy for supervised feature selection. Tsinghua Sci Technol 18(6):629–635. https://doi.org/10.1109/TST.2013.6678908
Fioretto S, Masciari E (2025) A comparative analysis of predictive process monitoring: object-centric versus classical event logs. Knowl Inf Syst, 1–44
Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res JMLR **20**
Folino F, Folino G, Guarascio M, Pontieri L (2024) Data- & compute-efficient deviance mining via active learning and fast ensembles. J Intell Inf Syst 62(4):995–1019

Folino F, Greco G, Guzzo A, Pontieri L (2011) Mining usage scenarios in business processes: outlier-aware discovery and run-time prediction. Data & Knowl Eng 70(12):1005–1029
Folino F, Guarascio M, Pontieri L (2017) A descriptive clustering approach to the analysis of quantitative business-process deviances. Proceedings of the ACM symposium on applied computing part F1280:765–770. https://doi.org/10.1145/3019612.3019660
Folino F, Folino G, Guarascio M, Pontieri L (2025) The force of few: boosting deviance detection in data scarcity scenarios through self-supervised learning and pattern-based encoding. Soft Comput, pp 1–16
Fraley C, Raftery A (2003) Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. J Classif 20:263–286. https://doi.org/10.1007/s00357-003-0015-3
Di Francescomarino C, Dumas M, Federici M, Ghidini C, Maggi FM, Rizzi W (2016) Predictive business process monitoring framework with hyperparameter optimization. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 9694:361–376
Francescomarino CD, Dumas M, Maggi FM, Teinemaa I (2019) Clustering-based predictive process monitoring. IEEE Trans Serv Comput 12(6):896–909. https://doi.org/10.1109/TSC.2016.2645153
Frank E, Witten IH (1998) Generating accurate rule sets without global optimization
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
Fu T, Zampieri G, Hodgson D, Angione C, Zeng Y (2021) Modeling customer experience in a contact center through process log mining. ACM Trans Intell Syst Technol 12(4):1–21
Galanti R, Leoni M, Monaro M, Navarin N, Marazzi A, Stasi B, Maldera S (2023) An explainable decision support system for predictive process analytics. Eng Appl Artif Intell 120:105904
Galanti R, Leoni M, Navarin N, Marazzi A (2023) Object-centric process predictive analytics. Expert Syst Appl 213:119173
Galanti R, Coma-Puig B, de Leoni M, Carmona J, Navarin N (2020) Explainable predictive process monitoring. In: Proceedings - 2020 2nd international conference on process mining, ICPM 2020, 1–8 arXiv:2008.01807. https://doi.org/10.1109/ICPM49681.2020.00012
Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK (2024) Bias and fairness in large language models: a survey. Comput Linguist 50(3):1097–1179
García-Bañuelos L, Van Beest NR, Dumas M, La Rosa M, Mertens W (2017) Complete and interpretable conformance checking of business processes. IEEE Trans Software Eng 44(3):262–290
Gedeon TD (1997) Data mining of inputs: analysing magnitude and functional measures. Int J Neural Syst 08(02):209–218. https://doi.org/10.1142/S0129065797000227
Gerlach Y, Seeliger A, Nolle T, Mühlhäuser M (2022) Inferring a multi-perspective likelihood graph from black-box next event predictors. In: International conference on advanced information systems engineering, pp 19–35. Springer
Gevrey M, Dimopoulos I, Lek S (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model 160(3), 249–264. https://doi.org/10.1016/S0304-3800(02)00257-0. Modelling the structure of acquatic communities: concepts, methods and problems
Gianola A, Montali M, Winkler S (2024) Object-centric conformance alignments with synchronization. In: International conference on advanced information systems engineering, pp 3–19 . Springer
Goldstein A, Kapelner A, Bleich J, Pitkin E (2013) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24. https://doi.org/10.1080/10618600.2014.907095
Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F (2019) Factual and counterfactual explanations for black box decision making. IEEE Intell Syst 34(6):14–23. https://doi.org/10.1109/MIS.2019.2957223
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
Guo N, Liu C, Li C, Zeng Q, Ouyang C, Liu Q, Lu X (2024) Explainable and effective process remaining time prediction using feature-informed cascade prediction model. IEEE Trans Serv Comput 17(3):949–962
Gupta N, Anand K, Sureka, A (2015) Pariket: Mining business process logs for root cause analysis of anomalous incidents. In: Databases in Networked Information Systems: 10th international workshop, DNIS 2015, Aizu-Wakamatsu, Japan, March 23-25, 2015. Proceedings 10, 244–263. Springer
Hall P, Gill N, Kurka M, Phan W (2017) Machine learning interpretability with H2O driverless AI. H2O. AI
Hanga KM, Kovalchuk Y, Gaber MM (2020) A graph-based approach to interpreting recurrent neural networks in process mining. IEEE Access 8:172923–172938. https://doi.org/10.1109/ACCESS.2020.3025999
Harl M, Weinzierl S, Stierle M, Matzner M (2020) Explainable predictive business process monitoring using gated graph neural networks. J Decis Syst 29(sup1):312–327. https://doi.org/10.1080/12460125.2020.1780780

He G, Aishwarya N, Gadiraju U (2025) Is conversational XAI all you need? Human-AI decision making with a conversational XAI assistant. In: Proceedings of the 30th international conference on intelligent user interfaces, pp 907–924
Hermann J, Rusche S, Moder L, Weibelzahl M (2024) Watt's next? leveraging process flexibility for power cost optimization. Bus Inf Syst Eng 66(5):541–563
Holmes G, Hall M, Prank E (1999) Generating rule sets from model trees. In: Foo N (ed) Advanced topics in artificial intelligence. Springer, Berlin, pp 1–12
Hoogendoorn T, Arachchige JJ, Bukhsh FA (2023) Survey of explainability within process mining: a case study of bpi challenge 2020. In: 2023 international conference on frontiers of information technology (FIT), 43–48. IEEE
Horita H, Hirayama H, Tahara Y, Ohsuga A (2016) Goal achievement analysis based on ltl checking and decision tree for improvements of pais. In: Proceedings of the 31st annual ACM symposium on applied computing, pp 1214–1218
Hsieh C, Moreira C, Ouyang C (2021) Dice4el: interpreting process predictions using a milestone-aware counterfactual approach. In: 2021 3rd international conference on process mining (ICPM), pp 88–95. IEEE
Huang TH, Metzger A, Pohl K (2022) Counterfactual explanations for predictive business process monitoring. Lecture notes in business information processing 437 LNBIP, pp 399–413
Irarrázaval ME, Maldonado S, Pérez J, Vairetti C (2021) Telecom traffic pumping analytics via explainable data science. Decis Support Syst 150:113559
Khemiri A, Hamri MEA, Frydman C, Pinaton J (2018) Improving business process in semiconductor manufacturing by discovering business rules. In: 2018 winter simulation conference (WSC), pp 3441–3448. IEEE
Kim S, Comuzzi M, Di Francescomarino C (2024) Explaining the impact of design choices on model quality in predictive process monitoring. J Intell Inf Syst, pp 1–26
King N (2012) Doing template analysis. Qualitative organizational research Core methods and current challenges 426:426–450
Krumeich J, Mehdiyev N, Werth D, Loos P (2015) Towards an extended metamodel of event-driven process chains to model complex event patterns. In: International conference on conceptual modeling, pp 119–130. Springer
Kubrak K, Milani F, Nolte A, Dumas M (2022) Prescriptive process monitoring: Quo Vadis? PeerJ Comput Sci 8:1097
Kumar NP, Rao MV, Krishna PR, Bapi RS (2005) Using sub-sequence information with knn for classification of sequential data. In: Distributed computing and internet technology: second international conference, ICDCIT 2005, Bhubaneswar, India, December 22-24, 2005. Proceedings 2, pp 536–546. Springer
Lakkaraju H, Kamar E, Caruana R, Leskovec J (2017) Interpretable & explorable approximations of black box models. arXiv. https://doi.org/10.48550/ARXIV.1707.01154
Lakshmanan GT, Duan S, Keyser PT, Curbera F, Khalaf R (2011) Predictive analytics for semi-structured case oriented business processes. Lecture notes in business information processing 66 LNBIP, pp 640–651. https://doi.org/10.1007/978-3-642-20511-8_59
Lapuschkin S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10:0130140. https://doi.org/10.1371/journal.pone.0130140
Leontjeva A, Conforti R, Di Francescomarino C, Dumas M, Maggi FM (2015) Complex symbolic sequence encodings for predictive monitoring of business processes. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 9253:297–313. https://doi.org/10.1007/978-3-319-23063-4_21
Levy D (2014). Production analysis with process mining technology
Li X-H, Shi Y, Li H, Bai W, Cao CC, Chen L (2021) An experimental study of quantitative evaluations on saliency methods. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 3200–3208
Liss L, Adams JN, van der Aalst WM (2023) Object-centric alignments. In: International conference on conceptual modeling, pp 201–219. Springer
Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2(1):56–67
Lundberg SM, Lee SI (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. Curran Associates Inc., Red Hook, NY, USA

Maggi FM, Di Francescomarino C, Dumas M, Ghidini C (2014) Predictive monitoring of business processes. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 8484 LNCS, pp 457–472https://doi.org/10.1007/978-3-319-07881-6_31
Maita ARC, Fantinato M, Peres SM, Maggi FM (2025) Interpretability in predictive process monitoring using process models: an expert evaluation of the visinter4ppm framework. KI-Künstliche Intelligenz, pp 1–21
Majlatow M, Shakil FA, Emrich A, Mehdiyev N (2025) Uncertainty-aware predictive process monitoring in healthcare: explainable insights into probability calibration for conformal prediction. Appl Sci 15(14):7925
Majumdar R, Takami K, Ogata H (2023) Learning with explainable AI-recommendations at school: extracting patterns of self-directed learning from learning logs. In: 2023 IEEE international conference on advanced learning technologies (ICALT), pp 245–249. IEEE
Malashin IP, Masich IS, Ageev DA, Evsyukov DA, Gantimurov AP, Nelyub VA, Borodulin AS, Tynchenko VS (2025) Automated detection of deviations in bankruptcy processes using process mining. IEEE Access
Malioutov DM, Varshney KR, Emad A, Dash S (2017) Learning interpretable classification rules with boolean compressed sensing. Transparent data mining for big and small data, 95–121
Mannhardt F, Koschmider A, Baracaldo N, Weidlich M, Michael J (2019) Privacy-preserving process mining: differential privacy for event logs. Bus Inf Syst Eng 61:595–614
Mannhardt F (2016) Sepsis cases - event log. Eindhoven University of Technology
Mathew DE, Ebem DU, Ikegwu AC, Ukeoma PE, Dibiaezue NF (2025) Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Process Lett 57(1):16
Mayer L, Mehdiyev N, Fettke P (2021) Manufacturing execution systems driven process analytics: a case study from individual manufacturing. Procedia CIRP 97:284–289
McDermid J, Porter Z, Habli I (2021) Ai explainability: the technical and ethical dimensions. Math Phys Eng Sci Philos Trans
Mehdiyev N, Krumeich J, Enke D, Werth D, Loos P (2015) Determination of rule patterns in complex event processing using machine learning techniques. Procedia Comput Sci 61:395–401
Mehdiyev N, Majlatow M, Fettke P (2024) Counterfactual explanations in the big picture: an approach for process prediction-driven job-shop scheduling optimization. Cogn Comput 16(5):2674–2700
Mehdiyev N, Majlatow M, Fettke P (2025) Augmenting post-hoc explanations for predictive process monitoring with uncertainty quantification via conformalized monte carlo dropout. Data Knowl Eng 156:102402
Mehdiyev N, Majlatow M, Fettke P (2025) Integrating permutation feature importance with conformal prediction for robust explainable artificial intelligence in predictive process monitoring. Eng Appl Artif Intell 149:110363
Mehdiyev N, Mayer L, Lahann J, Fettke P (2022) Deep learning-based clustering of processes and their visual exploration: an industry 4.0 use case for small, medium-sized enterprises. Expert Syst, 13139
Mehdiyev N, Majlatow M, Fettke P (2023) Communicating uncertainty in machine learning explanations: a visualization analytics approach for predictive process monitoring. arXiv preprint arXiv:2304.05736
Mehdiyev N, Majlatow M, Fettke P (2023) Quantifying and explaining machine learning uncertainty in predictive process monitoring: an operations research perspective. arXiv preprint arXiv:2304.06412
Mehdiyev N, Fettke P (2021) Local post-hoc explanations for predictive process monitoring in manufacturing. ECIS 2020 Research Papers
Mehdiyev N, Fettke P (2020) Explainable artificial intelligence for process mining: a general overview and application of a novel local explanation approach for predictive process monitoring. Stud Comput Intell 937, 1–28https://doi.org/10.1007/978-3-030-64949-4_1. arXiv:2009.02098
Mehdiyev N, Fettke P (2020) Prescriptive process analytics with deep learning and explainable artificial intelligence. In: European conference on information systems, pp 1–17
Mehdiyev N, Houy C, Gutermuth O, Mayer L, Fettke P (2021) Explainable artificial intelligence (xai) supporting public administration processes–on the potential of xai in tax audit processes. Innovation through information systems: volume i: a collection of latest research on domain issues, pp 413–428. Springer
Messalas A, Kanellopoulos Y, Makris C (2019) Model-agnostic interpretability with Shapley values. 2019 10th international conference on information, intelligence, systems and applications (IISA), pp 1–7
Miri N, Jalali A (2024) Uncovering patterns in object-centric process mining: an approach using drill-down and roll-up techniques. In: International conference on information integration and web intelligence, pp 49–54. Springer
Mitchell TM (1997) Machine learning. McGraw-Hill series in computer science, p 414. WCB/McGraw-Hill, Boston, MA

Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans Interact Intell Syst 11(3–4):1–45
Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR (2019) Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, 193–209
Montoya F, Astudillo H (2023) Causal graph: Interpretation of causal relationships in temporary deviations of business processes. In: 2023 XLIX Latin American computer conference (CLEI), pp 1–5. IEEE
Montoya F, Berríos E, Díaz D, Astudillo H (2023) Counterfactual explanability: An application of causal inference in a financial sector delivery business process. In: 2023 42nd IEEE international conference of the chilean computer science society (SCCC), pp 1–10. IEEE
Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT* '20, pp 607–617. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3351095.3372850
Mozolewski M, Bobek S, Ribeiro RP, Nalepa GJ, Gama J (2024) Towards evaluation of explainable artificial intelligence in streaming data. In: World conference on explainable artificial intelligence, pp 145–168. Springer
Myers D, Suriadi S, Radke K, Foo E (2018) Anomaly detection for industrial control systems using process mining. Computers & Security 78:103–125
Márquez-Chamorro AE, Resinas M, Ruiz-Cortés A (2017) Predictive monitoring of business processes: a survey. IEEE Trans Serv Comput 11(6):962–977
Márquez-Chamorro AE, Resinas M, Ruiz-Cortés A, Toro M (2017) Run-time prediction of business process indicators using evolutionary decision rules. Expert Syst Appl 87:1–14
Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, Schlötterer J, Keulen M, Seifert C (2023) From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput Surv 55(13s):1–42
Ouyang C, Sindhgatta R, Moreira C (2021) Explainable AI enabled inspection of business process prediction models arXiv:2107.09767
Padella A, de Leoni M, Dogan O, Galanti R (2022) Explainable process prescriptive analytics. In: 2022 4th international conference on process mining (ICPM), pp 16–23. IEEE
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE et al (2021) The prisma 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88:105906
Park G, Liss L, van der Aalst WM (2024) Learning recommendations from educational event data in higher education. J Intell Inf Syst, 1pp –20
Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: justifying decisions and pointing to the evidence. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8779–8788
Pasquadibisceglie V, Appice A, Castellano G, Malerba D (2023) Jarvis: joining adversarial training with vision transformers in next-activity prediction. IEEE Trans Serv Comput 17(4):1593–1606
Pasquadibisceglie V, Appice A, Ieva G, Malerba D (2024) Tsunami-an explainable ppm approach for customer churn prediction in evolving retail data environments. J Intell Inf Syst 62(3):705–733
Pasquadibisceglie V, Appice A, Malerba D (2024) Lupin: a LLM approach for activity suffix prediction in business process event logs. In: 2024 6th international conference on process mining (ICPM), pp 1–8. IEEE
Pasquadibisceglie V, Scaringi R, Appice A, Castellano G, Malerba D (2024) Prophet: explainable predictive process monitoring with heterogeneous graph neural networks. IEEE Trans Serv Comput
Pasquadibisceglie V, Castellano G, Appice A, Malerba D (2021) Fox: a neuro-fuzzy model for process outcome prediction and explanation. In: 2021 3rd international conference on process mining (ICPM), 112–119. IEEE
Pauwels S, Calders T (2019) Detecting anomalies in hybrid business process logs. ACM SIGAPP Appl Comput Rev 19(2):18–30
Pauwels S, Calders T (2019) An anomaly detection technique for business processes based on extended dynamic bayesian networks. In: Proceedings of the 34th ACM/SIGAPP symposium on applied computing, pp 494–501
Peeva V, Porsil M, van der Aalst WM (2024) Object-centric local process models. In: International conference on process mining, pp 376–388. Springer
Petsis S, Karamanou A, Kalampokis E, Tarabanis K (2022) Forecasting and explaining emergency department visits in a public hospital. J Intell Inf Syst. https://doi.org/10.1007/s10844-022-00716-6
Pfeiffer P, Rombach A, Majlatow M, Mehdiyev N (2025) From theory to practice: real-world use cases on trustworthy LLM-driven process modeling, prediction and automation. arXiv preprint arXiv:2506.03801

Polato M, Sperduti A, Burattin A, Leoni Md (2018) Time and activity sequence prediction of business process instances. Computing 1100(9):1005–1031
Polato M (2017) Dataset belonging to the help desk log of an italian company. University of Padova
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
Porouhan P (2024) Data-driven strategies for cardiovascular medication management: A dual approach using decision tree modeling and process mining. In: 2024 22nd international conference on ict and knowledge engineering (ICT &KE), pp 1–20. IEEE
Porsil VPM, van der Aalst WM (2025) Object-centric local process models. In: Process mining workshops: ICPM 2024 international workshops, Lyngby, Denmark, October 14–18, 2024, Revised Selected Papers, vol 533, p 376. Springer Nature
Prasidis I, Theodoropoulos N-P, Bousdekis A, Theodoropoulou G, Miaoulis G (2021) Handling uncertainty in predictive business process monitoring with bayesian networks. In: 2021 12th international conference on information, intelligence, systems & applications (IISA), pp 1–8. IEEE
Qafari MS, Van der Aalst W (2019) Fairness-aware process mining. In: On the move to meaningful internet systems: OTM 2019 conferences: confederated international conferences: CoopIS, ODBASE, C &TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings, pp 182–192. Springer
Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco. http://portal.acm.org/citation.cfm?id=152181
Rauch S, Frey CM, Zellner L, Seidl T (2024) Process-aware bayesian networks for sequential event log queries. In: 2024 6th international conference on process mining (ICPM), pp 161–168. IEEE
Rebmann A, Schmidt FD, Glavaš G, Aa H (2025) On the potential of large language models to solve semantics-aware process mining tasks. Process Sci 2(1):10
Rehse JR, Dadashnia S, Fettke P (2018) Business process management for Industry 4.0 - Three application cases in the DFKI-Smart-Lego-Factory. Inf Technol?: IT 60(3):133–141. https://doi.org/10.1515/itit-2018-0006
Rehse J-R, Mehdiyev N, Fettke P (2019) Towards explainable process predictions for industry 4.0 in the DFKI-Smart-Lego-Factory. KI - Künstliche Intell 33(2), 181–187https://doi.org/10.1007/s13218-019-00586-1
Reinkemeyer L (2020) Process mining in action. Process mining in action principles, use cases and outloook, 2
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, pp 1135–1144. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2939672.2939778
Rivera-Lazo G, Astudillo H, Ñanculef R (2023) Attention mechanisms in process mining: a systematic literature review. In: 2023 XLIX Latin American computer conference (CLEI), pp 1–10. IEEE
Rizzi W, Di Francescomarino C, Maggi FM (2020) Explainability in predictive process monitoring: When understanding helps improving. In: International conference on business process management, pp 141–158. Springer
Rizzi W, Di Francescomarino C, Ghidini C, Magg, FM (2024) Nirdizati: an advanced predictive process monitoring toolkit. J Intell Inf Syst, pp 1–33
Rosenfeld A (2021) Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th international conference on autonomous agents and multiagent systems, pp 45–50
Rutkowski L, Jaworski M, Pietruczuk L, Duda P (2014) The cart decision tree for mining data streams. Inf Sci 266:1–15
Saha A, Agarwal P, Ghosh S, Gantayat N, Sindhgatta R (2024) Towards business process observability. In: Proceedings of the 7th Joint international conference on data science & management of data (11th ACM IKDD CODS and 29th COMAD), pp 257–265
Saini V, Singh P, Sureka A (2020) Control-flow based anomaly detection in the bug-fixing process of open-source projects. In: Proceedings of the 13th innovations in software engineering conference (formerly Known as India Software Engineering Conference), pp 1–11
Salama A, Linke A, Rocha IP, Binnig C (2019) XAI: A middleware for scalable AI. In: Proceedings of the 8th international conference on data science, technology and applications. International conference on data science, technology and applications (DATA-2019), pp 109–120
Sarno R, Sinaga F, Sungkono KR (2020) Anomaly detection in business processes using process mining and fuzzy association rule learning. J Big Data 7(1):5
Savickas T, Vasilecas O (2018) Belief network discovery from event logs for business process analysis. Comput Ind 100:258–266
Savickas T, Vasilecas O (2014) Business process event log transformation into bayesian belief network

Sebin BL, Taskin NA, Mehdiyev N (2024) Exploring the intersection of large language models (LLMS) and explainable AI (XAI): A systematic literature review. In: Mediterranean conference on information systems (MCIS) 2024. https://aisel.aisnet.org/mcis2024/89, 89. AIS

Senderovich A, Di Francescomarino C, Ghidini C, Jorbina K, Maggi FM (2017) Intra and inter-case features in predictive process monitoring: a tale of two dimensions. In: Business Process management: 15th international conference, BPM 2017, Barcelona, Spain, September 10–15, 2017, Proceedings 15, 306–323. Springer

Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games II. Princeton University Press, Princeton, pp 307–317

Shneiderman B (2022) Human-centered AI. Oxford University Press, Oxford

Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: International conference on machine learning, pp 3145–3153. PMLR

Sindhgatta R, Ouyang C, Moreira C (2020) Exploring interpretability for predictive process analytics. In: International conference on service-oriented computing, pp 439–447. Springer

Sindhgatta R, Moreira C, Ouyang C, Barros A (2020) Exploring interpretable predictive models for business processes. In: Business process management: 18th international conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18, pp 257–272. Springer

Steeman W (2013) Bpi challenge 2013, incidents. Ghent University

Stevens A, De Smedt J (2022) Explainable predictive process monitoring: evaluation metrics and guidelines for process outcome prediction https://doi.org/10.48550/arxiv.2203.16073

Stevens A, De Smedt J, Peeperkorn J, De Weerdt J (2022) Assessing the robustness in predictive process monitoring through adversarial attacks. In: 2022 4th international conference on process mining (ICPM), pp 56–63. IEEE

Stevens A, De Smedt J, Peeperkorn J (2022) Quantifying explainability in outcome-oriented predictive process monitoring. Lecture notes in business information processing 433 LNBIP, pp 194–206. https://doi.org/10.1007/978-3-030-98581-3_15/FIGURES/3

Stierle M, Brunk J, Weinzierl S, Zilker S, Matzner M, Becker J (2021) Bringing light into the darkness - a systematic literature review on explainable predictive business process monitoring techniques. ECIS 2021 Research-in-Progress Papers. **8**

Su R, Guo H, Wang W (2024) Elastic online deep learning for dynamic streaming data. Inf Sci 676:120799

Tama BA, Comuzzi M, Ko J (2020) An empirical investigation of different classifiers, encoding, and ensemble schemes for next event prediction using business process event logs. ACM Trans Intell Syst Technol 11(6):1–34

Tavares GM, Oyamada RS, Junior SB, Ceravolo P (2023) Trace encoding in process mining: a survey and benchmarking. Eng Appl Artif Intell 126:107028

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

Teinemaa I, Dumas M, Magg, FM, Di Francescomarino C (2016) Predictive business process monitoring with structured and unstructured data. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)9850 LNCS,401–417. https://doi.org/10.1007/978-3-319-45348-4_23

Toh JX, Wong KJ, Agarwal S, Zhang X, Lu JJ (2022) Improving operation efficiency through predicting credit card application turnaround time with index-based encoding. Companion proceedings of the web conference 2022:615–620

Trescato I, Tavazzi E, Vettoretti M, Gatta R, Vasta R, Chiò A, Di Camillo B (2024) Dynamite: integrating archetypal analysis and process mining for interpretable disease progression modelling. IEEE J Biomed Health Inf

Tripathi A, Jadhav S, Singh S, Nandan SK, Vyas R, Vyas O (2024) Prom-ex: an explainable framework for anomaly detection in process mining using large language models. In: 2024 IEEE 8th international conference on information and communication technology (CICT), 1–6. IEEE

Turchi T, Malizia A, Paternò F, Borsci S, Chamberlain, A (2024) Adaptive xai: Towards intelligent interfaces for tailored AI explanations. In: Companion proceedings of the 29th international conference on intelligent user interfaces, 119–121

Unuvar M, Lakshmanan GT, Doganata YN (2016) Leveraging path information to generate predictions for parallel business processes. Knowl Inf Syst 47(2):433–461. https://doi.org/10.1007/S10115-015-0842-7

Velmurugan M, Ouyang C, Moreira C, Sindhgatta R (2021) Evaluating fidelity of explainable methods for predictive process analytics. In: International conference on advanced information systems engineering, pp 64–72. Springer

Velmurugan M, Ouyang C, Moreira C, Sindhgatta R (2021) Evaluating stability of post-hoc explanations for business process predictions. In: International conference on service-oriented computing, pp 49–64. Springer

Verenich I, Dumas M, La Rosa M, Nguyen H (2019) Predicting process performance: a white-box approach based on process models. J Softw Evolu Process 31(6):2170
Verenich I, Nguyen H, La Rosa M, Dumas M (2017) White-box prediction of process performance indicators via flow analysis. In: Proceedings of the 2017 international conference on software and system process, pp 85–94
Verenich I, Dumas M, La Rosa M, Maggi FM, Di Francescomarino C (2016) Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring. In: Business process management workshops: BPM 2015, 13th international workshops, Innsbruck, Austria, August 31–September 3, 2015, Revised Papers 13, pp 218–229. Springer
Visani G, Bagli E, Chesani F, Poluzzi A, Capuzzo D (2021) Statistical stability indices for lime: obtaining reliable explanations for machine learning models. J Oper Res Soc 73:1–11. https://doi.org/10.1080/01605682.2020.1865846
Vogt MW, van der Putten P, Reijers HA (2024) Providing domain knowledge for process mining with ReWOO-based agents. In: International conference on process mining, pp 663–676 . Springer
Völzer H, Zerbato F, Sulzer T, Weber B (2023) A fresh approach to analyze process outcomes. In: 2023 5th International conference on process mining (ICPM), pp 97–104. IEEE
Wang Q, Anikina T, Feldhus N, Ostermann S, Möller S (2024) Coxql: a dataset for parsing explanation requests in conversational XAI systems. arXiv preprint arXiv:2406.08101
Weinzierl S, Zilker S, Brunk J, Revoredo K, Matzner M, Becker J (2020) XNAP: Making LSTM-based next activity predictions explainable by using LRP. Lecture notes in business information processing 397, 129–141 arXiv:2008.07993. https://doi.org/10.1007/978-3-030-66498-5_10/COVER
Weytjens H, De Weerdt J (2022) Learning uncertainty with artificial neural networks for predictive process monitoring. Appl Soft Comput 125:109134
Wickramanayake B, Ouyang C, Moreira C, Xu Y (2022) Generating purpose-driven explanations: the case of process predictive model inspection. Lecture Notes Bus Inf Process 452:120–129. https://doi.org/10.1007/978-3-031-07481-3_14/FIGURES/3
Wickramanayake B, He Z, Ouyang C, Moreira C, Xu Y, Sindhgatta R (2022) Building interpretable models for business process prediction using shared and specialised attention mechanisms. Knowl Based Syst **248** https://doi.org/10.1016/j.knosys.2022.108773. arXiv:2109.01419
Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco
van Zelst SJ, Sani MF, Ostovar A, Conforti R, La Rosa M (2020) Detection and removal of infrequent behavior from event streams of business processes. Inf Syst 90:101451
Zhang T, Zhang M, Low WY, Yang XJ, Li BA (2025) Conversational explanations: discussing explainable AI with non-AI experts. In: Proceedings of the 30th international conference on intelligent user interfaces, pp 409–424
Zilker S, Weinzierl S, Zschech P, Kraus M, Matzner M (2023) Best of both worlds: combining predictive power with interpretable and explainable results for patient pathway prediction
Van Der Aalst W (2012) Process mining. Commun ACM 55(8):76–83
van der Aalst WM (2023) Object-centric process mining: unraveling the fabric of real processes. Mathematics 11(12):2691
van Dongen B (2011) Real-life event logs - hospital log. Eindhoven University of Technology
van Dongen B (2012) Bpi challenge 2012. Eindhoven University of Technology
van Dongen B (2014) Bpi challenge 2014: change details. Rabobank Nederland
van Dongen BFB (2015) Bpi challenge 2015. Eindhoven University of Technology
van Dongen B (2017) Bpi challenge 2017. Eindhoven University of Technology
van Dongen B, Borchert FF (2018) Bpi challenge 2018. Eindhoven University of Technology
van Dongen B (2019) BPI Challenge 2019. 4TU.Centre for Research Data. https://doi.org/10.4121/UUID:D06AFF4B-79F0-45E6-8EC8-E19730C248F1. https://data.4tu.nl/articles/_/12715853/1
van Dongen B (2020) BPI Challenge 2020. 4TU.Centre for Research Data. https://doi.org/10.4121/UUID:52FB97D4-4588-43C9-9D04-3604D4613B51. https://data.4tu.nl/collections/_/5065541/1