Predictive Process Monitoring Methods: Which One Suits Me Best?

Chiara Di Francescomarino[1], Chiara Ghidini[1],
Fabrizio Maria Maggi[2]*, and Fredrik Milani[2]*

[1] FBK-IRST, Via Sommarive 18, 38050 Trento, Italy
{dfmchiara, ghidini}@fbk.eu
[2] University of Tartu, Liivi 2, 50409 Tartu, Estonia
{f.m.maggi, milani}@ut.ee

Abstract. Predictive process monitoring has recently gained traction in academia and is maturing also in companies. However, with the growing body of research, it might be daunting for data analysts to navigate through this domain in order to find, provided certain data, what can be predicted and what methods to use. The main objective of this paper is developing a value-driven framework for classifying predictive process monitoring methods. This objective is achieved by systematically reviewing existing work in this area. Starting from about 780 papers retrieved through a keyword-based search from electronic libraries and filtering them according to some exclusion criteria, 55 papers have been finally thoroughly analyzed and classified. Then, the review has been used to develop the value-driven framework that can support researchers and practitioners to navigate through the predictive process monitoring field and help them to find value and exploit the opportunities enabled by these analysis techniques.

Keywords: Predictive Process Monitoring, Process Mining, Value-Driven Framework

1 Introduction

Process mining is a family of methods to analyze business processes based on their observed behavior recorded in *event logs*. In this setting, an event log is a collection of *traces*, each representing one execution of the process (a.k.a. a *case*). A trace consists of a sequence of timestamped events, each capturing the execution of an activity. Each event may carry a payload consisting of attribute-value pairs such as the resource(s) involved in the execution of the activity, or other data recorded with the event. Since process mining is a relatively young discipline, the open challenges in this field are still many [34,1]. In particular, one of these challenges defined in [34] is about "providing operational support" and, in particular, about the definition of techniques for supporting the three main operational support activities, i.e., *detect*, *predict* and *recommend*. Very recently,

* This research is supported by the Estonian Research Council Grant IUT20-55.

researchers have started focusing on the development of techniques supporting the operational support activity *predict* a.k.a. predictive process monitoring techniques.

Predictive process monitoring [42] is a branch of process mining that aims at predicting at runtime and as early as possible the future development of ongoing cases of a process given their uncompleted traces. As demonstrated in this paper, recently, a wide literature about predictive process monitoring techniques has become available. This large number of studies has triggered the need to order and classify what has already been done, as well as to identify gaps in the current literature, thus supporting and guiding researchers towards the future advances in this research field. In addition, although selected companies have integrated predictive methods in their business processes [33], the potential for greater impact is still very real. However, due to the large availability of techniques, it might be daunting for companies to navigate through this domain. In the quest to find, provided certain data, what can be predicted and what methods to use, companies can easily get lost.

As such, this paper has the objective to develop, based on the review of existing research, a value-driven framework for classifying existing predictive process monitoring methods, thus supporting, on the one hand, researchers who need to crystallize the existing work to identify the next challenges of this research area, and, on the other hand, companies that need to be guided to find the best solutions within the wide literature available.

In particular, the main research question we want to answer is (**RQ**): "*How can the body of relevant academic publications within the field of predictive process monitoring be classified as a framework?*". This question is answered through a systematic literature review that aims at identifying the state of the art of predictive process monitoring. We conducted the literature review based on a standard protocol defined in [36]. Starting from a keyword-based search from electronic libraries, we identified around 780 papers and, going through multiple filtering rounds, we selected 55 papers that were analyzed and categorized. The review was then used to develop the value-driven framework that supports researchers and practitioners to identify the methods that fit their needs. For example, one of the discriminative characteristics of the classified methods is the type of prediction. Once defined the type of prediction, the input data required is taken into consideration. The next step is understanding the family of algorithms applied, whether tools are available, the domain in which they have been applied and so on. We started from 4 main dimensions that were, in our opinion, the most relevant in the categorization of the identified methods (type of prediction, type of input, family of algorithms and tool availability) and we expanded them while progressing with the literature review.

The remainder of the paper is organized as follows. Section 2 describes the background about predictive process monitoring, while Section 3 presents the literature review protocol and its results. Section 4 provides a classification of the identified methods. In Section 5, the framework is presented and discussed. Finally, Section 6 concludes the paper.

2 Predictive Process Monitoring

The execution of business processes is generally subject to internal policies, norms, best practices, regulations, and laws. For this reason, compliance monitoring is an everyday imperative in many organizations. Accordingly, a range of research proposals have addressed the problem of assessing whether a process execution complies with a set of requirements [41]. However, these monitoring approaches are *reactive*, in that they allow users to identify a violation only *after it has occurred* rather than supporting them in *preventing* such violations in the first place.

Based on an analysis of historical execution traces, *predictive process monitoring* methods [42] continuously provide the user with predictions about the future of a given ongoing process execution. The forward-looking nature of predictive monitoring provides organizations with new frontiers in the context of compliance monitoring. Indeed, knowing in advance about violations, deviances and delays in a process execution would allow them to take preventive measures (e.g., reallocating resources) in order to avoid the occurrence of those situations and, as often happens, to avoid money loss.

3 Systematic Literature Review Protocol

The systematic review protocol of our literature review specifies the research questions, the search protocol, and the selection criteria predominantly following the guidelines provided by Kitchenham et al. [36]. The work was divided into two phases. In the first phase, two researchers designed the review protocol. Here, the research questions were formulated, electronic databases identified, inclusion and exclusion criteria defined, and data extraction strategy formulated. The second phase was conducted by two other researchers who reviewed the protocol, ran the searches, filtered the list of papers, produced the final list of papers and extracted the data. The data extraction was initially independently conducted by the two researchers. The results were then compared and discussed. If differences were noted, they were reconciled first by discussions, then with the participation of the researchers involved in the first phase and, if needed, by contacting the author team of the paper in question.

The main research question (**RQ**): "*How can the body of relevant academic publications within the field of predictive process monitoring be classified as a framework?*" is decomposed in the following four sub-questions. The first research question seeks to identify the different aspects of business processes that can be predicted by means of predictive process monitoring techniques. As such, the first sub-research question is formulated as **RQ1**: "*what aspects of business processes do predictive process monitoring techniques predict?*". The next two research questions concern the algorithms employed in predictive process monitoring techniques. The second is **RQ2**: "*what input data do predictive process monitoring algorithms require?*". The third research questions is **RQ3**: "*what are the main families of algorithms used in predictive process monitoring techniques?*". Finally, the last research question focuses on tools for predictive process monitoring i.e., **RQ4**: "*what are the tools that support predictive process monitoring?*".

The first research question is motivated by the necessity of knowing what can be predicted when guiding companies or researchers in the selection of a predictive process monitoring technique. The second research question is motivated by the fact that, to apply a certain technique, different types of inputs can be required. It is therefore crucial to know what information is needed to run each technique. Most of the techniques require an event log as input, but the information contained in it is often requested to be provided at different levels of granularity. The third research question is mainly relevant for researchers for classifying existing methods in different families, which could help to understand their strengths and limitations, and to identify the next challenges in this field. Finally, tool support, investigated in the fourth research question, is relevant for running the proposed methods in academy and industry contexts. Starting from these dimensions that were, in our opinion, the most relevant for categorizing the existing predictive process monitoring methods, while analyzing the literature, some additional relevant aspects were identified to discriminate among the methods under examination such as the type of output, the metrics used for evaluating the algorithms and the domains where the methods were evaluated.

To answer these research questions, we defined a search string to query some electronic libraries. We followed the guidance given by [36], which resulted in using the keywords "predictive", "prediction", "business process", and "process mining". The keywords "predictive" and "prediction" were derived from the research questions. However, the literature on predictive analysis is vast and encompasses areas outside of the domain of business processes. As such, we added the keyword "business process". Finally, predictive process monitoring concerns "process mining" and, therefore, this keyword was added. To retrieve as many results as possible, we intentionally left out additional keywords such as "monitoring", "technique", and "algorithm" so not to limit the search. The keywords were used to formulate the following boolean search string: ("predictive" OR "prediction") AND ("business process" OR "process mining").

Following the definition of the search string, the electronic libraries were chosen. The databases selected were Scopus, SpringerLink, IEEE Xplore, Science Direct, ACM Digital Library, and Web of Science. These were selected as they cover scientific publications within the field of computer science.

The results[1] were exported into an excel sheet for processing. The first filtering was for duplicates. Duplicate studies are those that appeared in more than one electronic database with identical title by the same author(s) [37]. Having performed this filtering, 779 papers were identified. Next, the list was filtered based on title of the study. Studies clearly out of scope were removed. In detail, we filtered (i) all documents that are not proper research papers, but rather editorial or book introductions (e.g., "25 Years of Applications of Logic Programming in Italy", or "Introduction to Process Mining"); (ii) all studies related to completely different research areas (e.g., "A RFID-based recursive process mining system for quality assurance in the garment industry", or "Process modeling and optimization of complex systems using Scattered Context Grammars"). After this filter, 186 papers remained. Position papers and papers published in workshops were excluded as their contribution is less mature as compared to

[1] The queries were run on October 20, 2017.

full paper published in conferences and journals. At this stage, 162 papers remained. We then filtered the papers by looking at the abstracts to further assess their relevance. Following this step, 77 papers remained. After this, we used the inclusion criterion "*does the study propose a novel algorithm or technique for predictive process monitoring?*". We found 50 papers that complied with this criterion. Finally, we added 5 relevant papers to these 50 papers, via a backward reference search.

From the final list of 55 papers, we extracted standard meta-data (title, authors, year of publication, number of citations, and type of publication). For each paper, the type of prediction considered was extracted in accordance with **RQ1**. Some methods can predict several aspects of a business process. In such cases, all aspects were recorded. Secondly, information about the input data was extracted. Process mining requires that logs contain at least a unique case id, activity names and timestamps. However, if a method requires additional data for analysis, this is extracted and noted (**RQ2**). Thirdly, we examined the type of algorithm. The underlying family of each identified predictive process monitoring method was extracted to address **RQ3**. We also extracted information about the validation of each method. The quality of a method depends indeed on its validation. As such, data about validation was extracted. In addition, if validated, data about if the log was synthetic or from real-life (including the industry domain) was extracted as well. Data about tool support was also extracted. This information considers if there is a plug-in or a standalone application for the proposed predictive process monitoring technique. This data relates to **RQ4**. In addition to the above data, the type of output, the metrics used for evaluating the algorithms and if the datasets used were public or not was annotated.[2]

4 Systematic Literature Review Results

The literature review reveals that the number of publications on predictive process monitoring has significantly increased in the last few years. By analyzing the meta-data extracted, we found that out of 55 papers analyzed 14 were published between 2006 and 2013, whereas 41 in the last 4 years (17 journals and 38 conferences in total).

The main dimension that is typically used to classify predictive process monitoring techniques is the type of prediction [21]. In the following sections, some of the research studies identified through our systematic literature review and characterizing three prediction type macro-categories, i.e., numeric, categorical and next activities predictions, are presented.

4.1 Numeric Predictions

We can roughly classify the studies dealing with numeric predictions in two groups, based on the specific type of predictions returned:

[2] The data extracted was entered into an excel sheet and is available for download at `https://docs.google.com/spreadsheets/d/1l1enKhKWx_3KqtnUgggrPl1aoJMhvmy9TF9jAM3snas/edit#gid=959800788`.

- time predictions;
- cost predictions;

Time predictions. The group of studies focusing on the time perspective is a rich group. Several studies, in this context, rely on explicit models. In [2], the authors present a set of approaches in which transition systems, which are built based on a given abstraction of the events in the event log, are annotated with time information extracted from the logs. In particular, information about elapsed, sojourn, and remaining time is reported for each state of the transition system. The information is then used for making predictions on the completion time of an ongoing trace. Further extensions of this approach are proposed in [53,54], where the authors apply machine learning techniques to annotate the transition systems. In detail, in [53], the transition systems are annotated with machine learning models such as Naïve Bayes and Support Vector Regression models. In [54], instead, the authors present, besides two completely new approaches based on Support Vector Regression, a refinement of the study in [53] that also takes into account data. Moreover, the authors evaluate the three proposed approaches both on stationary and non-stationary (i.e., characterized by evolving conditions) processes. Other extensions of the approach presented in [2] that also aim at predicting the remaining time of an ongoing trace, are the studies presented in [29,31]. In these studies, the annotated transition system is combined with a context-driven predictive clustering approach. The idea behind predictive clustering is that different scenarios can be characterized by different predictors. Moreover, contextual information is exploited in order to make predictions, together with control-flow [29] or control-flow and resources [31].

Another approach based on the extraction of (explicit) models (*sequence trees*) is presented in [10] to predict the completion time and the next future activity of a current ongoing case. Similarly to the predictive clustering approach, the sequence tree model allows for clustering traces with similar sequences of activities (control-flow) and to build a predictor model for each node of the sequence tree by leveraging data payload information. In [56], the authors use generally distributed transitions stochastic Petri nets (GDT-SPN) to predict the remaining execution time of a process instance. In detail, the approach takes as input a stochastic process model, which can be known in advance or inferred from historical data, an ongoing trace, and some other information as the current time in order to make predictions on the remaining time. In [57], the authors also exploit the elapsed time since the last event in order to make more accurate predictions on the remaining time and estimating the probability to miss a deadline.

Differently from the previous approaches, in [20], the authors only rely on the event log in order to make predictions. In detail, they develop an approach for predicting the remaining cycle time of a case by using non-parametric regression and leveraging activity duration and occurrences as well as other case-related data. In [4] and [11], the contextual clustering-based approach presented in [29,31] is updated in order to address the limitation of transition system-based approaches requiring the analyst to choose the log abstraction functions, by replacing the transition system predictor with standard regression algorithms. Moreover, in [11], the clustering component of the approach is further improved in order to address scalability and accuracy issues. In [49], Hidden Markov Models (HMM) are used for making predictions on the remaining time. A comparative evaluation shows that HMM provides more accurate results than annotated transition systems and regression models. In [59], *inter-case feature predictions* are introduced for predicting the completion time of an ongoing trace. The proposed approaches leverage not only the information related to the ongoing case, but also the status of other (concurrent) cases (e.g., the number of concurrent cases) in order to make predictions. The proposed encodings demonstrated an improvement of the results when applied to two real-life case studies.

A prediction type that is very close to time, but slightly different is the prediction of the delay of an ongoing case. In [62], queuing theory is used to predict possible online delays in business process executions. The authors propose approaches that either enhance traditional approaches based on transition systems, as the one in [2], to take queueing effects into account, or leverage properties of queue models.

Cost predictions. A second group of studies focuses on cost predictions. Also in this group, we can find studies explicitly relying on models as the study in [66]. In such a study, cost predictions are provided by leveraging a process model enhanced with costs (i.e., a frequent-sequence graph enhanced with costs) taking into account information about production, volume and time.

4.2 Categorical Predictions

The second family of prediction approaches predicts categorical values. In this settings, two main specific types of predictions can be identified:

- risk predictions;
- categorical outcome predictions.

Risk predictions. A first large group of studies falling under the umbrella of outcome-oriented predictions, deals with the prediction of risks.

Also in this case an important difference among state-of-the-art approaches is the existence of an explicit model guiding the prediction. For example, in [15], the authors present a technique for reducing process risks. The idea is supporting process participants in making risk-informed decisions, by providing them with predictions related to process executions. Decision trees are generated from logs of past process executions, by taking into account information related to data, resources and execution frequencies provided as input with the process model. The decision trees are then traversed and predictions about risks returned to the users. In [12] and [13], two extensions of the study in [15] are presented. In detail, in [12], the framework for risk-informed decisions is extended to scenarios in which multiple process instances run concurrently. In particular, in order to deal with the risks related to different instances of a process, a technique that uses integer linear programming is exploited to compute the optimal assignment of resources to tasks to be performed. In [13], the study in [15] is extended so that the process executions are not considered in isolation anymore, but, rather, the information about risks is automatically propagated to similar running instances of the same process in real-time in order to provide early runtime predictions.

In [45], three different approaches for the prediction of process instance constraint violations are investigated: machine learning, constraint satisfaction and QoS aggregation. The authors, beyond demonstrating that all the three approaches achieve good results, identify some differences and propose to combine them. Results on a real case study show that combining these techniques actually allows for improving the prediction accuracy.

Other studies, devoted to risk prediction, do not take into account explicit models. For instance, in [50], the authors make predictions about time-related process risks by identifying and leveraging process risk indicators (e.g., abnormal activity execution time or multiple activity repetition) by applying statistical methods to event logs. The indicators are then combined by means of a prediction function, which allows for highlighting the possibility of transgressing deadlines. In [51], the authors extend their previous study by introducing a method for configuring the process risk indicators. The method learns from the outcomes of completed cases the most suitable thresholds for the process risk indicators, thus taking into account the characteristics of the specific process and, therefore, improving the accuracy.

Categorical outcome predictions. A second group of predictions relates to the fulfillment of predicates. Almost all studies falling under this category do not rely on any explicit model. For example, in [42] a framework for predicting the fulfillment (or the violation) of a predicate in an ongoing execution, is introduced. Such a framework makes predictions by leveraging: (i) the sequence of events already performed in the case; and (ii) the data payload of the last activity of the ongoing case. The framework is able to provide accurate results, although it demands for a high runtime overhead. In order to overcome such a limitation, the framework has been enhanced in [18] by introducing a clustering preprocessing step in which cases sharing a similar behaviour are clustered together. A predictive model - a classifier - for each cluster is then trained with the data payload of the traces in the cluster. In [17], the framework is enhanced in order to support users by providing them with a tool for the selection of the techniques and the hyperparameters that best suit their datasets and needs.

In [40], the authors consider traces as complex symbolic sequences, that is, sequences of activities each carrying a data payload consisting of attribute-value pairs. By starting from this assumption, the authors focus on the comparison of different feature encoding approaches, ranging from traditional ones, such as counting the occurrences of activities and data attributes in each trace, up to more complex ones, relying on HMM. In [67], the approach in [40] is enhanced with clustering, by proposing a two-phase approach. In the first phase, prefixes of historical cases are encoded as complex symbolic sequences and clustered. In the second phase a classifier is built for each of the clusters. At runtime, (i) the cluster closest to the current ongoing process execution is identified; and (ii) the corresponding classifier is used for predicting the process instance outcome (e.g., whether the case is normal or deviant).

In [65], in order to improve prediction accuracy, unstructured (textual) information, contained in text messages exchanged during process executions, is also leveraged, together with control and data flow information. In particular, different combinations of text mining (bag-of-n-grams, Latent Dirichlet Allocation and Paragraph Vector) and classification (Random Forest and logistic regression) techniques have been proposed and exercised. In [48], an approach based on evolutionary algorithms is presented. The approach, which uses information related to a window of events in the event log, is based on the definition of process indicators (e.g., whether a process instance is completed on time, whether it is reopened). At training time, the process indicators are computed, the training event log encoded and the evolutionary algorithms applied for the generation of a predictive model composed of a set of decision rules. At runtime, the current trace prefix is matched against the decision rules in order to predict the correct class for the ongoing running instance.

4.3 Next Activities Predictions

A third more recent family of studies deals with predicting the sequence of the future activities and their payload given the activities observed so far, as in [54,64,22,19]. In [54], the authors propose an approach for predicting the sequence of future activities of a running case by relying on an annotated data-aware transition system, obtained as a refinement of the annotated transition system proposed in [2].

Other approaches, e.g., [22,24,64], make use of RNNs with LSTM (Recurrent Neural Networks with Long Short-Term Memory) cells. In particular, in [22,24], an RNN with two hidden layers trained with back propagation is presented, while, in [64], an LSTM and an encoding based on activities and timestamps is leveraged to provide predictions on the next activities and their timestamps. Finally, the study in [19] investigates how to take advantage of possibly existing a-priori knowledge for making predictions on the sequence of future activities. To this aim, an LSTM approach is equipped with the capability of taking into account also some given knowledge about the future development of an ongoing case.

5 Value-Driven Framework for Selecting Predictive Process Monitoring Methods

Table 1 and 2 report the devised framework.[3] By reading it from left to right, in the first column, we find the **`prediction type`**. Our review shows that the algorithms can be categorized according to six main types of high level categories of prediction types. The first category, `time prediction`, encompasses all the different aspects of process execution time such as `remaining time` or `delay`. The second main category of identified prediction types is related to `categorical outcome`(s). Such methods predict the probability of a certain predefined outcome, such as if a case will lead to a disruption, to the violation of a constraint, or whether it will be delayed. The third type of prediction type is related to `sequence of next outcomes/values`. These predictions focus on the

[3] For space limitations, in this article, an abridged version of the framework is presented. The complete version of the framework includes additional data and is available for download at `https://docs.google.com/spreadsheets/d/1l1enKhKWx_3KqtnUgggrPl1aoJMhvmy9TF9jAM3snas/edit#gid=959800788`.

Pred. type	Det. Pred. type	Input 1	Input 2	Input 3	Tool	Domain	Family of algorithm 1	Family of algorithm 2	Family of algorithm 3	Refer.
time	maint. time	event log (with timestamps)			N	automotive	time series	probabilistic model		[58]
	activity delays					financial telecomm.	queueing theory	transition system		[61,62]
	rem. time					telecomm.	stat. analysis			[6]
					ProM plugin	public admin.	transition system			[3,2]
						customer supp.	stochastic Petri net			[55]
			process model			financial customer supp.	regression	classification		[69]
		event log (with timestamps) with data			N	unspecified	pattern mining			[10]
					Y but unavail.	unspecified	classification	time series		[9]
					Y	financial public admin.	regression	classification		[68]
						healthcare	stochastic Petri net			[60]
					ProM plugin	public admin.	regression			[20]
						public admin.	transition system	regression	classification	[53]
						customer supp. public admin. financial	transition system	regression		[54]
						financial logistics	stochastic Petri net			[56,57]
			inter-case metrics		Y	healthcare manufacturing	regression			[59]
		event log (with timestamps) with data and contextual information			Y but unavail.	logistics	clustering	regression		[32]
					Y	logistics	clustering	pattern mining		[5]
								transition system		[31]
					ProM plugin	logistics	clustering	transition system		[30]
			labelling funct.	proc. model	ProM plugin	no validation	classification			[39]
categorical outcome	outcome	act. durations and routing probab.	threshold(s)	proc. model	N	synthetic	simulation	stat. analysis		[63]
		event log	labelling funct.		Y implem.	financial automotive	prob. automata			[7]
		event log (with timestamps) with data			N	logistics	neural network	constraint-sat.	QoS aggregation	[45]
			labelling funct.			healthcare	probab. automata	classification		[40]
						logistics	classification			[8]
						synthetic	classification			[35]
					Y but unavail.	synthetic	probab. automata			[38]
						synthetic	classification	neural network		[43]
						healthcare	clustering	classification		[18]
					Y	no valid.	stat. analysis			[25]
						logistics	stat. analysis			[47]
						financial public admin.	classification			[68]
					ProM pl.	healthcare	classification			[42]
						healthcare	clustering	classification		[17]
					ProM and Camunda pl.	automotive healthcare	evol. algorithm			[48]
			threshold(s)		Y but unavail.	unspecified	classification			[9]
		event log (with timestamps) with data and contextual information			Y	domotic	stat. analysis			[26]
			labelling funct.		Y but unavail.	healthcare	clustering	classification		[16]
				proc. model	ProM plugin	no validation	classification			[39]
			threshold(s)		ProM plugin	logistics	clustering	transition system		[29]
			clusters of behav.		N	logistics manufacturing	clustering	classification		[28]
		event log (with timestamps) with data and unstructured text	labelling funct.		N	financial	classification	text mining		[65]
	next activity	event log (with timestamps) with data			N	unspecified	pattern mining			[10]
					Y	domotic	stat. analysis			[27]
		event log (with timestamps) with data and contextual information			Y	domotic	stat. analysis			[26]
			labelling funct.	proc. model	ProM plugin	no valid.	classification			[39]
	last value of an attribute	event log (with timestamps) with data and contextual information	labelling funct.	proc. model	ProM plugin	no valid.	classification			[39]

Table 1. Predictive process monitoring framework: `time` and `categorical outcome` predictions

Pred. type	Det. Pred. type	Input 1	Input 2	Input 3	Tool	Domain	Family of algorithm 1	Family of algorithm 2	Family of algorithm 3	Refer.
sequence of outcomes/values	sequence of future activities	event log (with timestamps)			Y implem.	customer supp. financial public admin.	neural network			[64]
					Y but unavail.	financial automotive customer supp.	neural network			[44]
					Y	financial automotive	neural network			[23]
			backgr. knowledge		Y impl.	healthcare automotive financial public admin. customer supp.	neural network			[19]
					Y	financial automotive	neural network			[24]
		event log (with timestamps) with data			ProM plugin	customer supp. public admin. financial	neural network			[54]
		event log (with timestamps) with data and contextual information			Y but unavail.	logistics	clustering	regression		[32]
	sequence of future activity timestamps	event log (with timestamps)			N	customer supp. financial public admin.	neural network			[64]
			backgr. knowledge		Y impl.	healthcare automotive financial public admin. customer supp.	neural network			[19]
					Y	financial automotive	neural network			[24]
risk	risk	event log (with timestamps)			Y but unavail.	logistics	clustering	classification		[11]
			labelling funct.		Camunda pl.	financial	similarity-weight. graph	stat. analysis		[13]
		event log (with timestamps) with data	threshold(s)		N	transport logistics	neural network			[46]
					ProM plugin	financial logistics	stochastic Petri net			[57]
			labelling funct.	proc. model	Yawl pugin	no valid.	classification			[15]
						logistics unspecified	evol. algorithm			[14]
inter-case metr.	inter-case metrics	event log (with timestamps)			Y but unavail.	logistics	clustering	regression		[11]
		event log (with timestamps) with data	labelling funct.		ProM plugin	unspec.	regression			[52]
			threshold(s)		N	transport logistics	neural net.			[46]
						no valid.	classification	regression	time series	[71]
					Y but unavail.	unspec.	classification	time series		[71]
	workload	event log (with timestamps) with data and contextual information	labelling funct.		ProM plugin	no valid.	classification			[9]
cost	cost	event log (with timestamps) with data	threshold(s)		N	transport logistics	neural net.			[46]
		event log (with timestamps) with resources	cost schema		ProM plugin	no valid.	trans. system	stat.analysis		[70]

Table 2. Predictive process monitoring framework: `sequence of outcomes/values`, `risk`, `inter-case metrics`, `cost`

probability that future sets of events will occur in the execution of a case. The fourth prediction type is `risk`. When elimination of risks is not feasible, reducing and managing risks becomes important. The fifth prediction type pertains to `inter-case metrics`. The final category is related to `cost` predictions.

The next step (second column) in the framework concerns the **input data**. Event logs containing different types of information should be provided as input to the different methods (e.g., `event logs (with timestamps)`, `event logs (with timestamps) with data`. In some cases, together with the event log, other inputs are required. For instance, in case of the outcome-based predictions, the `labelling function`, e.g., the specific predicate or category to be predicted, is usually required.

The framework also considers the existence of **tool support**. If a tool has been developed, using, evaluating, and understanding the applicability, usefulness and potential benefits of a predictive process monitoring technique becomes easier. Given that tool support is provided, the framework captures the type of support provided such as whether the tool is a stand-alone application or a plug-in of a research framework, e.g., a `ProM plug-in`.

The validation of the algorithms on logs can take different forms. For instance, it can be achieved by using `synthetic logs`. Such validations can be considered as "weaker" as they do not necessarily mirror the complexity and variability of `real-life logs`. Algorithms tested on real-life logs reflect industry logs the best and are therefore, considered as "stronger". Furthermore, the suitability of an algorithm is better if validated on logs from the same industry domain as the one of the company seeking to use it. As such, the framework makes note of the **domain** from which the logs originate. When a domain is specified, it indicates that the algorithm has been tested on a log from that domain. If no domain is specified, the algorithm has not been validated on a real-life log.

At the heart of each predictive process monitoring method lies the specific algorithm used to implement it. The **family of algorithm** might matter when assessing advantages and limitations of an approach and as such, it is incorporated in the framework. The specific algorithm is not listed in the framework, but rather the foundational technique it is based on, such as `regression`, `neural networks`, or `queuing theory`.

The proposed framework has two main benefits. First, it can be used by companies to identify, along the above outlined parameters, the most suitable predictive process monitoring method(s) to be used in different scenarios. Second, it can be used by researchers to have a clear structuration and assessment of the existing techniques in the predictive process monitoring field. This assessment is crucial to identify gaps in the literature and relevant research directions to be further investigated in the near future. For example, only one paper discusses techniques that take advantage of possibly existing a-priori knowledge for making predictions [19]. Further investigation is also needed for what concerns the use of incremental learning algorithms as a way to incrementally construct predictive models by updating them whenever new cases become available, which is a crucial topic, but only discussed in [43]. Another direction for future research is to further investigate the use of inter-case features for constructing a predictive model. This means that scenarios should be taken into consideration where not only the information related to the ongoing case, but also the status of other (concurrent) cases (e.g., the number of concurrent cases) are considered in order to make predictions (this type of techniques is only discussed in [59]).

The main threat to validity of our work refers to the potential selection bias and inaccuracies in data extraction and analysis typical of literature reviews. In order to minimize such issues, our systematic literature review carefully adheres to the guidelines outlined in [36]. Concretely, we used well-known literature sources and libraries in information technology to extract relevant works on the topic of predictive process monitoring. Further, we performed a backward reference search to avoid the exclusion of potentially relevant papers. Finally, to avoid that our review was threatened by insufficient reliability, we ensured that the search process could be replicated by other researchers. However, the search may produce different results as the algorithm used by source libraries to rank results based on relevance may be updated.

6 Conclusion

Predictive process monitoring approaches have been growing quite fast in the last few years. If, on the one hand, such a spread of techniques has provided researchers and practitioners with powerful means for analyzing their business processes and making predictions on their future, on the other hand, it could be difficult for them to navigate through such a complex and unknown domain. By means of a systematic literature review in the predictive process monitoring field, we provide data analysts with a framework to guide them in the selection of the technique that best fit their needs.

In the future, we plan to empirically evaluate the proposed framework with users in order to assess its usefulness in real contexts. Furthermore, we would like to extend the existing framework with other dimensions of interest for the academic and the industrial world. The presented literature review can indeed be considered as a basis where to incorporate broader theoretical perspectives and concepts across the predictive process monitoring domain that might help future research endeavors to be well-directed.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016)
2. van der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Information Systems 36(2), 450–475 (2011)
3. van der Aalst, W.M.P., Pesic, M., Song, M.: Beyond process mining: From the past to present and future. In: CAiSE 2010 (2010)
4. Bevacqua, A., Carnuccio, M., Folino, F., Guarascio, M., Pontieri, L.: A data-adaptive trace abstraction approach to the prediction of business process performances. In: ICEIS (1). SciTePress (2013)
5. Bevacqua, A., Carnuccio, M., Folino, F., Guarascio, M., Pontieri, L.: A data-driven prediction framework for analyzing and monitoring business process performances. In: Enterprise Information Systems. pp. 100–117 (2014)
6. Bolt, A., Sepúlveda, M.: Process remaining time prediction using query catalogs. In: Lohmann, N., Song, M., Wohed, P. (eds.) BPM Workshops (2014)

7. Breuker, D., Matzner, M., Delfmann, P., Becker, J.: Comprehensible predictive models for business processes. MIS Q. 40(4), 1009–1034 (Dec 2016)
8. Cabanillas, C., Di Ciccio, C., Mendling, J., Baumgrass, A.: Predictive Task Monitoring for Business Processes. Springer International Publishing (2014)
9. Castellanos, M., Salazar, N., Casati, F., Dayal, U., Shan, M.C.: Predictive Business Operations Management. Springer (2005)
10. Ceci, M., Lanotte, P.F., Fumarola, F., Cavallo, D.P., Malerba, D.: Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining. Springer International Publishing (2014)
11. Cesario, E., Folino, F., Guarascio, M., Pontieri, L.: A Cloud-Based Prediction Framework for Analyzing Business Process Performances. Springer (2016)
12. Conforti, R., de Leoni, M., La Rosa, M., van der Aalst, W.M.P., ter Hofstede, A.H.M.: A recommendation system for predicting risks across multiple business process instances. Decision Support Systems 69 (2015)
13. Conforti, R., Fink, S., Manderscheid, J., Röglinger, M.: Prism – a predictive risk monitoring approach for business processes. In: BPM. pp. 383–400. Cham (2016)
14. Conforti, R., ter Hofstede, A.H.M., La Rosa, M., Adams, M.: Automated risk mitigation in business processes. In: On the Move to Meaningful Internet Systems: OTM 2012. pp. 212–231 (2012)
15. Conforti, R., de Leoni, M., La Rosa, M., van der Aalst, W.M.P.: Supporting risk-informed decisions during business process execution. In: CAiSE (2013)
16. Cuzzocrea, A., Folino, F., Guarascio, M., Pontieri, L.: A multi-v2016a multi-view multi-dimensional ensemble learning approach to mining business process deviances. In: IJCNN (2016)
17. Di Francescomarino, C., Dumas, M., Federici, M., Ghidini, C., Maggi, F.M., Rizzi, W.: Predictive business process monitoring framework with hyperparameter optimization. In: CAiSE 2016. pp. 361–376 (2016)
18. Di Francescomarino, C., Dumas, M., Maggi, F.M., Teinemaa, I.: Clustering-based predictive process monitoring. IEEE Trans. on Services Computing PP(99) (2016)
19. Di Francescomarino, C., Ghidini, C., Maggi, F.M., Petrucci, G., Yeshchenko, A.: An eye into the future: Leveraging a-priori knowledge in predictive business process monitoring. In: BPM. Springer (2017)
20. van Dongen, B.F., Crooy, R.A., van der Aalst, W.M.P.: Cycle time prediction: When will this case finally be finished? In: OTM 2008. pp. 319–336 (2008)
21. Dumas, M., Maggi, F.M.: Enabling process innovation via deviance mining and predictive monitoring. In: BPM - Driving Innovation in a Digital World, pp. 145–154 (2015), https://doi.org/10.1007/978-3-319-14430-6_10
22. Evermann, J., Rehse, J.R., Fettke, P.: A deep learning approach for predicting process behaviour at runtime. In: PRAISE-2016 (2016)
23. Evermann, J., Rehse, J.R., Fettke, P.: A deep learning approach for predicting process behaviour at runtime. In: BPM Workshops. Springer (2017)
24. Evermann, J., Rehse, J.R., Fettke, P.: Predicting process behaviour using deep learning. Decision Support Systems (2017)
25. Feldman, Z., Fournier, F., Franklin, R., Metzger, A.: Proactive event processing in action: A case study on the proactive management of transport processes (industry article). In: ACM DEBS (2013)
26. Ferilli, S., Esposito, F., Redavid, D., Angelastro, S.: Predicting process behavior in WoMan. In: AI*IA 2016 Advances in Artificial Intelligence (2016)
27. Ferilli, S., Esposito, F., Redavid, D., Angelastro, S.: Extended process models for activity prediction. In: Foundations of Intelligent Systems. Springer (2017)
28. Folino, F., Greco, G., Guzzo, A., Pontieri, L.: Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction. Data & Knowledge Engineering 70(12) (2011)

29. Folino, F., Guarascio, M., Pontieri, L.: Discovering context-aware models for predicting business process performances. In: OTM. Springer (2012)
30. Folino, F., Guarascio, M., Pontieri, L.: Context-aware predictions on business processes: An ensemble-based solution. In: New Frontiers in Mining Complex Patterns. pp. 215–229 (2013)
31. Folino, F., Guarascio, M., Pontieri, L.: Discovering High-Level Performance Models for Ticket Resolution Processes. Springer (2013)
32. Folino, F., Guarascio, M., Pontieri, L.: Mining predictive process models out of low-level multidimensional logs. In: CAiSE (2014)
33. Halper, F.: Predictive analytics for business advantage. TDWI Research (2014)
34. IEEE Task Force on Process Mining: Process mining manifesto. In: Proc. BPM Workshops. Lecture Notes in Business Information Processing, vol. 99, pp. 169–194. Springer (2011)
35. Kang, B., Kim, D., Kang, S.H.: Real-time business process monitoring method for prediction of abnormal termination using knni-based lof prediction. Expert Syst. Appl. 39(5) (Apr 2012)
36. Kitchenham, B.: Procedures for performing systematic reviews. Keele, UK, Keele University 33(2004), 1–26 (2004)
37. Kofod-Petersen, A.: How to do a structured literature review in computer science. Ver. 0.1. October 1 (2012)
38. Lakshmanan, G.T., Shamsi, D., Doganata, Y.N., Unuvar, M., Khalaf, R.: A markov prediction model for data-driven semi-structured business processes. Knowl. Inf. Syst. 42(1) (Jan 2015)
39. de Leoni, M., van der Aalst, W.M.P., Dees, M.: A general framework for correlating business process characteristics. In: Business Process Management. Springer (2014)
40. Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., Maggi, F.M.: Complex symbolic sequence encodings for predictive monitoring of business processes. In: BPM 2015. Springer (2015)
41. Ly, L.T., Maggi, F.M., Montali, M., Rinderle-Ma, S., van der Aalst, W.M.P.: Compliance monitoring in business processes: Functionalities, application, and tool-support. Inf. Syst. 54, 209–234 (2015)
42. Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: Proc. CAiSE 2014 (2014)
43. Maisenbacher, M., Weidlich, M.: Handling concept drift in predictive process monitoring. In: IEEE SCC. pp. 1–8. IEEE Computer Society (2017)
44. Mehdiyev, N., Evermann, J., Fettke, P.: A multi-stage deep learning approach for business process event prediction. In: CBI. vol. 01 (July 2017)
45. Metzger, A., Leitner, P., Ivanovi, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S., Pohl, K.: Comparing and combining predictive business process monitoring techniques. IEEE Trans. on Systems, Man, and Cybernetics: Systems 45(2) (2015)
46. Metzger, A., Föcker, F.: Predictive business process monitoring considering reliability estimates. In: Dubois, E., Pohl, K. (eds.) Advanced Information Systems Engineering. pp. 445–460. Springer International Publishing, Cham (2017)
47. Metzger, A., Franklin, R., Engel, Y.: Predictive monitoring of heterogeneous service-oriented business networks: The transport and logistics case. In: Proc. of SRII. SRII '12 (2012)
48. Mrquez-Chamorro, A.E., Resinas, M., Ruiz-Corts, A., Toro, M.: Run-time prediction of business process indicators using evolutionary decision rules. Expert Systems with Applications 87 (2017)
49. Pandey, S., Nepal, S., Chen, S.: A test-bed for the evaluation of business process prediction techniques. In: 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom) (Oct 2011)

50. Pika, A., van der Aalst, W.M.P., Fidge, C.J., ter Hofstede, A.H.M., Wynn, M.T.: Predicting Deadline Transgressions Using Event Logs (2013)
51. Pika, A., van der Aalst, W.M.P., Fidge, C.J., ter Hofstede, A.H.M., Wynn, M.T.: Profiling Event Logs to Configure Risk Indicators for Process Delays (2013)
52. Pika, A., van der Aalst, W.M.P., Wynn, M.T., Fidge, C.J., ter Hofstede, A.H.M.: Evaluating and predicting overall process risk using event logs. Information Sciences 352-353, 98 – 120 (2016)
53. Polato, M., Sperduti, A., Burattin, A., de Leoni, M.: Data-aware remaining time prediction of business process instances. In: 2014 International Joint Conference on Neural Networks (IJCNN) (July 2014)
54. Polato, M., Sperduti, A., Burattin, A., de Leoni, M.: Time and activity sequence prediction of business process instances. Computing (Feb 2018)
55. Rogge-Solti, A., Vana, L., Mendling, J.: Time series Petri net models - enrichment and prediction. In: SIMPDA 2015. CEUR Workshop Proceedings (December 2015)
56. Rogge-Solti, A., Weske, M.: Prediction of remaining service execution time using stochastic Petri nets with arbitrary firing delays. In: ICSOC. pp. 389–403 (2013)
57. Rogge-Solti, A., Weske, M.: Prediction of business process durations using non-markovian stochastic Petri nets. Information Systems 54 (2015)
58. Ruschel, E., Santos, E.A.P., de Freitas Rocha Loures, E.: Mining shop-floor data for preventive maintenance management: Integrating probabilistic and predictive models. Procedia Manufacturing 11 (2017)
59. Senderovich, A., Di Francescomarino, C., Ghidini, C., Jorbina, K., Maggi, F.M.: Intra and Inter-case Features in Predictive Process Monitoring: A Tale of Two Dimensions. Springer (2017)
60. Senderovich, A., Shleyfman, A., Weidlich, M., Gal, A., Mandelbaum, A.: P$\hat{3}$-folder: Optimal model simplification for improving accuracy in process performance prediction. In: Business Process Management. Springer (2016)
61. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining – predicting delays in service processes. In: CAiSE 2014. pp. 42–57 (2014)
62. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining for delay prediction in multi-class service processes. Inf. Syst. 53 (2015)
63. Si, Y.W., Hoi, K.K., Biuk-Aghai, R.P., Fong, S., Zhang, D.: Run-based exception prediction for workflows. Journal of Systems and Software 113 (2016)
64. Tax, N., Verenich, I., La Rosa, M., Dumas, M.: Predictive business process monitoring with LSTM neural networks. In: Proc. CAiSE 2017 (2017)
65. Teinemaa, I., Dumas, M., Maggi, F.M., Di Francescomarino, C.: Predictive business process monitoring with structured and unstructured data. In: BPM 2016 (2016)
66. Tu, T.B.H., Song, M.: Analysis and prediction cost of manufacturing process based on process mining. In: ICIMSA (May 2016)
67. Verenich, I., Dumas, M., La Rosa, M., Maggi, F.M., Di Francescomarino, C.: Complex Symbolic Sequence Clustering and Multiple Classifiers for Predictive Process Monitoring. Springer (2016)
68. Verenich, I., Dumas, M., La Rosa, M., Maggi, F.M., Di Francescomarino, C.: Minimizing overprocessing waste in business processes via predictive activity ordering. In: CAiSE 2016. Springer (2016)
69. Verenich, I., Nguyen, H., La Rosa, M., Dumas, M.: White-box prediction of process performance indicators via flow analysis. In: Proceedings of the 2017 International Conference on Software and System Process. ICSSP 2017 (2017)
70. Wynn, M.T., Low, W.Z., ter Hofstede, A.H.M., Nauta, W.: A framework for cost-aware process management: Cost reporting and cost prediction. Journal of Universal Computer Science 20(3), 406–430 (2014)
71. Zeng, L., Lingenfelder, C., Lei, H., Chang, H.: Event-Driven Quality of Service Prediction, pp. 147–161 (2008)