Information Systems and e-Business Management (2023) 21:193–227
https://doi.org/10.1007/s10257-022-00577-0

ORIGINAL ARTICLE

Beyond descriptive taxonomies in data analytics: a systematic evaluation approach for data-driven method pipelines

Patrick Zschech[1]

Received: 4 August 2021 / Revised: 22 August 2022 / Accepted: 12 October 2022 /
Published online: 11 November 2022

Abstract
Taxonomies can serve as a valuable tool to capture dimensions and characteristics of data analytics solutions in a structured manner and thus create transparency about different design options of the technical solution space. However, previous taxonomic approaches often remain at a purely descriptive level without leveraging morphological structures to investigate the mechanisms between different combinatorial options given in data analytics pipelines. To this end, we propose a taxonomic evaluation approach to evaluate and construct the technical core of analytical information systems more systematically. Specifically, we present a rough guidance model consisting of four steps, which we subsequently instantiate with two application scenarios from the fields of industrial maintenance and predictive business process monitoring. In this way, we demonstrate how taxonomic frameworks can guide the creation of structured evaluation studies to consider the construction and assessment of data analytics pipelines in a multi-perspective and holistic manner. Our approach is sufficiently generic to be applied to various domains, scenarios, and decision support tasks.

Keywords Data analytics · Taxonomy development · Evaluation framework · Ablation and substitution studies · Predictive maintenance · Predictive business process monitoring

1 Introduction

Over the past decades, analytical information systems (IS) have become an indispensable anchor for many organizations and our daily life. They support medical staff in diagnosing hard-to-find diseases (McKinney et al. 2020), prevent failures

✉ Patrick Zschech
patrick.zschech@fau.de

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

and breakdowns in manufacturing environments (Kraus and Feuerriegel 2019), monitor business processes for proactive resource allocation (Heinrich et al. 2021), and recommend products and services based on customer preferences (Li et al. 2020)—just to name a few examples.

The technical core of analytical information systems consists of *data-driven method pipelines*. They specify, for example, (i) which datasets are selected, processed, and analyzed, (ii) which data preparation steps are performed, and (iii) which data-driven methods from the fields of statistics and machine learning (ML) are used to build analytical models for data-driven decision support (Janiesch et al. 2021; Michalczyk and Scheu 2020). However, the construction and evaluation of such pipelines is a challenging endeavor as there are often multiple options to choose from. This may involve the choice of data types and data attributes depending on the domain peculiarities or the choice and specification of different data pre-processing techniques. Likewise, the field is characterized by continuous algorithmic innovations from computer science and engineering disciplines, which constantly produce new analytical models and methods, such as deep neural networks, for which a variety of network architectures have been proposed (Janiesch et al. 2021; Leijnen and Veen 2020; Manyika et al. 2011). When facing a multitude of design options, it is crucial to understand their impact on the overall pipeline and to identify interaction effects when combining different pipeline components in order to develop and deploy effective information systems. Besides, there is often no "one-fits-all" approach that proves to be the best solution across different circumstances (Flath and Stein 2018). Instead, it requires a profound evaluation to determine a promising pipeline with its specific components for each distinct situation.

In practice, however, the construction and evaluation of data-driven method pipelines are often performed in several trial-and-error cycles (Janardhanan 2020). Although it is nowadays a widely accepted standard to follow structured procedure models such as CRISP-DM to divide the pipeline development into structured sub-components (Mariscal et al. 2010; Michalczyk and Scheu 2020), this still does not allow for a transparent representation of the different design options and their impact on the pipeline. At the same time, it is becoming increasingly common in the field of ML to conduct so-called *ablation and substitution studies.* They are performed to systematically examine the effect of individual building blocks in ML-based pipelines (Cohen and Howe 1988; Sheikholeslami et al. 2021). Nevertheless, their scope is often restricted to the examination of limited model-centric parameters (e.g., architectural components of neural networks), without considering broader contextual aspects (e.g., domain specifics and data properties). Furthermore, running experiments with multiple different pipelines based on various ML models and methods is costly as it requires large computing times and resources, especially in data-intensive domains. Therefore, a well-defined setting is crucial to systematically investigate the effects of different design options along the overall pipeline in a sustainable manner without wasting resources in redundant experiments. Another challenge is that relevant knowledge for the pipeline construction is often spread over multiple involved stakeholders, such as domain experts, data engineers, and modeling specialists

(Hesenius et al. 2019; Zschech et al. 2020). As a result, there is a high risk that pipeline components and specifications are chosen based on individual experience and the subjective background of the respective pipeline developer.

As a remedy, we propose a systematic evaluation approach for data-driven method pipelines to construct and evaluate the technical core of analytical information systems more comprehensively and systematically. The goal of our approach is to establish structured frameworks which can capture different design options along data-driven pipelines and guide the preparation and execution of well-structured evaluation studies based on different framework configurations. With this approach, we adopt the general ideas of *data analytics procedure models* as well as *ablation and substitution studies*, which we combine on a more holistic level. This can provide orientation to researchers and practitioners alike by organizing a broad solution space in a systematic and transparent way. More specifically, our proposed approach aims at supporting data analytics teams consisting of data science scholars as well as ML engineers and pipeline developers. The former group of data science scholars is primarily supported by providing guidance for structured framework developments to organize the solution space. The latter group of ML engineers and pipeline developers can use the derived framework elements to perform systematic evaluation studies based on different framework configurations to obtain prescriptive insights for promising pipeline specifications.

As a methodical basis, our research draws on the pivotal instrument of *taxonomic frameworks*, which are currently gaining momentum in the IS community (Szopinski et al. 2019). In general, taxonomies serve as a viable approach for organizing knowledge in a structured manner so that researchers and practitioners can study the relationship among concepts to analyze and understand complex domains (Gregor 2006). In this context, there have also been several research efforts in data analytics that use taxonomies to systematize components, methods, and applications of analytical information systems in various domains and contexts (e.g., Krieger and Drews 2018; Wambsganss et al. 2021; Wanner et al. 2022). However, previous approaches mostly remain at a purely descriptive level and do not leverage taxonomic structures to investigate the mechanisms between different combinatorial options given in data-driven analytical pipelines. In other words, they do not use different configurations of the taxonomy elements to investigate their impact on the overall pipeline performance using quantitative evaluation metrics to derive prescriptive insights about promising pipeline configurations. This is where we contribute to the field by addressing the following research question:

RQ How can we create and apply taxonomic evaluation frameworks that guide the preparation and execution of systematic evaluation studies for data-driven method pipelines based on different pipeline configurations?

To address this question, we integrate the approach of *taxonomy development* (Nickerson et al. 2013) into a broader methodical framework so that it can be used

for the systematic evaluation of data analytics pipelines. For this purpose, we propose a rough sequence of four guiding steps, which we subsequently instantiate for demonstration purposes to showcase our approach's overall feasibility. More specifically, we use two application scenarios. The first one covers a scenario in the realm of *industrial maintenance* and the second one stems from the field of *business process monitoring*.

The remaining paper is organized as follows. In Sect. 2, we introduce the necessary foundations and refer to related work. In Sect. 3, we reflect on our research approach and describe the process of how our method proposal was assembled. Subsequently, we outline our proposal in Sect. 4, followed by a thorough demonstration based on the mentioned application scenarios in Sect. 5. We then proceed to discuss the results in terms of the achieved contribution and limitations in Sect. 6. Finally, we conclude our paper and provide an outlook for future work in Sect. 7.

2 Foundations and related work

In this section, we describe the necessary foundations for our method proposal. Thus, we first provide a brief understanding of systematic procedure models in the field of data analytics and introduce the idea behind ablation and substitution studies. Subsequently, we refer to previous work on taxonomy developments in IS research and data analytics.

2.1 Construction and evaluation of data analytics pipelines

Procedure models generally organize tasks or activities of construction and implementation processes into structured, logically arranged steps in which corresponding methods and techniques are applied. In the realm of analytical information systems, several such procedure models have been developed to provide instructions for all relevant phases specific to the construction of data analytics pipelines (Mariscal et al. 2010). Prominent examples are the CRISP-DM methodology (cross-industry standard process for data mining) (Wirth and Hipp 2000) and the KDD (knowledge discovery in databases) process model (Fayyad et al. 1996). They offer generic guidance across different domains and basically consist of the following steps: (i) *domain understanding* (i.e., gathering task and domain characteristics), (ii) *data understanding* (i.e., gathering data-related context characteristics), (iii) *data preparation* (i.e., applying methods to bring data assets into a suitable form), (iv) *modeling* (i.e., applying analytical methods/models), and (v) *evaluation* (i.e., assessing the quality of the overall pipeline with suitable metrics). Due to their domain independence, such procedure models can be applied as structural guidance in a wide variety of contexts. At the same time, however, they can be criticized for being too generic. Hence, they do not provide sufficient guidance which design options need to be considered when constructing analytical information systems for specific decision support tasks.

Ablation and substitution studies are another useful aid for the construction and evaluation of data analytics pipelines, especially when working with ML models. Their goal is to examine the contribution and effects of individual building blocks on the performance of complex systems by removing or replacing these building blocks (Cohen and Howe 1988). This examination usually involves model-centric components, such as architectural layers or neurons of deep neural networks, as well as data-centric components in the form of dataset features that a model is being trained on. Beyond that, however, basically any design choice or module of a pipeline can be considered as an ablatable or substitutable component (Meyes et al. 2019; Sheikholeslami et al. 2021). Even though conducting ablation and substitution studies seems to be an intuitive and simple practice to identify and assess critical design choices in data analytics pipelines, it is still not part of standard practices and has only recently begun to attract increasing interest in research and industry (Sheikholeslami et al. 2021).

2.2 Taxonomy development in information systems research and data analytics

Taxonomies play an essential role in IS research. They provide a structure to organize knowledge of a specific field, help to understand and analyze complex domains, and enable researchers to study the relationship among concepts (Nickerson et al. 2013). For this reason, a growing number of IS researchers are dedicated to the development of taxonomies in different sub-disciplines (see Oberländer et al. 2019 and Szopinski et al. 2019 for an overview).

Similarly, there is a growing interest in the subfield of data analytics to organize the technological and organizational facets of analytical information systems, methods, and applications into structured sub-components. For example, Wanner et al. (2022) developed a taxonomy based on a corpus of 904 data analytics articles to structure dimensions and characteristics of smart manufacturing applications. Subsequently, they used the resulting framework elements to identify and describe different archetypes using a cluster analysis. A similar approach was pursued by Matschak et al. (2022) for the field of ML-based fraud detection systems. Based on 54 publications, the authors derived a taxonomic scheme with salient design characteristics, which were subsequently used to identify archetypal design patterns using a cluster analysis. By contrast, Krieger and Drews (2018) proposed a taxonomic framework for classifying big data analytics applications in auditing. They examined twelve use cases to devise the taxonomy development and subsequently applied the results to describe two exemplary cases. A comparable approach was taken by Heinrich et al. (2019). They investigated deep learning approaches for object counting and derived a corresponding taxonomy based on 99 object counting publications, which was subsequently discussed by using exemplary cases. Further examples of data analytics taxonomies can be found in natural language processing (Wambsganss et al. 2021), adversarial machine learning (Heinrich et al. 2020), 3D object detection (Fernandes et al. 2021; Friederich and Zschech 2020), business process monitoring (Rama-Maneiro et al. 2021; Wolf et al. 2021), and many other areas. Nevertheless, most of the existing approaches develop taxonomic frameworks only

for descriptive systematization and classification purposes. As a result, they often do not use the full potential of taxonomic framework structures, such as for exploring relationships between different combinatorial options given in data-driven analytical pipelines. Put differently, they do not use different configurations of the taxonomy elements to assess and compare their impact on the overall pipeline performance to derive insights about promising pipeline configurations. This is where we contribute to the field by proposing a novel methodical approach.

From a development perspective, there are different procedures applicable for taxonomy development (Oberländer et al. 2019). Most recent contributions in the IS and data analytics community are often based on the method proposed by Nickerson et al. (2013) as it provides systematic guidance for the overall development process (Szopinski et al. 2019). Their method basically consists of the following three phases: (i) determining a meta-characteristic, (ii) specifying ending conditions, and (iii) identifying dimensions and characteristics towards the taxonomy creation. The actual step of identifying dimensions and characteristics can then be carried out either with an empirical-to-conceptual or a conceptual-to-empirical path. It is recommended to combine both paths to integrate different perspectives (Nickerson et al. 2013). Moreover, for collecting relevant taxonomy objects, researchers in the realm of data analytics often combine the taxonomy development with systematic literature search processes (e.g., vom Brocke et al. 2009; Webster and Watson 2002) to draw on the broad body of existing knowledge archived in various academic databases and other source systems (e.g., Heinrich et al. 2019; Matschak et al. 2022; Nadj and Schieder 2017; Wambsganss et al. 2021; Wanner et al. 2022). An overview of different data analytics taxonomies, along with their scope, their development approaches, and their application purposes is provided in Online Appendix I.

3 Research approach

Our method proposal is the result of a cumulative, multi-stage research project in which the findings of individual stages were critically reflected in separate publications and finally led to the composition of the overall approach. Table 1 provides an overview of the individual stages and summarizes (i) the related publication projects, (ii) the pursued objectives, (iii) the scope of the projects, (iv) the applied methodical approaches, (v) the roughly estimated efforts, (vi) the produced results, (vii) a synthesized set of key observations during development and evaluation activities, and (viii) the implications for the new method proposal. In the following, we reflect on the some of the key aspects of each stage and describe the reasoning behind the composition of our method proposal in a compact manner. A more detailed description of the individual stages can be found in Online Appendix II and in the respective publications.

Initially, the project started with the objective to develop a framework that can capture dimensions and characteristics of data analytics applications in the particular field of industrial maintenance (*stage 1*, Zschech 2018). As the field is characterized by many different design options when constructing analytical solutions (e.g., different analysis tasks, varying data types, multi-faceted analysis

Table 1 Overview of individual research stages

	Stage 1: Taxonomy development for data analytics solutions	Stage 2a: Taxonomy development for data analytics pipeline based on specific decision support task	Stage 2b: Evaluation and ablation studies for data analytics pipelines	Stage 3: Combination of taxonomy development and ablation studies for systematic pipeline evaluation
(i) Related publications	Zschech (2018)	Zschech et al. (2019)	Heinrich et al. (2021)	Zschech (2020) + present article
(ii) Objectives	• Development of framework to capture salient aspects and design properties of data analytics solutions	• Development of framework to capture components and configurations of task-specific analytics pipeline	• Comprehensive evaluation of data analytics pipelines based on ML models	• Systematic evaluation of data analytics pipelines based on ML models using a taxonomic framework
(iii) Scope of the projects	• Focus on broad spectrum of analytical solutions in the field of industrial maintenance	• Focus on prognostic decision support task for machine failure prediction of turbofan engines	• Focus on deep neural networks for predictive business process monitoring	• Focus on prognostic decision support task for machine failure prediction of turbofan engines
(iv) Methodical approaches	• Systematic literature search for knowledge retrieval guided by vom Brocke et al. (2009) • Systematic taxonomy development approach for framework creation guided by Nickerson et al. (2013) • Interviews with industry partners (e.g., domain experts, data analysts) for framework refinement and evaluation	• Systematic literature search for knowledge retrieval guided by vom Brocke et al. (2009) • Systematic taxonomy development for framework creation guided by Nickerson et al. (2013) (incl. adjustments made during stage 1) • Integration of structural phases of CRISP-DM (Mariscal et al. 2010)	• Literature search for the identification of related studies and reviews • Development of pipes-and-filters architecture to examine and compare different pipeline configurations (Buschmann 1996) • Quantitative evaluation studies based on computational experiments	• Composition of new method proposal consisting of method fragments applied during Stage 1–2b (cf. Sect. 4, Fig. 1)
(v) Roughly estimated efforts	• Literature search: ~ 45 h • Initial taxonomy development: ~ 70 h • Interviews + taxonomy refinement: ~ 25 h	• Literature search: ~ 85 h • Taxonomy development: ~ 160 h	• Literature search: ~ 15 h • Pipeline implementation and (re-) construction of neural networks: ~ 180 h • Computing time: ~ 300 h	• Literature search + taxonomy development: ~ 245 h (from Stage 2a) • Pipeline implementation and (re-) construction of ML models: ~ 170 h • Computing time: ~ 60 h

Table 1 (continued)

	Stage 1: Taxonomy development for data analytics solutions	Stage 2a: Taxonomy development for data analytics pipeline based on specific decision support task	Stage 2b: Evaluation and ablation studies for data analytics pipelines	Stage 3: Combination of taxonomy development and ablation studies for systematic pipeline evaluation
(vi) Produced results	• Tripartite taxonomy distinguishing between analysis objectives, data properties, and analysis methods (cf. Online Appendix II)	• Taxonomy for task-specific data analytics pipeline distinguishing between rough steps of CRISP-DM model (cf. Table 2)	• Comprehensive evaluation results for empirical model comparison (cf. Table 5)	• Comprehensive evaluation results for systematic pipeline comparison based on taxonomic evaluation framework (cf. Table 3)
(vii) Key observations during development and evaluation activities	• Small adjustments of the applied taxonomy development approach were necessary to guarantee transparency and parsimony • Resulting framework was perceived as useful for systematization and communication purposes • Initial framework lacked characteristics related to data preparation and evaluation aspects for more comprehensive pipeline assessment	• Narrower focus on the framework's scope was necessary to retain comprehensibility and parsimony • Resulting framework offers structured access to assess and compare individual pipeline configurations • Previous studies often evaluate overall pipeline as a whole without conducting holistic ablation studies by distinguishing between different pipeline configurations	• Development and evaluation of data analytics pipelines requires consideration of multiple design choices with individual impacts on pipeline performance • Confirmation that previous studies often evaluate overall pipeline as a whole without conducting holistic ablation studies by distinguishing between different pipeline configurations	• Morphological structures of the taxonomic framework from Stage 2a can be leveraged to guide the preparation and execution of systematic evaluation studies based on different pipeline configurations • Resulting evaluation studies are helpful (i) to assess the suitability of alternative design options for different decision contexts, and (ii) to verify the adequacy of combining different pipeline configurations

Table 1 (continued)

	Stage 1: Taxonomy development for data analytics solutions	Stage 2a: Taxonomy development for data analytics pipeline based on specific decision support task	Stage 2b: Evaluation and ablation studies for data analytics pipelines	Stage 3: Combination of taxonomy development and ablation studies for systematic pipeline evaluation
(viii) Implications and lessons learned for method proposal	• Allowance of non-exclusive characteristics and creation of sub-dimensions for taxonomy development • Combination of systematic literature search + systematic taxonomy as a useful approach to organize multi-faceted solution space of data analytics solutions • Extension of framework dimensions by covering CRISP-DM steps	• Focus on a specific decision support task and allowance of further refinement criteria to keep complexity manageable • Use of CRISP-DM steps to distinguish between different domain/data properties specifying the decision context, different pipeline configurations specifying design choices, and different evaluation properties specifying performance criteria	• Use of taxonomic frameworks in conjunction with quantitative evaluation metrics to guide systematic ablation and substitution studies • Use of a pipes-and-filters architecture to implement and execute systematic evaluation studies	*Implications for further improvement (cf. Sect . 6)*: • Provision of recommendations for defining appropriate abstraction level for framework development and evaluation studies • Creation of reference cards/tables with reusable evaluation results to avoid redundant experiments of computationally expensive ML models

methods), the aim was to organize the solution space in a structured manner so that involved stakeholders like domain experts and developers can quickly grasp salient domain characteristics and design properties.

To carry out the research, we used a taxonomy development approach inspired by the procedure model of Nickerson et al. (2013) and chose to define a tripartite meta-characteristic covering *analysis objectives*, *data characteristics*, and *analysis methods* to distinguish between output, input, and throughput dimensions of data analytics solutions (Tsai et al. 2014). By reflecting on the development process, we could see that small adjustments of the original procedure model were necessary for our purpose. Nickerson et al. (2013) postulate the fundamental requirement that dimensions should not be redundant and that characteristics should be mutually exclusive. However, due to hierarchical and combinatorial relationships in data analytics solutions, we realized that both criteria led to an inflated set of characteristics within individual dimensions. As a remedy, we considered to allow non-exclusive characteristics and the creation of sub-dimensions to guarantee transparency and parsimony within the resulting taxonomic framework.

For the identification of taxonomy objects, a literature-based approach was chosen by conducting a systematic literature search (vom Brocke et al. 2009) in several digital libraries from the fields of computer science, engineering, and IS. The combination of a systematic literature search and a taxonomy development approach for data analytics solutions turned out to be a useful way to retrieve and organize a multi-faceted solution space which is typically spread across a large number of academic publications.

In addition, the taxonomy development was carried out in cooperation with a medium-sized IT service provider offering data analytics solutions for various industrial branches, such as semiconductor industry, automotive, and plant engineering. The involvement of expert knowledge from industry ensured the practical relevance of the research endeavor. Furthermore, it contributed to the iterative refinement of the taxonomy and enabled a reflection of the results in terms of an external evaluation. As such, it could be revealed that the developed taxonomy was perceived as a useful systematization framework and a viable communication tool for bringing together different actors (e.g., domain experts and data analysts) to collectively discuss a multidisciplinary problem space.

Additionally, it was also possible to obtain valuable feedback from the industry partner for the further development of the framework results. Instead of only distinguishing between the three meta-characteristics of *analysis objectives*, *data characteristics*, and *analysis methods*, they suggested to apply a broader view by considering all relevant steps of commonly applied data analytics procedure models such as CRIS-DM (cf. Sect. 2.1). In this way, the industry partner saw an opportunity to enrich overly generic procedure models with more domain specificity by using the framework elements to capture and organize different design options along each step of such procedure models. In order to meet this demand and enable a more comprehensive pipeline assessment, we obtained the necessity of extending our initial framework structure by covering additional dimensions related to *data preparation* and *evaluation* aspects.

The idea of this structural extension was subsequently examined in a second subproject to evaluate its feasibility (*stage 2a*, Zschech et al. 2019). However, to capture meaningful dimensions and characteristics that could be applied at a comparable level, a much more specific focus had to be set than in the first run. Especially the additional dimension of *data preparation* would have resulted in too much variability in the solution space if the scope had been kept too general, thereby impairing the comprehensibility and parsimony of the taxonomic framework. Thus, instead of a broad domain consideration (e.g., industrial maintenance), a narrower focus on a specific decision support task (e.g., prediction of machine failures) had to be chosen. Additionally, it turned out to be beneficial to further refine the scope of the taxonomic framework by concentrating on a certain type of setting (e.g., choice of maintenance scenario) or a certain type of prediction methods (e.g., deep neural networks) to set a suitable focus. For testing purposes, we chose a frequently cited predictive maintenance scenario based on degrading turbofan engines and run through another process of taxonomy development. This second run was guided by the main steps of the CRISP-DM procedure model for the extraction of suitable framework elements. All other practices that proved to be effective in the first run were retained. Further details on the exact execution are given in Sect. 5, as the case also serves as one of demonstration examples in this article.

By reflecting on the second development process, it was found that the resulting framework was indeed able to position and compare different components/configurations of data-driven method pipelines in a structured manner. At the same time, it was found that such a fine-grained, taxonomic consideration of pipeline components was rarely used within the examined literature corpus to conduct systematic ablation and substitution studies. Instead, most authors or developers only consider their pipeline as a whole and evaluate the overall approach against a single metric. Thus, it could be observed that it generally lacks a broader distinction between different design options along the analytical pipeline to measure the impact of certain building blocks, such as specific data preparation or modeling steps.

A similar observation could be made in another parallel project (*stage 2b*, Heinrich et al. 2021). The goal of this project was to examine and compare different deep neural networks for prediction tasks in the field of business process monitoring. During the reflection of manifold design options and the reconstruction of various networks from related work, it could be confirmed that most authors only evaluate their solution as a whole based on a single prediction score—without taking into account a more nuanced view, such as testing their network's effectiveness for different domain conditions and data properties.

Based on these observations, the idea emerged that taxonomic frameworks might not only be used for descriptive systematization purposes to capture design options of data analytics solutions. Instead, the morphological structures of the taxonomy could be leveraged in conjunction with quantitative evaluation metrics to investigate the mechanisms between different combinatorial options given in data analytics pipelines, taking the idea of ablation and substitution studies to a more holistic level. On this basis, evaluation studies can be defined and performed more comprehensively and systematically in order (i) to assess the suitability of alternative design options for different contexts, and (ii) to verify the adequacy of combining different

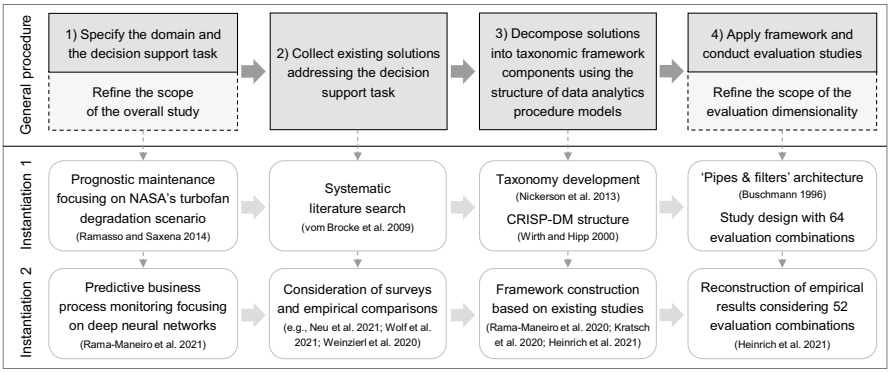

Fig. 1 Method proposal for the development of taxonomic evaluation frameworks

pipeline components. The implementation of this idea was subsequently tested with the predictive maintenance case mentioned above (Zschech 2020) and resulted in the composition of the corresponding method proposal for this article.

4 Proposal of a systematic evaluation approach for data-driven method pipelines

Based on the reflection in the previous section, it was possible to derive a novel approach for the systematic assessment of data-driven method pipelines based on taxonomic evaluation frameworks. Our proposal's general procedure consists of four rough guiding steps, summarized in Fig. 1 (upper part). In the following, we introduce each step with a brief description. Subsequently, we instantiate our method proposal with two exemplary application scenarios for a more detailed illustration of the individual steps (cf. Fig. 1 lower part).

In the *first* step, the domain and the decision support task need to be specified. The task must be sufficiently well delimited, and it needs to allow for support from data-driven methods that can be evaluated using quantitative evaluation metrics. This may include diagnostic, predictive, or even prescriptive decision support tasks (Stefani and Zschech 2018) in which the task performance can be directly measured. For instance, exemplary tasks in sales could be predicting customer churn (Chou et al. 2021) or allocating sales representatives (Bischhoffshausen et al. 2015), while in manufacturing typical applications could be visual quality inspection (Yang et al. 2020) or predictive/prescriptive performance modeling (Brodsky et al. 2015). Furthermore, the option could be considered to refine the overall taxonomic evaluation study's scope to focus on specific settings or circumstances, which will be demonstrated in both application scenarios in Sect. 5.

In the *second* step, the existing knowledge base needs to be screened for the collection of analytical solutions based on data-driven methods that address the specified decision support task. In this way, an overview can be obtained of the alternative design options for building data analytics pipelines for the specified decision

support task. To realize this step, it is advisable to draw on established research methods for conducting systematic literature searches (e.g., vom Brocke et al. 2009; Webster and Watson 2002).

In the *third* step, the identified solutions need to be decomposed into modular components to obtain the taxonomic structure of the evaluation framework. For this step, it is advisable to adopt the guidelines proposed by Nickerson et al. (2013). However, as a crucial extension, the extraction of dimensions and characteristics is supposed to follow the general structure of data analytics procedure models, which are basically organized into the previously mentioned steps of (i) domain understanding, (ii) data understanding, (iii) data preparation, (iv) modeling, and (v) evaluation (Mariscal et al. 2010). Furthermore, due to hierarchical and combinatorial relationships between different pipeline components, we propose to consider non-exclusive characteristics and the creation of sub-dimensions to guarantee transparency and parsimony within the taxonomic framework.

In the *fourth* step, the taxonomic evaluation framework is used to define and conduct quantitative evaluation studies by reconstructing the identified solution components in different combinations. In this way, the extracted framework elements serve as evaluation options that are iteratively modified under ceteris paribus conditions. This follows the general idea of ablation and substitution studies, in which the effects of individual pipeline components are examined by systematically removing and replacing these components. Thus, by using a pipes-and-filters architecture (Buschmann 1996), all conceivable combinations of data preparation and modeling methods can be studied based on different data properties concerning their impact on multiple evaluation criteria. However, instead of using the entire evaluation framework, the option could be considered to refine the scope of the study design to focus on specific aspects. Such options are also chosen in both demonstration cases by focusing on a subset of combinations to keep the complexity of the demonstration examples manageable.

5 Demonstration of the proposed approach

To demonstrate our proposed approach, we apply it to two different application scenarios, which were also part of the investigations in our multi-stage research project (cf. Sect. 3, Table 1). The first scenario covers the predictive maintenance case focusing on a turbofan degradation setting as a frequently discussed scenario within the industrial maintenance community. Here, the proposed steps are carried out in a detailed manner to illustrate their implementation exemplarily. The second scenario is located in the field of business process monitoring with a particular focus on the task of next event prediction. In this example, we do not perform all four steps ourselves from scratch but draw in some parts on the results of existing work. In other words, we reuse taxonomic structures from existing systematization frameworks and rely on the computational results from an existing evaluation study. In doing so, we aim to show that the approach we propose can also be used to reconstruct existing evaluation results from a taxonomic and thus more systematic and holistic perspective.

Both scenarios, i.e., industrial maintenance and business process monitoring, cover central issues that receive a lot of attention in research and practice alike. In addition, they are representative examples of data-intensive applications in which (i) high-dimensional data collections (i.e., condition monitoring data vs. event log data) are used for central decision support tasks and (ii) for which the analytical solution space is characterized by a broad variety of design options along the development of data-driven method pipelines. Thus, in both scenarios it is beneficial to consider the construction and evaluation of the analytical solution space in a systematic and structured manner by applying our proposed taxonomic evaluation approach in order to derive prescriptive insights about promising pipeline specifications.

5.1 Application in industrial maintenance

Industrial maintenance plays a crucial role in manufacturing as it helps production sites to guarantee high reliability, human safety, and low environmental risks. For this purpose, modern production environments increasingly focus on proactive maintenance strategies like predictive maintenance (PdM) based on data-driven prognostic solutions to efficiently use given resources and avoid redundant expenditures (Bousdekis et al. 2018). In this course, the systematic construction and evaluation of data analytics pipelines embedded in corresponding maintenance information systems are of utmost importance.

- *Step 1*: Specification of decision support task and refinement of scope.

The main goal of anticipatory maintenance approaches is to predict faults and failures before they occur and determine the remaining useful life (RUL) of technical assets by identifying relationships between extensive monitoring data and critical events (Bousdekis et al. 2018). Therefore, we concentrate on the decision support task of RUL prediction for our demonstration.

Furthermore, we refine the overall scope to keep the study's complexity manageable. To this end, we looked into different technical settings that are commonly discussed within the PdM community, such as milling machines, bearings, turbofan engines, or battery charging cycles (Eker et al. 2012; Lei et al. 2018). For our study, we chose a turbofan degradation scenario. More specifically, we used NASA's *commercial modular aero-propulsion system simulation* (CMAPSS) that provides a realistic scenario with several publicly available datasets that can be used for development and evaluation purposes. In this scenario, the NASA Ames Research Center replicated the behavior of turbofan engines under a variety of operating conditions and a continuous degradation due to varying fault injection parameters. The resulting four datasets with varying degrees of complexity (i.e., FD001-FD004) consist of multivariate time series containing parameters and condition monitoring measurements of operating cycles from different turbofan engines (Saxena et al. 2008). Due to the realistic properties, hundreds of researchers from various disciplines have already used the scenario, bringing forth a wide variety of prognostic solution approaches (Ramasso and Saxena 2014).

- *Step 2* Collection of existing solutions.

To examine the existing knowledge base and identify the large number of studies developing prognostic solutions based on NASA's turbofan degradation scenario, we followed the guidelines proposed by vom Brocke et al. (2009) for conducting and documenting a systematic literature search[1]. More specifically, we applied a database search using the following libraries: *AIS Electronic Library, EBSCOhost, IEEE Xplore, ScienceDirect*, and *SpringerLink*. As search terms, we used the keywords 'NASA turbofan degradation' and several synonyms (e.g., 'C-MAPSS'), leading to 128 unique items. Additionally, we performed a forward search based on the C-MAPSS introduction provided by Saxena et al. (2008) (+52 items), searched the websites of the PHM Society and the NASA Prognostics Center of Excellence (+40 items), and performed a backward search based on a review conducted by Ramasso and Saxena (2014) (+7 items). Thus, it was possible to obtain 227 unique hits (day of search: 2018-09-24), which had to be further reduced by appropriate filter criteria. For this purpose, we defined four inclusion criteria, which we applied in sequential order. More specifically, we ensured that the studies (i) were written in English (-1 item, 226 items remaining), (ii) were based on one of the datasets (-68 items, 158 items remaining), (iii) dealt with a prognostic approach (-30 items, 128 items remaining), (iv) applied a data-driven (-4 items, 124 items remaining), and (v) proposed a previously unknown solution (-18 items, 106 items remaining). The resulting subset of 106 studies was then used for the subsequent step of the taxonomy development (cf. Online Appendix IV). A list of full references for each study can be found in Online Appendix V.

- *Step 3* Decomposition into taxonomic framework components.

In the next step, the vast corpus of studies proposing prognostic solutions was used to develop the structure of the taxonomic evaluation framework. Following the guidelines proposed by Nickerson et al. (2013), the development process was structured into several steps and iterations, as briefly outlined in Sect. 2.2. The meta-characteristic—as the central root element—was defined as *distinct components of a data-driven method pipeline*. Concerning the ending conditions, Nickerson et al. (2013) define certain subjective criteria that must be fulfilled, e.g., that a taxonomy is sufficiently *robust* to contain enough dimensions and characteristics to separate between the objects of interest, while it is sufficiently *concise* to not exceed the cognitive load of the taxonomy user. Moreover, the method requires the specification of objective ending conditions, e.g., that every characteristic within its dimension is unique and not repeated. At this point, we adopted the following four criteria for our approach to determine the end of the iterative

[1] Please note that we explicitly refer to the guidelines of phase 3 (i.e., literature search) of the proposed framework by vom Brocke et al. (2009). The authors also provide further advice on the systematic preparation of the literature search as well as on the analysis and synthesis of identified literature, which is beyond the scope of this paper.

development process: (i) all objects have been examined, (ii) at least one object can be assigned for each characteristic across all dimensions, (iii) no new dimensions or characteristics were added in the last iteration, and (iv) no dimensions or characteristics were modified in the last iteration.

After specifying these criteria, the actual step of extracting dimensions and characteristics was carried out. At this stage, the procedure proposed by Nickerson et al. (2013) was refined as specified in Sect. 4 by additionally taking into account the general structure of the CRISP-DM procedure model (Wirth and Hipp 2000) to distinguish between distinct components of data-driven method pipelines. Moreover, we allowed non-exclusive characteristics and the creation of sub-dimensions. As recommended by Nickerson et al. (2013), the extraction process covered both empirical as well as conceptual knowledge. Empirical knowledge was directly obtained when analyzing each individual pipeline in the corpus and extracting elemental parts of prognostic solutions. Conceptual knowledge, on the other hand, was derived from existing survey papers and systematizations that were identified during the literature search above (e.g., Ramasso and Saxena 2014; Saxena et al. 2008). In total, we went through eight iterations to identify suitable framework elements of a taxonomic evaluation framework.

In a first iteration, we looked into salient properties related to the domain and data understanding of the decision scenario. By examining the characteristics of the different datasets used for the prognostic solutions in our literature corpus, we could identify different levels of complexity. More specifically, we identified one sub-dimension to distinguish between a varying number of *fault modes* and another sub-dimension to distinguish between a varying number of *operational conditions*, both of which can be seen as crucial influencing factors for the performance of prognostic solutions (Ramasso and Saxena 2014). In the next four iterations, we identified a broad range of data preparation methods. They could be organized into the four sub-dimensions of *normalization, noise reduction, dimensionality reduction*, and *feature selection*. In the sixth iteration, we considered all design choices for the modeling step. Here, we could distinguish between the following three fundamental groups of *prognostic modeling approaches* (Ramasso and Saxena 2014): (i) *direct RUL-mapping* (in which a functional mapping between the multidimensional feature space and the RUL is developed), (ii) *indirect RUL-mapping* (in which two functional mappings via a health index are established), and (iii) *similarity-based matching* (in which a library of trajectories with known failure times is created that are subsequently used for curve matching). In a seventh iteration, we extracted a series of *performance metrics* for prognostic model assessment which could be organized into accuracy-based, precision-based, and prognostics-specific metrics (Saxena et al. 2008). Finally, in a last iteration, all solutions were screened again and since no more modifications occurred, all ending conditions were met to complete the taxonomy development process. The results of the taxonomy development are summarized in Table 2. For a more comprehensive description of the framework's elements and further details on the overall process of the taxonomy development, please refer to Zschech et al. (2019). Furthermore, a list of all 106 examined studies with their respective components can be found in Online Appendix IV.

Table 2 Taxonomic evaluation framework for PdM focusing on NASA's turbofan scenario

CRISP-DM	Dimension	Characteristics		
Domain and data understanding	Fault modes	Single fault mode	Multiple fault modes	
	Operational conditions	Single condition	Multiple conditions	
Data preparation	Normalization	Standardization	Rescaling	
	Noise reduction	Moving average	Exponential smoothing	Polynomial smoothing
	Feature selection	Manual selection	Filter	Wrapper
	Dimensionality reduction	Hierarchical	Non-hierarchical	
Modeling	Prognostic approach	Direct RUL-mapping	Indirect RUL-mapping via health index	Similarity-based matching
Evaluation	Performance metric	Accuracy-based	Precision-based	Prognostic-specific

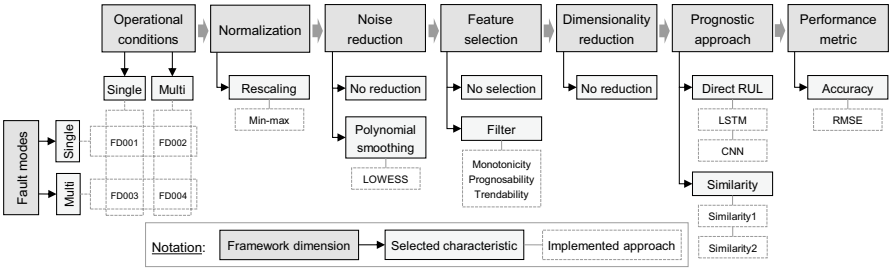

Fig. 2 Exemplary study design derived from the taxonomic PdM evaluation framework

- *Step 4* Framework application and quantitative evaluation studies.

After the extraction of the taxonomic framework, the derived elements can be used to create a systematic study design for different evaluation purposes. Thus, the derived elements can be considered as design options when implementing data-driven prognostic solutions in similar settings. More specifically, the first two dimensions (i.e., *domain and data understanding*) specify the decision scenario's context in which data analytics pipelines are constructed and tested. That is, it can be distinguished between different levels of complexity in terms of existing fault modes and operational conditions. The last dimension (i.e., *evaluation*) covers multiple evaluation options for measuring the pipeline's overall performance. In other words, different types of evaluation metrics are offered for a quantitative assessment. The remaining dimensions in between specify the configuration of the *data preparation* and *modeling* pipeline. Hence, there are several design choices conceivable with different effects on the pipeline's performance, depending on the domain/data properties and the combination of multiple pipeline components.

For our demonstration, we focus only on a subset of framework elements to keep the scope and complexity of the evaluation study manageable. So instead of using the entire framework by considering all conceivable design options from each dimension in Table 2, we only choose an exemplary selection, which we considered promising for our demonstration purposes. Our selection is highlighted with colored cells in Table 2.

The selected elements are implemented with concrete approaches that are described in the following paragraphs. Furthermore, please note that some dimensions can also be skipped in the given scenario, which is possible for all four data preparation dimensions. Thus, it can be evaluated how the performance of the pipeline is affected by removing those ablatable components. Fig. 2 summarizes our exemplary study design with the selected design choices to conduct a systematic evaluation study.

In our exemplary evaluation study, we consider different complexity levels of the turbofan degradation scenario. That is, we evaluate the performance of different pipelines for four different scenarios, which result from the combined consideration of varying fault modes and operational conditions. The four scenarios are also represented by the inherent properties of the four frequently applied C-MAPSS datasets (cf. Fig. 2). In other words, we consider one scenario with a single fault mode and a single operational condition (i.e., FD001), a second scenario with a single fault mode and multiple operational conditions (i.e., FD002), a third scenario with multiple fault modes and a single operational condition (i.e., FD003), and a fourth scenario with multiple fault modes and multiple operational conditions (i.e., FD004).

Concerning the construction of the data preparation pipeline, a normalization step is realized by using a rescaling approach through a min-max transformation (Tao et al. 2016). Subsequently, in a first variant, noise reduction is implemented via locally weighted scatterplot smoothing (LOWESS) as a concrete approach for polynomial smoothing (Khelif et al. 2017). In a second variant, the step of noise reduction is skipped to examine its specific impact on the overall performance. A similar approach is carried out for the step of feature selection. In a first path, all input features are used without any selection procedure. In a second path, a filter method is applied based on a weighted combination of the metrics "monotonicity", "prognosability", and "trendability" (Coble 2010). The next step of dimensionality reduction is skipped without any implementation (cf. Fig. 2).

For the prognostic modeling step, the two categories of direct RUL-mapping and similarity-based matching are chosen. The direct RUL-mapping is realized with two different kinds of deep neural networks, which are commonly applied for this type of RUL modeling. More specifically, a long short-term memory (LSTM) network (Zheng et al. 2017) and a convolutional neural network (CNN) (Babu et al. 2016) are implemented. The similarity-based approach is also realized through two implementations. While both share the same procedure for constructing the library of trajectories (Khelif et al. 2017), they differ in the applied approach for curve fitting and the type of similarity score (Malhotra et al. 2016; Wang et al. 2017). Finally, for performance evaluation, the root mean square error (RMSE) is used as a standard accuracy-based metric to assess the quality of the RUL estimation task (Lim et al. 2016).

The implementation[2] of the individual approaches described above is organized in modules using the programming language Python. The general structure of the taxonomic evaluation framework allows modules from different framework

[2] For the technical realization of the described methods and models, we mostly used the settings recommended by the respective developers/authors or applied default configurations. Further details on each implemented approach, such as the choice of hyperparameters, can be found in Online Appendix III: Implementation Details.

Table 3 Evaluation results for the selected elements of the PdM evaluation framework (the stronger the color intensity, the lower the prediction error measured by RMSE)

Metric: RMSE		Single operational condition				Multiple operational conditions			
		No noise reduction		Polynom. smoothing		No noise reduction		Polynom. smoothing	
		No feat. selection	Filter	No feat. selection	Filter	No feat. selection	Filter	No feat. selection	Filter
Single fault mode	LSTM	15.19	16.02	13.85	15.40	32.15	32.40	30.68	31.68
	CNN	15.23	17.77	14.86	17.31	30.67	30.64	30.67	30.58
	Similarity1	18.37	19.21	19.85	19.95	29.55	29.84	28.77	28.56
	Similarity2	14.42	16.31	16.03	18.37	23.88	24.44	24.32	24.38
Multiple fault modes	LSTM	18.84	18.59	17.81	32.47	34.56	38.41	33.59	39.52
	CNN	18.20	22.83	15.83	25.32	31.79	32.49	32.36	32.72
	Similarity1	28.57	27.89	29.90	30.22	32.55	33.41	33.42	33.93
	Similarity2	20.31	22.25	22.52	22.74	27.24	27.36	27.18	27.94

dimensions to be stacked in sequential processing steps using a pipes & filters architecture (Buschmann 1996). In this way, modular pipelines can be constructed in which the output of one module represents the input of the subsequent one. For this purpose, a dictionary is created to check the combinability of different modules with each other. In the present example of the turbofan degradation scenario, the developed framework allows the combination of all dimensions without any restrictions to obtain a fully populated evaluation matrix. However, it is also conceivable that some cells of the matrix remain unoccupied in the case of limited combinability. To automatically generate the evaluation results, conditional statements are used to execute those modules that correspond to a particular combination, while all predefined combinations are executed using loop constructs.

For demonstration purposes, the resulting evaluation matrix is illustrated in Table 3. The framework dimensions and the implemented approaches cover row and column elements, while the cells of the matrix reflect the results of the chosen evaluation metric (i.e., RMSE values). The evaluation matrix is organized into four quadrants for better readability according to the four C-MAPSS datasets FD001–FD004. They cover the scenario's different complexity levels (cf. grid-like scheme in Fig. 2). Alternative configurations of the data preparation pipeline are reflected by columns (i.e., noise reduction and feature selection), whereas alternative prognostic models are organized in rows (i.e., deep neural networks and similarity-based models). A color scheme, adjusted for each quadrant, highlights the differences in performance. The lower the RMSE values, the stronger the color intensity, indicating that an individual pipeline performs better than another.

Based on the quantitative results of the evaluation matrix in Table 3, it is possible to draw several conclusions about the mechanisms behind different combinatorial options given in data analytics pipelines. Thus, on the one hand, it is possible to assess the suitability of alternative data-driven methods in different settings. For example, it can be observed that direct prognostic models based on deep neural networks (i.e., LSTM and CNN) tend to perform slightly better than similarity-based approaches in settings with single operational conditions, especially when multiple

fault modes are present. By contrast, similarity-based models tend to perform better than direct approaches in scenarios with multiple operational conditions. This observation is particularly true for the second similarity-based model (Similarity2), which generally shows low prediction errors across all settings.

On the other hand, it is also possible to assess the adequacy of combining different method components. For example, it can be noted that neural networks without explicit feature selection, in most cases, achieve much better results compared to their variants with feature selection using the filter approach. This observation confirms the assumption that deep neural networks are generally capable of automatically extracting relevant features without the need for additional feature engineering (Janiesch et al. 2021). Similarly, it can be noted that polynomial smoothing, except in the case of FD002 (i.e., single fault, multiple operational conditions), generally reduces the performance of similarity-based approaches. One explanation could be that noise reduction removes essential information from the signals that would have been relevant for matching similar curve segments. Therefore, such method combinations should be avoided in comparable settings.

Overall, the few analysis examples illustrate which useful insights can be gained by applying such a taxonomic evaluation framework. For demonstration purposes, the scope has been kept deliberately small, so even more dimensions, characteristics, and concrete implementations are conceivable to expand the scope and conduct more in-depth analyses. In the next section, we demonstrate how our approach can also be applied to existing frameworks and evaluation studies to gain insights from a different angle.

5.2 Application in business process monitoring

Business process management is generally concerned with the identification, discovery, analysis, improvement, implementation, monitoring, and controlling of business processes (Dumas et al. 2018). The specific subfield of process monitoring has gained increasing importance in recent years. It leverages data-driven approaches to analyze business processes at runtime and predict their future behavior, performance, and outcome. This helps companies identify problems and risks before they occur and derive recommendations for managing and controlling processes at an early stage (Kratsch et al. 2020).

- *Step 1* Specification of decision support task and refinement of scope.

Predictive process monitoring (PPM) supports various decision support tasks, such as forecasting remaining cycle times, detecting business rule violations, anticipating process outcomes, or predicting next events and sequences in running instances. For our demonstration example, we concentrate on the task of next event prediction as it is one of the most frequently researched tasks within the PPM community (Evermann et al. 2017; Heinrich et al. 2021).

Furthermore, we refine the scope of the taxonomic evaluation and concentrate on a specific set of prediction methods. It can be observed that early PPM approaches chiefly focused on methods that required explicit process representations in terms of previously known process models (Marquez-Chamorro et al. 2018). By contrast, recent work steadily moves towards deep neural networks due to their capability of automated representation learning and their superior prediction results. Thus, we exemplify our approach with this type of prediction method.

- *Step 2* Collection of existing solutions.

Instead of collecting individual solutions ourselves for this demonstration example, we screened the existing knowledge base for survey articles and papers that empirically compare deep neural networks for PPM applications. More specifically, we used Google Scholar and applied the keywords "predictive (business) process monitoring" and "deep learning" in combination with the keywords "review" OR "survey" (date of search: April 2021). After screening the first 50 search results, we could identify several survey papers that summarize the field with a specific focus on deep neural networks (e.g., Harane and Rathi 2020; Neu et al. 2021; Rama-Maneiro et al. 2021; Stierle et al. 2021; Wolf et al. 2021). Likewise, we could identify several quantitative evaluation studies in which various deep neural networks are examined and compared in computational experiments (e.g., Kratsch et al. 2020; Rama-Maneiro et al. 2021; Weinzierl et al. 2020). Among these studies, there are also parts in which the authors do not only describe and compare existing deep neural networks but also extract characteristic pipeline components and classify them using systematization frameworks. This prior knowledge could be used to build a taxonomic evaluation framework in the next step.

- *Step 3* Decomposition into taxonomic framework components.

The most comprehensive survey of deep neural network approaches for PPM is provided by Rama-Maneiro et al. (2021). The authors systematically structure existing solutions into different pipeline components. More specifically, they systematize the following aspects: (i) input data, (ii) prediction task, (iii) type of neural network, (iv) sequence encoding, (v) event encoding, and (vi) performance metrics. These aspects, together with their distinct options as identified by the authors, can be translated directly into framework elements of a corresponding taxonomic evaluation framework (cf. Table 4). For an in-depth description of each dimension and characteristic, please refer to the full article by Rama-Maneiro et al. (2021).

By using the taxonomic elements from Rama-Maneiro et al. (2021), the evaluation framework covers central aspects for the related dimensions of *domain understanding*, *data preparation*, *modeling*, and *evaluation*. Nevertheless, it neglects variational factors concerning the dimension of *data understanding* which was not directly discussed by Rama-Maneiro et al. (2021). For this purpose, we can extend the framework by the considerations of Kratsch et al. (2020) and Heinrich et al. (2021), who looked into crucial process data properties of real-life event logs with a substantial impact on the overall quality of prediction

Table 4 Taxonomic evaluation framework for PPM focusing on deep neural networks

CRISP-DM	Dimension	Characteristics				
Domain understanding	Prediction task	Activity-related	Time-related	Outcome-related	Attribute-related	
Data understanding	Process variation	Low	Medium	High		
	Event repetitiveness	Low	Medium	High		
	Sparsity	Low	Medium	High		
Data preparation	Choice of input data	Activities	Time features	Attributes	Linear temporal logic	Process model
	Sequence encoding	Continuous	Prefix padded	N-grams	Single event	Timed state sample
	Event encoding	Embedding	One-hot encoding	Frequency-based		
Modeling	Type of neural network	Feedforward neural network	Autoencoder	Recurrent neural network	Convolutional neural network	
Evaluation	Performance metric	Classification-based	Regression-based	String-based		

pipelines. As a result, the taxonomic evaluation framework in Table 4 additionally covers three central data characteristics of process variation (i.e., variant-to-instance ratio), event repetitiveness (i.e., event-to-activity ratio), and sparsity (i.e., activity-to-instance ratio). For further details on all three characteristics, please refer to the full articles.

- *Step 4* Framework application and quantitative evaluation studies.

After establishing the framework structure, it can be used to guide the preparation and execution of a systematic study design for different evaluation purposes, similar to the example of the previous demonstration. However, as described above, instead of reconstructing different deep neural network pipelines from scratch, we draw on the empirical results of an already existing evaluation study and reframe the results with the aid of the derived taxonomy structure. More specifically, we draw on the empirical results from Heinrich et al. (2021), in which a key-value-predict (KVP) network and a gated convolutional neural network (GCNN) are introduced as two novel deep neural network for the task of next event prediction. Within the study, the novel networks are compared to two baseline networks, i.e., an LSTM network and stacked autoencoders (SAE). The evaluation of all four networks is based on eleven real-life benchmark datasets with varying properties, for which multiple evaluation metrics are used to assess the predictive performance. For our demonstration, we extract the empirical results of selected pipelines and map their characteristics to our taxonomic evaluation framework. In the following paragraphs, we describe how the selected evaluation pipelines can be classified using our derived framework structure (cf. colored cells in Table 4). Moreover, we outline how the empirical results can be considered from a more holistic and systematic perspective through the lens of the taxonomic framework. Figure 3 provides an overview of the reconstructed study design based on the selected evaluation pipelines from Heinrich et al. (2021).

In accordance with the decision support task specified in step 1, the given evaluation study focuses on the task of predicting the next event in running instances. This

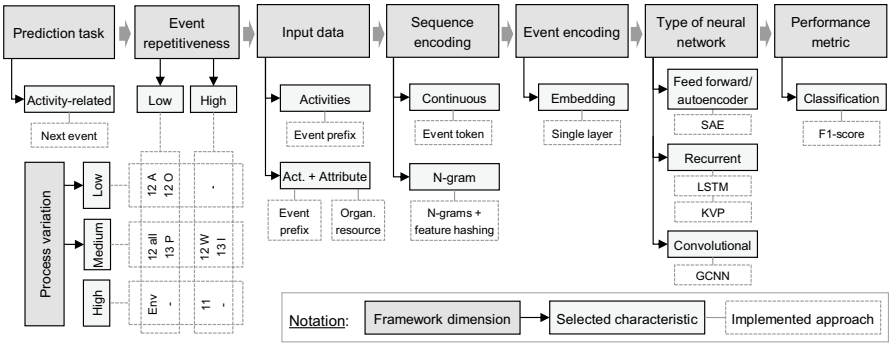

Fig. 3 Exemplary study design derived from the taxonomic PPM evaluation framework

constitutes a typical activity-related *prediction task*. From the perspective of the data properties, the eleven datasets used within the study cover a broad spectrum of event log characteristics and complexity levels. Through the combined consideration of the three properties *process variation*, *event repetitiveness*, and *sparsity*, different complexity levels can be expressed to describe the nature of the prediction scenario. For example, the applied event log BPI'11 is characterized by a high process variation (0.85), a high event repetitiveness (3.97), and a high sparsity (0.55), which constitutes a highly complex prediction scenario. In contrast, the event log BPI'12 A shows a low process variation (0.01), a low event repetitiveness (1.00), and a low sparsity (0.001), which constitutes a rather simple prediction scenario.

For our exemplary reconstruction of the evaluation results, we focus in the following on the combined consideration of only two crucial data properties, which are represented by a sufficient number of datasets. This step serves to ensure an illustrative presentation of the reconstructed results on a two-dimensional grid for our demonstration purposes. More specifically, we look into prediction scenarios with a low (0.01–0.04), medium (0.19–0.34), and high (0.85–0.99) level of *process variation* in combination with a low (1.00–1.66) and high (2.25–3.97) level of *event repetitiveness*. Thus, based on these specified properties, we reconstruct the empirical results from the following eight datasets: BPI'11 (van Dongen 2011), BPI'12 all, BPI'12 A, BPI'12 O, BPI'12 W (van Dongen 2012), BPI'13 P, BPI'13 I (Steeman 2014), and EnvLog (Buijs 2014). The grid-like scheme in Fig. 3 illustrates which combinations of data properties are covered by the individual datasets. Beyond that, we neglect an additional consideration of the *sparsity* property due to a lack of sufficient data combinations. Implicitly, however, the property still has an effect on the prediction performance, as we will see later in the reconstructed evaluation results.

Concerning the construction of the data preparation pipeline, only two different options for the *choice of input data* were originally considered by Heinrich et al. (2021). On the one hand, next events were predicted solely based on previous activities without any further features (i.e., event prefixes). On the other hand, activities were enriched with additional event attributes (i.e., organizational resource information) to evaluate their combined impact on the prediction performance. Considering the choice of *sequence* and *event encoding*, a continuous

Table 5 Evaluation results for selected elements of the PPM evaluation framework (the stronger the color intensity, the better the prediction performance measured by F1-score)

Metric: F1-score		Low event repetitiveness				High event repetitiveness			
		Example 1		Example 2		Example 1		Example 2	
		No attr.	With attr.	No attr.	With attr.	No attr.	With attr.	No attr.	With attr.
Low process variation	SAE	0.602	0.764	0.732	0.821	–	–	–	–
	LSTM	0.848	0.808	0.833	0.819	–	–	–	–
	KVP	0.853	0.822	0.847	0.820	–	–	–	–
	GCNN	0.880	0.882	0.857	0.882	–	–	–	–
Medium process variation	SAE	0.684	–	0.428	0.458	0.795	–	0.463	0.477
	LSTM	0.867	–	0.699	0.629	0.715	–	0.734	0.742
	KVP	0.876	–	0.702	0.603	0.723	–	0.706	0.741
	GCNN	0.844	–	0.631	0.683	0.710	–	0.613	0.677
High process variation	SAE	0.579	0.581	–	–	0.380	–	–	–
	LSTM	0.683	0.649	–	–	0.665	–	–	–
	KVP	0.787	0.599	–	–	0.702	–	–	–
	GCNN	0.595	0.601	–	–	0.528	–	–	–

encoding was applied using event tokens that are fed into a single embedding layer of the corresponding network architecture. One exception is the architecture of the SAE network. This specific approach uses a sequence encoding based on n-grams and feature hashing. Due to these mixed pipeline configurations, it is hardly possible to identify the contribution of a single design choice to the pipeline's overall performance (e.g., type of event/sequence encoding vs. type of network).

Considering the *type of neural networks*, the KVP and LSTM networks are representatives of recurrent network architectures; the GCNN is a convolutional network, and the SAE is a combination of stacked autoencoders and a feedforward neural network. For the quantitative assessment, eleven different classification metrics were considered in the original evaluation study. However, for simplicity, we concentrate on a single metric using the F1-score as a commonly applied *performance metric*. For further details on the specific implementation of each approach, please refer to the original study.

In analogy to the first demonstration example, we can now relate the components of the reconstructed study design in relation to each other to set up a systematic evaluation matrix (Table 5). That is, the domain and data properties specify the decision scenario's context for which the different pipelines are reconstructed and evaluated. This time, it can be distinguished between different levels of complexity in terms of process variation and event repetitiveness. As a result, the evaluation matrix is divided into six areas covering different combinations of both process properties. The chosen evaluation metric (i.e., F1-score) offers a quantitative assessment of the prediction performance for each pipeline. Again, we use a color scheme to highlight the differences in performance. The higher the value of the F1-score, the stronger the color intensity, indicating that an individual pipeline performs better than another. Alternative configurations of the data preparation pipeline are reflected by columns (i.e., choice of input data), whereas alternative types of neural networks are organized in rows (i.e., SAE, LSTM, KVP, GCNN).

Since the original study did not follow a strictly systematic study design—and some of the combinations simply cannot be filled (e.g., due to missing data attributes)—there are several empty cells in the evaluation matrix. By contrast, some combinations of process properties are represented by two datasets, as reflected by distinct columns (i.e., Examples 1 and 2).

Even though the evaluation matrix is not entirely filled, several multi-perspective insights can be derived from the taxonomic evaluation approach. For example, it can be observed that the increase in complexity of both process properties (i.e., process variation and event repetitiveness) indeed has an impact on the prediction quality. Thus, we see a general tendency of decreasing performance with increasing repetitiveness and increasing process variation across all models. The lowest performance can be obtained for all models on the combination of high repetitiveness and high variation. Here, the KVP network performs best and shows its strengths compared to other models because the advanced attention mechanisms help to capture relevant patterns in complex process structures. The GCNN, on the other hand, has difficulties with more complex settings, which is particularly expressed when process variation is high. In contrast, the GCNN shows the best results on both sample datasets with low event repetitiveness and low process variation, proving its suitability for less complex process environments. The LSTM model generally shows solid prediction results across all circumstances without any remarkable performance leaps or drops in a particular combination.

Furthermore, it is noticeable that the SAE exhibits by far the lowest prediction qualities. As noted within the original evaluation study, this might be due to some implementation issues when reconstructing the original network architecture. Interestingly, however, the SAE dominates all other architectures on a single dataset; that is, BPI12w with medium variation and high repetitiveness (F1-score: 0.795). In contrast to similar datasets showing this complexity level (e.g., BPI 13 I), this specific event log is also characterized by an exceptionally low level of sparsity—a property that is not reflected by the current evaluation matrix. Thus, the specific combination indicates a promising context for the application of the SAE, which should be investigated more thoroughly in future evaluation studies.

Considering the performance differences between prediction pipelines with and without additional attributes, we can also observe several tendencies. In the case of the GCNN and the SAE, the data augmentation leads to higher prediction performance. These architectures seem to be better suited to deal with additional information. By contrast, the two recurrent networks (i.e., KVP and LSTM) reveal an opposite effect. Here, the additional attributes impair the prediction qualities since both networks have difficulties in processing the increased number of unique event tokens, resulting in overfitting. However, the latter effect is only true in contexts with low event repetitiveness. A statement about the opposite case (i.e., high repetitiveness) would require more empirical results as it is currently only reflected by a single dataset.

Overall, the exemplary results show that the taxonomic evaluation approach allows deriving various multi-perspective insights—similarly to those retrieved in the previous demonstration example. Of course, this consideration may not replace the original study with all its in-depth examinations. Nevertheless, it provides a

structured procedure to systematically identify and present relevant relationships, patterns, and trends while uncovering conspicuous outliers that require further investigations.

6 Discussion

This section discusses the merits and limitations of the proposed approach and outlines implications for further research and practical applications.

Given a multitude of alternative design options when building analytical information systems, taxonomies offer a viable approach to organize the solution space of data-driven method pipelines in a structured manner. The resulting framework elements (i.e., dimensions and corresponding characteristics) can then guide the creation and execution of structured evaluation studies to consider the construction and assessment of data analytics pipelines more comprehensively and systematically. We illustrated the benefits of this approach by proposing a generic guidance model and instantiating the approach with two demonstration examples from data-intensive application domains.

Although state-of-the-art prediction models in the form of advanced deep neural networks were used in both exemplary instantiations, there was no single approach that showed dominating performance values across all situations in any of the two examples. This result underlines the need for a structured evaluation approach that considers different design options from a more holistic and multi-perspective view. Thus, with the presented method, fine-grained evaluation studies could be performed (or reconstructed) in order (i) to assess the suitability of alternative design options for different contexts, and (ii) to verify the adequacy of combining specific pipeline components.

Our proposed approach is the result of a cumulative research project. As such, it integrates key concepts and ideas from several adjacent areas of research and practice, and combines them into a new method proposal. Conversely, our approach thereby also makes several contributions to those areas from which it was assembled, including (i) data analytics procedure models, (ii) ablation and substitution studies, and (iii) taxonomy developments.

Let us start with *data analytics procedure models*. As outlined before, such procedure models are generally considered helpful as they provide structural guidance for the systematic development of data analytics pipelines. At the same time, they have been criticized for being too generic as they do not capture relevant characteristics of specific solution spaces. In response, there have been some recent efforts to offer procedure models that are more tailored towards domain-specific particularities (e.g., Huber et al. 2019). However, even such models may not adequately capture crucial design options for specific decision support tasks. At this point, we see a valuable contribution of our approach in enriching generic procedure models with more domain specificity by using the framework elements to capture and organize different design options of the solution space for each step of the procedure model. By additionally incorporating the quantitative results of systematic evaluation studies, such enrichments may not only be limited to purely descriptive systematization

purposes. Instead, we generally see an opportunity that the morphological structures of the taxonomic frameworks can be used to derive prescriptive design knowledge (Kundisch et al. 2021; Möller et al. 2021) to inform the construction of future analytical information systems.

Likewise, our approach provides a contribution to the field of *ablation and substitution studies*. The general idea of such studies originally stems from the field of ML, where certain types of models (e.g., deep neural networks) are composed of multi-layered components for which the impact on the overall performance is investigated. Therefore, current practices are often very model-centric and focus mainly on architectural aspects. In contrast, little attention is paid to the contextual conditions under which new methods and models are either more or less appropriate (e.g., Heinrich et al. 2021; Kratsch et al. 2020). Against this background, our approach provides new incentives to consider contextual circumstances more holistically in terms of domain-specific circumstances and data properties that might have an impact on the suitability of data analytics pipelines. Put differently, we could say that current practices in conducting ablation and substitution studies tend to follow a bottom-up strategy, focusing mainly on small-scale model and method components, while our approach rather propagates a top-down strategy, investigating effects from a broader contextual solution space. Moreover, since there are hardly any guidance models or standardized procedures available in this area (Sheikholeslami et al. 2021), our approach is one of the first of its kind.

From a *taxonomy development* perspective, our proposal can be seen as a contextualized development approach for the area of data analytics and data-driven method pipelines. While our approach largely follows the guidance model by Nickerson et al. (2013) for the central step of constructing a taxonomic framework, it also required some crucial modifications and extensions for our specific application area. This includes, for example, an orientation along the structure of data analytics procedure models for the extraction of relevant framework dimensions, or the admission of non-exclusive characteristics. As such, we follow the example of other modified taxonomy development methods (e.g., Notheisen et al. 2019) which required an application-specific adjustment. In this respect, we see great potential that our contextualized approach can be reused or even further developed by other researchers and practitioners for similar application areas.

Moreover, with our approach we offer one of the few examples, especially in the realm of data analytics, where the goal of taxonomy development goes beyond purely descriptive systematization purposes. As outlined above, this is achieved during the taxonomy's usage phase by drawing conclusions about the combinations of different framework elements in conjunction with quantitative evaluation metrics. This allows to derive prescriptive insights about promising pipeline configurations for different contexts. Although some researchers already combine taxonomy developments with subsequent cluster analyses to identify archetypal groups (e.g., Matschak et al. 2022; Wanner et al. 2022), to the best of our knowledge, there is no other competing approach yet that leverages morphological structures of data analytics taxonomies to guide the preparation and execution of systematic evaluation/ablation studies. Thus, our approach can also be seen as a valuable direction for a new and contextualized taxonomy purpose, which has not yet been discussed as such by

other taxonomy developers and/or users within the IS discipline (Schoormann et al. 2022).

On top of that, the two different demonstration examples in the previous section have shown that our proposed approach is not limited to a specific data analytics scenario. Instead, our method allows a high degree of flexibility with respect to different application domains, decision support tasks, problem classes, and particular settings. Nevertheless, some limitations were applied in both demonstration examples to keep the complexity manageable. In the first example, the focus was set on a specific technical scenario (i.e., turbofan engines) while the methodical basis was considered broadly. In the second example, on the other hand, a narrower focus was placed on a specific type of method (i.e., deep neural networks), but the field of application was kept flexible. Beyond these examples, the complexity may increase remarkably if the scope is chosen too broadly. This could possibly result in too many domain- and method-specific dimensions and characteristics, from which the combinability of components - but also the comparability of corresponding data analytics pipelines - may suffer. However, we have not explicitly considered such constraints so far, which will be the subject of future work.

Furthermore, as a side product, we were able to derive two valuable taxonomies that can be considered as reusable artifacts to research and practice. Thus, both taxonomic frameworks can be recycled to guide the creation of new study designs for further evaluation aspects. Similarly, they can be leveraged to systematize and differentiate future work that is concerned with the development of novel data analytics pipelines. In this course, we were also able to show that our proposed approach is not only suitable to guide the creation of new evaluation studies but can also be helpful for reconstructing already existing ones. As illustrated in our second demonstration example, interesting anomalous spots (e.g., performance leap of SAE) could be detected that require further investigations. As already mentioned, this cannot replace in-depth considerations but could provide new complementary insights.

As with any research, our work is not free of limitations. Currently, our approach offers a rough orientation on how to obtain a taxonomic evaluation framework via a top-down strategy. That is, in the first step, the overarching decision support task is defined and then the supporting pipeline is divided into individual components along the structure of data analytics procedure models. One level further below, however, it is hardly possible to make any further concrete recommendations as to which level of abstraction should be chosen for deriving suitable framework elements. As exemplified in our two demonstration examples, it is generally advisable to distinguish between rough types of models and methods (e.g., different architectural topologies of neural networks, different types of feature encodings, rough types of data preparation steps, etc.) to assess their performance and general suitability. For each chosen type, however, representative implementations must be selected for their technical realization. These implementations, in turn, have several design options in the sense of configurations that need to be specified. That is, on an extreme end, it is even possible that every single hyperparameter of a data analytics pipeline could constitute a taxonomic dimension on its own (e.g., choice of activation functions in neural networks). However, this would increase the framework's complexity drastically while limiting our approach's benefits of providing transparency. Against this background,

it is necessary to choose an appropriate level of abstraction and decide in advance which pipeline components and properties are likely to have a crucial impact on the corresponding evaluation metrics, without risking that the scope is chosen too broad or too narrow. Admittedly, this requires a certain level of data science expertise and sufficient experience in developing data analytics pipelines, which needs to be taken into account when setting up the team for the application of our proposed method.

Another limitation arises from the amount of time and effort required to carry out all four steps of our proposed approach. In our second demonstration example, it was possible to draw on existing systematization frameworks to create the taxonomic evaluation framework. In such a case, the critical foundation for conducting systematic evaluation studies is already given. If, on the other hand, one starts without preliminary work and carries out all the proposed steps from scratch—as shown in the first demonstration example—this involves a considerable amount of time and effort. More specifically, as outlined in Table 1, the total effort required for the first demonstration example was about 475 hours, including 85 hours for knowledge retrieval and systematic literature search to cover the analytical solution space, 160 hours for taxonomic framework development, 170 hours for technical implementation and pipeline development, and 60 hours computing time for model training. Thus, especially the initial steps of knowledge retrieval and literature search as well as the subsequent step of the taxonomy development can be very time-consuming. For practitioners who need to develop data analytics pipelines in companies under time pressure with limited resources, the effort and benefit of this approach may not be in a justifiable ratio. Against this background, we see the main responsibility for carrying out these steps primarily with researchers in the respective decision support domains. Once appropriate evaluation frameworks have been developed and initial quantitative evaluation results are available, practitioners can recycle them for their own purposes and enrich them with further results from additional evaluation studies. At this point, we pursue the vision that our approach can be leveraged to create *reference cards* or *reference tables* that provide reusable insights into which pipeline constellations work well or poorly under certain conditions. In the long run, this would have the advantage that computational experiments for similar contexts and decision support tasks would not have to be executed repeatedly from scratch, but the results could be reused in a sustainable manner to avoid redundant experiments, especially in the case of computationally intensive ML models such as deep neural networks (cf. Table 1, stage 2b).

A last limitation concerns the evaluation of our approach. Since the proposed method was developed incrementally by reflecting on the findings of multiple subprojects and individual publications, it has already gone through several phases of internal and external evaluation (i.e., together with project members, the industrial collaboration partner, reviewers of peer-review processes). This ensured practical relevance and methodical stringency. Furthermore, in this article, we have shown the feasibility and usefulness of our approach with two exemplary demonstration cases. Nevertheless, there is still a lack of an application-oriented evaluation of our proposal by applying it under real conditions with different user groups and assessing its usefulness in an external environment. In this context, it is particularly important to assess the effort required in relation to the benefits achieved, which requires a

carefully defined study design over a longer period of time, taking into account several assessment criteria, such as the duration of development and evaluation activities, the human and IT resources required, the value of intellectual findings as well as the transferability and scalability of results to other contexts. To this end, a larger longitudinal study is planned as a next step, in which our proposal is applied under real circumstances together with a mixed team of researchers and practitioners to support the construction and evaluation of data analytics pipelines for various decision support tasks.

7 Concluding remarks

In this paper, we proposed a taxonomic evaluation approach for data analytics pipelines to evaluate and construct the technical core of analytical information systems more comprehensively and systematically. To this end, we presented a rough guidance model consisting of four subsequent steps. Our approach adopts the general ideas of data analytics procedure models as well as ablation and substitution studies. As a methodical basis, we draw on a well-established taxonomy development method by Nickerson et al. (2013), which we contextualized for our specific application purpose. By instantiating our proposal in two exemplary application scenarios from the fields of industrial maintenance and business process monitoring, we demonstrated the suitability and usefulness of conducting systematic evaluation studies with the help of taxonomic frameworks. With our approach, we generally see an opportunity on how to leverage descriptive morphological taxonomies to derive prescriptive design knowledge (Kundisch et al. 2021; Möller et al. 2021) for the development of more domain- and context-specific analytical information systems in the realm of data-driven decision-making. In future steps, it is planned to apply our proposed approach to further application scenarios covering other domains and decision supports tasks to verify the transferability of the results. The findings will be used to improve the initial method proposal and provide a stronger formalization for better applicability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10257-022-00577-0.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was partially supported by the Federal Ministry of Education and Research (BMBF) within the project "White-Box-AI" (Grant Number 01IS22080) and managed by the project management agency Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR), DLR-Projektträger.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Babu GS, Zhao P, Li XL (2016) Deep convolutional neural network based regression approach for estimation of remaining useful life. In: Database systems for advanced applications. Lecture notes in computer science. Springer, Cham, pp 214–228. https://doi.org/10.1007/978-3-319-32025-0_14).

Bousdekis A, Magoutas B, Apostolou D, Mentzas G (2018) Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance. J Intell Manuf 29:6. https://doi.org/10.1007/s10845-015-1179-5

Brodsky A, Shao G, Krishnamoorthy M, Narayanan A, Menasce D, Ak R (2015) Analysis and optimization in smart manufacturing based on a reusable knowledge base for process performance models. In: IEEE international conference on big data, Santa Clara, CA, USA: IEEE, pp 1418–1427. https://doi.org/10.1109/BigData.2015.7363902

Buijs JCAM (2014) Environmental permit application process ('WABO'), CoSeLoG project—municipality 4, Media types: application/x-gzip, text/xml. Eindh Univ Technol. https://doi.org/10.4121/UUID:E8C3A53D-5301-4AFB-9BCD-38E74171CA32)

Buschmann F (ed) (1996) Pattern-oriented software architecture: a system of patterns. Wiley, Chichester, New York

Chou P, Chuang HH-C, Chou Y-C, Liang T-P (2021) Predictive analytics for customer repurchase: interdisciplinary integration of buy till you die modeling and machine learning. Eur J Oper Res S0377221721003350. https://doi.org/10.1016/j.ejor.2021.04.021)

Coble JB (2010) Merging data sources to predict remaining useful life—an automated method to identify prognostic parameters. University of Tennessee, Knoxville

Cohen PR, Howe AE (1988) How evaluation guides AI research: the message still counts more than the medium. AI Magazine 9(4):35–35. https://doi.org/10.1609/aimag.v9i4.952

Dumas M, La Rosa M, Mendling J, Reijers HA (2018) Fundamentals of business process management. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-56509-4

Eker OF, Camci F, Jennions IK (2012) Major challenges in prognostics: study on benchmarking prognostics datasets. In: European conference of the prognostics and health management society, Dresden, Germany, pp 148–155

Evermann J, Rehse J-R, Fettke P (2017) Predicting process behaviour using deep learning. Decis Support Syst 100:129–140. https://doi.org/10.1016/j.dss.2017.04.003)

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37–54. https://doi.org/10.1609/aimag.v17i3.1230)

Fernandes D, Silva A, Névoa R, Simões C, Gonzalez D, Guevara M, Novais P, Monteiro J, Melo-Pinto P (2021) Point-cloud based 3D object detection and classification methods for self-driving applications: a survey and taxonomy. Inform Fusion 68:161–191. https://doi.org/10.1016/j.inffus.2020.11.002

Flath CM, Stein N (2018) Towards a data science toolbox for industrial analytics applications. Comput Ind 94:16–25. https://doi.org/10.1016/j.compind.2017.09.003

Friederich J, Zschech P (2020) Review and systematization of solutions for 3D object detection. In: Proceedings of the 15th international conference on Wirtschaftsinformatik (WI), Potsdam, Germany: GITO Verlag, pp 1699–1711. https://doi.org/10.30844/wi_2020_r2-friedrich

Gregor S (2006) The nature of theory in information systems. MIS Q 30:3. https://doi.org/10.2307/25148742

Harane N, Rathi S (2020) Comprehensive survey on deep learning approaches in predictive business process monitoring. In: Gunjan VK, Zurada JM, Raman B, Gangadharan GR (eds) In: Modern approaches in machine learning and cognitive science: a walkthrough. Springer International Publishing, Cham, p 885. https://doi.org/10.1007/978-3-030-38445-6_9).

Heinrich K, Zschech P, Janiesch C, Bonin M (2021) Process data properties matter: introducing gated convolutional neural networks (GCNN) and key-value-predict attention networks (KVP) for next event prediction with deep learning. Decis Support Syst 143:113494. https://doi.org/10.1016/j.dss.2021.113494

Heinrich K, Roth A, Zschech P (2019) Everything counts: a taxonomy of deep learning approaches for object counting. In: Proceedings of the 27th European conference on information systems (ECIS), Stockholm-Uppsala, Sweden. https://aisel.aisnet.org/ecis2019_rp/63

Heinrich K, Graf J, Chen J, Laurisch J, Zschech P (2020) Fool me once, shame on you, fool me twice, shame on me: a taxonomy of attack and defense patterns for AI security. In: Proceedings of the 28th European conference on information systems (ECIS), Marrakesh, Morocco. https://aisel.aisnet.org/ecis2020_rp/166/

Hesenius M, Schwenzfeier N, Meyer O, Koop W, Gruhn V (2019) Towards a software engineering process for developing data-driven applications. In: Proceedings of the 7th international workshop on realizing artificial intelligence synergies in software engineering, Montreal, Quebec, Canada: IEEE Press, pp 35–41. https://doi.org/10.1109/RAISE.2019.00014

Huber S, Wiemer H, Schneider D, Ihlenfeldt S, Model (2019) Procedia CIRP(79),403–408. https://doi.org/10.1016/j.procir.2019.02.106.

Janardhanan P (2020) Project repositories for machine learning with tensorflow. Procedia Comput Sci (171), pp 188–196. https://doi.org/10.1016/j.procs.2020.04.020

Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electron Mark 31(3):685–695. https://doi.org/10.1007/s12525-021-00475-2

Khelif R, Chebel-Morello B, Malinowski S, Laajili E, Fnaiech F, Zerhouni N (2017) Direct remaining useful life estimation based on support vector regression. IEEE Trans Ind Electron 64(3):2276–2285. https://doi.org/10.1109/TIE.2016.2623260)

Kratsch W, Manderscheid J, Röglinger M, Seyfried J (2020) Machine learning in business process monitoring: a comparison of deep learning and classical approaches used for outcome prediction. Bus Inform Syst Eng. https://doi.org/10.1007/s12599-020-00645-0)

Kraus M, Feuerriegel S (2019) Forecasting remaining useful life: interpretable deep learning approach via variational bayesian inferences. Decis Support Syst 125:113100. https://doi.org/10.1016/j.dss.2019.113100)

Krieger F, Drews P (2018) Leveraging big data and analytics for auditing: towards a taxonomy. In: Proceedings of the 39th international conference on information systems (ICIS), San Francisco, USA, p 9. https://aisel.aisnet.org/icis2018/datascience/Presentations/16/

Kundisch D, Muntermann J, Oberländer AM, Rau D, Röglinger M, Schoormann T, Szopinski D (2021) An update for taxonomy designers: methodological guidance from information systems research. Bus Inform Syst Eng. https://doi.org/10.1007/s12599-021-00723-x)

Lei Y, Li, Naipeng, Guo L, Li, Ningbo, Yan T, Lin J (2018) Machinery health prognostics: a systematic review from data acquisition to RUL prediction. Mech Syst Signal Process 104:799–834. https://doi.org/10.1016/j.ymssp.2017.11.016

Leijnen S, van Veen F (2020) The neural network zoo. Proceedings (47:1), p 9. https://doi.org/10.3390/proceedings47010009

Li L, Chen J, Raghunathan S (2020) Informative role of recommender systems in electronic marketplaces: a boon or a bane for competing sellers. MIS Q 44:4. https://doi.org/10.25300/MISQ/2020/14614)

Lim P, Goh CK, Tan KC (2016) A time window neural network based framework for remaining useful life estimation. In: International joint conference on neural networks, pp 1746–1753. https://doi.org/10.1109/IJCNN.2016.7727410

Malhotra P, Ramakrishnan TVV, Anand A, Vig G, Agarwal L, Shroff G (2016) "Multi-Sensor Prognostics Using an Unsupervised Health Index Based on LSTM Encoder-Decoder," in *1st ACM SIGKDD Workshop on Machine Learning for Prognostics and Health Management*, San Francisco, CA, USA. (http://arxiv.org/abs/1608.06154)

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity | McKinsey. Technical Report, Technical Report, McKinsey Global Institute. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Mariscal G, Marbán Ó, Fernández C (2010) A survey of data mining and knowledge discovery process models and methodologies. Knowl Eng Rev 25:2. https://doi.org/10.1017/S0269888910000032

Marquez-Chamorro AE, Resinas M, Ruiz-Cortes A (2018) Predictive monitoring of business processes: a survey. IEEE Trans Serv Comput 11(6):962–977. https://doi.org/10.1109/TSC.2017.2772256

Matschak T, Trang S, Prinz C (2022) A taxonomy of machine learning-based fraud detection systems. In: Proceedings of the 30th European conference on information systems (ECIS). https://aisel.aisnet.org/ecis2022_rp/173

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S (2020) International evaluation of an AI system for breast cancer screening. Nature 577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6

Meyes R, Lu M, de Puiseau CW, Meisen T (2019) Ablation studies in artificial neural networks. ArXiv:1901.08644 [Cs, q-Bio]. http://arxiv.org/abs/1901.08644

Michalczyk S, Scheu S (2020) Designing an analytical information systems engineering method. In: Proceedings of the 28th European conference on information systems (ECIS), AIS virtual conference, June 15. https://aisel.aisnet.org/ecis2020_rip/57

Möller F, Haße H, Azkan C, Valk H, van der, Otto B (2021) Design of goal-oriented artifacts from morphological taxonomies: progression from descriptive to prescriptive design knowledge. In: Proceedings of 16th international conference on wirtschaftsinformatik (WI). https://aisel.aisnet.org/wi2021/ZMethods/Track01/1

Nadj M, Schieder C (2017) Towards a taxonomy of real-time business intelligence systems. In: Proceedings of the 25th European conference on information systems (ECIS), Guimarães, Portugal, June 10. https://aisel.aisnet.org/ecis2017_rp/33

Neu DA, Lahann J, Fettke P (2021) A systematic literature review on state-of-the-art deep learning methods for process prediction. Artif Intell Rev. https://doi.org/10.1007/s10462-021-09960-8

Nickerson RC, Varshney U, Muntermann J (2013) A method for taxonomy development and its application in information systems. Eur J Inform Syst 22(3):336–359. https://doi.org/10.1057/ejis.2012.26)

Notheisen B, Willrich S, Diez M, Weinhardt C (2019) Requirement-driven taxonomy development – a classification of blockchain technologies for securities post-trading, presented at the Hawaii international conference on system sciences. https://doi.org/10.24251/HICSS.2019.558

Oberländer AM, Lösser B, Rau D (2019) Taxonomy research in information systems: a systematic assessment. In: Proceedings of the 27th European conference on information systems (ECIS), Stockholm-Uppsala, Sweden. https://aisel.aisnet.org/ecis2019_rp/144

Rama-Maneiro E, Vidal J, Lama M (2021) Deep learning for predictive business process monitoring: review and benchmark. IEEE Trans Serv Comput. https://doi.org/10.1109/TSC.2021.3139807

Ramasso E, Saxena A (2014) Performance benchmarking and analysis of prognostic methods for CMAPSS datasets. Int J Prognostics Health Manage 5(2):1–15

Saxena A, Celaya J, Balaban E, Goebel K, Saha B, Saha S, Schwabacher M (2008a) Metrics for evaluating performance of prognostic techniques. In: International conference on prognostics and health management, Denver, USA, pp 1–17. https://doi.org/10.1109/PHM.2008a.4711436

Saxena A, Goebel K, Simon D, Eklund N (2008b) Damage propagation modeling for aircraft engine run-to-failure simulation. In: International conference on prognostics and health management, Denver, USA, pp 1–9. https://doi.org/10.1109/PHM.2008b.4711414

Schoormann T, Möller F, Szopinski D (2022) Exploring purposes of using taxonomies. In: Proceedings of the 17th international conference on Wirtschaftsinformatik (WI), Nürnberg, Germany. https://aisel.aisnet.org/wi2022/wi_interdisciplinary/wi_interdisciplinary/5

Sheikholeslami S, Meister M, Wang T, Payberah AH, Vlassov V, Dowling J (2021) AutoAblation: automated parallel ablation studies for deep learning. In: Proceedings of the 1st workshop on machine learning and systems, Online United Kingdom: ACM, April 26, pp 55–61. https://doi.org/10.1145/3437984.3458834

Steeman W (2014) BPI Challenge 2013, Ghent University. https://doi.org/10.4121/UUID:A7CE5C55-03A7-4583-B855-98B86E1A2B07

Stefani K, Zschech P (2018) Constituent elements for prescriptive analytics systems. In: Proceedings of the 26th European conference on information systems (ECIS), Portsmouth, UK. https://aisel.aisnet.org/ecis2018_rp/39

Stierle M, Brunk J, Weinzierl S, Zilker S, Matzner M, Becker J (2021) Bringing light into the darkness—a systematic literature review on explainable predictive business process monitoring techniques.

In: Proceedings of the 29th european conference on information systems (ECIS), Portsmouth, UK. https://aisel.aisnet.org/ecis2021_rip/8
Szopinski D, Schoormann T, Kundisch D (2019) Because your taxonomy is worth it: towards a framework for taxonomy evaluation. In: Proceedings of the 27th European conference on information systems (ECIS), Stockholm-Uppsala, Sweden. https://aisel.aisnet.org/ecis2019_rp/104
Tao M, Man Z, Zheng J, Cricenti A, Wang W (2016) A new dynamic neural modelling for mechatronic system prognostics. In: International conference on advanced mechatronic systems, pp 437–442. https://doi.org/10.1109/ICAMechS.2016.7813487
Tsai C-W, Lai C-F, Chiang M-C, Yang LT (2014) Data mining for internet of things: a survey. IEEE Commun Surv Tutorials 16(1):77–97. https://doi.org/10.1109/SURV.2013.103013.00206
van Dongen B (2012) BPI Challenge 2012, Media types: application/x-gzip, text/xml. Eindh Univ Technol. https://doi.org/10.4121/UUID:3926DB30-F712-4394-AEBC-75976070E91F
vom Brocke J, Simons A, Niehaves B, Riemer K, Plattfaut R, Cleven A (2009) Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: Proceedings of the 17th European conference on information systems (ECIS), Verona, Italy
von Bischhoffshausen JK, Paatsch M, Reuter M, Satzger G, Fromm H (2015) An information system for sales team assignments utilizing predictive and prescriptive analytics. In: 2015 IEEE 17th conference on business informatics, Lisbon, Portugal: IEEE, July, pp 68–76. https://doi.org/10.1109/CBI.2015.38
van Dongen B (2011) Real-life event logs—Hospital log, media types: application/x-gzip, text/xml, Eindhoven University of Technology. https://doi.org/10.4121/UUID:D9769F3D-0AB0-4FB8-803B-0D1120FFCF54
Wambsganss T, Engel C, Fromm H (2021) Improving explainability and accuracy through feature engineering: a taxonomy of features in NLP-based machine learning. In: Proceedings of the 42nd international conference on information systems (ICIS), Austin, Texas, December 12. https://aisel.aisnet.org/icis2021/data_analytics/data_analytics/1
Wang Z, Tang W, Pi D (2017) Trajectory similarity-based prediction with information fusion for remaining useful life. In: Intelligent data engineering and automated learning. Lecture Notes in Computer Science. Springer, Cham, pp 270–278. https://doi.org/10.1007/978-3-319-68935-7_30).
Wanner J, Wissuchek C, Welsch G, Janiesch C (2022) A taxonomy and archetypes of business analytics in smart manufacturing. The data base for advances in information systems. http://arxiv.org/abs/2110.06124
Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Q 26:2
Weinzierl S, Zilker S, Brunk J, Revoredo K, Nguyen A, Matzner M, Becker J, Eskofier B (2020) An empirical comparison of deep-neural-network architectures for next activity prediction using context-enriched process event logs. ArXiv:2005.01194 [Cs]. http://arxiv.org/abs/2005.01194
Wirth R, Hipp J (2000) CRISP-DM: towards a standard process model for data mining. In: Proceedings of the fourth international conference on the practical application of knowledge discovery and data mining, pp 29–39
Wolf F, Brunk J, Becker J (2021) A framework of business process monitoring and prediction techniques. In: Proceedings of the 16th international conference on wirtschaftsinformatik (WI), Duisburg-Essen, Germany, p 13
Yang J, Li S, Wang Z, Dong H, Wang J, Tang S (2020) Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. Materials 13:24. https://doi.org/10.3390/ma13245755)
Zheng S, Ristovski K, Farahat A, Gupta C (2017) Long short-term memory network for remaining useful life estimation. In: IEEE international conference on prognostics and health management, pp 88–95. https://doi.org/10.1109/ICPHM.2017.7998311
Zschech P (2018) A Taxonomy of Recurring Data Analysis Problems in Maintenance Analytics. In: Proceedings of the 26th European Conference on Information Systems (ECIS), Portsmouth, UK. https://aisel.aisnet.org/ecis2018_rp/197
Zschech P (2020) Data Science and Analytics in Industrial Maintenance: Selection, Evaluation, and Application of Data-Driven Methods," Doctoral Thesis. Dresden, Germany: Technische Universität Dresden. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-723182
Zschech P, Bernien J, Heinrich K (2019) Towards a Taxonomic Benchmarking Framework for Predictive Maintenance: The Case of NASA's Turbofan Degradation. In: Proceedings of the 40th International Conference on Information Systems (ICIS), Munich, Germany. https://aisel.aisnet.org/icis2019/data_science/data_science/4

Zschech P, Horn R, Höschele D, Janiesch C, Heinrich K (2020) Intelligent user assistance for automated data mining method selection. Bus Inform Syst Eng 62(3):227–247. https://doi.org/10.1007/s12599-020-00642-3