Process Science

RESEARCH Open Access

# Explainable predictive process monitoring: a user evaluation

Williams Rizzi[1†], Marco Comuzzi[2*†], Chiara Di Francescomarino[3†], Chiara Ghidini[4†], Suhwan Lee[5†], Fabrizio Maria Maggi[4†] and Alexander Nolte[6,7†]

[†]Williams Rizzi, Marco Comuzzi, Chiara Di Francescomarino, Chiara Ghidini, Suhwan Lee, Fabrizio Maria Maggi, and Alexander Nolte contributed equally to this work.

*Correspondence: mcomuzzi@unist.ac.kr

[1] Fondazione Bruno Kessler, Trento 38123, Italy
[2] Ulsan National Institute of Science and Technology, Ulsan 44919, Korea
[3] Universita' di Trento, Povo 38123, Italy
[4] Free University of Bozen-Bolzano, Bolzano 39100, Italy
[5] Utrecht University, Utrecht 3584, Netherlands
[6] Eindhoven University of Technology, Eindhoven 5612, Netherlands
[7] Carnegie Mellon University, Pittsburgh 15213, PA, USA

**Abstract**

Explainability is motivated by the lack of transparency of black-box machine learning approaches, which do not foster trust and acceptance of machine learning algorithms. This also happens in the predictive process monitoring field, where predictions, obtained by applying machine learning techniques, need to be explained to users, so as to gain their trust and acceptance. In this work, we carry on a user evaluation on explanation approaches for predictive process monitoring aiming at investigating whether and how the explanations provided (i) are understandable; (ii) are useful in decision making tasks; (iii) can be further improved for process analysts with different predictive process monitoring expertise levels. The results of the user evaluation show that, although explanation plots are overall understandable and useful for decision making tasks for business process management users — with and without experience in predictive process monitoring — differences exist in the comprehension and usage of different plots, as well as in the way users with different predictive process monitoring expertise understand and use them.

**Keywords:** Predictive process monitoring, Process mining, Explainable artificial intelligence, Explanation plots, Qualitative observational study

## Introduction

Predictive Process Monitoring (PPM) is a branch of Process Mining that aims at providing predictions on the future of an ongoing process execution by leveraging past historical execution traces. An increasing number of PPM approaches leverage machine and deep learning techniques in order to learn from past historical execution traces the outcome of an ongoing process execution, the time remaining till the end of an ongoing execution, or the next activities that will be performed.

In many of these applications, users are asked to trust a Machine Learning (ML) model that supports them in making decisions. If ML algorithms lack explainability, users may take incorrect actions due to the difficulty in understanding the outputs provided by the algorithms, or they may avoid decisions altogether because they do not trust the algorithms. Therefore, cognitive insights are needed in addition to ML-based results to ensure reliable decision-making. In recent years, eXplainable Artificial Intelligence (XAI) has been investigating the problem of explaining ML models so as to foster trust and acceptance of these

models. Some of the recent XAI post-hoc approaches have also been applied and investigated in the field of PPM (Galanti et al. 2020; Velmurugan et al. 2021b; Rizzi et al. 2020) in order to make a predictor returning, besides predictions, also prediction explanations (for instance in the form of *explanation plots*). None of these works, however, has so far investigated whether users actually understand and use these plots when making decisions.

With this work, we aim at providing a first investigation with users on whether explanation plots at *event*, *trace* and *event log* level in PPM are understood by users and can support users in making decisions. Note that we do not deal here with the evaluation or comparison of (prediction and XAI) techniques. Although this and other aspects can be considered, in this work, we evaluate the effectiveness of the plots for users independently of how the plots are generated. In particular, we focus on the following research questions:

| |
|---|
| **RQ**$_1$. *How do users make sense of explanation plots in PPM?* |
| **RQ**$_2$. *How can explanation plots support users in decision making tasks in PPM?* |
| **RQ**$_3$. *How can PPM explanation plots be improved?* |

In order to answer these research questions, we carried out a qualitative evaluation with users working in the field of BPM and with different levels of PPM expertise. We provided them with a problem, as well as with some predictions and explanation plots at *event*, *trace* and *event log* level based on the (post-hoc) explainers most frequently used in PPM: LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017) and ICE (Goldstein et al. 2015). Specifically, we used the explanation plots collected in Nirdizati (Rizzi et al. 2019), a dedicated PPM tool.

We asked the users (i) some questions for checking their comprehension of the plots, (ii) to make a decision by leveraging the plots, and (iii) how they would change the explanation plots to make them more useful and understandable. The results of the user evaluation show that, although explanation plots are overall understandable and useful for decision making tasks for all participants, differences exist in the comprehension and usage of different plots, as well as in the way users with different PPM expertise understand and use them. In particular, the study reveals that, while on the one side for PPM experts understanding the explanation plots is easier than for participants without PPM experience, on the other hand, using the plots for making decisions is easier for participants who do not have PPM experience. PPM experts, indeed, tend to be more conservative in their decisions, feeling that explanation plots, mainly showing correlations and not causalities, do not provide them with enough evidence to make specific recommendations. Moreover, the evaluation carried out revealed interesting suggestions and desiderata for explanation plots, such as the need for interactive elements in the user interface as well as for what-if analysis support.

## Background

In this section, we report the main concepts discussed in this paper, i.e., explainability approaches (Explainability approaches section), and explainability applied to the PPM field (Explainability in predictive process monitoring section).

### Explainability approaches

The lack of transparency of the black-box ML approaches, which do not foster trust and acceptance of ML algorithms, motivates the need of explainability approaches. In the literature, there are two main groups of techniques used to develop explainable systems, a.k.a. *explainers*: post-hoc and ante-hoc techniques. Post-hoc techniques allow models to be trained as usual, with explainability only being incorporated at testing time. Ante-hoc techniques entail integrating explainability into a model since the training phase. In this work, we mainly focus on post-hoc explainers since we want to use these instruments to improve the state-of-the-art approaches for PPM available in the literature without altering them, but instead building new solutions on top of them.

An example of post-hoc explainer are the Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al. 2016). LIME learns an interpretable model locally around the prediction and explains the predictive models providing individual explanations for individual predictions. The explanations are generated by approximating the underlying model with an interpretable one, learned using perturbations of the original instance. In particular, each feature is assigned with an *importance value* that represents the influence of that feature on a particular prediction. Another post-hoc explainer are the SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). SHAP is a game-theoretic approach explaining the output of any ML model. It connects optimal credit allocation with local explanations using the classic Shapley Values (Shapley 2016) from game theory and their related extensions. SHAP provides local explanations based on the outcome of other explainers. It identifies a new class of additive feature importance measures, for which there exists a unique solution with a set of desirable properties. Other post-hoc explainers (Friedman 2001) show the marginal effect of some features using partial dependence plots. In Goldstein et al. (2015), the authors refine the partial dependence plots definition providing the visualization of functional dependencies for individual observations through Individual Conditional Expectation (ICE) plots. Ribeiro et al. (2018) also introduces a model-agnostic approach to explain complex systems using rules called anchors. Anchors, when available, are local conditions that can explain predictions.

### Explainability in predictive process monitoring

PPM (Di Francescomarino 2019) is a branch of process mining (van der Aalst et al. 2011) that aims at exploiting event logs of past historical process execution traces to predict how ongoing (uncompleted) process executions will unfold up to their completion. Typical examples of predictions on the future of an execution trace relate to its completion time, to the fulfilment or violation of a certain predicate, or to the next sequence of activities that will be executed.

PPM approaches are typically characterized by two phases: a *training phase*, in which a predictive model is learned from historical traces, and a *prediction phase*, in which the predictive model is queried for predicting the future developments of an ongoing case. Recent works addressing PPM challenges mainly leverage machine learning or statistical models, i.e., implicit models of the process rather than explicit

process models. The interested reader is referred to Márquez-Chamorro et al. (2018); Di Francescomarino et al. (2018) for recent surveys covering several PPM concepts and techniques.

Recently, explainability approaches have been applied and investigated in the field of PPM (Harl et al. 2020; Sindhgatta et al. 2020; Weinzierl et al. 2020b; Galanti et al. 2020; Velmurugan et al. 2021b; Rizzi et al. 2020).

Some works focus on applying model-specific explainability approaches to provide explanations for predictions obtained through neural network predictive models, e.g., gated graph neural networks (Harl et al. 2020), attention-based LSTM models (Sindhgatta et al. 2020), layer-wise relevance propagation to LSTM (Weinzierl et al. 2020b).

Other works focus on generic or model-agnostic post-hoc explanation approaches (Galanti et al. 2020; Velmurugan et al. 2021b; Rizzi et al. 2020). For example, in Rizzi et al. (2020), explanations are used to identify the features leading to wrong predictions in order to improve the accuracy of the predictive model. In Galanti et al. (2020, 2021), Shapley values (Shapley 2016) are leveraged for providing users with explanations in the form of tables explaining the predictions related to a specific ongoing process execution. In addition, they use plots relying on heatmaps to specify for each feature and at each point in time the impact of the feature on a prediction.

In Nirdizati (Rizzi et al. 2019), different types of explanation plots are implemented to provide explanations in the context of binary classifications. In particular, explanation plots at *event*, *trace* and *event log* levels are provided. For event-based explanations, LIME (Ribeiro et al. 2016) and SHAP (Lundberg and Lee 2017) explanation plots, the post-hoc explainers more frequently used in PPM, are adapted to the PPM scenario to measure the importance of each feature for a prediction provided at a specific trace prefix. In order to provide trace-based explanations, a temporal stability (Velmurugan et al. 2021a; Teinemaa et al. 2018) plot is used. This plot allows users to visualize the importance of each feature at different trace prefixes, thus revealing also how stable the importance of the features is for predictions returned at different prefix lengths. Finally, ICE (Goldstein et al. 2015) explanation plots have been implemented in Nirdizati to provide log-based explanations. This type of plot reports, for a specific feature, information about the average value of the predicted label (between 0 and 1 with 0 meaning *false* and 1 meaning *true*) for the different feature values, as well as the number of traces in the event log containing each value. In this paper, we leverage the PPM explanation plots implemented in Nirdizati for carrying out the user evaluation.

### XAI user studies classification

Due to the rapid growth of the field of XAI and to the key role of users for XAI systems, several empirical studies with human subjects have been conducted to investigate and evaluate these systems.

We can classify evaluations for these XAI systems based on the levels of tests used (test of *satisfaction*, test of *comprehension* or test of *performance*), as well as the type of tasks the participants carried out (*verification*, *forced choice*, *forward simulation*, *counterfactual simulation*, *system usage* or *annotation*) (Chromik and Schuessler 2020). An evaluation can indeed be focused on evaluating the satisfaction of the users with the explanations or their subjective assessment of the understanding of the system; their

comprehension of the system in terms of the mental model they have of it; or the overall performance in terms of human-XAI system performance.

In addition, in verification tasks, participants are asked about their satisfaction with the explanations; in forced choice tasks, participants are asked to choose among different explanations; in forward simulation tasks, participants are provided with data and explanations and asked to predict the system's output; in counterfactual simulation tasks, participants should predict what input changes are required for getting an alternative output from the system; in system usage tasks, participants are asked to use the system for its original purpose, e.g., for decision making tasks; annotation tasks require participants to provide an explanation based on the input provided to the system and the produced output.

Some of the works in the literature focus only on the evaluation of the *satisfaction* level. For instance, in Krause et al. (2016), an interactive interface (Prospector) is proposed to the users to understand how the features of a dataset affect the prediction of a model overall. A team of 5 data scientists was asked to interact with this tool for 4 months to debug a set of models and, at the end of the experiment, the data scientists were interviewed on whether they felt that the provided support was beneficial for their work. In Spinner et al. (2020), *explAIner*, a visual analytic system for interactive XAI, is evaluated by 9 participants with different levels of expertise. The users were asked to use the system and their feedback was then collected. Results reveal that the system is intuitive and helpful in the daily analysis routine.

Other works focus not only on the evaluation of the satisfaction level, but also on the users' *comprehension*. For instance, in Lundberg and Lee (2017), users were asked to carry out an *annotation* task, i.e., they were asked to provide explanations on simple models. The human explanations were then compared with LIME, DeepLIFT and SHAP explanations and their consistency with human intuition evaluated. The results show that the SHAP explanations are closer to human ones than the ones of the other approaches.

Besides satisfaction and comprehension, a last group of works focuses instead on evaluating the *performance* of the user (and the system). For instance in Wang and Yin (2021), users were asked to carry out *forward-simulation* tasks. The work compares and evaluates four common model-agnostic XAI methods on these tasks. The results show that the effect of AI explanations depends on the level of domain expertise of the users. In Hase and Bansal (2020), *counterfactual simulation* tasks were given to participants, as well as *forward-simulation tasks*. In this work, the authors propose a user study to evaluate the *simulatability* of different explanation approaches, where an approach is simulatable if a user is able to predict its behavior on new instances.

Finally, in some works, participants are asked to carry out *system usage* and, more specifically, decision-making tasks. For instance, in Malhi et al. (2020), 65 participants were asked to carry out a decision-making task to evaluate the human understanding of the behavior of explainable agents. Three user groups were considered: a group was provided with an agent without explanations and the other two with explanations generated by LIME and SHAP, respectively. The results show that notable (though not statistically significant) differences exist between the groups without and with explanations in terms of bias in human decision making. Decision-making tasks are also used in Krause et al.

(2018) to evaluate the effect of using aggregated rather than single data point explanations. The results show that using aggregated explanations has a positive impact on the detection of biases in data.

As far as XAI in PPM is concerned, Elkhawaga et al. (2024) have defined quantitative measures to assess the extent to which XAI methods in PPM reflect the facts learned during the model development process. Galanti et al. (2023) have provided the only evaluation of how users understand and use explanation plots in the field of PPM. This was conducted as a post-hoc questionnaire evaluating the user satisfaction and comprehension. Users were not observed while using the plots. In this work, we evaluate explanation plots at *satisfaction*, *comprehension* and *performance* level using forward simulation tasks.

## Methodology

To study how individuals make sense of explanation plots in PPM ($\mathbf{RQ}_1$), how they use them for PPM decision making tasks ($\mathbf{RQ}_2$) and identify characteristics and aspects to improve them ($\mathbf{RQ}_3$), we conducted a qualitative observational study. This approach is suitable because it allows us to draw insights into "how" individuals interact with prediction explanations in the context of PPM (Lazar et al. 2017). Our study aims at demonstrating that explainability instruments in the considered tasks improve the comprehensibility of the predictions and support users in making decisions.

Here, we first provide an overview of our study setting (Setting section) before discussing our data collection strategy (Data collection section) and analysis methodology (Analysis procedure section).

### Setting

For our study, we selected individuals with varying degrees of general, research and real life experience as business process management researchers and process analysts. We focused on these individuals because decisions regarding how processes are conducted are commonly made by individuals with extensive expertise in Business Process Management (BPM). We selected a total of eight individuals as participants.

This sample size is comparable to other works which focused on studying how individuals understood, utilized and interacted with graphical process models (Nolte and Prilla 2013), a graphical user interface (Nielsen and Landauer 1993), and an audio component that was embedded in a graphical user interface (van der Aa et al. 2020). While the subject under study of these works is different compared to our work, the study goals are similar in that they focus on how users understand what is presented to them, decide how to use what is presented, decide whether or not they perceive it to be useful, and discuss their perception of how it can be improved. Moreover, when analyzing the collected data we reached a point of saturation (Mason et al. 2010) in that the concepts that we identified remained stable.

Out of the eight individuals, we selected four that had expertise in PPM in addition to their expertise in BPM. We made this differentiation since it can be expected that knowledge related to PPM would affect whether and how individuals understand explanation plots and how they use them for decision making. In particular, we classified them as process analysts (PPM) and process management (BPM) researchers,

according to how they are considered by knowing their works and looking at their profiles. Moreover, we asked them to declare the years of experience they had in both the disciplines, by distinguishing between general experience, research experience and practical (especially implementation) experience. Table 1 summarizes the answers provided by the subjects, revealing varying degrees of experience in the two fields among them. Within these two groups of four participants, we selected individuals from different domains to foster the applicability of our findings beyond the confines of a specific use case.

For our study, we created two separate scenarios: one from the medical domain (Domain M (medical) section) and one from the financial domain (Domain F (financial) section). We selected these domains because they are common domains where PPM techniques are applied (van Dongen 2011, 2012, 2017). While the former is inspired by the realistic *X Ray* process model described in Pesic (2008) and leveraged in a number of works (Maggi et al. 2018; Di Ciccio et al. 2015), the latter is inspired by the real scenario investigated in Galanti et al. (2020). We assigned individuals with varying degrees of expertise related to BPM and PPM (B = BPM expert and P = PPM expert) to both scenarios to mitigate potential bias (M = medical and F = financial) in Table 1). In the following, we describe the two scenarios we used (Domain M (medical) and Domain F (financial) sections) before elaborating on the plots we used (Plots section) and the tasks we asked participants to complete within the two scenarios (Tasks section).

#### *Domain M (medical)*

We consider a process pertaining to the treatment of patients with fractures, for which we want to predict patients who will recover soon or late. Every process execution starts with examining the patient. If an X-ray is performed, then the X-ray risk must be assessed before it. Treatments reposition, cast application and surgery require that an X-ray is performed before. If a surgery is performed, then a rehabilitation must be prescribed eventually after it. Finally, after every cast application, a cast removal must be performed. Each process execution refers to a patient whose age is known. Moreover, for each patient, the type of treatment prescribed is also known and whether or not the patient should carry on a rehabilitation after the treatment. We predict whether the patient will recover quickly (1) or not (0).

**Table 1** Participants and their years of experience related to BPM and PPM. The experience in the field is characterised in terms of general experience, research experience, and practical usage experience

| Expertise | | BM1 | BM2 | BF1 | BF2 | PM1 | PM2 | PF1 | PF2 |
|---|---|---|---|---|---|---|---|---|---|
| BPM | General | 12 | 4 | 15 | 13 | 14 | 10 | 12 | 7 |
| | Research | 12 | 2 | 15 | 10 | 14 | 8 | 12 | 7 |
| | Practical | 8 | 2 | 10 | 13 | 14 | 0 | 12 | 0 |
| PPM | General | 2 | 0 | 2 | 3 | 7 | 2 | 8 | 7 |
| | Research | 2 | 0 | 2 | 3 | 7 | 2 | 8 | 7 |
| | Practical | 0 | 0 | 0 | 1 | 6 | 0 | 8 | 2 |

#### *Domain F (financial)*

We consider a process of a bank to handle the closing of a bank account. A request is first created, by either the owner of the account or a 3rd-party, such as an attorney. Then, the request is evaluated. As part of the evaluation, a risk assessment may also occur, which involves checking for abnormal transactions in the account history, or whether the account has been involved in illicit or suspicious activities. Risk assessment is optional and may also be executed later in the process. Then the outstanding balance of the account is determined, including pending payments. Before the closing can be finalized, an investigation of the account owner's heirs has to be executed, in order to understand to whom the outstanding balance should be transferred. The outcome of the process can be either of the following: the request will be executed (1) or the request will be sent to credit collection (0) for a more thorough assessment. The latter may happen, for instance, when there are irregularities in a request, pending payments, or the heirs are unreachable. Since this procedure takes long and involves a high amount of resources, the bank would like to minimise the number of requests sent to credit collection.

### *Plots*

For our study, we selected three plots in order to investigate three levels of explanations for predictions: (i) at *event*, (ii) *trace* and (iii) *event log* level. We chose plots implemented in Nirdizati (Rizzi et al. 2019), since they are based on the post-hoc explainers more frequently used in the PPM context; in addition, they provide explanations for the three main elements characterizing the PPM data. The standardized interface provided by Nirdizati (see https://user-evaluation-mock.web.app/) offers a cohesive framework for visualizing and analyzing the plots consistently and facilitates accurate comparisons.

The first plot (*Plot P1*) - the plot at *event* level - is an adaptation of the SHAP explanation plots to the PPM field. Figure 1 shows an example of this type of plot. For all features, which in the case of an encoding based on the event position in the trace (index-based encoding (Leontjeva et al. 2015)) correspond to the activities executed at different positions of the trace, the plot shows the impact on the prediction returned by the predictive model, i.e., the correlations in terms of SHAP values of each feature (and associated value) with the prediction. For instance, in *Plot P1*, reported in Fig. 1, the most impacting feature-value pair is the pair composed of the event at position

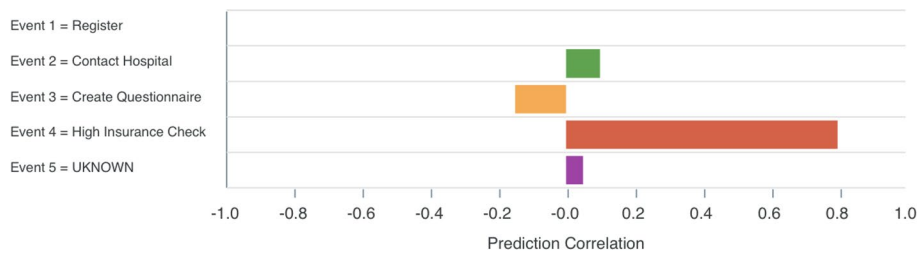



**Fig. 1** SHAP values related to the correlation between a (feature, value) pair and the returned prediction (*Plot P1*)

four (*event_4*) and activity `High Insurance Check`, i.e., activity `High Insurance Check` occurring at position four has a high correlation with the returned prediction (expressed as a SHAP value of 0.8). Such a correlation means that this feature-value pair strongly affects the returned prediction.

The second plot (*Plot P2*) - the plot at *trace* level - shows how the correlation of each feature (and associated value) with the returned prediction (expressed in terms of SHAP values) changes at different positions in the ongoing trace. The plot shows hence how *stable* the importance of a feature for a prediction is as the prefix length at which the prediction is carried out increases. This plot is an adaptation of the *temporal stability* plot (Teinemaa et al. 2018) used in PPM to check the stability of a prediction at increasing prefix lengths to the case of the prediction explanations. Differently from the original temporal stability plots, in which the prediction values are plotted at different prefix lengths, in these plots, the SHAP values of the features (and corresponding values) are plotted at different prefix lengths. For instance, in *Plot P2*, reported in Fig. 2, feature *event_2* corresponding to the event at the second position of the trace (green line) is rather unstable, as it positively or negatively correlates with the returned prediction based on the specific point of the trace in which the prediction is made. Instead, feature *event_4* corresponding to the event at position four of the trace (red line) positively correlates with the prediction starting from the prediction carried out at prefix length four in a stable way. This means that, starting from prefix of length four, the returned prediction highly depends on the event occurred at position four, which is actually known only from prefix length four onwards. Note that, in the plot in Fig. 2, the specific values of the features are not reported, as they may change for different prefix lengths. For example, *event_4* at prefixes shorter than four is set to null. The values of the features are however visible by hovering the

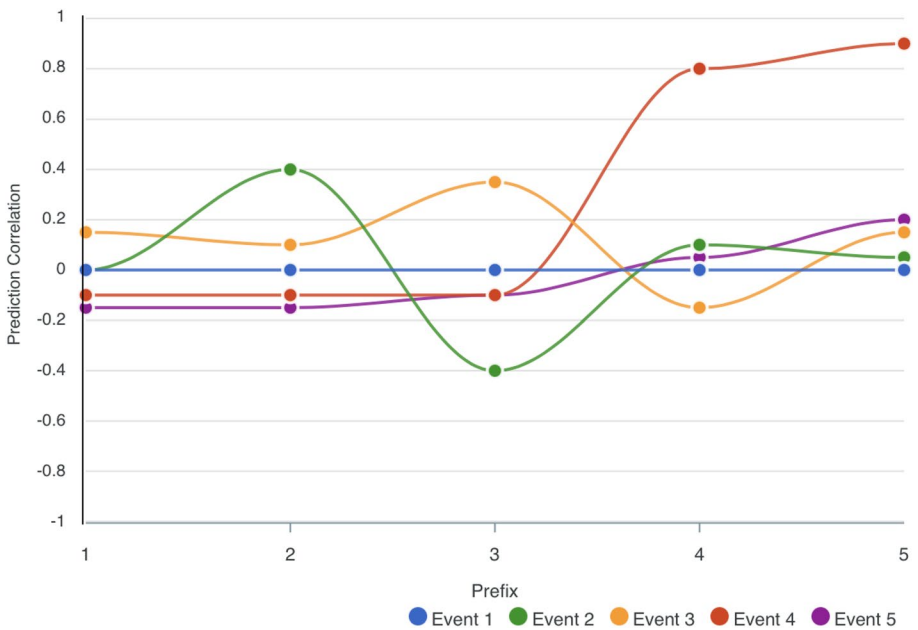



**Fig. 2** Temporal stability of the SHAP values related to the correlation between each feature and the returned prediction (*Plot P2*)

mouse on the dynamic version of the plot (see Appendix A3 in Rizzi et al. (2022)). The plot allows the user to get an idea of the stability of the importance of a certain feature as the trace evolves, thus increasing the confidence of the user in making decisions based on explanations using that feature (Goldstein et al. 2015).

The third plot (*Plot P3*) - the plot at the *log* level - shows, for a specific feature, the average value of the predicted label (between 0 and 1 with 0 meaning *false* and 1 meaning *true*) for the different feature values, as well as the number of traces in the event log containing each value. The plot, which mainly collects some statistics on the training data, is based on the ICE plots (Goldstein et al. 2015). In a binary setting, in which two possible values are associated to the predicted label, the average value of the label is the average percentage of traces for which the label value is *true*. For example, *Plot P3*, reported in Fig. 3, shows the average value of the predicted label for the different values of feature *event_4*, as well as the frequency of the values of that feature on the whole event log. For instance, value `High Insurance Check` for feature *event_4* has an average label value of 0.7, i.e., in 70% of the traces of the training set having as value for feature *event_4* activity `High Insurance Check`, the associated label is *true*, while, for the remaining 30%, the label is *false*. Moreover, the plot also shows that such a value occurs in around 750 traces in the training set. Therefore, the plot allows the user to get an overall idea of the distribution of the labels for a specific feature value, as well as of the frequency of that specific value in the whole event log.

### *Tasks*

As first task, the participants carried out a comprehension task where they were asked to interpret the three types of plots described in Plots section (the details of this task are provided in Appendix A1 of Rizzi et al. (2022)). Then, for each of the scenarios described before (Domain M (medical) and Domain F (financial) sections), we developed a set of tasks that required participants to use the plots for decision making. In the following, we describe the tasks that we developed for domain M (Domain M (medical) section). The full list of tasks including the domain description, the plots and the questions asked can be found in Appendix A2 of Rizzi et al. (2022).

**Task Description M1.a:** Consider an incomplete trace of a patient who has carried out one of the treatments reposition, cast application, or surgery. For this patient, the

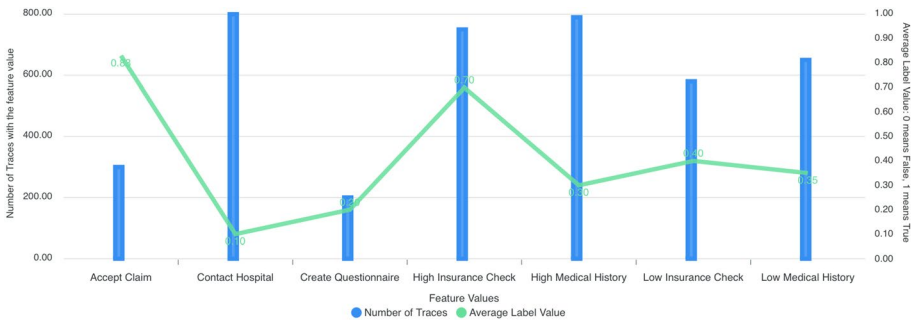
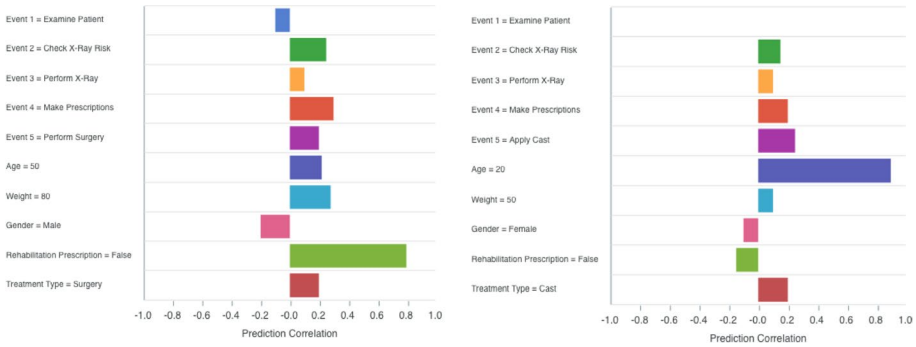



**Fig. 3** Average label value for different values of a feature on the whole event log (*Plot P3*)




(a) Predicted outcome for late recovery. (b) Predicted outcome for quick recovery.

**Fig. 4** SHAP values related to the correlation between features, their value and the predicted outcome (*Plot P1*)

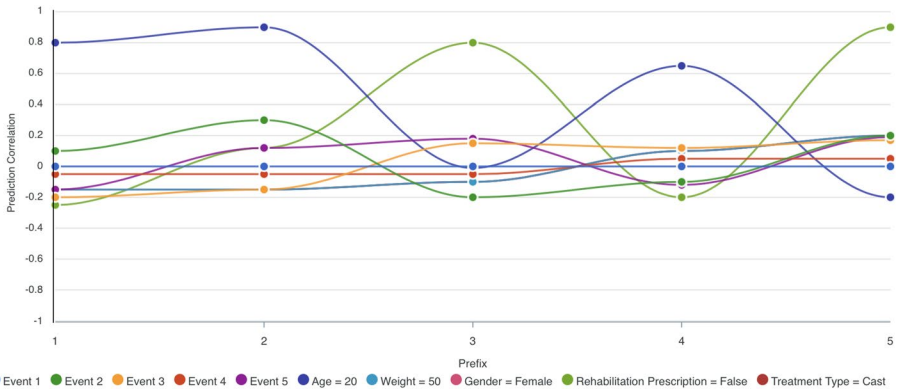



**Fig. 5** Temporal stability of the SHAP values related to the correlation between each feature and the predicted outcome for a quick recovery (*Plot P2*)

prediction is that she will not recover soon from the fracture (0). The explanation of the prediction is reported in Fig. 4a. Would you recommend carrying on the rehabilitation? Why?

**Task Description M1.b:** Consider now the incomplete trace of another patient who has carried out one of treatments reposition, cast application, or surgery. For this patient the prediction is that she will recover soon from the fracture (1). The explanation of the prediction is reported in Fig. 4b. Based on the information presented in the plot, what course of action would you suggest?

**Task Description M2:** Consider an incomplete trace of a patient who has carried out one of treatments reposition, cast application, or surgery. For this patient, the prediction is that she will recover soon from the fracture (1). The explanation of the predictions from the beginning of the trace up to the current point is reported in Fig. 5. Would you recommend carrying on the rehabilitation? Why?

**Task Description M3**: Consider a set of process executions related to patients who have carried out one of treatments reposition, cast application, or surgery. For some of these patients, the prediction is that they will recover soon from the fracture (1), for others the prediction is that it will take time for them to recover (0). The explanation of the

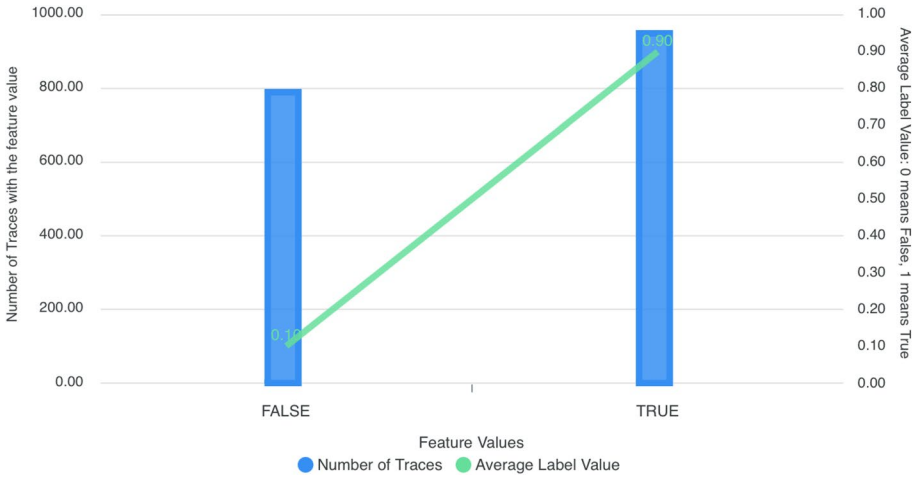



**Fig. 6** Average recovery time from a fracture through different values of a feature on the whole event log (*Plot P3*)

**Table 2** Participants, their expertise and provided order of domains

| Expertise | Domain Order | Code |
|---|---|---|
| BPM | M then F | BM1 |
| | | BM2 |
| | F then M | BF1 |
| | | BF2 |
| PPM | M then F | PM1 |
| | | PM2 |
| | F then M | PF1 |
| | | PF2 |

predictions for different patients in the training set is reported in Fig. 6. Would you recommend rehabilitation as a general practice for all patients?

### Data collection

Before the main study we conducted three pre-test interviews with individuals that were similar to our study population. They all had experience related to BPM and two of them also had experience related to PPM. Based on these pre-test interviews we refined our study procedure and interview protocols.

For our main study we conducted semi-structured observational interviews with individuals within our study population which lasted between 68 and 109 minutes each. To guide the participants through the study procedure, we created a web interface that contained the description of the domains as well as the related plots and tasks.[1][2]

The first part of the web interface contained explanation plots and tasks for the comprehension task. The second and third parts contained explanation plots and tasks for

[1] The interface can be accessed at https://user-evaluation-mock.web.app/

[2] Some interaction capabilities of the interface are summarized in Appendix A3 of Rizzi et al. (2022).

the decision making tasks in domains M and F. We presented the two scenarios in different orders within the two groups of subjects to be able to mitigate the effect of individuals learning how to interpret the plots (c.f. Table 2).

The interviews were conducted via Zoom by a team consisting of a *facilitator* and an *observer*, with the facilitator guiding the participant and the observer serving in a supporting role. Participants were encouraged to think aloud, to ask questions and to point out interesting aspects related to the different plots during the interview. All interviews were video-recorded.

Each interview started with the facilitator introducing the study procedure and the application to the participants who were then asked to open the link to the comprehension task and share their screen. The facilitator then asked the participants a set of predefined questions regarding their understanding (e.g., "*What do the different plots show?*") and interpretation (e.g., "*Which are the feature(s) influencing most the prediction related to the visualized trace?*") of the different plots ($\mathbf{RQ}_1$).

After the comprehension task the participants were asked to open the web interface containing explanation plots and tasks for the decision making tasks in domains M and F. The facilitator then introduced domains and tasks and asked the participants to explain the plots (e.g., "*How do you interpret the plot?*", $\mathbf{RQ}_1$) and to decide what to do next based on their interpretation of the plot (e.g., "*Based on the information presented in the plot, what course of action would you suggest?*", $\mathbf{RQ}_2$). In addition, the facilitator also asked specific questions related to each plot (e.g., "*Would you recommend carrying on the rehabilitation? Why?*", for task M1.a) to further examine the decision making process leading the participant to suggest a certain course of action ($\mathbf{RQ}_2$).

After finishing the tasks, the facilitator conducted a short follow-up interview. The questions focused on the participant's understanding of the different plots (e.g., "*Which visualization(s) was/were the hardest to interpret? Why do you think you were struggling with these visualizations in particular?*", $\mathbf{RQ}_1$) and the perceived usefulness for decision making (e.g., "*Which of these visualizations helped you to make more informed decisions?*", $\mathbf{RQ}_2$). In addition, the facilitator also asked each participant suggestions on how to improve the plots to better support their ability of making informed decisions based on them (e.g., "*Is there any additional information that would have helped you to understand the different visualizations and make more informed decisions during the study?*", $\mathbf{RQ}_3$)

Finally, the participants were asked to answer a questionnaire after the interview. The questionnaire focused on the perceived ease of use ($\mathbf{RQ}_1$) (Davis 1989) and perceived usefulness of each plot ($\mathbf{RQ}_2$) (Davis 1989).

#### Analysis procedure

Our qualitative analysis mainly focused on the video recordings, and the observational and follow-up interviews. The collected questionnaire data served as an additional qualitative data point.

For our analysis, we combined deductive and inductive coding. Starting with deductive coding, we first utilized a set of eight predefined codes that were based on our first two research questions. We focused on whether participants arrived at the *correct interpretation* or *wrong interpretation* and if they *required additional information* to understand

the different plots ($\mathbf{RQ}_1$). In addition, we assessed whether participants arrived at the *correct decision* or the *wrong decision* based on the different plots and identified whether the participants indicated some form of *reasoning* that drove their decision ($\mathbf{RQ}_2$). Finally, we also assessed how participants interacted with the visualizations (*interaction with visualizations*) as a way to identify means for improvement ($\mathbf{RQ}_3$).

A group of three researchers, all of which are co-authors of this paper, collaboratively applied these codes to all interviews and observations thus creating an initial clustering of participant mentions and observations. The coding was done by plot, meaning that first, all responses related to plot 1 for all interviewees were coded before moving to plot 2. These codes allowed us to create an initial structure for further investigation.

Afterward, the same researchers conducted a second round of inductive coding, during which we focused on understanding why individuals might have interpreted the visualizations correctly or incorrectly, why they might have arrived at a correct or wrong decision, and how visualizations could be improved. This round of coding was thus similar to open coding in that we allowed for new concepts to emerge. During this process, ten additional codes emerged that pointed, e.g., toward a *learning curve* when aiming to understand the visualizations ($\mathbf{RQ}_1$), towards participants having varying levels of *confidence* in their decision while sometimes making a *correct decision with wrong reasoning* ($\mathbf{RQ}_2$), and adding interactive elements including a *what-if-analysis* to the visualization ($\mathbf{RQ}_3$). The concepts that emerged were thus mainly related to sources of misunderstanding or to expanding the capabilities of the visualizations we utilized as a basis for our study.

## Findings

In this section, we discuss our findings related to how participants made sense of explanation plots (RQ1: How do users make sense of explanation plots in PPM? section, $\mathbf{RQ}_1$) and how they utilized them to make decisions (RQ2: How can explanation plots support users in decision making tasks in PPM? section, $\mathbf{RQ}_2$). We also outline suggestions on how plots can be improved (RQ3: How can PPM explanation plots be improved? section, $\mathbf{RQ}_3$). For each RQ, the themes identified derive directly from the consolidation of the pre-defined and emerging codes identified during the data analysis.

### RQ1: How do users make sense of explanation plots in PPM?

Our interviewees generally found the different plots to be reasonably easy to use. This is evident by their responses related to the perceived ease of use of the different plots (Fig. 7) as well as by statements of different participants during the interviews (*"I found [these] quite easy to interpret"*, BM2, *"It's easy to get"*, BF2, *"the interpretation is clear"*, PF1). Out of the three plots we used, the participants perceived P1 to be the easiest and P2 to be the hardest to use. The PPM experts generally perceived the plots to be easier to use than the BPM experts. This difference is minimal for plots P1 and P2 while albeit slightly larger for P3.

The fact that plots were generally easy to use, however, does not mean that the experts did not have to climb a learning curve to fully master the use of the plots. The main themes regarding the difficulties in understanding and using the plots emerged during the interviews with both types of experts are discussed next.

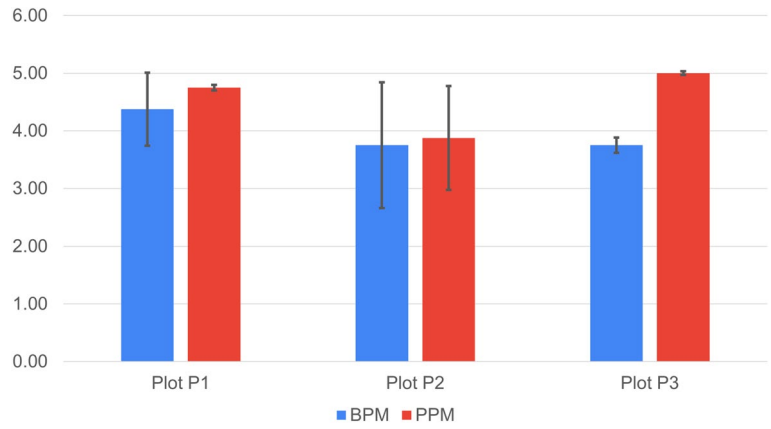



**Fig. 7** Perception of BPM and PPM experts of the ease of use of the different explanation plots (1=very low, 5=very high). We report in the plot median and variance

***Struggling at first sight of the plots***

Most users struggled at the beginning with understanding specific aspects of each plot. This could have been expected, considering that many experts, in particular the BPM ones, were seeing such plots for the first time.

In P1, the main misunderstanding from experts concerned the meaning of the Y-axis. The Y-axis, in this plot, simply lists the features (and associated values) that are considered by the model at a certain prefix and the order in which these are shown does not matter. During the interviews, however, some experts associated a meaning with the ordering of the features on the Y-axis. BM2 and BF2 associated this with a notion of feature importance, i.e., the higher features on the Y-axis were considered somehow more important in the predictive model (*"So this is from top to bottom, I'm assuming based on, okay, some kind of value for each event of the event log"*, BM2). PM2 interpreted this as a chronological order, i.e., the higher features on the Y-axis were assuming the corresponding values earlier (*"I think that after event one, there is nothing yet.... And then after event [...] it seems that there is a relatively low, but definitely some positive correlation"*, PM2). Such misinterpretations may also have been due to the fact that the comprehension task only used features defined by event labels.

Regarding P2, two PPM experts (PM2 and PF1) struggled to understand the fact that the data shown are part of a time-series and that, therefore, they have to be considered as a whole. That is, the correlation values are valid at a given prefix length, and they may vary if the same plot is regenerated for a different prefix length. These issues were not related to the stability/instability of the data shown but to the difficulty of the experts in grasping correlation of time-series data.

Finally, the BPM experts considered P3 too complex. For BM1 and BF2, the plot was hard to understand at the first glance because of its different graphical appearance and different semantics compared to plots P1 and P2. The double Y-axis, representing the average label value and the number of traces containing a specific feature value, also was deemed as confusing by the same experts ("*Expert: So, trying to find out what for instance, the green value means it is average labor value or contact hospitals which means all traces that had an event contact hospitals. Interviewer: that have contact hospital at*

*event 4 is the average value of the label. Expert: Okay, and now I get it.",* BM1; "*Expert: With the average label value, average label value. [...] What is an average label value? Interviewer: Labels can be one or zero. Expert: Yeah, the outcome. Right."* BF2).

Apart from these issues, after struggling initially, the experts managed to correctly interpret the information in the plot without explicit help from the facilitator.

#### *Requiring additional explanations*

In many cases, the experts explicitly required additional explanations from the facilitator to understand specific aspects of each plot.

Some experts, in the BPM group, required an explicit explanation of the meaning of term "correlation" in plot P1 ("*The prediction correlation to be honest, it doesn't tell me anything. So, there is a set of events and correlation with respect to what?*" BF1). Here, the lack of experience with XAI techniques of the BPM experts plays a key role in their understanding of this term appearing on the X-axis of P1.

After having seen P1, all experts were able to understand P2. However, some experts struggled to understand the meaning of what a "stable feature" in P2 is and required the explicit guidance of the facilitator. Intuitively, some experts associated feature stability with a *flat* line in the plot. However, they also realized immediately that a flat line at 0 identifies an irrelevant feature that has no correlation with the trace outcome at any prefix length ("*The stability suggests me [one feature] because it's always zero. But it also means that doesn't affect at all the the prediction. It is stable but it's useless*" BF1). After this initial hesitation, most experts (BF1, BM2,PM2, PF1, and PM1) correctly considered as stable the feature(s) for which the line in the plot was flat after having taken a positive value. Some of the experts, however, considered other aspects when determining stable features in P2, such as the interplay of different features (BM2, PM2), e.g., "*The correlation is moving to a negative value when [another feature] takes a known value*" (BM2), or the change in polarity of the correlation value (PF1). Finally, one expert (BM1) incorrectly identified as stable only the feature with the highest correlation value at the last prefix shown in P2.

#### *Recognising the bias of P3 bias due to data imbalance*

During the interviews, the experts generally understood correctly the information provided in P3. For instance, BM1, who required explicit explanations on both P1 and P2, could easily understand the content of P3, even though the information conveyed by this plot is far more than the one provided in P1 and P2 ("*For me, it was clear that the correlation between the values of a feature that we are evaluating [...] There is like less noise, so to say, like less information that I'm not using to make my assumptions.",* BM1).

However, two experts (BF1 and PF1) assumed that P3 could point to a predictive model that is unreliable or biased when the labels in the event log are not balanced ("*It could be more reliable whenever I have a set of traces that has an equal number of true and false.",* BF1; "*If you have a very skewed data set where you have too high number of traces having the same label and where the other label in a binary setting are fewer [...] then you can have some bias in the prediction.",* PF1). One possible interpretation is that these experts over-analyzed the information shown in the plot.

**RQ2: How can explanation plots support users in decision making tasks in PPM?**

Similar to the findings reported in the previous section (RQ1: How do users make sense of explanation plots in PPM? section) our interviewees generally found the different plots to be reasonably useful for solving the decision making tasks. This is evident by their responses related to the perceived usefulness of different plots (Fig. 8) as well as by statements of different participant during the interviews ("*I clearly understood what I should recommend*", BM1, "*all of [the plots] serve a purpose*", BM2, "*So it's very intuitive the way to represent this part here,*", BF2). The general perception of the usefulness of the plots is lower than the perception of their ease of use though. The survey also showed that the PPM experts who participated in our study found all three plots equally useful. The BPM experts, however, perceived P2 to be considerably less useful than the other plots. This is possibly due to the fact that this is a more technical plot and understanding how it can be used in practice is more difficult for people without a PPM expertise.

To answer ($\mathbf{RQ}_2$), we have identified several themes regarding how the experts have used the plots in the decision making tasks. These themes are reported in the following sections.

***Identifying the correct decision, but through the wrong reasoning***

In some cases, experts ended up making the correct decision in a task, but interpreting the information in the plots in the wrong way. A typical example of this situation is the one of experts (BM1, BM2, PM2, and PF1) using only the information associated with the last prefix shown in plot P2 in Task A2. It is only by coincidence that this led the experts to make the correct decision. The reason behind this behavior could be that experts anchored their reasoning on the principle that the more information is given to a predictive model, the more accurate it is likely to be ("*The latest you make the prediction, the more accurate it is because then you have more information*", BM1). Hence, they considered reasonable to trust only the information associated with the longer prefix in plot P2.

Another example concerns experts suggesting to prescribe the rehabilitation in task A1 because it looked like the only option available, even though other options were

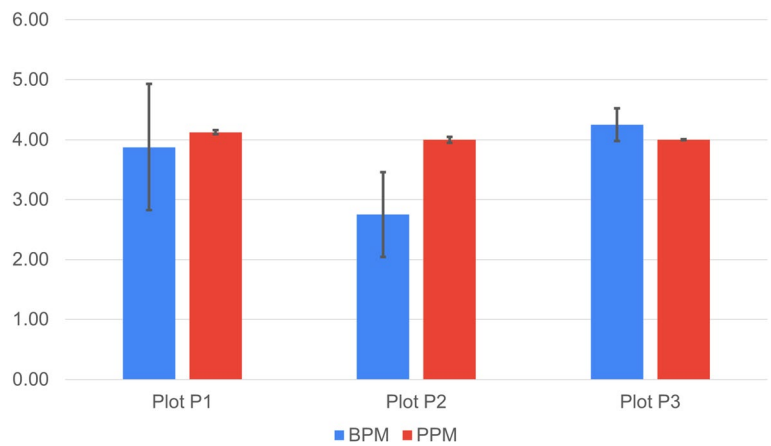



**Fig. 8** Perception of BPM and PPM experts of the usefulness of the different explanation plots (1=very low, 5=very high). We report in the plot median and variance

available, like proceeding without taking any action ("*I suggest doing [..] for the reason that it's the only thing that the [doctor] can change, or affect.*", BM2).

#### *Reaching a decision, but with limited confidence*

For different reasons, the experts were not always confident in the decisions they made using the plots. This was particularly the case of tasks in which plot P1 was used (Tasks A1 and B1). Right at the end of task A1, when asked for general comments, PF2 was concerned by the possibility of making a wrong choice ("*What is the real business risk and setting? What kind of decisions? [...] how do I need to act on the information I see?*", PF2) and, in the subsequent tasks, could not make decisions based on the information shown by the plots.

PM2, in both task A1 and B1, remarked that her lack of confidence on the decision made was due to the lack of clear options among which to choose and to the lack of knowledge about the possible decisions that could be made ("*There's definitely no guarantee that [this decision] will then always have a positive [outcome].*", PM2).

BM1, in task A1, lamented a low confidence in the decision made due to lack of confidence in the effect that a specific intervention (to perform the rehabilitation in this particular case) may have ("*I don't know what I can assume [...] now that rehabilitation is false, rehabilitation has -0.15 prediction correlation, then if rehabilitation is true, the prediction correlation will be +0.85?*" BM1).

In other cases, however, the experts were more confident about the interpretation of plot P1 and the decision. This happened when they could easily identify the possible decisions and their effects on a trace execution exploiting their own domain knowledge. For instance, in task M1.b, BF2 felt confident to recommend no action because, according to plot P1, the trace was already predicted to have a positive outcome ("*I'm more positive towards this interpretation, I feel more confident because age is nothing we have under control.*", BF2) .

#### *Enabling further investigation to take decisions*

A common theme emerging from the interaction with both groups of experts is that often, even when declaring to be confident about the decision made, experts still mentioned the need to do more investigation in order to fully validate the decision. In some cases, experts (PM1, PM2, PF2, BM2, and BF2) mentioned the necessity to further investigate a particular decision using additional analysis or data gathering (e.g., "*Yeah, prescribe [..] but again, I would first investigate why*", BF2). The PPM experts were more explicit in this regard mentioning specific additional data analysis techniques, requiring "*to see more data*" (PM2) by performing additional "*data collection, [..] full randomized trials, or bandit tests*" (PF2).

#### *Combining information from different plots to confirm decisions*

Experts showed to be able to integrate the different perspectives offered by the different plots and often used the information provided in plots shown earlier to confirm the decision made in the current task ("*If I forget about the previous plot, I would say no. If I consider a combination between the two, yes*", BF1). In particular, BM1, BM2, BF1, BF2, PM1, PM2, and PF1 were able to spot the different perspectives shown by the plots

and to infer a more complete overview of the situation presented in the decision making tasks using more than one plot at the same time ("*each single plot provides you with a view of the reality.*", PF1). We can then infer that showing all the plots at the same time would have been beneficial for the comprehension of the situation at hand, "*to make more informed decision.*" (PF1), and could have even improved the experts' confidence in their decisions.

#### *Providing option enumeration and related what-if analysis*

Most of the experts (BM1, BM2, BF1, BF2, PF1, PM2, and PF2) felt the necessity to receive additional information regarding the choices available when asked to suggest a course of action using plot P1 or P2 ("*what are my degrees of freedom?*", PM2). In one case (BM1), the expert had also to be reminded that taking no action was one of the possible options.

Once the possible options were enumerated, experts (BF2, PM1, PF1, and PF2) also mentioned that tools supporting a prescriptive (what-if) analysis of the different options could have been useful to make more informed decisions. According to the experts, these could take the form of trace matching, i.e., matching the current trace to similar one(s) for which the outcome is known ("*I don't have a counter proof. So if I had another trace in which the third event was [same value]*", BF2, "*you could do matching here and you could try to see to match this to the to the closest case [..] to see what happened to that.*", PF2), or simulation, i.e., if the observational data implementing a particular strategy is not available then the expert can simulate the new behavior to check how the correlation values will vary after the new strategy is implemented (*we don't know what happens after this point, then you could add some actionable attribute that I can change in order to change the course of the future.*", PF1).

#### *Accounting for the characteristics of the domain*

The domain in which tasks are situated has clearly influenced the experts decisions ("*So it's not like in the previous domain, [..] here the domain influences a lot the reasoning*", BF1). Specifically, for tasks in Domain M, experts (PM1, PM2, BF1, and BF2) mentioned that depending on the severity of the case at hand, which was not one of the available attributes in the event log, their choice could change ("*one might be tempted to say that the best way to make [..] is to [..]. However, it's hard to tell, because there could be some confounding attribute aspect that is not recorded in the log*", PM1 or "*assess the situation on a case by case basis*", PM2). We argue that this is due to the medical domain being perceived as one where process analysis is particularly challenging because special cases are frequent and process users (physicians, nurses) often have a higher amount of freedom in deciding the course of action, e.g., on the treatments to administer (Munoz-Gama et al. 2022). The domain F was considered by experts more driven by clear and standardized procedures and therefore less challenging as far as predicting the outcome of cases is concerned.

#### *Highlighting that correlation is not causation*

The experts, in particular the PPM ones, mentioned multiple times that the plots they were using were showing a correlation, e.g., between feature values and predicted

outcome, but that such a correlation should not be directly interpreted as a causal relation ("*Yes, there is a correlation, of course, but it doesn't mean there is a causation*" PM1). One of them in particular (PF2) could not make a decision for this reason ("*Obviously, we do see that the fact that the patient was young contributed a lot to the fact that the positive outcome for the patient. I guess that makes sense. Young people recover more easily. That's a correlation. But I would say just from domain knowledge that's even likely to be a causation. Even though we're not sure about that. And so, that's something you can easily read in this plot. But we don't have anything to make recommendations here*", PF2).

Other PPM experts took a lighter stance in this regard, for instance, suggesting, after having mentioned the correlation v. causation theme, that "*there's definitely no guarantee*" (PM2) that the course of action that they suggested was going to yield the desired outcome, or that their suggested course of action was simply a "*suggestion to the domain expert*" (PM1).

#### *Quantifying the cost of the suggested course of action as decision variable*

The experts often mentioned that understanding the cost of a suggested course of action is a fundamental variable when making a decision ("*I would probably suggest not to do [..] because it would incur in an additional cost*" PM1). In tasks that used plot P3 (Tasks A3 and/or B3), BM1, BM2, BF1, PM1, PM2, and PF2 decided to avoid making a definite decision for the process "*depending on the cost*" (BF1), even if they could make a decision that was perceived as beneficial for the outcome of the process. Specifically, BM1, BM2, BF1, PM1, PM2, and PF2 suggested to make the risk assessment a mandatory task for the process in task B3 of Domain F only after having considered the cost of this task. Similarly, PM1, PM2, and PF2 suggested to make the rehabilitation mandatory in the process in task A3 of Domain M, only if its cost is reasonable.

#### *Considering the expert background in the decision making process*

The background of BPM and PPM influenced, for different aspects, the decision making process.

First, we noticed that the PPM experts tended to be more conservative in their decisions, often mentioning that the plots did not give them enough evidence to make specific recommendations and that they would rather delegate these to domain experts.

Regardless the task or plot, PF2 felt that the plots were not conveying enough evidence to make any specific recommendation ("*I can do a little bit of diagnostics [..] but I can't make any recommendations*", PF2). This feeling was partly echoed also by PM2 ("*there's definitely no guarantee that [this decision] will always have a positive [outcome]*", PM2), PM1 and PF1, who consistently showed to have concerns regarding any choice they made, even mentioning that they deemed necessary that someone else had the final word (e.g., requiring a "*suggestion to the domain expert*", PM1; or "*assessing the situation on a case by case basis*", PM2).

Conversely, BPM experts BM1, BM2, BF1, and BF2 were more willing to make clear and specific recommendations based on the information available, often relying on their domain knowledge ("*I have a feeling that those cases are very rare [therefore] rehabilitation should be a general practice*" BM2).

The BPM experts generally experienced a linear interaction pattern throughout the interview, answering the questions that they were asked, without showing any particular additional concern regarding how the plots were produced. The questions they raised during the interviews were focused exclusively on understanding how to perform the task at hand. For BPM experts, we also witnessed a steep learning curve effect, which lead to the tasks on the second domain taking a considerably lower amount of time than the ones on the first domain, regardless of the order in which the domains were presented.

The PPM experts instead showed a deeper understanding of the proposed plots, but they also had a tendency to over-analyze the information showed in them. This, at times, hindered the flow of the interview and led their interviews to take on average longer than the ones with BPM experts. However, the quality of the decisions made by PPM expert appeared not to have been affected by this tendency to over-scrutinize the information in the plots.

#### RQ3: How can PPM explanation plots be improved?

Several suggestions for improving explanation plots have also emerged during the interviews. These are related to the plots used in this investigation, which are based on the post-hoc explainers more frequently used in the PPM context. However, the suggestions provided apply to PPM explaianability plots in general. These suggestions revolve around three main themes, which are discussed below.

##### *Improving the interface*

Several comments received during the interviews were directed towards improving the interface through which the plots were shown. For example, the interface could be improved by showing all the plots together in a dashboard setting. This would help giving the user multiple perspectives regarding the process at the same time. We have drawn this conclusion by analyzing the behavior of the experts during the interviews. Both the BPM and PPM experts used multiple plots at the same time to attain a more precise idea regarding their decisions. For instance, many experts used plot P3 to find a confirmation, at the global event log level, of the decisions they made based on the information found in P1 or P2. In alternative, experts mentioned that it would be effective to show P1 and P2 together or, at least, P2 before P1, since plot P2 represents a more general overview of the feature contributions for different prefixes, whereas P1 is a snapshot of P2 at a specific prefix length.

In addition, a number of specific improvements have emerged regarding the elements of individual plots. Regarding P2, the wavy lines associated with individual features were considered generally confusing and, more specifically, hinting at the visualization of a continuous variable ("*Curving is great when the domain is continuous, here is discrete, there is no 1.5 event*", BF2). In fact, the correlation values shown by P2 are discrete, because they are defined only at discrete prefix length values. Therefore, the continuous lines may be substituted by discrete plots, e.g., by markers or bars. Furthermore, P2 does not show information about the prefix length at which a feature takes its value. For instance, feature "Rehabilitation prescription", in Domain A, does not have a value before the activity in which the doctor decides whether to prescribe the rehabilitation

or not is executed. This information can be helpful for decision making tasks and can be captured in the plot, for instance, by greying out the lines and/or labels associated with a particular feature until the feature takes a known value. Regarding P3, the green line connecting the bars of the histogram (indicating for each feature value the number of traces containing that value) has been considered to be misleading, since it hints at a connection between the feature values, which, in fact, does not exist ("*I ended up looking at data points [...] what's the point of having a line?*", BM2).

#### *Adding interactive elements*

Experts have highlighted the lack of interactive elements in the plots they were asked to use. A first level of missing interaction concerns the filtering of specific attributes, traces, or events dynamically, instead of showing one static view of them all. While, in P2, it is possible to hide the lines associated with attributes that are not deemed important for a decision (for instance, because they are associated with a constantly oscillating correlation and therefore not helpful to make decisions regarding the case outcome), a similar level of interactivity is not available in P3. Experts suggested that, in P3, it should be possible to filter only traces with certain characteristics, such as as traces similar to the one on which a decision has to be made ("*In this third plot [...] I just see these numbers, should I ask what is happening behind? [...] I can divide [the data] in small groups [...] checking what it is happening behind the scene, how the data varies, how [the data] depends on the events and the variables*", BF1).

A second level of interaction concerns supporting what-if analysis. As highlighted earlier, this is a requirement expressed in particular by the PPM experts, who have a more hands-on experience with predictive models and explainability methods. What-if analysis represents a means to understand what would happen to the outcome of a process when the decision regarding a certain course of action is actually put in practice. Plots P1 and P2 are particularly suited to provide this sort of actionability, whereas this is not the case for Plot P3, which does not represent the course of actions of a single trace, but a global overview of the information available in the entire event log.

#### *Including additional information fields*

In several cases, the experts mentioned, during the interviews, the need to enrich the plots with more information. From a technical standpoint, experts requested more information to gauge how much they could trust the performance of the predictive model from which the plots were derived. This may involve, for all plots, showing the accuracy measures of the predictive models, like F-score or AUC. Regarding P3, experts mentioned that showing the relative frequency of the feature values in the event log would have been more useful for supporting the decision making tasks with respect to the absolute number of traces including them. The PPM experts also pointed out the need to show additional statistical information in P3, such as box plots providing the confidence interval of the label value for a given feature value.

More generally, the experts lamented the lack of information explaining the plots, such as more text describing their semantics, explanations of the acronyms used, different color codes for different types of features (e.g., to distinguish features associated with the

occurrence of activities in a trace from other features) and a more clear explanations of the term "correlation" in plots P1 and P2.

## Discussion and conclusion

In this section, we analyze and discuss the results obtained from the user study by drawing the main implications for practice (Implications for practice section), for future research (Implications for research section), as well as by highlighting its limitations (Limitations section).

### Implications for practice

Our findings revealed how different individuals utilize explanation plots for decision making and can thus serve as a basis for suggestions on how to employ them in the context of PPM. While individuals appreciated the plots as a valuable source of information, they often did not perceive the information presented to be rich enough to make a decision confidently. It is thus important to provide additional information to decision makers in the plots. Our findings are in line with those of Lim et al. (2009) in which different explanation plots were employed to improve the intelligibility of context-aware intelligent systems.

First, the experts asked for domain-related information, for details about each specific case that were not present in the explanation plots, for options to filter the information that was available in the plots, and for the possibility to perform additional analysis like on-the-fly what-if analysis and statistical significance tests. Domain-related information and case details can be easily added, as well as suitable filters. Performing additional analysis as suggested by the participants might be more challenging. The need for a what-if analysis has indeed also been reported in Davis and Kottemann (1994). However, in this work, the authors also found that providing such analysis can create an 'illusion of control' that causes the users to overestimate its effectiveness. When providing more information, it is thus necessary to carefully consider which information to provide and to whom.

The BPM experts suggested to improve the labels and captions in the plots to better fit the decision making tasks. They also had issues in deciding which actions to take and asked which actions would even be possible. We would thus suggest to provide examples of possible actions that were taken in past process executions similar to the one of interest, as explored in Lu and Sadiq (2007).

The PPM experts asked for information regarding predictions and the way they were computed including the data distribution of the original data and measures representing the prediction accuracy of the predictive model. This information can be easily provided, and would make the decision making process more informed. We also observed that the PPM experts avoided reading the text in the captions. So we would suggest to make these elements of the interface more prominent or use pop-ups when the user interacts with them. Finally, the PPM experts suffer from the information overload coming from their expertise often resulting in getting lost in the details of the different plots (Buchanan and Kock 2001). We therefore suggest to complement these plots with plots that provide information at a higher level of abstraction that focus more on the business value of

the information provided and on the decision that needs to be made rather than on the model-specific information provided by the plots analyzed in this user evaluation.

#### Implications for research

The conducted study provided a set of suggestions on possible improvements of the presented plots. A first future research direction would be conducting an extensive literature review to identify further requirements complementing the collected improvement suggestions. Besides the improvements of the plots, the experts also highlighted several avenues of research related with improving decision making in process monitoring through XAI techniques.

Future research should investigate the most effective way of implementing what-if analysis tools to complement the existing plots, e.g., whether focusing on finding similarities between historical traces and the one for which a decision should be made, or implementing full-fledged simulations tools. Reinforcement learning (Sutton and Barto 1998) and bandit algorithms (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020) may also be considered to enable decision support tools to learn the best course of action (policy) that maximizes an expected gain.

Future research should also address explainability from the standpoint of establishing causality between feature values and process outcomes, rather than simply correlation. This need has also been reported by Lipton (2001) who found causality to be one of the four key aspects fostering explanation. In this direction, anchors (Ribeiro et al. 2018) are a post-hoc explanation technique that yields if-then rules explaining the behavior of the underlying model, together with an indication of the precision and coverage of the rules. In process mining, causal reasoning has been applied recently to control flow decision points (Leemans and Tax 2022).

From a human factor standpoint, more research is needed to understand which skills are required by decision makers to perform well in the decision making tasks with the support of explanation plots. Counter-intuitively, experts in our study lacking deep ML knowledge felt more comfortable using the plots. On the one hand, this can be due automation bias (Özalp et al. 2023), which occurs when non-experts trust AI decisions even more than their own judgment, and has been linked with higher task complexity and increased time pressure. On the other hand, more contextual information and opportunities to explore the suggested decisions, which also emerged as a need of our users, is likely to improve the understanding of the decision task yielding less biased user choices (Bove et al. 2022; Özalp et al. 2023). Future research should investigate to what extent other variables such as decision making experience, process modeling expertise, or process domain knowledge may influence the ability to make correct decisions in PPM scenarios. It could be also challenging to determine at which level of ML expertise a line can be drawn between experts and non-experts users. Along this line, the improvements of the plots identified in this work should be validated with different types of decision makers and decision making tasks.

Finally, future research should investigate to what extent the vision of self-optimizing business processes can be implemented with the support of AI tools within the business process lifecycle, i.e., whether the decision making tasks, possibly exploiting the feedback produced by XAI techniques, can be fully delegated to automated

tools, bypassing human decision makers. In this direction, recent research on prescriptive monitoring of business processes (Bozorgi et al. 2021; Weinzierl et al. 2020a) has dealt with (semi-automatically) incorporating the effects of what occurs during the execution of a business process (for instance as the result of human actions) while predicting the value of aspects of interest in the long term.

### Limitations

The goal of our study was to investigate how users make sense of explanation plots and how they use them for decision making. Furthermore, we aimed at exploring how explanation plots can be improved. It is thus reasonable to conduct a qualitative observational study (Lazar et al. 2017). There are, however, innate limitations associated with this study design. We interviewed individuals from different backgrounds, domain expertise and expertise related to BPM and PPM. Despite making a reasonable selection of the participants, it is not possible to generalize findings beyond our study context since studying different individuals with different backgrounds from different domains and different levels of expertise might yield different results.

Moreover, the study was conducted by a team of researchers, which poses a threat to validity since different researchers might perceive the reactions of study participants differently. To minimize this threat, we ensured that throughout the process of the study, which included the preparation of the study material, the conduction of the study and the analysis of the study results, at least two individuals from the research team collaborated on each step to avoid depending on the perception of individual researchers.

We also opted for studying a specific artificial setting utilizing specific plots and asking predefined questions. This can lead to observations and interpretations that might not have happened or might have happened in a different way in a real-life setting. To mitigate this threat, we made a state-of-the-art founded selection of the explanation plots, study domains and decision tasks. We acknowledge though that there is a remaining risk associated with studying an artificial rather than a real setting. We are willing to accept this threat because it allowed us to compare study findings across subjects which would not be possible when studying a real case.

Finally, we abstain from making causal claims providing instead a rich description of the observed behavior and reported perceptions of the study participants based on which we discuss differences in how different individuals made sense of explanation plots and used them for decision making.

**Authors' contributions**
W.R., M.C. and S.L. conducted the experiment. W.R., M.C., S.L. and A.E. analysed the results. All authors contributed to developing, writing and reviewing the paper.

**Funding**
This work was supported by the NRF Korea, Grant Number 2022R1F1A1072843.

**Availability of data and materials**
Not applicable.

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

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 20 February 2024 Accepted: 12 September 2024
Published online: 02 October 2024

**References**

Bove C, Aigrain J, Lesot MJ et al (2022) Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In: 27th international conference on intelligent user interfaces. ACM, New York, p 807–819

Bozorgi ZD, Teinemaa I, Dumas M et al (2021) Prescriptive process monitoring for cost-aware cycle time reduction. In: Di Ciccio C, Di Francescomarino C, Soffer P (eds) 3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021. IEEE, New York, pp 96–103. https://doi.org/10.1109/ICPM53251.2021.9576853

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found Trends Mach Learn 5(1):1–122. https://doi.org/10.1561/2200000024

Buchanan J, Kock N (2001) Information overload: A decision making perspective. Multiple criteria decision making in the new millennium. Springer, Berlin, Heidelberg, pp 49–58

Chromik M, Schuessler M (2020) A taxonomy for human subject evaluation of black-box explanations in XAI. In proceedings of the IUI workshop on explainable smart systems and algorithmic transparency in emerging technologies (ExSS-ATEC'20) Cagliari, Italy, p. 7

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quart 13(3):319–40

Davis FD, Kottemann JE (1994) User perceptions of decision support effectiveness: Two production planning experiments. Decis Sci 25(1):57–76

Di Ciccio C, Bernardi ML, Cimitile M et al (2015) Generating event logs through the simulation of declare models. In: Barjis J, Pergl R, Babkin E (eds) Enterprise and Organizational Modeling and Simulation - 11th International Workshop, EOMAS 2015, Held at CAiSE 2015, Stockholm, Sweden, June 8-9, 2015, Selected Papers, Lecture Notes in Business Information Processing, vol 231. Springer, Cham, pp 20–36. https://doi.org/10.1007/978-3-319-24626-0_2

Di Francescomarino C (2019) Predictive business process monitoring. In: Encyclopedia of Big Data Technologies. https://doi.org/10.1007/978-3-319-63962-8_105-1

Di Francescomarino C, Ghidini C, Maggi FM et al (2018) Predictive process monitoring methods: Which one suits me best? In: Weske M, Montali M, Weber I et al (eds) Business Process Management - 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings, Lecture Notes in Computer Science, vol 11080. Springer, Cham, pp 462–479. https://doi.org/10.1007/978-3-319-98648-7_27

van Dongen B (2011). Real-life event logs - hospital log. https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54

van Dongen B (2012) Bpi challenge 2012. https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f

van Dongen B (2017) Bpi challenge 2017. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b

Elkhawaga G, Elzeki OM, Abu-Elkheir M (2024) Reichert M why should I trust your explanation? an evaluation approach for XAI methods applied to predictive process monitoring results. in IEEE Transactions on Artificial Intelligence, p 1458–1472

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–232

Galanti R, Coma-Puig B, de Leoni M et al (2020) Explainable predictive process monitoring. In: van Dongen BF, Montali M, Wynn MT (eds) 2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020. IEEE, New York, pp 1–8. https://doi.org/10.1109/ICPM49681.2020.00012

Galanti R, de Leoni M, Marazzi A et al (2021) Integration of an explainable predictive process monitoring system into IBM process mining suite. In: Proceedings of ICPM Doctoral Consortium and Tool Demonstration. CEUR Workshop Proceedings, Aachen

Galanti R, de Leoni M, Monaro M et al (2023) An explainable decision support system for predictive process analytics. Eng Appl Artif Intell 120:105904

Goldstein A, Kapelner A, Bleich J et al (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65

Harl M, Weinzierl S, Stierle M et al (2020) Explainable predictive business process monitoring using gated graph neural networks. J Decis Syst 312–327. https://doi.org/10.1080/12460125.2020.1780780

Hase P, Bansal M (2020) Evaluating explainable AI: which algorithmic explanations help users predict model behavior? In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, Online, pp 5540–5552. https://doi.org/10.18653/V1/2020.ACL-MAIN.491

Krause J, Perer A, Ng K (2016) Interacting with predictions: Visual inspection of black-box machine learning models. In: Kaye J, Druin A, Lampe C, et al (eds) Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016. ACM, New York, pp 5686–5697. https://doi.org/10.1145/2858036.2858529

Krause J, Perer A, Bertini E (2018) A user study on the effect of aggregating explanations for interpreting machine learning models. In: ACM KDD Workshop on Interactive Data Exploration and Analytics. ACM, New York

Lattimore T, Szepesvári C (2020) Bandit algorithms. Cambridge University Press, Cambridge
Lazar J, Feng JH, Hochheiser H (2017) Research methods in human-computer interaction. Morgan Kaufmann, Cambridge, MA
Leemans SJ, Tax N (2022) Causal reasoning over control-flow decisions in process models. In: International Conference on Advanced Information Systems Engineering. Springer, Cham, p 183–200
Leontjeva A, Conforti R, Di Francescomarino C et al (2015) Complex symbolic sequence encodings for predictive monitoring of business processes. In: Proc. of BPM 2015, LNCS, vol 9253. Springer, Cham, p 297–313
Lim BY, Dey AK, Avrahami D (2009) *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In: Proc. of CHI 2009. ACM, New York, pp 2119–2128. https://doi.org/10.1145/1518701.1519023
Lipton P (2001) What good is an explanation? In: Hon G, Rakover S (eds) Explanation. Springer Verlag, Heidelberg, Germany, pp 43–59
Lu R, Sadiq SW (2007). On the discovery of preferred work practice through business process variants. https://doi.org/10.1007/978-3-540-75563-0_13
Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S et al (eds) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017. Springer, Cham, p 4765–4774
Maggi FM, Di Ciccio C, Di Francescomarino C et al (2018) Parallel algorithms for the automated discovery of declarative process models. Inf Syst 74(Part):136–152. https://doi.org/10.1016/j.is.2017.12.002
Malhi A, Knapic S, Främling K (2020) Explainable agents for less bias in human-agent decision making. In: Calvaresi D, Najjar A, Winikoff M et al (eds) Explainable, Transparent Autonomous Agents and Multi-Agent Systems - Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9-13, 2020, Revised Selected Papers, Lecture Notes in Computer Science, vol 12175. Springer, Cham, pp 129–146. https://doi.org/10.1007/978-3-030-51924-7_8
Mason M, et al (2010) Sample size and saturation in PhD studies using qualitative interviews. In: Forum qualitative Sozialforschung/Forum: qualitative social research, vol. 11, n. 3
Márquez-Chamorro AE, Resinas M, Ruiz-Cortés A (2018) Predictive monitoring of business processes: A survey. IEEE Trans on Services Comp 11(6):962–977. https://doi.org/10.1109/TSC.2017.2772256
Munoz-Gama J et al (2022) Process mining for healthcare: Characteristics and challenges. J Biomed Inform 103994. https://doi.org/10.1016/j.jbi.2022.103994
Nielsen J, Landauer TK (1993) A mathematical model of the finding of usability problems. In: Human-Computer Interaction, INTERACT '93, IFIP TC13 International Conference on Human-Computer Interaction, 24-29 April 1993, Amsterdam, The Netherlands. ACM, New York, pp 206–213
Nolte A, Prilla M (2013) Anyone can use models: Potentials, requirements and support for non-expert model interaction. Int J e-Collab (IJeC) 9(4):45–60
Özalp E, Hartwig K, Reuter C (2023) Trends in explainable artificial intelligence for non-experts. KI-Kritik/AI Critique 4:223
Pesic M (2008) Constraint-based workflow management systems : shifting control to users. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR638413
Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola AJ et al (eds) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, New York, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
Ribeiro MT, Singh S, Guestrin C (2018) Anchors: High-precision model-agnostic explanations. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press, Whashington D.C, pp 1527–1535. https://doi.org/10.1609/AAAI.V32I1.11491
Rizzi W, Simonetto L, Francescomarino CD et al (2019) Nirdizati 2.0: New features and redesigned backend. In: Depaire B, Smedt JD, Dumas M et al (eds) Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 co-located with 17th International Conference on Business Process Management, BPM 2019, Vienna, Austria, September 1-6, 2019. Springer, Cham, p 154–158
Rizzi W, Francescomarino CD, Maggi FM (2020) Explainability in predictive process monitoring: When understanding helps improving. In: Fahland D, Ghidini C, Becker J et al (eds) Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings, Lecture Notes in Business Information Processing, vol 392. Springer, Cham, pp 141–158. https://doi.org/10.1007/978-3-030-58638-6_9
Rizzi W, Comuzzi M, Di Francescomarino C et al (2022) Explainable predictive process monitoring: A user evaluation. https://arxiv.org/abs/2202.07760. Accessed 1 June 2024
Shapley LS (2016) 17. A Value for n-Person Games. Princeton University Press, Princeton, pp 307–318. https://doi.org/10.1515/9781400881970-018
Sindhgatta R, Moreira C, Ouyang C et al (2020) Exploring interpretable predictive models for business processes. In: Fahland D, Ghidini C, Becker J et al (eds) Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings, Lecture Notes in Computer Science, vol 12168. Springer, Cham, pp 257–272. https://doi.org/10.1007/978-3-030-58666-9_15
Spinner T, Schlegel U, Schäfer H et al (2020) explainer: A visual analytics framework for interactive and explainable machine learning. IEEE Trans Vis Comput Graph 26(1):1064–1074. https://doi.org/10.1109/TVCG.2019.2934629
Sutton RS, Barto AG (1998) Reinforcement learning - an introduction. Adaptive computation and machine learning. MIT Press, Cambridge
Teinemaa I, Dumas M, Leontjeva A et al (2018) Temporal stability in predictive process monitoring. Data Min Knowl Discov 32(5):1306–1338. https://doi.org/10.1007/s10618-018-0575-9
van der Aa H, Balder KJ, Maggi FM et al (2020) Say it in your own words: Defining declarative process models using speech recognition. In: International Conference on Business Process Management. Springer, Cham, pp 51–67

van der Aalst WMP et al (2011) Process mining manifesto. In: Daniel F, Barkaoui K, Dustdar S (eds) Business Process Management Workshops - BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I, Lecture Notes in Business Information Processing, vol 99. Springer, Berlin, pp 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

Velmurugan M, Ouyang C, Moreira C et al (2021a) Evaluating fidelity of explainable methods for predictive process analytics. In: Nurcan S, Korthaus A (eds) Intelligent Information Systems - CAiSE Forum 2021, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings, Lecture Notes in Business Information Processing, vol 424. Springer, Cham, pp 64–72. https://doi.org/10.1007/978-3-030-79108-7_8

Velmurugan M, Ouyang C, Moreira C et al (2021b) Evaluating stability of post-hoc explanations for business process predictions. In: Hacid H, Kao O, Mecella M et al (eds) Service-Oriented Computing - 19th International Conference, ICSOC 2021, Virtual Event, November 22-25, 2021, Proceedings, Lecture Notes in Computer Science, vol 13121. Springer, Cham, pp 49–64. https://doi.org/10.1007/978-3-030-91431-8_4

Wang X, Yin M (2021) Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In: Hammond T, Verbert K, Parra D et al (eds) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. ACM, New York, pp 318–328. https://doi.org/10.1145/3397481.3450650

Weinzierl S, Dunzer S, Zilker S et al (2020a) Prescriptive business process monitoring for recommending next best actions. In: Fahland D, Ghidini C, Becker J et al (eds) Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings, Lecture Notes in Business Information Processing, vol 392. Springer, Cham, pp 193–209. https://doi.org/10.1007/978-3-030-58638-6_12

Weinzierl S, Zilker S, Brunk J et al (2020b) XNAP: making lstm-based next activity predictions explainable by using LRP. In: del-Río-Ortega A, Leopold H, Santoro FM (eds) Business Process Management Workshops - BPM 2020 International Workshops, Seville, Spain, September 13-18, 2020, Revised Selected Papers, Lecture Notes in Business Information Processing, vol 397. Springer, Cham, pp 129–141. https://doi.org/10.1007/978-3-030-66498-5_10

## Publisher's Note