Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVIIICycle

Thesis for the Degree of Doctor of Philosophy

A Process-Aware Perspective on the Challenges and Solutions of Public Administration Digital Transformation

by

Simona Fioretto

Advisor: Prof. Elio Masciari

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

I fiumi lo sanno: non c'è fretta.
Arriveremo laggiù alla fine.

A Process-Aware Perspective on the Challenges and Solutions of Public Administration Digital Transformation

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

Simona Fioretto

December 2025

Approved as to style and content by

Prof. Elio Masciari, Advisor

Università degli Studi di Napoli Federico II
Ph.D. Program in Information Technology and Electrical Engineering
XXXVIIIcycle - Chairman: Prof. Stefano Russo

itee PhD
information technology
electrical engineering

http://itee.dieti.unina.it

Finanziato dall'Unione europea
NextGenerationEU

MUR
Ministero dell'Università e della Ricerca

La borsa di dottorato è stata cofinanziata con risorse del
Piano Nazionale di Ripresa e Resilienza, D.M. 351/2022
Missione 4 – Componente 1 – Investimento 4.1 "Dottorati Pubblica Amministrazione"
CUP: E66E22000190006

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 10, 2025

Simona Fioretto

Abstract

The European Union (EU) is paying increasing attention to new technologies, particularly the Digital Transformation (DT) of its member states. Yet Public Administrations (PAs) struggle to meet expectations and keep pace with technological change. To track progress, the EU monitors DT through initiatives such as the State of the Digital Decade report. Despite policy efforts and investments, many PAs still face fragmented data, outdated procedures, and weak integration between processes and Information Technology (IT) systems, resulting in inefficiencies, reduced transparency, and limited citizen-centricity in public services. This dissertation is motivated by the persistent gap between the strategic objectives of DT and the operational challenges faced by PAs. To address this, three complementary approaches are investigated. First, Business Process Management (BPM) was applied in an on-site analysis of judicial offices, which revealed fragmented data sources, the lack of a centralized Information System (IS), and missing end-to-end process visibility, showing the value of aligning process modeling with IT implementation. Second, a Systematic Literature Review (SLR) synthesized the state of the art in Predictive Process Monitoring (PPM), highlighting the challenges and opportunities of transitioning from classical to Object Centric Event Logs (OCELs). Third, experimental applications of Process Mining (PM) and PPM to the Land Matrix dataset on large-scale land acquisitions showed that clustering and advanced encoding strategies Artificial Intelligence (AI)-based can reduce process complexity and improve predictive performance, even in irregular and semi-structured domains. Together, these contributions provide a methodological framework to support the DT of PAs by combining top-down process modeling and redesign with bottom-up, data-driven insights.

Keywords: Digital transformation, public administration, business process management, process mining, predictive process monitoring.

Sintesi in lingua italiana

L'Unione Europea (UE) sta dando crescente attenzione alle tecnologie emergenti, in particolare alla trasformazione digitale (TD) degli Stati membri. Le Pubbliche Amministrazioni (PA) però faticano a tenere il passo con il cambiamento tecnologico. Per monitorare i progressi della TD, l'UE ha promosso iniziative come il rapporto State of the Digital Decade. Nonostante politiche e investimenti, molte PA affrontano ancora dati frammentati, procedure obsolete e scarsa integrazione tra processi e sistemi informativi, con conseguenti inefficienze, minore trasparenza e limitata centralità del cittadino. Questa tesi nasce dal divario tra gli obiettivi strategici della DT e le sfide operative delle PA e adotta tre approcci complementari. Il primo, il BPM, è stato applicato all'analisi di uffici giudiziari, rivelando dati frammentati, la mancanza di sistemi informativi centralizzati e di visibilità dei processi end-to-end, sottolineando la necessità di allineare la modellazione dei processi con l'implementazione IT. Poi, una revisione sistematica della letteratura ha sintetizzato lo stato dell'arte del PPM, evidenziando sfide e opportunità del passaggio dai tradizionali event log ai log object-centrici. Infine, applicazioni di PM e PPM al dataset Land Matrix su acquisizioni di terreni su larga scala hanno mostrato che tecniche di clustering e strategie di codifica avanzate basate su Intelligenza Artificiale, possono ridurre la complessità dei processi e migliorare le prestazioni predittive anche in domini irregolari e semi-strutturati. I risultati offrono una metodologia per supportare la TD delle PA, basata sull'integrazione tra la modellazione e il ridisegno dei processi in ottica top-down e l'analisi bottom-up fondata sui dati.

Parole chiave: Trasformazione digitale, pubblica amministrazione, business process management, process mining, predictive process monitoring.

Acknowledgements

The research presented in this dissertation has been supported by a PhD scholarship financed by the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4 "Education and Research", Component 1 "Enhancement of the offer of educational services: from nurseries to universities" - Investment 3.4 "Advanced teaching and university skills" OR Investment 4.1 "Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage".

I would like to express my sincere gratitude to my supervisor, Professor Elio Masciari, for his trust, guidance, and for giving me the opportunity to conduct this research during my PhD.

I would also like to thank CIRAD Montpellier and Dr. Roberto Interdonato for welcoming me and allowing me to carry out part of my research during my PhD period.

Finally, I wish to thank the Court of Appeal of Naples for giving me the opportunity to gain a deeper understanding of the Public Administration context.

Finanziato dall'Unione europea NextGenerationEU

MUR Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Contents

Abstract i
Sintesi in lingua italiana iii
Acknowledgements v
List of Acronyms xi
List of Figures xv
List of Tables xix
List of Symbols xxi

1 Introduction **1**
1.1 Motivation 2
1.2 Research Goal 3
1.3 Thesis Contribution 4

2 Background **7**
2.1 Business Process Management 7
2.2 Event Logs 9
2.2.1 Clarifying Example 12
2.3 Process Mining 15
2.4 Predictive Process Monitoring 16

3 An Empirical Analysis of Digital Transformation in Judicial Offices **21**
3.1 The Status of Digital Transformation in Public Administration 21
3.2 The Judicial System 23
3.3 Case Study: Judicial Offices 24
3.3.1 Process Identification 25
3.3.2 Process Discovery 27
3.3.3 Process Analysis 34
3.3.4 Process Redesign 37
3.3.5 Process Implementation 39
3.3.6 Process Evaluation 41
3.4 Considerations 42
3.4.1 An Adapted BPM Framework for the Judicial Domain 43
3.4.2 An Emerging Object-Centric Perspective 44
3.5 Conclusion 45

4 Predictive Process Monitoring in Real-Life Contexts: A Comparative Review of Classical and Object-Centric Approaches **47**
4.1 Predictive Process Monitoring 47
4.2 Related Work on PPM Surveys and Benchmark 49
4.3 Search Methodology 55
4.3.1 Research Questions 55
4.3.2 Study Selection 56
4.4 Methods Classification 60
4.5 Approaches Analysis and Classification 62
4.5.1 Object-Centric Event Logs 63
4.5.2 Classical Event Logs 67
4.6 Discussion 77
4.6.1 Event Log 77

4.6.2 Prediction Task 80
4.6.3 Method 82
4.6.4 Contribution Area 84
4.6.5 Domain 86
4.7 Conclusion and future work 87

5 Process Mining and Predictive Process Monitoring in a new context 91
5.1 Motivation 91
5.2 Background and Related Work 93
5.2.1 Background 93
5.2.2 Related Work 95
5.3 Dataset Description 97
5.3.1 Raw Dataset Description 97
5.3.2 Data Preprocessing and Event Log Preparation . . . 98
5.4 Event log Analysis 100
5.4.1 Land Acquisition Process 100
5.4.2 Event log Inspection 101
5.4.3 Directly Follows Graph 103
5.5 Experimental Setup for Complexity Reduction and Predictive Analysis 104
5.5.1 Experimental Plan 104
5.5.2 Log Decomposition 105
5.5.3 Index-Based Encoding and Categorial Variable Encoding 106
5.5.4 Clustering Algorithms 108
5.5.5 Classification 109
5.5.6 Evaluation Metric 110
5.6 Results 111
5.6.1 Log Decomposition 111

5.6.2 Complexity Reduction 112
5.6.3 Outcome Prediction 114
5.7 Conclusion 116

6 Conclusions 119
6.1 Conclusion 119
6.1.1 Limitations and Future Directions 121

A Appendix 123
A.1 Predictive Process Monitoring 123
A.2
Dataset Overview 132
A.3
LSLA Process Overview 133
A.4
Pre-Clustering Directly Follow Graphs (DFGs) of Sub-Logs 136

Bibliography 139

Author's publications 153

List of Acronyms

The following acronyms are used throughout the thesis.

AHC Agglomerative Hierarchical Clustering

AI Artificial Intelligence

BERT Bidirectional encoder representations from transformers

BP Business Process

BPM Business Process Management

BPMN Business Process Model and Notation

BPIC Business Process Intelligence Challenge

DESI Digital Economy and Society Index

DFG Directly Follow Graph

DL Deep Learning

DNN Deep Neural Network

DT Digital Transformation

EPC Event-driven Process Chain

ERP	Enterprise Resource Planning
EU	European Union
ICT	Information and Communication Technology
IS	Information System
IT	Information Technology
KPI	Key Performance Indicator
LSLA	Large Scale Land Acquisition
LSTM	Long Short Time Memory
MAE	Mean Absolute Error
ML	Machine Learning
NN	Neural Network
OCEL	Object Centric Event Log
PA	Public Administration
PM	Process Mining
PoC	Proof of Concept
PPM	Predictive Process Monitoring
RQ	Research Question
SLR	Systematic Literature Review
XAI	Explainable Artificial Intelligence

List of Figures

2.1 BPM life cycle 8

2.2 Example of Building Permit Process utilizing ER-diagram, representing the cardinality between the different object types in the considered object-centric event log. For each object type, the cardinality with the subsequent or the previous object type is represented as (min cardinality, max cardinality). 14

2.3 Predictive Process Monitoring 18

3.1 The AS-IS BPMN model of the Leave Management process describe the sequence of activities for managing staff leaves requests, including submission, verification, approval, and update of records. The process is currently supported by manual procedures and basic spreadsheet tools, ensuring operational continuity but with limited system integration. . 33

3.2 The AS-IS BPMN model of the Workload Assignment process describes how cases are distributed among the staff members. The process involves several steps, including the collection of case information, verification of staff availability, and manual case allocation. It partially depends on information from the Leave Management process to verify personnel availability and ensure balanced task distribution. The process currently relies on static data and limited system support for these checks. 34
3.3 Adapted BPM life cycle for constrained public-sector environments 44
3.4 Conceptual representation of the object-centric nature of judicial personnel processes. The judge's availability acts as a shared object updated by events from the *Leave Management* and *Workload Assignment* processes, linking heterogeneous event streams into a unified life-cycle. 45

4.1 Documents for year trend 58
4.2 Citations for year trend 58

A.2 Dotted chart of the event log. The x-axis represents timestamps; the y-axis indicates case indices. The color legend on the right maps activities to dots. 134
A.1 BPMN representation of Land Acquisition Process 135
A.3 Directly Follow Graph of the event log showing high complexity and low interpretability. 135
A.4 Directly Follow Graph of the failed sub-log showing high complexity and low interpretability. 136
A.5 Directly Follow Graph of the in production sub-log showing high complexity and low interpretability. 137

A.6 Directly Follow Graph of the work in progress sub-log showing high complexity and low interpretability. 137

List of Tables

1.1 Structure of the Thesis and Corresponding Research Contributions . . . 6

2.1 Prediction Task based on predicted value and horizon . . . 19

2.2 Method based on encoding . . . 19

3.1 Core Processes . . . 28

3.2 Actors and Responsibilities . . . 29

3.3 Supporting Systems and Tools . . . 29

3.4 Data Objects and Documents . . . 30

3.5 Activities of the *Leave Management* Process . . . 31

3.6 Activities of the *Workload Assignment* Process . . . 32

3.7 Summary of Identified Process Issues within the OJP Office 37

3.8 Summary of To-Be Redesign Actions and Expected Impacts 38

3.9 Summary of Functional and Additional Requirements . . . 42

4.1 Comparison of related works based on selected variables, ordered by year and name. Specifically, **Focus** specifies whether the study concentrated on a particular area (e.g., only Deep Learning (DL) methods, next activity prediction, etc.). **Query** details the research query employed in the study. **Database** identifies the research databases where the queries were executed. #**Works** represents the number of studies included in the analysis. **Year** indicates the publication year of the study. If the information is not provided in the study, it is marked as Not Available (NA). 52

5.1 Statistical description of the DFG. event log measures refer to the characteristics of the input data, while DFG measures describe properties of the resulting graph used for process modeling. 104
5.2 Failed Sub-Log description before and after clustering . . . 113
5.3 Work In Progress Sub-Log description before and after clustering 113
5.4 In Production Sub-Log DFG description before and after clustering 113
5.5 Classification results for the attributes and activities dataset encoded using three encoding techniques and evaluated across five metrics 116

A.1 Classification of Object-Centric Event Logs (Ordered by year and author) based on **Study, Prediction Task, Method, Metrics, Contribution, Domain**, where the number of object types (ot) of the dataset is reported in brackets. . . 123
A.2 Classification of Classic Event Log (Ordered by year and author) based on **Study, Prediction Task, Method, Metrics, Contribution, Domain** 126

A.3 Excerpt of selected columns from the LSLA raw dataset, extracted from publicly available data. Each row corresponds to a Deal ID. The columns Negotiation status and Implementation status describe the chronological evolution of each deal. The Intention of Investment column is shown as an example of other deal attributes, providing information about the type of intended investment. 133
A.4 **Excerpt of selected columns of the LSLA event log. Each row represents an event belonging to a case, which is identified by the Deal ID.** 134

List of Symbols

The following symbols are used within the thesis

$\emptyset$	Empty set
$\forall, \exists$	Universal and existential quantifiers
$\in$	Membership relation ("element of")
$\leq$	Total order relation
$\nrightarrow$	Partial (non-total) mapping
π	Mapping or associating function
$\setminus$	Set difference operator
$\subseteq$	Subset (inclusion) relation
$P(U)$	Power set of U (set of all subsets of U)
U	Universal set (universe)

Chapter 1

Introduction

If you can't describe what you are doing as a process, you don't know what you're doing.

William Edwards Deming

Society, as we know it, is experiencing an incredible change. Data are everywhere, in our mobile phones, in our cars, in our cooking machines, and especially in our computers. The amount of produced and stored data is increasing day by day, thereby creating an ever-growing volume of information [54]. The advent of AI solutions in daily activities along with the expansion of high-speed networks (4G/5G) has fundamentally transformed the Big Data landscape: data volumes, variety, and complexity have dramatically increased, surpassing the capabilities of traditional analytic systems [67]. Also, since the pandemic, we have increasingly acknowledged that information itself is necessary to gain insights into worldwide experiences and to support societal needs [49].

However, turning data into information and information into insights is far from easy: it requires extensive data manipulation, contextual knowledge, and appropriate tools. In this context, organizations approaches have strongly diverged. In fact, while some recognized the importance of analyzing data and processes implementing advanced technical solutions, others, such as small and medium enterprises and PAs remained a step behind.

This distinction, together with the growing adoption of AI, has created a significant gap between organizations and citizens used to increasingly advanced technological tools, and PAs struggling to keep up with them. This sheds light on the necessity of accelerating the DT of PAs, which entails the coordinated application and integration of ITs and ISs.

1.1 Motivation

To achieve a realistic DT and generate a significant impact on an organization, it is first essential to clarify the meaning of the term, which is often misunderstood or used incorrectly.

In everyday language, the word *digital* is frequently employed as a synonym for technological or electronic. However, its meaning originates in computer science, where it qualifies any system or device in which quantities are represented through discrete numerical values. In this sense, digital technologies enable information to be encoded, stored, and processed automatically [135].

This shift to the digital world, has produced disruptive changes in markets. Whereas in the past data had to be preserved in large paper archives, today information can be stored and processed far more efficiently, generating increasing amount of information that, if properly managed, can significantly increase organization value. The transition from physical to digital records provides additional benefits, including increased opportunities for knowledge acquisition, time and costs savings, and greater accuracy in data recording and processing [60].

Over the years, digital technologies have become increasingly sophisticated, gradually penetrating every aspect of daily life. For example, mobile networks connect us continuously to the rest of the world, and through smartphones we can access information anywhere and share it instantly with friends, colleagues, or even strangers. This remarkable driver of progress has also transformed consumer habits and expectations, placing growing pressure on firms to adapt and realign with these evolving needs [60, 135].

In order to meet these needs, organizations must implement and strategically leverage digital technologies to modify and improve fundamental aspects of their operations. This process is commonly referred to as DT. If correctly implemented, DT enables small ventures and start-ups with limited capital and experience to disrupt markets and challenge established players. However, while organizations that anticipate societal shifts can reshape business models and profoundly transform entire industries, those that do not understand and embrace these changes risk obsolescence.

The reason for this failure often lies in the misunderstanding of the distinction between DT and the related concepts of digitization and digitalization. In the past, these terms were sometimes used interchangeably [89, 137]; however, their definitions are now relatively well established. *Digitization* refers to the conversion of analog information into digital form, enabling its storage and processing by computers. *Digitalization* builds on this foundation by applying digital technologies to improve workflows—essentially automating and streamlining existing processes. DT, by contrast, as previously stated, represents a broader and more fundamental change, leveraging digital technologies and knowledge to enable end-to-end process integration and to reconfigure how an organization operates, creating new value. In particular, the notion of *transformation* highlights the disruptive changes triggered by new technologies [37, 95], which contribute to the reshaping of the entire industries and societies.

So, while technology is central, DT requires addressing organizational, cultural, and societal dimensions [66, 80].The only deployment of advanced systems does not guarantee success; in fact many firms adopt state-of-the-art solutions without achieving real impact. The effective transformation is achieved through a *process-aware* approach, which begins with process identification and modeling, and recognizes all data and resources flowing through organizational activities, ensuring end-to-end process visibility and orchestration [114].

This holistic perspective can be addressed leveraging BPM and data-driven insights to redesign workflows and integrate process thinking into decision-making [43, 27]. The alignment of technology with process-oriented knowledge is the only way for organizations to achieve sustainable DT.

1.2 Research Goal

To implement an effective DT, it is essential to start from the understanding, modeling, and redesigning of Business Processs (BPs) [122].

More in detail, a multidisciplinary, process-aware approach is required.

This encompasses both BPM and PM [82]. The first one provides the means to understand and structure organizational processes by relying on standardized management tools that work from the top down. At the same time, PM derives insights directly from data and enables a bottom-up perspective on how processes are actually executed. In other words, the BPM approach is employed to extract knowledge on organizational BPs from the people involved into their execution. On the other hand, PM is applied directly to data, deriving insights from IS executions, reconstructing process models, and supporting simulations. Together, these approaches can be adopted by organizations still in the early stages of DT to better understand their BPs and ISs, and to implement a holistic transformation that addresses multiple aspects.

The above context, and the aforementioned challenges, shape the research goal that drives this dissertation.

Research Goal (RG): To support the DT of PAs by bridging the gap between organizational processes and digital technologies through a process-aware, data-driven perspective.

From the RG, the following **Research Questions (RQs))** are derived:

- **RQ1:** How can the core processes and resources within PA be identified, modeled, and optimized to support DT?
- **RQ2:** What are the challenges and opportunities in the application of PM and in particular PPM to real-world data coming from IS, and how can knowledge from traditional approaches be transferred to this new real data paradigm?
- **RQ3:** How can PM and PPM be applied to unstructured or semi-structured data to extract actionable insights and demonstrate their applicability in new domains such as land governance?

1.3 Thesis Contribution

To answer to the RQs, the dissertation is structured around three main research lines:

1. **Exploring the PA context:** empirical analysis through a BPM approach of how BP are currently managed, what data and IS are

available, and identifying discrepancies, integration issues, and opportunities for improved data–system alignment.

2. **Investigating the transition from classical to object-centric PPM:** SLR to identify research advancement in Classical Event Logs PPM and how this knowledge can be transferred to OCEL paradigms, which better reflect the multi-entity and interconnected nature of real data in IS.

3. **Extending and validating PM and PPM techniques in new domains:** experimental evaluation of PM and PPM techniques in a non-native event log environment characterized by high variability and noise, to validate their applicability, robustness, and adaptability in complex, real-world data domains.

The rest of the thesis is structured as follows.

Chapter 1 introduces the thesis, by giving to the reader an overview of the topics described in the thesis, and by offering an organized and wide understanding of the motivation and research lines and contribution.

Chapter 2 provides the background for the main topics of the dissertation. More specifically, it outlines the theoretical foundations of the methods employed, namely BPM, PM, and PPM.

Chapter 3 presents an empirical case study within the context of the Italian PA, illustrating the application of DT and BPM principles. This chapter provides practical insight into the real organizational environment in which the research is performed.

Chapter 4 investigates through a SLR the methodological shift of PPM approaches from classical event logs to OCELs, providing insights into recent innovations and emerging research directions in the field.

Chapter 5 focuses on the application of PM and PPM techniques within the domain of Large-Scale Land Acquisitions (LSLAs). This chapter explores a novel context characterized by unstructured data, with the aim of evaluating the methodological boundaries of PM approaches in real-world environments.

Chapter 6 presents the overall conclusions drawn from the analyses conducted in the previous chapters. It offers a comprehensive discussion of the main findings, highlighting the critical issues that emerged throughout the research and reflecting on their methodological and practical implications.

In the following Table it is presented the structure of the thesis.

Table 1.1. Structure of the Thesis and Corresponding Research Contributions

Contribution / Focus	Addressed RQs	Chapter	Description
Exploring the PA context through a BPM approach	RQ1	Chapter 3	Empirical analysis of judicial processes to identify inefficiencies and misalignments between data, systems, and workflows.
Investigating the transition from Classical to OCELs PPM	RQ2	Chapter 4	SLR comparing Classical and OCELs models, outlining methodological shifts and open research gaps.
Applying PM and PPM in unstructured data domains	RQ3	Chapter 5	Experimental assessment of PM and PPM methods on LSLA data to test their applicability in noisy, real-world contexts.

Chapter 2

Background

The important thing is not to stop questioning. Curiosity has its own reason for existence.

Albert Einstein

This chapter introduces the core concepts supporting this work. It first outlines the principles of BPM and the structure of event data. Then, PM is formally defined and its main domains are discussed. Finally, notions of PPM are presented.

2.1 Business Process Management

In the book [139] a BP is defined as a set of activities that are performed in coordination to achieve a business goal. BPM is a research field for process-aware system design and optimization, that includes formal modeling techniques and performance evaluation tools to support design, administration, configuration, enactment, and analysis of BPs [36].

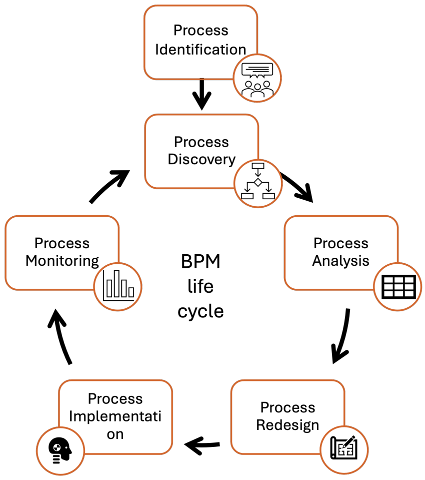

Figure 2.1. BPM life cycle

The BPM life cycle defines a loop management system composed of phases that generate formalized artifacts used for continuous optimization [35]:

- *Process Identification*: selection of critical processes and definition of their scope; *output*: prioritized processes.

- *Process Discovery*: documentation and modeling of the *as-is* process; *output*: process models in standard notations (e.g., Business Process Model and Notation (BPMN), Event-driven Process Chains (EPCs), UML Activity Diagrams, Petri Nets).

- *Process Analysis*: identification of inefficiencies, bottlenecks, and deviations; *output*: diagnostic insights and performance indicators (e.g., Key Performance Indicators (KPIs)).

- *Process Redesign*: definition of the *to-be* model and specification of required organizational and technological changes; *output*: redesigned processes and improvement plan.

- *Process Implementation*: deployment of the redesigned process through process automation supported by IT systems; *output*: operationalized process and enabling technologies.

- *Process Monitoring*: continuous observation and control of performance to ensure compliance with strategic and operational goals; *output*: performance dashboards and feedback metrics sustaining the iterative nature of the cycle.

This structured approach integrates top–down process design and bottom–up performance feedback, ensuring alignment between BP and supporting technologies, which is an essential prerequisite for effective DT in the context of PAs.

Since design and documentation are fundamental phases of the BPM methodology, it is essential to communicate process-related information clearly and unambiguously. The most widely used formalisms for process modeling are graph-based representations, where nodes typically represent process tasks (or, in some notations, states and possible events), while arcs denote ordering relations among tasks. For instance, an arc from node n_1 to node n_2 indicates an execution dependency, meaning that n_2 can be performed only after n_1 is completed [18] .

Among the most significant graph-based languages are Petri nets and BPMN. However, BPMN has become the de facto standard for representing BPs in a structured and machine-interpretable form. It provides a formal graphical syntax for modeling control flows, actors, and data dependencies while ensuring readability across both business and IT domains [136]. Specifically, BPMN defines four main element classes: (i) flow objects (events, activities, and gateways); (ii) connecting objects (sequence and message flows); (iii) swimlanes (pools and lanes); and (iv) artifacts (data objects and annotations). This structure enables the formal specification of process behavior in conjunction with the allocation of responsibilities.

2.2 Event Logs

In this section, both the formal definition and a practical example of event logs are provided, as they represent the input data for all PM approaches introduced so far in the literature.

Formal Definition Event logs are data containing information about the execution of BPs, where each event records the occurrence of a specific activity together with its related attributes, such as timestamp, resource, and case identifier. In the literature, when referring to event logs, it is generally implied that the representation is *case-centric*, which here is referred to as *Classical Event Logs*, to help the reader better understand the difference with *OCELs*.

For the definition of event logs we decided to refer to the definitions offered in [10, 13, 11, 127].

The core of the formal definition of event logs is the definition of *Universes.*

Definition 1 (Universes for Event Logs). *In set theory the Universal Set or Universe is defined as a super set, denoted by U, of all sets without any repetition, that contains all the elements within a given context, where a set is defined as a collection of distinct, well-defined objects. Let U_c denote the universe of case identifiers, U_e denotes the universe of event identifiers, U_{act} denotes the universe of activities, U_{time} denotes the universe of timestamps, U_{att} denotes the universe of attribute names, U_{val} denotes the universe of attribute values, U_{typ} denotes the universe of attribute types, U_{ot} denotes the universe of object types, U_{ob} denotes the universe of object identifiers.*

A notion that plays a crucial role in every approach is the notion of *Classical Event Log.*

Definition 2 (Classical Event Log). *A Classical Event Log is a tuple TL = $(E, \pi_{act}, \pi_{time}, \pi_{case}, \leq)$ where:*

- $\pi_{act} : E \to A$ *is the function associating an activity to each event;*
- $\pi_{time} : E \to T$ *is the function associating a timestamp to each event;*
- $\pi_{case} : E \to P(U_c) \setminus \{\emptyset\}$ *is the function associating a non-empty set of cases to each event, where P stands for power set of U_c;*
- $\leq \subseteq E \times E$ *is a total order on E;*

However, in the context of real-life processes, data frequently assume a graph structure, characterized by the interaction and synchronization of multiple objects. The aforementioned complexity implies that a single BP can not be associated with a single object, such as a product or a service.

Instead, a multitude of objects (e.g. client orders, shipment orders, product IDs) are linked through a shared set of activities. The interconnection of these case notions gives rise to the formation of OCELs.

Definition 3 (Object Centric Event Log). *Let $E \subseteq U_e$ be the set of event identifiers, $A \subseteq U_{act}$ be the set of activity names, $AN \subseteq U_{att}$ be the set of attributes names and $AV \subseteq U_{val}$ be the set of attribute values under the requirement that $AN \cap AV = \emptyset$, $AT \subseteq U_{typ}$ be the set of attribute types, $OT \subseteq U_{ot}$ be the set of object types, $O \subseteq U_{ob}$ be the set of object identifiers. An object-centric event log is defined as a tuple $L = (E, A, AN, AV, AT, OT, O, \pi_{typ}, \pi_{act}, \pi_{time}, \pi_{vmap}, \pi_{omap}, \pi_{otyp}, \pi_{ovmap}, \leq)$ such that are defined the following operations:*

- $\pi_{typ} : AN \cup AV \to AT$ *is the function associating an attribute name or value to its corresponding type;*

- $\pi_{act} : E \to A$ *is the function associating an event identifier to its activity;*

- $\pi_{time} : E \to U_{time}$ *is the function associating an event (identifier) to a timestamp;*

- π_{vmap}*:* $E \to (AN \nrightarrow AV)$ *such that*

$$\pi_{typ}(n) = \pi_{typ}(\pi_{vmap}(e)(n))$$
$$\forall e \in E \forall n \in dom(\pi_{vmap}(e))$$

  is the function associating an event identifier to its attribute value assignments;

- $\pi_{omap} : E \to P(O)$ *is the function associating an event identifier to a set of object identifiers;*

- $\pi_{otyp} : O \to OT$ *is the function assigning a unique object type to each object identifier;*

- $\pi_{ovmap} : O \to (AN \nrightarrow AV)$ *such that*

$$\pi_{typ}(n) = \pi_{typ}(\pi_{ovmap}(o)(n))$$
$$\forall n \in dom(\pi_{ovmap}(o)) \forall o \in O$$

is the function associating an object to its attribute value assignments;

- $\leq$ *is a total ordering (i.e., it respects the anti-symmetry, transitivity, and connexity properties).*

Given the above Def. 3 it is possible to make further considerations about Classical Event Logs.

An Event Log is considered *Classical* if there exists an ot $\in U_{ot}$ such that for any e $\in$ E: | π_{omap} (e)(ot)|= 1 and for all ot' $\in U_{ot} \setminus$ ot: π_{omap}(e)(ot') = $\emptyset$. In other words, an Event Log is considered Classical if it satisfies two proprieties: there exists one object type such that for every event in E it is associated one unique object identifier i.e. the case ID, and that for every object type belonging to the same Universe of object types, it does not exist a function associating an Event to another Object Identifier.

However, as stated before, processes in real life organizations do not respect the properties of Classical Event Logs, exhibiting a complex graph structure that is modeled by OCELs.

2.2.1 Clarifying Example

In order to understand the meaning of the formal definition above, we provide an example of OCELs.

In the context of PA many processes are run, thus providing a good scenario for OCELs analysis.

Example 1. *A key example is provided by the process of applying for and granting building permits, which is a process typically managed by local government or municipal authorities to ensure that construction projects comply with local building laws and safety regulations. Considering the following description of a typical* building permit application process*, it is possible to better understand OCELs and the associated challenges of the flattening operation. The process involves several key entities, including* applications, reviews, inspections, decisions, and permits*. These entities interact with each other via a series of events, thus contributing to the collective workflow.*

On June 10th, 2024, an applicant, J. Smith, submits a building permit application with an ID number, i.e., #789, to the city planning office. The *submission includes essential documents (e.g., site plan, construction plan) and is an event, labeled as Application Submitted, linked to the specific* application *case ID (#789).*

Subsequently, on June 15th, 2024, the submitted documents undergo a thorough review *process by a city planner, J. Doe. The Documents Reviewed event is recorded in association with the application ID #789 and the specific review ID #0044.*

The next step in the process involves scheduling a site inspection*, which occurs on June 20th, 2024. The Inspection Scheduled event links the application #789 and the planned inspection (e.g., identified in the inspection process as #2024_001). The Inspection Conducted event documents the completion of this critical process and is associated with the application (#789) and the specific inspection instance (#2024_001).*

Then, a decision *is made in the decision process, to approve or reject the request, which is an administrative decision. The result of this process is the Permit Granted event, which connects the application #789 and the decision process instance, identified as #456). The ID associated with this process tracks the approval phase before the permit is issued.*

Once the decision has been made, the permit *process begins, focusing on the delivery and activation of the permit. Finally, the building permit is officially issued to the applicant on July 1st, 2024. This event, labeled Permit Issued, is associated with the permit process, identified as #0874, the decision process #456 and the application itself #789.*

As construction begins on July 15th, 2024, the building permit becomes active. The Permit Activated event marks this transition and is associated with the permit process #0874, the decision process #456 and the application itself #789.

For instance, in the previous example it is possible to see that the "building permit process" is the result of many processes, such as the "review process", the "inspection process", the "decision-making process", which represent the object types or multiple case notions. Every object type is distinguished by an object identifier, for instance, the review process is characterized by a unique case ID, delivering the Documents Reviewed; the inspection process having a unique case ID delivers the Inspection Conducted document, and then the decision making process delivers the Permit Granted with a new ID. The review process, characterized by an assigned reviewer and a specific protocol used, might be linked to both the case ID for tracking the application being reviewed and the review process (e.g., Application #789 and the specific review instance #0044). The inspection process will have a separate case ID (e.g., Inspection #456), which is linked to the inspection event and associated inspector. The decision-making process will have its own case ID (e.g., Permit #456), linking the decision event (Permit Granted) to the applicant and the permit issued.

Therefore, there is not only one process related to a single ID, but there are many processes related to different IDs collaborating with each other. This is a clear example of OCELs, which provides an example of the interactions among various entities within the "building permit application" process, offering valuable insights into the sequence of events and the relationships among different objects. The objects of the process are modeled in Figure 2.2, following the same representation as proposed in [46].

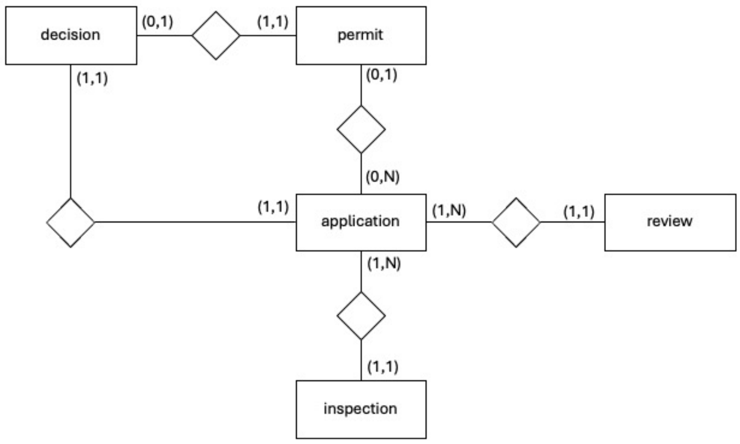

Figure 2.2. Example of Building Permit Process utilizing ER-diagram, representing the cardinality between the different object types in the considered object-centric event log. For each object type, the cardinality with the subsequent or the previous object type is represented as (min cardinality, max cardinality).

However, current PM applications and well-established techniques often require data to be *flattened* (i.e., reduced to a sequential structure) to align with the assumptions of Classical Event Logs. This *flattening* operation consists in the choice of a specific object type to be used for creating a specific event log view from the complete dataset [124] in order to obtain Classical Event Logs from OCELs.

Yet, this flattening operation raises three principal problems [2] known as *deficiency*, *convergence* and *divergence*. The first issue concerns event disappearance, meaning that during the flattening operation, important contextual relationships between entities can be lost. The second issue relates to event duplication, where an event may be associated with multiple cases, in fact, an event such as *Permit Granted* is associated with multiple cases: the application, the decision, and the permit. If a traditional PM approach is used with *applications* as the case notion, this event would be duplicated for each related entity (e.g., for every permit and decision). This duplication inflates the frequency of *Permit Granted* and misrepresents the actual number of events. The last issue involves the presence of independent, repeated executions of a group of activities within a single case, potentially creating loops or ordering unrelated events, if *decisions* are used as the case notion, the sequence of events within a single decision case may appear disordered. For example, within the same case, multiple *Documents Reviewed* and *Inspection Conducted* events might interleave without a clear order. This loss of sub-instance ordering creates artificial loops or parallel behaviors in the resulting process model, even though such behaviors do not exist in the real-world process.

Therefore, to effectively utilize real-life data without resorting to the flattening operation, it is essential to preserve the original data structure of Event Logs by using OCELs. This approach maintains the interrelationships between different object types, allowing for a more accurate representation of the complex, real-world processes and their interactions.

2.3 Process Mining

At the intersection of process science and data science we find PM. PM is a data-driven discipline that bridges process modeling and data mining, enabling the extraction of process knowledge from event data recorded in ISs [123]. PM reconstructs the actual behavior of processes as executed in reality, providing an empirical foundation for process understanding, compliance checking, and optimization.

PM encompasses three main analytical branches [124]:

- *Process Discovery*: automatic construction of a process model from event log data, typically based on DFGs and algorithms such as Alpha Miner, Heuristic Miner, or Inductive Miner, which infer control-flow relations between activities.
- *Conformance Checking*: comparison between an existing process model and the real behavior recorded in the log, measuring deviations through quantitative metrics such as fitness, precision, and generalization.
- *Process Enhancement*: enrichment or refinement of process models by incorporating empirical performance data (e.g., bottleneck identification, resource utilization, or time-based patterns).

By combining these perspectives, PM provides an integrated view of process execution that supports evidence-based decision making and continuous process optimization. It establishes the methodological and analytical foundations upon which predictive approaches, such as PPM, are built.

2.4 Predictive Process Monitoring

PPM is a specific research area of PM that employs historical complete process execution traces in order to make predictions about the future evolution of features of an uncompleted execution trace of an ongoing process [72, 32]. By leveraging PPM it is possible to predict the performance indicators of a running process, such as the outcome of a process execution, its completion time, or the sequence of its future activities, to cite a few.

The starting point of every PM application is the *Event Log* [69] (i.e., the record of information related to the execution of a BP). Based on the Event Log notion provided before, we can understand the Event Stream. In the online setting, when the process is running, it may happen that the order in which events appear is not grouped by case, resulting in a stream of events relating to different cases. The resulting event record is called *Event Stream*, and it turns out to be of interest in the context of PPM, since the predictions are performed in the online setting. If the length of each trace in the Event Log is known, it is possible to define the *Prefix Log* as a part of a trace, i.e. the partial trace, where each partial trace is a subsequence of events starting from the beginning of a trace, up to a specific point, reflecting the initial segments of the full traces in the original event log. It should be noted that the prefix log can be of a different length until the complete trace, and it is used to simulate a process that is still in execution, for which the events are not already all executed [134]. In PPM, Event Log traces (i.e. completed process executions) are used to train a *Predictive Model*, which is in general a prediction algorithm trained on historical data (indicated as X) to assign a predicted label on new data (indicated as Y). The primary objective of this process is to generalize the pattern identified in the training data, thereby providing a mapping X $\rightarrow$ Y [134].

Given the previous concepts, it is possible to define the general PPM framework, as shown in Figure 2.3. This framework starts from Event Logs Data containing information on the history of the process (i.e., historical complete process executions) which are given as input to the specific prediction algorithm implemented. The algorithm is trained to discover patterns in data and then generalize them. Finally, during the execution of the process, the previously trained model takes as input the Event Stream or the Prefix Log, i.e. live data representing the ongoing trace of a runtime process, to perform the Prediction Task, thus supporting decisions about the ongoing and future execution of the process.

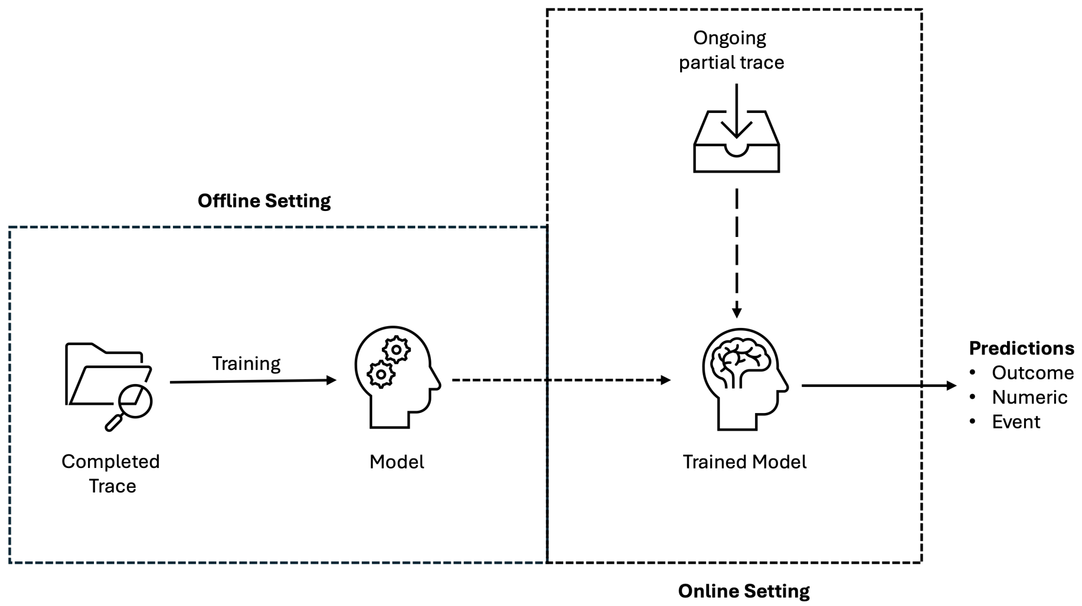

Figure 2.3. Framework of PPM implementation of 3 blocks. The Event Log data are used to train the prediction model which then perform three types of predictions at runtime on given event stream.

In this general framework, it is important to make a distinction between the *Prediction Point* and the *Predicted Point*, as suggested by [134]. Indeed, while the Prediction Point indicates the time at which the algorithm performs the prediction, the Predicted Point is the point in the future for which we want to make the prediction. Predictions are performed to a specific point in the future, which is referred to as the time horizon h. The choice of h is determined by the specific prediction task.

More in detail, PPM techniques can be categorized based on three main elements [32]:

- *Prediction task*: the type of prediction as shown in Table 2.1 can be outcome-based predictions (categorical or boolean outcome values, i.e. in the context of a permit application process, an outcome-based prediction could determine whether an application will be approved or rejected), numeric value predictions (i.e., time required to complete the execution of the whole process or of the single activity, process performance, costs) or event predictions (concerning the future control-flow behavior, i.e. next event classes, sequences of future activities with data payloads as timestamp or resources). In addition, based on the horizon, it is possible to distinguish between predictions of the case up to its completion (suffix prediction or remaining time prediction), and predictions of a specific point in the future different from the end of the case (next required time referring to the time necessary to complete the execution of some specific activities, and next event class obtaining the classes of activities different from the end of the process);

		Predicted Value		
		Outcome	Numeric	Event
Horizon	Point	categorical	next required time	next event class
	End	boolean	remaining time prediction	suffix prediction

Table 2.1. Prediction Task based on predicted value and horizon

- *Method*: the choice of the prediction algorithm is fundamental for guaranteeing accurate predictions. As regards the predictive model choice, as shown in Table 2.2, a distinction must be made between two macro categories based on the type of encoding of the Event Log: model-based or process-aware approaches, and non process-aware approaches. The former use an explicit representation of the process model in various notations (i.e., Petri Nets, Transition Systems, etc.), whereas the latter uses Machine Learning (ML) or DL techniques on the features representation of Event Logs;

		Predictive Model
Encoding	Event Log	model-based
	Process Model	process-aware

Table 2.2. Method based on encoding

- *Input information*: the input type is defined by the perspective from which the process is analyzed. Indeed, a process model can be analyzed from different points of view, with respect to the data chosen for analysis [32]. In the literature, the perspective analyzed can be of the following types: control-flow, meaning that the sequence of events is of interest for the predictions [42]; event payload, meaning that data attributes play an important role, such as the time to execute an activity; unstructured information, such as information external to the process that is textual and that can provide additional information about the process; and contextual information about the process context, for instance the scheduled working time of the organizational resources involved in the execution of the activities of the process.

Chapter 3

An Empirical Analysis of Digital Transformation in Judicial Offices

It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.

Charles Darwin

3.1 The Status of Digital Transformation in Public Administration

The adoption of digital solutions in PAs has been slow and fragmented, hindered by institutional and regulatory complexities, independent agencies, bureaucratic procedures, legacy IT systems, and a general lack of digital skills. Research conducted before the COVID-19 pandemic revealed that nearly 70% of governments considered their digital capabilities to stay behind those of the private sector, leaving PAs unprepared for such a disruptive event [26].

In fact, the urgency of digitalizing public services became evident especially when the COVID-19 pandemic forced governments to ensure service continuity in contexts where face-to-face interactions were severely restricted [83]. In this sense, the pandemic acted as a catalyst, requiring accelerated investments in digital infrastructure and process redesign aimed at strengthening resilience, ensuring continuity, and improving service delivery. Since then, substantial efforts have been devoted to advance the DT of the public sector, reflected in new policy initiatives, scientific research, and complementary activities.

Within EU, the European Commission explicitly identified DT of governments as a strategic priority. From 2014 to 2022, Digital Economy and Society Index (DESI) provided annual assessments of Member States across multiple dimensions, including connectivity, digital skills, technological integration, and the availability of digital public services.

Since 2022, DESI has been incorporated into the *State of the Digital Decade* report, which currently represents the main instrument to monitor progress towards the 2030 programme objectives. This report includes a comprehensive set of indicators summarizing performance of European Countries across the four dimensions of the *Digital Decade Policy Programme 2030*: *digital skills*, *digital infrastructure*, *digitalisation of business*, and *digitalisation of public services.*

The DT status of public services is further examined through the *eGovernment Benchmark 2025* [38], which evaluates three key dimensions: *Online Service Delivery* assessing governments on the provision of public services online, *Interoperability Signifiers* evaluating cross-border collaboration, and *User-Friendly Portals* assessing the maturity of government portals in terms of accessibility, navigation, online support, and completeness of information.

In 2024, Italy achieved a score of 76 in *Online Service Delivery* compared to EU average of 81, indicating persistent shortcomings in the full digitalization of administrative procedures, particularly for cross-border users. In *Interoperability Signifiers*, Italy scored 64 against the EU average of 67, confirming that data exchange across administrations remains critic. Finally, in *User-Friendly Portals*, the score of 67 (EU average: 76) reflects ongoing limitations in user experience and service continuity.

Although the performance of Italy is not that far from the European average, the country remains characterized by persistent limitations in interoperability, data integration, and institutional coordination.

In addition, although the pandemic catalyzed accelerated investments, problems such as low digital literacy, fragmented governance, and infrastructural bottlenecks continue to limit sustainable DT [116, 8].

The scientific literature has investigated the barriers that hinder the DT of PAs, aiming to model them in a systematic and objective manner. These studies converge in identifying a recurring set of constraints that consistently affect the implementation of DT across public institutions, mainly due to organizational [26], regulatory [9], and technical [81] issues.

3.2 The Judicial System

The judicial system constitutes a critical domain of public service delivery in the context of DT. Unlike sectors such as healthcare or finance, where digital technologies have reshaped organizational processes, the courtroom procedures are largely unchanged since the 1980 [106]. This structural inertia has generated persistent inefficiencies, producing delays that undermine public trust and increase costs for citizens and businesses. Recent analysis highlights a gap between the growing demand for legal services and the outdated organizational structures of judicial offices, resulting in longer processing times, case backlogs, and reduced effectiveness [51]. Also, according to the Organisation for Economic Co-operation and Development (OECD), the EU, and the International Monetary Fund, the Italian judicial system ranks among the least efficient in Europe [51]. Several studies underscore the urgency of reform, as citizens increasingly reject lengthy proceedings and high legal expenses when there are faster and more accessible alternatives [106]. In addition, the judicial system plays a pivotal role in society, directly impacting on national economies with its performance. In fact, in jurisdictions with stronger protection of shareholders and creditors (which are typical of common law systems) the confidence in investments and the development of the financial market result to be higher [51].

The inefficiency of the judicial system can be attributed to several factors; however, one of the most critical activities concerns case assignment. Judges are required to analyze heterogeneous proceedings (civil, criminal, labor, administrative, etc.), performing both case classification and judge selection manually, relying primarily on individual expertise and static information. Moreover, the availability of judges, which is often affected by institutional duties, absences, and leaves, is not effectively managed through digital systems. Although recent innovations such as multi-agent simulations and scheduling techniques have been successfully applied in various domains, their adoption within judicial contexts remains limited [34]. Consequently, manual workflows continue to perpetuate inefficiencies throughout the judicial process.

These characteristics make it a challenging domain, but also a fertile ground for transformative change, where process analysis and digital technologies can play a pivotal role in improving transparency, accountability, and efficiency.

In this context, BPM is introduced as a structured, process-oriented methodology for analyzing, modeling, and redesigning such workflows.

3.3 Case Study: Judicial Offices

This section presents the results of an empirical analysis conducted within the Italian judicial system, following the BPM life-cycle methodology described in Chapter 2. The study was carried out over a six-month period in 2023 at the Court of Appeal of Naples, through alternating phases of stakeholder participation, on-site process observation, and data-driven analysis. This chapter presents work aligned with the research conducted in [5, 6] within judicial offices.

Context: The Court of Appeal of Naples The Court of Appeal of Naples is one of the largest and most complex judicial institutions in Italy, with jurisdiction over a broad portion of the Campania region. It reviews the decisions issued by lower courts to ensure consistency in the application of the law and the protection of the rights of citizens. The Court is structured into specialized divisions (civil, criminal, labour, and family), each handling a large and diverse volume of cases. From an organizational point of view, the Court performs a wide range of judicial and administrative activities, including case registration, document management, hearing scheduling, and compliance with procedural deadlines.

Our analysis focuses on the *personnel administration domain*, which includes the management of judicial staff availability and assignments. These activities are coordinated by the *Ordinary Judiciary Personnel (OJP) Office*, responsible for staff allocation, validation of leave requests, and communication of assignments to ensure continuity of operations. The application of the BPM life-cycle to the identified area enables the identification, modeling, and assessment of administrative workflows, providing a foundation for process analysis and redesign.

For ethical and analytical purposes, all names used in this document, including organizational roles, file names, data objects, and system identifiers, are abstracted or renamed, with no direct correspondence to actual people or files. This abstraction allows for a clear and generalizable representation of processes while protecting sensitive information.

3.3.1 Process Identification

The first phase of the BPM life cycle, *Process Identification*, focuses on the identification of key processes, whose improvement can effectively enhance the operational performance of the Court.

To this end, a series of semi-structured interviews and a set of exploratory questions (see Table 3.3.1), were conducted with administrative personnel of the Court of Appeal of Naples. These activities enabled to understand the operational landscape of the Court and provided qualitative evidence on how daily activities are coordinated, which tasks are perceived as critical, and where procedural or technological constraints occur.

1.	Which core activities are performed by your office on a regular basis?
2.	How can these activities be represented as processes, in terms of inputs, outputs, and responsibilities?
4.	Which organizational roles are involved in each process, and how do they interact?
5.	Which process steps are perceived as the most time-consuming, error-prone, or repetitive?
6.	How is process performance monitored within the office?

Then, on-site observation of workflows, information exchanges, and existing digital systems provided further insight into how administrative and judicial activities interact on a daily basis.

Selection Criteria Following the interview phase, processes were evaluated according to three main criteria adapted from [36]:

- **Strategic Relevance:** processes directly contributing to the core mission of the Court, i.e., ensuring continuity and balance in judicial activities;
- **Frequency and Impact:** processes executed frequently and affecting the daily functioning of multiple organizational roles;
- **Improvement Potential:** processes showing inefficiencies that could be addressed through redesign and digital integration.

Results of Identification Two processes emerged as the most problematic and resource-intensive from both an operational and managerial perspective:

- **Leave Management:** covers the processes for recording, approving, and monitoring staff absences. The feedback obtained from the interview indicated difficulties in maintaining timely and accurate updates, together with the lack of an integrated overview of staff availability, affecting daily scheduling and coordination.
- **Workload Assignment:** governs the distribution of cases among judges and sections. Participants reported difficulties in maintaining balanced workload allocation, mainly due to fragmented data sources and manual coordination practices.

Justification of Scope Improving the above BPs can promote greater internal efficiency, supporting both the strategic objectives of the Court and the goals of DT in the Italian PA.

3.3.2 Process Discovery

Following the identification phase, the *Process Discovery* phase reconstructed and formally modeled the as-is BPs previously identified, i.e. *Leave Management* and *Workload Assignment*. Based on the exploratory questions formulated during the identification phase, more detailed questions (see Table 3.3.2) were used to examine the operational aspects of each process, including resources, ISs, data flows, and constraints.

1.	What resources or ISs are required at each stage of process execution?
2.	How are staff absences and leave requests currently recorded, validated, and communicated?
3.	What criteria guide the allocation of cases and the distribution of workloads among judges?
4.	Which IT systems or tools support these processes, and what limitations or redundancies do they present?

As-Is Processes

Through iterative collaboration with the involved personnel, the two core processes were analyzed, decomposed into their elementary activities, and represented using the BPMN standard introduced in Chapter 2.

Table 3.1 provides an overview of the processes under study.

Table 3.1. Core Processes

Process ID	**Process Name**	**Description**
1	Leave Management	Manages the submission, verification, approval, and recording of judges' leave requests and absences.
2	Workload Assignment	Supports the allocation of cases among judges and sections, considering competence, current workload, and availability.

In line with BPM, the following tables summarize the key findings from the discovery phase. Each table represents one of the layers of process knowledge: organizational (roles, responsibilities, and involved units), informational (data, documents, and ISs), and operational (activities, workflows, and tools).

Actors and Responsibilities The processes involve three primary organizational roles, collaborating through formal and informal communication channels.

Their interactions are summarized in Table 3.2.

Table 3.2. Actors and Responsibilities

Actor	Main Responsibilities	Processes
Judge	Submits leave requests or communicates temporary unavailability.	1
OJP Office	Manages leave records and staff registers; supports the Decision Authority by verifying availability of judges and communicating assignments when required.	1, 2
Decision Authority	Validates and approves workload allocation decisions based on section requirements.	2

Supporting Systems and Data Objects Process execution relies on a limited set of tools, primarily spreadsheets and a legacy IS. These systems operate in isolation and require manual data handling, as shown in Tables 3.3 and 3.4.

Table 3.3. Supporting Systems and Tools

System	Functional Description	Processes
Excel-Based Management System	Local spreadsheet environment used for recording and updating judges' leave, availability, and section assignments; supports daily monitoring and operational reporting.	1, 2
Judicial IS	Institutional legacy platform used to manage judges' profiles, assignments, and historical workload data; provides official records for institutional reporting.	2

Table 3.4. Data Objects and Documents

Data Object	Description	Processes
Leave Request Form	Paper or verbal request submitted by a judge to communicate an upcoming absence or leave period.	1
Leave Register (Excel)	Structured file used as input to workload allocation and updated during leave management; records absences, leave periods, and current availability of judges.	1, 2
Section Assignment Table (Excel)	Spreadsheet showing the distribution of judges across court sections.	2
Judicial Record	Digital record stored in the Judicial IS containing workload, assignments, and historical activities for each judge.	2

Activities per Process Tables 3.5 and 3.6 summarize the main activities, their executors, and corresponding artifacts.

Table 3.5. Activities of the *Leave Management* Process

Step	Activity Description	Responsible Actor	Input / Output Artifact (Data Object or Communication Flow)
1	Submit leave request	Judge	Leave Request Form — Communication Flow to OJP Office (output)
2	Verify leave details and check overlaps	OJP Office	Leave Register (input)
3	Approve or reject request	OJP Office	Approval Record (output)
4	Record leave data and update register	OJP Office	Updated Leave Register (output)
5	Request modification of leave period (if overlap detected)	OJP Office	Communication Flow to Judge (output)
6	Communicate final approval or rejection to the judge	OJP Office	Communication Flow to Judge (output)

Table 3.6. Activities of the *Workload Assignment* Process

Step	Activity Description	Responsible Actor	Input / Output Artifact (Data Object or Communication Flow)
1	Identify new case requiring assignment	Decision Authority	Case Record (input)
2	Review section distribution of judges	Decision Authority	Section Assignment Table (input)
3	Check eligibility (workload, experience, prior assignments)	Decision Authority	Judicial Record (input)
4	Request updated report of judges' programmed and past leaves	Decision Authority	Communication Flow to OJP Office (output)
5	Provide report of judges' leaves and availability	OJP Office	Leave Register (output)
6	Select judge–case match	Decision Authority	—
7	Communicate selected matches to the OJP Office	Decision Authority	Communication Flow to OJP Office (output)
8	Update workload data and notify judge	OJP Office	Updated Judicial Record (output)

Process Models The resulting BPMN models (Figures 3.1 and 3.2) depict the current *As-Is* configuration of the two processes.

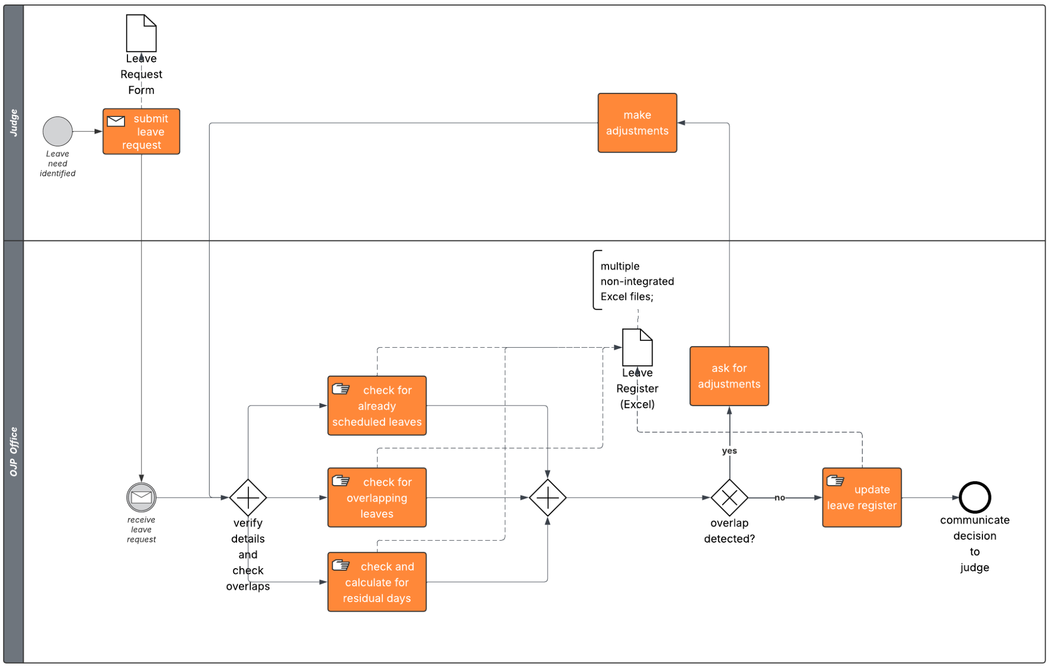

Figure 3.1. The AS-IS BPMN model of the Leave Management process describe the sequence of activities for managing staff leaves requests, including submission, verification, approval, and update of records. The process is currently supported by manual procedures and basic spreadsheet tools, ensuring operational continuity but with limited system integration.

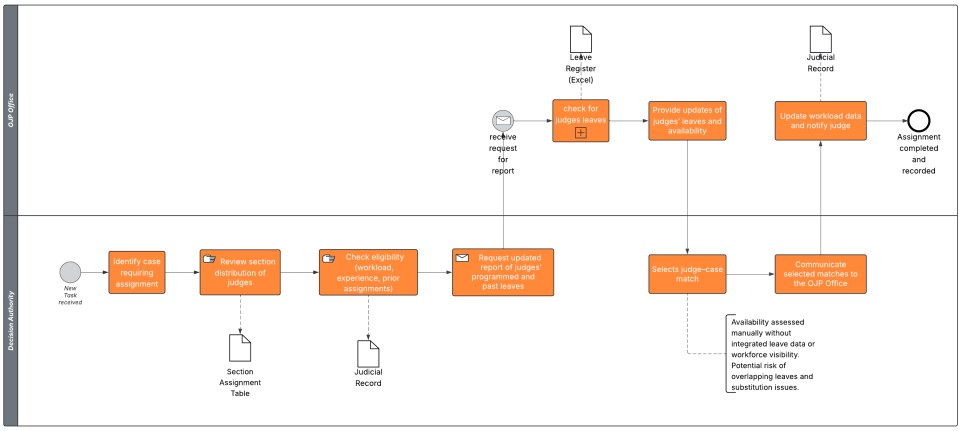

Figure 3.2. The AS-IS BPMN model of the Workload Assignment process describes how cases are distributed among the staff members. The process involves several steps, including the collection of case information, verification of staff availability, and manual case allocation. It partially depends on information from the Leave Management process to verify personnel availability and ensure balanced task distribution. The process currently relies on static data and limited system support for these checks.

3.3.3 Process Analysis

Following the Process Discovery, the modeled BPs are analyzed to identify inefficiencies, determine inter dependencies among activities, and establish requirements for process redesign and optimization.

Analytical Framework In line with the BPM life-cycle methodology [36, 131], the *Process Analysis* phase aims to identify inefficiencies and redesign requirements across four key dimensions:

- **Organizational:** roles, responsibilities, and communication mechanisms within and across processes;
- **Informational:** data objects, information flows, and integration of repositories;
- **Technological:** IT systems and tools supporting process execution;
- **Performance:** efficiency, timeliness, and monitoring of process outcomes.

These dimensions provide a structured perspective for diagnosing weaknesses and defining improvement requirements.

Analysis of the Leave Management Process The *Leave Management* process begins when a judge submits a leave request to the OJP Office, typically via email or in paper form. Requests are then recorded in a local Excel register and verified through a series of checks, such as overlapping dates or remaining leave balances, performed using separate files. The current configuration is mainly based on manual procedures and locally maintained spreadsheets, ensuring operational continuity but limiting automation and real-time validation. Unplanned events, such as illness or institutional duties, are managed through ad-hoc coordination, as no dedicated digital tool currently supports real-time updates or substitution planning.

Main process issues identified:

- **Manual verification across multiple files:** checks on leaves, balances, and updates are carried out manually, requiring coordination effort and increasing the possibility of inconsistencies.
- **Dispersed and non-integrated data sources:** information on requests, approvals, and staff availability is maintained in separate Excel sheets.
- **Absence of a unified leave management environment:** there is no single system supporting the full cycle of requests, approvals, and record management in an integrated way.
- **Limited visibility of real-time availability:** currently, no immediate view of availability of judges is provided, making scheduling and substitution planning less straightforward.
- **Reactive handling of unforeseen events:** no digital mechanism supports quick updates or reassignments.

Analysis of the Workload Assignment Process The *Workload Assignment* process concerns the distribution of judicial cases among judges and sections within the Court of Appeal. Its main objective is to ensure timely and equitable allocation of cases, taking into account individual expertise, current commitments, and availability of judges. Currently case assignment is managed directly by the Decision Authority and OJP Office. Decisions are primarily based on contextual knowledge supported by locally maintained information rather than digital procedures. While this approach offers flexibility and adaptability to specific circumstances, it also makes workload balancing less capable of quickly adjusting to unexpected changes or emerging needs.

Main process issues identified:

- **Manual cross-checks between different sources:** information on judges' previous roles, current assignments, and case relevance is verified manually across separate files and records.
- **Fragmented and non-integrated data:** case details, staff availability, and section composition are maintained in distinct sheets and static extracts from the judicial IS, requiring frequent updates and checks.
- **Limited visibility of staff availability:** there is no real-time view of who is currently present, or temporarily assigned to a case, complicating scheduling and planning.
- **Limited monitoring and historical overview:** the workload evolution of a case is not systematically tracked, limiting the possibility of trend analysis or performance evaluation.

Influence of the Leave Management Process The limitations identified in the *Leave Management* process directly affect the efficiency of the *Workload Assignment* process. The lack of synchronization between leave data and workload records can generate scheduling overlaps or delayed reassignments.

Main inter-process issues identified:

- **Limited integration between leave and workload data:** information on availability, absences, and assignments is maintained separately, without automated synchronization or shared access.

- **Partial visibility of current and planned availability:** the OJP Office cannot easily obtain a consolidated overview of judges present or on leave.

Findings The analysis of the *Leave Management* and *Workload Assignment* processes revealed several challenges influencing the daily operations of the offices. Table 3.7 summarizes the main *process issues* identified.

Table 3.7. Summary of Identified Process Issues within the OJP Office

Process Issue	Description	Analytical Dimension
Manual management of operational data	Activities such as recording, verification, and updates are carried out manually using spreadsheets, requiring constant coordination among staff.	Organizational Technological
Dispersed and non-integrated data sources	Information on leaves, workload, and section composition is stored across separate repositories (e.g., Excel files, paper forms), with limited interoperability and synchronization.	Informational
Limited visibility of availability and workload	No consolidated dashboard provides an immediate overview of judges' presence, absences, or current assignments, making planning and monitoring less efficient.	Informational Technological
Absence of automated or assisted allocation tools	Case distribution is managed without digital mechanisms supporting automated assignment suggestions.	Technological Performance
Lack of integration between related processes	*Leave Management* and *Workload Assignment* operate on separate datasets, with no real-time data exchange or automated updates between them.	Organizational Informational

3.3.4 Process Redesign

Based on the issues identified during process analysis, redesign actions were defined to enhance the digital maturity of the BPs.

The interventions on the BPs include:

- **Leave Management:** introduction of a digital module to centralize leave requests, automate verification steps (e.g., overlap detection, entitlement checks), and update availability data in real time.

- **Workload Assignment:** development of a complementary tool for recording and visualizing judicial assignments, supporting tasks allocation and ensuring historical traceability.

These interventions aim to change activities from fragmented and manual operations toward a structured and semi-automated process landscape aligned with BPM principles.

The redesign actions are summarized in Table 3.8, along with the corresponding critical issues and expected impacts.

Table 3.8. Summary of To-Be Redesign Actions and Expected Impacts

Identified Issue (As-Is)	**Redesign Action (To-Be)**	**Expected Impact**
Manual management of operational data	Introduction of digital forms and automated validation for leave requests and workload updates	Reduced manual effort and improved data accuracy in daily operations.
Dispersed and non-integrated data sources	Creation of a shared database connecting leave, workload, and section data	Improved consistency and easier access to updated information across processes.
Limited visibility of availability and workload	Development of dashboards providing real-time and planned availability views	Better overview of judges' availability, supporting proactive planning and scheduling.
Absence of automated allocation support	Introduction of digital tools assisting case distribution, based on up-to-date data on workload and availability	Simplifies allocation decisions, improves workload balance, and shortens assignment times.
Lack of integration between related processes	Connection of *Leave Management* and *Workload Assignment* modules through shared and synchronized data	Ensures consistent updates across processes and avoids overlaps or scheduling conflicts.

3.3.5 Process Implementation

The *Process Implementation* phase corresponds to the transition from the redesigned process model to its executable form [36]. In this case-study, direct deployment within the judicial information infrastructure was not feasible due to confidentiality, governance, and procurement constraints.

To address these limitations, the implementation was carried out in the form of proposed Proof of Concept (PoC) functionalities within the same existing IT environment of the Court of Appeal of Naples.

Accordingly, a set of functional prototypes was proposed using Excel-based automation and interactive Business Intelligence dashboards (e.g., Power BI mock-ups) to simulate the execution of digital processes.

Both tools were proposed as prototype components compatible with the current Information and Communication Technology (ICT) environment of the Court, serving as a *PoC* for gradual DT.

Leave Management Process Within the implementation phase, a set of *Leave Management functionalities* were proposed within the existing Excel environment of the *OJP Office*, serving as a PoC for DT.

The main functionalities proposed are summarized below:

- **Digital recording of leave requests:** structured data-entry interface for the registration of leave and absence requests, including automated checks on data completeness and potential overlaps.
- **Automated data validation:** embedded control functions verify input accuracy and detect duplicates in real time, improving reliability and reducing manual verification.
- **Centralized historical archive:** automatic generation of a structured log of all recorded leaves, ensuring traceability and enabling consultation or export for analytical purposes.
- **Summary reporting functions:** dynamic extraction of aggregated data (e.g., identification of judges with extended or accumulated leaves) to support monitoring and planning.
- **Integrated data update procedures:** coordinated update of records across sheets to maintain consistency between monthly summaries,

residual balances, and new entries.

- **Configurable calendar view:** parameterized visualization of months and reference years to align the tool with institutional planning periods and reporting needs.

The features support the recording, validation, and monitoring of leave requests of judges through automated functions for data entry control, overlap detection, and calendar-based visualization of planned and ongoing absences.

Workload Assignment Process Similarly, a set of **Workload Assignment visual dashboards** were proposed within the existing Microsoft environment of the Court of Appeal of Naples. These dashboards serve as a PoC, demonstrating how integrated data visualization can support fairness, transparency, and operational control in case allocation. The prototype consolidates information on judicial personnel by combining:

- *Personal data:* biographical attributes, tenure, and roles held within the Court;

- *Work data:* real-time indicators of availability, current assignments, and temporary absences due to illness, missions, or other justified reasons.

The main functionalities proposed are summarized below:

- **Management of judicial personnel data:** structured interface for recording and updating judges' profiles, including role, section, and assignment details, with built-in consistency checks.

- **Workforce overview and availability status:** visual panels providing a real-time view of active, assigned, and temporarily unavailable judges, supporting immediate understanding of workforce composition.

- **Assignment management:** functions enabling the registration, modification, and monitoring of assigned tasks or institutional duties, with progress tracking through status indicators (e.g., open, ongoing, completed).

- **Integrated list and monitoring views:** consolidated tables showing all current and past assignments, filterable by section, judge, or assignment status, facilitating supervision and coordination.

- **Reporting and summary visualization:** dashboards summarizing workload distribution and judge availability across sections, providing a foundation for data-informed workload balancing and planning.

These proposals are demonstrative, functioning as an experimental PoC that validates the feasibility and operational relevance of process automation, highlights system requirements, and defines the conceptual framework for a future, fully integrated IS.

3.3.6 Process Evaluation

Following the design of the proposed PoC artifacts, a **Process Evaluation** was conducted to assess the perceived usefulness within the operational context of the Court of Appeal of Naples. The evaluation focused on validating their conceptual soundness and practical relevance through stakeholder feedback. The assessment aimed to verify the coherence of the proposed functionalities with process needs and to identify the most valuable and feasible components for future adoption. Then, the validated functionalities were consolidated into a set of **functional requirements**, defining the specifications for an integrated IS.

The evaluation served as a validation and requirements deployment step within the BPM life cycle, transforming PoC design into actionable specifications for digital process integration.

Table 3.9. Summary of Functional and Additional Requirements

ID	Requirement	Description
Functional Requirements		
FR1	Unified Data Repository	Integration of leave, workload, and availability information within a shared environment to eliminate data silos and ensure consistency across processes.
FR2	Automated Data Validation	Automatic checks for missing, inconsistent, or duplicated entries during data entry and update operations.
FR3	Real-Time Data Synchronization	Automatic propagation of changes across all relevant modules to maintain real-time data coherence.
FR4	Rule-Based Validation and Approval Workflows	The system automatically enforces configurable rules for validating and approving operations, such as detecting overlapping leave requests, verifying judge availability, and applying authorization hierarchies.
FR5	Role-Based Access Management	User profiles (e.g., administrative staff, section president, system administrator) determine scope of access, editing rights, and data visibility.
FR6	Analytical and Visualization Dashboards	Interactive dashboards summarizing workload distribution, staff availability, and leave balances to support managerial supervision.
FR7	Historical Traceability	Automatic logging of operations (insertions, updates, validations) to ensure process traceability.

3.4 Considerations

The application of the BPM life cycle to the Court of Appeal of Naples demonstrated its suitability for identifying improvement opportunities and guiding digital transition, even in contexts with limited technological maturity. The proposed PoC artifacts served as conceptual demonstrations of how process-aware functionalities could enhance coordination and data consistency across administrative units. Although not implemented, these prototypes helped visualize the potential for greater process transparency, operational continuity, and resource traceability, which are essential foundations for future data-driven management within judicial administrations.

3.4.1 An Adapted BPM Framework for the Judicial Domain

This empirical research helped in the definition of an **adapted BPM life cycle** tailored to constrained environments, such as the judicial domain. The proposed framework builds upon the classical BPM life cycle [36], introducing more flexibility in the execution and evaluation of its phases. Specifically, the traditional *Process Monitoring* phase is replaced by a **Process Evaluation** phase. This adaptation acknowledges the empirical constraints of the judicial context, i.e. data confidentiality, and procedural rigidity, which hinder the operationalization of redesigned processes within the institutional IS landscape. Rather than focusing on performance monitoring through KPIs, the *Process Evaluation* phase allows for a qualitative assessment of the redesigned workflows and the identification of *functional requirements* emerging from the PoC functionalities. It serves as a structured mechanism for capturing user feedback and validating process improvements in the absence of an operational IS, thereby obtaining functional requirements for future software developments.

Under these conditions, the BPM life cycle 3.3 concludes with the development of a PoC, which serves simultaneously as:

- a validation of the redesigned processes in a controlled environment;
- an instrument for deploy technical and organizational requirements for future digital implementation;
- a feedback loop that informs the next iteration of process improvement and system design.

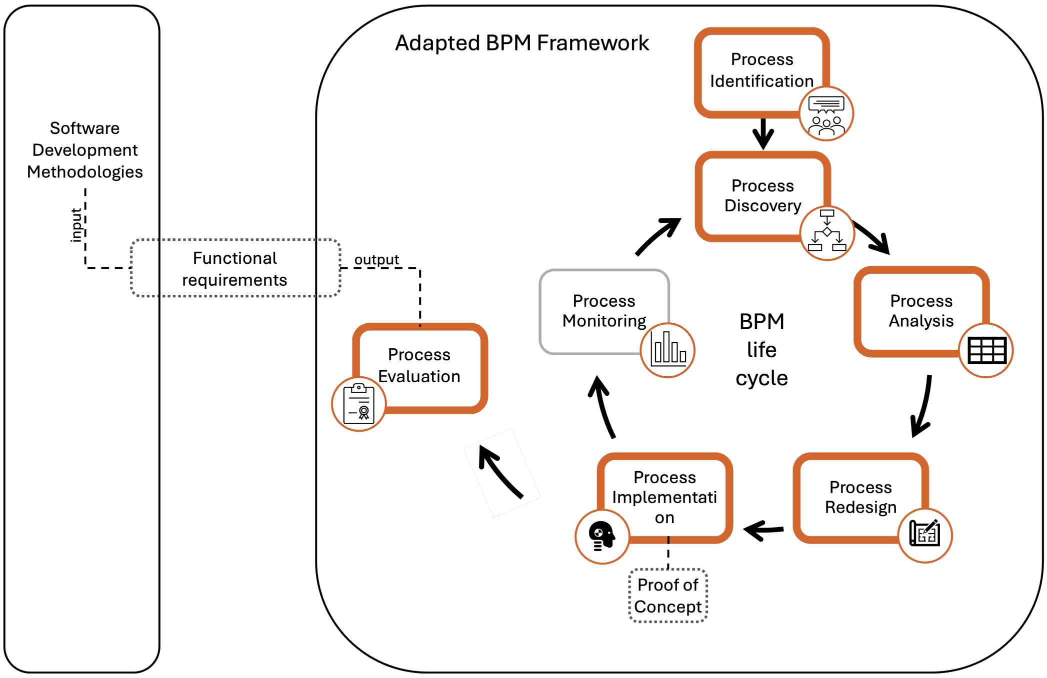

Figure 3.3. Adapted BPM life cycle for constrained public-sector environments

This framework enables a gradual and sustainable approach to DT, ensuring methodological continuity even in data-sensitive contexts. By iterating between redesign, PoC implementation, and evaluation, organizations can progressively evolve toward digital maturity without requiring disruptive technological change.

3.4.2 An Emerging Object-Centric Perspective

The analysis conducted so far highlighted a fundamental conceptual feature: the organizational processes of judicial personnel management are inherently **object-centric**, as observed previously in the *building permit process* (see Chapter 2). From the discovery of the processes, it became clear that both *Leave Management* and *Workload Assignment* revolve around the **availability of judges**. In fact, leave requests update periods of unavailability, while case assignments increase active workload. Each event modifies the same underlying object state, generating a network of interdependent process instances rather than independent workflows.

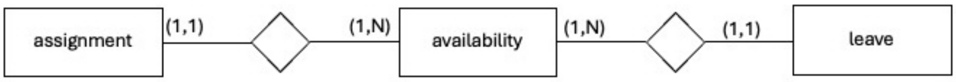

Figure 3.4. Conceptual representation of the object-centric nature of judicial personnel processes. The judge's availability acts as a shared object updated by events from the *Leave Management* and *Workload Assignment* processes, linking heterogeneous event streams into a unified life-cycle.

This observation aligns with the concept of OCELs [127, 129]. In fact 2 shows that in complex organizational domains, multiple processes often operate concurrently over shared business objects.

3.5 Conclusion

The case study demonstrated the applicability and effectiveness of the BPM life-cycle approach in a judicial setting, demonstrating how process-aware methodologies can support DT in highly regulated and low-maturity PAs. The findings provided insights into the organizational and technological barriers that hinder transformation and established a foundation for future process automation and continuous improvement initiatives within the domain.

Chapter 4

Predictive Process Monitoring in Real-Life Contexts: A Comparative Review of Classical and Object-Centric Approaches

I can't change the direction of the wind, but I can adjust my sails to always reach my destination.

Jimmy Dean

4.1 Predictive Process Monitoring

PPM techniques are emerging as part of the more general research scenario of PM. They play a crucial role in the continuously evolving process of DT by constantly supporting the organizational decision-making processes providing (accurate) predictions on the future behavior of processes. As introduced in Chapter 2, PPM extends traditional PM techniques by integrating predictive analytics to anticipate process behaviour and support proactive decision-making. Existing methodologies predominantly rely on *classical event logs*, which represent processes executions through a single case identifier. Recently, the emergence of OCELs enabled new PPM approaches capable of capturing interactions among multiple processes and shared objects. Such object-centric perspectives are particularly relevant for complex administrative scenarios, such as the *Building Permit Process* discussed in Chapter 2, where multiple entities (applications, inspections, payments, documents) coexist and interact. A similar multi-object dynamic was also observed in the judicial context analyzed in Chapter 3, where the availability of the judge is simultaneously influenced by *Leave Management* and *Workload Assignment* processes. These examples illustrate how PA workflows inherently exhibit object-centric characteristics, making OCEL-based PPM approaches especially suitable for modeling and predicting their behavior.

Therefore, based on the aforementioned considerations, in this chapter PPM approaches are analyzed, to classify newly introduced techniques. Indeed, even if PPM approaches used so far in the literature can be distinguished based on several characteristics (such as the application domain, addressed prediction task, employed methodology, chosen evaluation metrics, performances results etc..), in this survey it is added a new classification feature related to the distinction of PPM approaches based on the different input Event Logs being processed.

The main goal of this chapter is to provide a SLR on PPM, discussing the research on Event Log type by distinguishing Classical Event logs and OCELs, and then discussing the most widely used methods, the different prediction tasks, the innovative aspect such as methods, concepts or assumptions that characterize the approaches proposed so far and the application domain. By providing a deep distinction between Classical Event Logs and OCELs we aim to highlight the possibility of using the already known methodologies in a new context by providing the important details and findings of the approaches in a structured way. In addition, to the best of our knowledge this is the first survey distinguishing between the different Event Log type. Moreover, this chapter aims to contribute to the PPM research field by providing a rather complete starting point for researchers approaching the methodology for the first time.

4.2 Related Work on PPM Surveys and Benchmark

A plethora of surveys and benchmarks have been proposed for the PM research area, with the aim of providing a contextual framework and structure. The book [125] represents the most comprehensive handbook on PM, covering the subject in detail from theoretical definition to practical use. Also, the authors in [7] conducted a survey on the process discovery problem, and [20] offered an overview of different approaches in the context of conformance checking.

Despite being a relatively new research field, many studies have been carried out also in the area of PPM. The studies proposed in the field of PPM can be divided into three main categories according to the achieved objective:

- *surveys* which are classical literature reviews categorizing papers based on a range of characteristics;
- *surveys and benchmarking analysis* where both a classification and assessment of the identified methods are done;
- *benchmarking analysis* where only an assessment of the methods is done;

The purpose of this section is to highlight any study which may provide the reader with a comprehensive introduction to the field of PPM, and which may distinguish our SLR from those previously undertaken, as Table 4.1 shows.

The literature review conducted by the authors in [76] evaluates 39 different research studies in PPM, provided with data-sets and quality metrics. The approaches are categorized into process-aware and non-process-aware, based on the output variable type (numerical or categorical), and further distinguishing between regression and classification methods. They highlight the lack of dataset availability, challenges in data encoding, and limited comparisons between methods. The key finding is that regression is linked to process-aware methods and classification to non-process-aware methods. Most methods use ML (especially decision trees), with common tasks focused on predicting remaining time and delays. Performances are evaluated using Root-Mean-Square Error (RMSE), accuracy, and precision.

The literature review conducted by the authors in [33] represents a comprehensive survey of PPM approaches, analyzing 51 studies. The analyzed studies are categorized by prediction goals (e.g., time, risk, cost) and input data, with a focus on Event Logs and contextual information. The authors also assess predictive techniques such as regression, Neural Network (NN), and clustering. They emphasize the need to tailor predictive models to business requirements and contexts, and evaluate model performance using precision, recall, and F-measure. Their work provides a framework for selecting appropriate methods to improve operational efficiency and guide future research.

In [120] the authors make both a review and a benchmark on outcome-oriented approaches. A SLR on 7 primary studies and subsumed studies (representing more recent versions that do not contribute substantially compared to primary studies) was conducted. They categorize the offline phase based on steps like prefix selection, trace bucketing, and sequence encoding, and the online phase for real-time prediction. They highlight that prefix selection and classification algorithms are independent and focus their taxonomy on trace bucketing and encoding methods. The benchmark, conducted on 11 approaches across 7 studies, evaluates performance in terms of prediction accuracy, earliness, and computation time (CT).

The study in [134] provides a comprehensive overview and benchmark analysis on remaining time prediction approaches. The analysis is done on 24 relevant studies on remaining time prediction, categorizing methods into discriminative (non-process-aware) and generative (process-aware). They recommend prefix bucketing and encoding for better performance and benchmark 16 real-life datasets. The evaluation uses accuracy, Mean Absolute Error (MAE) and earliness with prefixes up to 20 events, performing hyper-parameter optimization via grid search. Their findings emphasize the importance of clustering event logs and using appropriate encoding for improved prediction accuracy.

In the work [115] the authors explore a benchmark analysis to compare ML techniques for classification tasks, with a specific focus on next event prediction approaches. The analysis is performed across 6 publicly available datasets. They test 20 classifiers with various validation methods and sampling procedures, finding that the Credal Decision Tree (C-DT) consistently outperforms others in accuracy, showing robustness across event log variations. C-DT is particularly effective in low-variability logs but also performs well with high-variability logs.

The survey conducted in [55] provides an overview of DL approaches utilized in PPM. In particular the focus is on Long Short Time Memory (LSTM) networks and Stacked Autoencoders. They explore various encoding techniques, including Word Embedding, One-Hot Encoding, and N-grams, and categorize predictions into numerical, categorical, and next event types. The survey evaluates models using precision, recall, accuracy, and F-measure, and discusses the effectiveness of different DL models, highlighting areas for future research in improving PPM performance.

The study proposed by the authors in [99] sets out a review analysis and benchmarking experimentation of DL techniques for PPM. They categorize methods based on input data types, prediction tasks, NN features, sequence encoding, and event encoding. Sequence encoding involves transforming traces into tensors, while event encoding represents events as feature vectors. The authors evaluate 10 approaches across 12 publicly available datasets. They apply Bayesian analysis and hierarchical models to rank and compare the performances of the methods, providing insights into their applicability and effectiveness for predictive models in PPM.

The SLR provided by the authors in [113] is focused on explainable predictive BP monitoring (XPBPM) approaches. They analyze 19 studies, classifying the approaches using concepts from explainable AI XAI, such as intrinsic models, post-hoc explanations, and evaluation methods. The review highlights the shift from intrinsically interpretable models to DL-based approaches, which offer superior predictive performance. However, post-hoc explanations remain underdeveloped, and the authors stress the need for rigorous evaluation to assess the effectiveness of explainability techniques.

The work in [86] performs a SLR on 32 papers about DL methods in PPM. They categorize methods based on Network Architecture, Data Preprocessing and Feature Engineering, and Prediction Target. They identify Feed-Forward Networks (FFNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) as popular architectures. They stress the importance of balancing model simplicity and generalization, especially regarding prefix length and parameter selection. For preprocessing, one-hot encoding and embedding are common for categorical data. The paper also evaluates prediction levels and types, identifying process-level and time-step predictions, with regression and classification tasks. They outline five major challenges in input data and preprocessing.

The following Table 4.1, summarizes some relevant features of the analysed survey that could be used for a quick pick when searching for an entry point to dive into the topic.

Table 4.1. Comparison of related works based on selected variables, ordered by year and name. Specifically, **Focus** specifies whether the study concentrated on a particular area (e.g., only DL methods, next activity prediction, etc.). **Query** details the research query employed in the study. **Database** identifies the research databases where the queries were executed. #**Works** represents the number of studies included in the analysis. **Year** indicates the publication year of the study. If the information is not provided in the study, it is marked as Not Available (NA).

Study	Focus	Query	Database	#Works	Year
Marquez et al. [76]	all	(predictive monitoring AND business process) (business process AND prediction)	Scopus WOS Google Scholar	41	2017
Di Francesco-marino et al. [33]	all	(predictive OR prediction) AND (business process OR process mining)	Scopus SpringerLink IEEE Xplore Science Direct ACM Digital Library Web of Science	51	2018

Continued on next page

Study	Focus	Query	Database	#Works	Year
Teinemaa et al. [120]	outcome prediction	(predictive process monitoring) (predictive business process monitoring) (business process prediction)	Google Scholar	7	2019
Verenich et al. [134]	remaining time prediction	(predictive process monitoring) (predictive business process monitoring) (predict (the) remaining time) (remaining time prediction) (predict (the) remaining * time)	Google Scholar	24	2019
Harane et al. [55]	DL approaches	NA	NA	3	2020
Rama et al. [99]	DL approaches	NA	NA	21	2021
Stierle et al. [113]	Explainable Artificial Intelligence (XAI)	(business process AND predict* AND interpretable OR explainable OR comprehensible)	Scopus EbscoHost Web of Science	19	2021

Continued on next page

Study	Focus	Query	Database	#Works	Year
Neu et al. [86]	DL approaches	business process * predict* DL OR neural networks*	Web of Science Science Direct IEEE Explor ACM Digital Library Springer Link	32	2022
Our study	all	(predictive process monitoring AND process mining AND business process) (predictive monitoring OR predictive business process monitoring OR prescriptive process monitoring) (predictive process monitoring object centric event log) (object centric event log AND predictive process monitoring)	Scopus IEEE Google Scholar	47	2024

In addition, the study [21] identifies key challenges within the field of PPM and explores future directions. This work serves as a valuable resource for advancing the field and addressing the limitations of existing contributions.

From Table 4.1, we can observe some of the main differences between our SLR and the ones analyzed in the current study. Specifically, our work is not limited to a particular approach, thereby avoiding restrictions to specific prediction tasks or methods applied, as we can see from the "Focus" column. Our proposed SLR falls under the category of surveys that analyze PPM approaches addressing all prediction tasks (i.e., numerical, categorical, and outcome-based) and comparing all methods implemented (both model-based and BPM-model-based), without performing a benchmarking analysis. This study belongs to the category of research that covers all types of PPM, without focusing on just one kind of prediction task or method.

Another key distinction is in the research query used, from "Query" column, which includes relations with OCELs. To the best of our knowledge, no other review study has considered classifying PPM methods based on the type of Event Log input, specifically distinguishing between OCELs and Classical Event Logs.

Finally, another distinction obviously lies in the more recent period of investigation considered in our study.

4.3 Search Methodology

The aim of this section is to present the methodology to perform a deep overview and classification of methods for PPM. To this purpose, we leverage the SLR methodology proposed by [62] to structure our research.

4.3.1 Research Questions

The initial step driving our investigation concerns the formulation of the research questions. While numerous studies have explored PPM approaches from various perspectives, there is limited research on the impact of incorporating information from OCELs to enhance predictions. The relevance and effectiveness of this information depend on the specific prediction task, the method employed, the advancements in research and the application domain. Despite the shared goal of advancing research, different proposals contribute in diverse ways, making it challenging to structure and discern the actual improvements and the remaining gaps. These considerations led to the formulation of the following research questions:

RQ_1: What is the nature of the input Event Logs in PPM approaches?

RQ_2: How do the prediction tasks, the methods used, and the application domains differ based on the type of Event Log?

RQ_3: To which phase of the PPM framework does the scientific contribution of each research proposal specifically relate?

These research questions are the basis for the comprehensive exploration and analysis carried out in our study.

4.3.2 Study Selection

The study selection process involved the formulation of queries aligned with the proposed research questions. These queries are generally elaborated to address the research questions without limiting the scope to specific prediction tasks or methodologies. To address the identified gaps and formulate our research questions effectively, we leveraged specific search queries. These queries are designed to retrieve relevant literature on the integration of OCELs in PPM within the domain of PM and BP.

The following search queries were designed to identify relevant studies for our SLR, with the aim of ensuring coverage of all aspects related to PPM and OCELs. More in details:

- "predictive process monitoring object centric event log" and "object centric event log AND predictive process monitoring": these queries specifically target studies utilizing OCELs in the context of PPM, directly addressing the first research question and ensuring the inclusion of relevant works.

- "predictive process monitoring AND process mining AND business process" focuses on literature dedicated to PPM, ensuring that the retrieved studies are directly aligned with the main research topic. Including PM broadens the scope to encompass studies related to the analysis, improvement, and optimization of BP through the extraction of insights from Event Logs. BP further expands the scope by incorporating literature on BP, providing a comprehensive understanding of the broader context in which PPM and PM techniques are applied.

- "predictive monitoring OR predictive business process monitoring OR prescriptive process monitoring" focus on capturing studies that use synonyms or related terms, addressing the risk of excluding relevant research due to variations in terminology.

The combination of search terms and adoption of a structured approach ensures the systematic identification of studies across interconnected domains. This methodological approach is designed to facilitate a thorough and reliable review while capturing a broad range of research exploring various aspects of PPM.

These queries were executed across three platforms: Scopus, IEEE, and Google Scholar, selected as they cover scientific publications within both the fields of management and computer science.

In order to focus the research on relevant studies, an analysis was conducted on the Scopus platform to identify PPM-related studies, focusing on the number of publications over the years (Figure 4.1) and their citation trends (Figure 4.2). The analysis revealed an increase in both the number of publications and citations starting from 2019, highlighting significant advancements in the field during this period.

This tendency initially prompted the decision to concentrate the research on studies published from 2019 on wards. Nevertheless, to ensure a comprehensive review, it was also decided to include significant studies published prior to 2019. This decision was taken in order to capture foundational contributions and provide a complete overview of the evolution of the research domain.

Inclusion criteria comprise studies focused on PPM comprehending different predictive modeling approaches such as ML, data mining, AI, or SL, using different Event Logs (OCELs and Classical Event Logs), using both real-world data or artificial data.

In order to consider only relevant studies, we designed a range of exclusion criteria to assess the relevance of the studies

EX1 The study does not actually propose a PPM method. Studies that do not propose or discuss methods, techniques, or applications explicitly focused on PPM were excluded. With this criterion, all studies focused on other PM techniques which do not include PPM were not included.

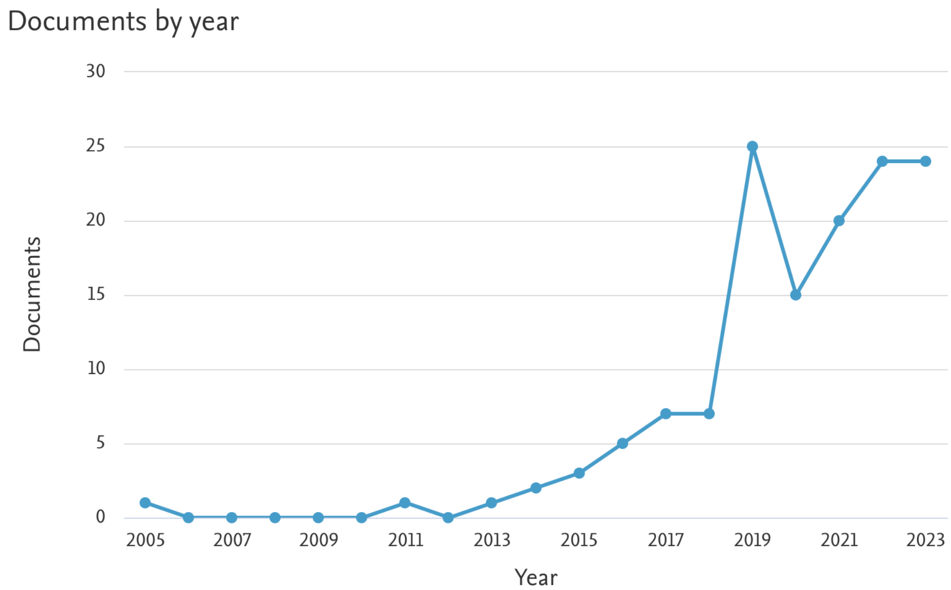

Figure 4.1. Chart reporting the number of published papers in the field for year on scopus database

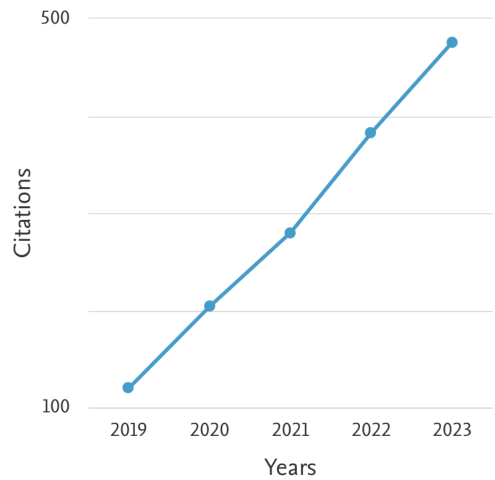

Figure 4.2. Chart reporting the number of citations in the field for year (from year 2019 to 2023) on scopus database

EX2 The study perform general predictions, which are not in the field of PM. In this way we excluded all the studies that perform predictions, but are not focused on leveraging specific PM techniques.

EX3 The study does not leverage PM techniques. Studies focusing on BP without employing specific PM techniques for prediction were exclude.

To apply the research methodology outlined above, queries were executed on three selected platforms: Scopus, Google Scholar, and IEEE Xplore. Starting with Scopus, the scope limitation filter was applied to the areas of Computer Engineering, Engineering, and Business Management, ensuring the inclusion of papers that focus on the technological and managerial aspects of PPM.

A similar query was run on Google Scholar, limiting the results to the first 10 pages, based on the assumption that the most relevant and influential papers are typically placed within the first few pages, thus ensuring the capture of the majority of pertinent literature. The same methodology was applied to IEEE Xplore, using the same query strategy along with the corresponding inclusion and exclusion criteria.

Subsequently, by applying the inclusion and exclusion criteria, the selection of studies was refined based on their relevance, adherence to the research topic, and influence. Non-relevant studies, identified through a preliminary analysis of keywords, titles, and other metadata (e.g., studies focusing on different PM approaches, decision-making, or unrelated metrics), were removed. This process was applied to the results from all platforms, and duplicates were eliminated.

The remaining papers were then further assessed for their relevance to the research objectives. A more in-depth evaluation of the content and abstracts was conducted using the inclusion and exclusion criteria. Papers that were not aligned with the research focus were excluded from the final selection.

After removing duplicates and non-relevant studies, the final set of papers, based on their alignment with the research focus, resulted in a collection of 47 articles.

4.4 Methods Classification

According to the results and the research questions, the obtained studies can be categorized by the following dimensions:

- *Event Log Type*: the type of event log considered for the application of the PPM is used to distinguish among the approaches;
- *Prediction Task*: we classify the considered papers based on the type of prediction;
- *Method*: this dimension takes into account the predictive model being leveraged;
- *Contribution Area*: this parameter takes into account the innovation gathered by the specific approach;
- *Domain*: the domain or context in which the framework has been tested or applied.

The selected papers will be discussed w.r.t. the above mentioned dimensions in what follows.

Event Log Type Most existing PPM approaches rely on the assumption that process executions can be represented through a single case identifier, as in Classical Event Logs. However, in real-life scenarios, where business processes are often interdependent, this assumption rarely holds. To overcome this limitation, OCELs were introduced (see Chapter 2), enabling the representation of multiple interacting objects and concurrent process instances, which affects the design of prediction algorithms. OCELs provide richer information about process executions compared to Classical Event Logs, as they are structured as graphs that require specific handling for effective utilization. In this context, it is important to differentiate between PPM approaches that deal with Classical Event Logs and those that handle OCELs. This distinction is essential to understand how researchers addressed the challenges posed by different types of event data, with OCELs requiring specific modelling and processing strategies. Examining the relationship between these two perspectives also provides valuable insight into how the evolution of event-log representations can inform and extend current research in the field.

Prediction Task According to [32], there are three types of prediction tasks: outcome-based, numeric value prediction and next event prediction, which heavily affect the prediction method to be used. The incorporation of this feature into the classification process enables the identification of the prevalence of specific prediction tasks, determination of their association with particular methods, and exploration of their relationship to the distinction between different types of Event Logs. This approach facilitates a more in-depth comprehension of the manner in which prediction tasks interact with the characteristics of Event Logs, thereby offering insights into the tailored methods and strategies that are required for effective PPM.

Method The methodologies implemented in PPM can be categorized into two main families as discussed in Chapter 2: model-based approaches, that leverage explicit representation of the process and non-process aware approaches, relying on the implicit representation of the process. Among the non-process aware methods, ML and DL are the most commonly used. While ML methods are related to Statistical Models able to make predictions based on traditional algorithm, DL approaches are based on DNN, therefore making predictions with an higher level of complexity. Research proposals often employ various methods influenced by factors like the Event Log type, Prediction Task, and application Domain. The method choice also considers factors such as training time and interpretability of the predictions. DL methods are often referred to as "black box" due to their lack of transparency; they can provide highly accurate predictions which in turn require significant training time and lack explainability. ML methods, on the other hand, offer more explainable results and faster training, but may sacrifice prediction accuracy.

Each approach balances accuracy, explainability, and training time based on the requirements of the task. Including the "Method" feature in the classification allows us to examine the most commonly used methods, identify priorities (accuracy, explainability, or training time), and whether the Event Log type influences the choice of method, if certain methods are more prevalent, and how changes in the Event Log can affect the applied method. This analysis will shed light on the interaction between Event Log characteristics and method selection, providing insights for more tailored and effective approaches in PPM.

Contribution Area The contribution area is specifically linked to the PPM framework, addressing the particular contributions of research papers, such as encoding techniques or modifications to prediction methods. Each approach proposed in the literature presents distinctive features that make the studies innovative. Introducing a specific contribution area ensures that each approach can be differentiated by categorizing it into a particular pattern. This initially makes it challenging to directly compare results across studies using different approaches.

However, understanding and categorizing contributions based on their innovative aspects allows researchers to identify domains where significant improvements have been made. This framework serves as an essential starting point for newcomers to the field, helping them quickly pinpoint areas where their contributions can have the most impact.

Domain The domain in which the PPM framework is applied plays a crucial role in determining the challenges and opportunities associated with PPM. Different domains impose unique requirements on Event Logs, Prediction Tasks, and Performance Metrics. Including the "Domain" feature in the classification allows the investigation of connection among domains, Event Log structures, performance metrics and methods. The analysis of the domain of application provides insights into how the proposed frameworks generalize across diverse real-world scenarios and how the type of Event Log is linked to the domain. For instance, the public administration domain may handle OCELs differently compared to Traditional Event Logs, reflecting the unique requirements and characteristics of this domain.

4.5 Approaches Analysis and Classification

The choice made so far in the previous section about the classification dimensions, led to the distinction of the analyzed approaches in the following subsections. Specifically, the examined proposals are shown in detail by dividing them according to the Event Log Type.

4.5.1 Object-Centric Event Logs

The application of traditional PPM techniques in real-life scenario pose several challenges. In fact, classical PPM approaches need to be adapted to gain insights from object interactions stored with OCELs. Furthermore, PPM approaches focusing on OCELs are recent and in constant evolution. As a result of the above considerations, the proposals in the PPM area that deal with OCELs are still few and employ a variety of different approaches to manage OCELs. This section presents a detailed overview of the PPM approaches that deal with OCELs, including the chosen method, the prediction task, and the contribution area.

Gherissi et al. [50] propose a PPM methodology for predicting the next activity, next event time, and remaining time. This methodology utilises LSTM networks on OCELs, with the introduction of the enriched log notion, which considers object interactions, and the addition of a layer to the LSTM network, resulting in improved prediction accuracy. The performance of the approach is measured using accuracy for next activity prediction and MAE for time prediction. Experiments conducted on a publicly available dataset concerning customer order management with three object types have demonstrated the significant impact of feature enrichment on next activity prediction and the impact of the LSTM layer on the accuracy of remaining time predictions.

Rohrer et al. [104] propose a PPM methodology using Generative Adversarial Networks (GANs) for OCELs. The GAN architecture, incorporating a Sequence-to-Sequence model with LSTM layers, predicts next activity sequences and timestamps. The utilization of rich object-attribute information enhances prediction accuracy, which is evaluated using sequence similarity and MAE, outperforming traditional event log-based methods. The experiments are conducted on a synthetic dataset, about an order process with four different object types.

Galanti et al. [48] propose a framework that integrates information about *object interactions* through feature representation. The authors propose a framework using CatBoost and feature aggregation to integrate object interactions in OCELs. It predicts KPIs relevant to stakeholders, evaluated with MAE and F1-score and demonstrating significant improvements over traditional single-id approaches. The application domain is on confidential real data of an utility provider company about production and distribution of electricity and gas having five object types.

Adams et al. [1] combine Graph Neural Networks (GNNs), Graph Embeddings, and Flattened Event Logs for predicting next activities, timestamps, and remaining times. The use of GNNs preserves graph structures, and outperformes traditional flattened logs. The evaluation metrics include MAE for regression and accuracy for classification. Experiments are conducted on real-life event log of a loan application business process matched with an offer from a financial institution, having only two object types.

Galanti et al. [46] investigate whether the enhanced prediction performance of GNNs in PPM justifies their extensive training time. In details, they compare GNNs, LSTMs, and CatBoost for KPI prediction in OCELs. According to the experiments evaluated with MAE for numerical KPIs and F1-score for categorical KPIs, while GNNs effectively capture object interactions, CatBoost outperforms in training time and prediction accuracy. Experimentations are conducted in both an Italian utility-provider company (energy company) focused on the production/extraction of electricity and gas with 5 object types, and in a well-known American technology company with 3 object types.

Knopp et al. [63] develop object-centric process simulation models on OCELs by combining object graphs, Petri nets and stochastic predictor with a log frequency-based classifier for next activity prediction. In the prediction they rely on single objects predictions with objects being flattened and then enriched with object features. The evaluation metrics include Earth Mover's Distance (EMD) to assess the accuracy of simulations and quality of generated objects, behavioral conformance and timing behavior. For the experimentation two datasets are considered, an artificial log of an order management process (OM), and artificial procure-to-pay process (P2P).

Niro et al. [87] propose a framework for anomaly detection in using Graph Convolutional Autoencoder (GCNAE) on OCELs. The approach is compared to standard autoencoder (AE) and a LSTM autoencoder (LSTMAE) flattening the OCELs. The efficacy of this approach is evaluated using two distinct datasets, encompassing the loan application process with two object types and a synthetic DS2 dataset that simulates an order management process with three object types. The evaluation metrics employed include F1-score, Precision-Recall (PR) AUC and Receiver Operating Characteristic (ROC) Area Under the Curve, and Recall, with the results indicating enhanced detection capabilities.

Ruffini et al. [105] in their preliminary proposal set out to investigate the use of GNN on OCELs, and the representation of additional relations among process activities (i.e., relations between event data, or ordering relations different from the directly following ones) with a view to developing tailored XAI techniques to analyse the impact of the different relations on the predicted outcome. The experimental results will be evaluated using well-established classification performance metrics, such as the Area Under the Curve (AUC) and the F1-score, which are typically applied to publicly available datasets.

Smit et al. [108] present an encoding technique for OCELs leveraging multi-dimensional data structure for remaining time prediction, surpassing flattening operation. This method, based on Heterogeneous Object Event Graphs (HOEGs), employs a graph-based encoding that preserves the relationships between objects. A heterogeneous GNN processes the data and is evaluated using MAE, MSE, and MAPE, showing improvements in MAE and MAPE compared to existing methods. Scalability is assessed by measuring training and prediction times, and generalizability is tested using a train/validation/test split. The experiments use a loan application OCEL with three object types, an order management process with five object types (three selected for analysis), and a real-life operational system from a large financial organisation with three object types.

Van der Aalst et al. [130] address the inclusion of Object-Centric Event Data (OCED) in many PM tasks, by showing the flattening operation, the convergence and diverge problems. They introduce Object-Centric PM (OCPM) to enhance traditional methods by capturing multi-object interactions in Event Logs. The contribution is a conceptual framework and formalization of OCED, enabling comprehensive process insights without proposing specific algorithms. The experimentation is done on both OC-PM and Celonis Process Sphere, having six object types.

Berti et al. [12] propose a graph-based feature extraction method for anomaly detection and throughput time prediction using OCELs transformed into interaction graphs. The features representing OCELs are extracted and fed into ML models, including isolation forests for anomaly detection and regression models for throughput time prediction. XAI techniques, such as RIPPER, interpret model results. Anomaly scores from isolation forests identify outliers, while throughput time prediction is evaluated using MAPE and RMSPE, demonstrating improved accuracy. Experiments are conducted on several logs, i.e. order management, order to cash (O2C) and procure to pay (P2P) logs, and a recruiting process, each with multiple object types.

Chifeng et al. [23] propose the Object-Centered Predictive Process Prediction (OCPPP) method for business process prediction in intelligent healthcare. This method combines Colored Petri Nets (CPNs) for explainability with AI models for analysing multimodal data. Specifically, YOLOv7 is employed for the detection of queue lengths in image data, while LSTM is utilised for the prediction of queue sizes using time-series data. The performance of YOLOv7 is evaluated using metrics such as Precision, Recall, and mAP, while the accuracy of LSTM is assessed through MSE, RMSE, MAE, and R-squared. The effectiveness of the framework in providing accurate, explainable predictions for complex processes is demonstrated by experiments conducted on healthcare datasets, including COCO2017 images, hospital surveillance images, and registration logs.

Le et al. [68] in prescriptive PM propose a PhD project aiming to create a concept-drift-aware prescriptive module using counterfactual reasoning and CatBoost. The system aims to predict KPIs and recommends next activities, leveraging business simulation and A/B testing for evaluation, showcasing adaptability to dynamic processes.

Smit et al. [109] propose HOEG, which integrates events and objects into a graph structure with diverse node types, enhancing the encoding for predicting remaining times in OCELs using heterogeneous GNNs. This method avoids data aggregation, improving predictive performance in terms of MAE and MSE by emphasizing object attributes and interactions. Three datasets are used for experimentation: Business Process Intelligence Challenge (BPIC)17 (loan application process with two object types), OTC about an order management process from a WfMS, and FI (WfMS from a financial institution, using a single case notion).

For further details on the characteristics of the analyzed studies, please refer to the more comprehensive extensive Table A.1 in which all the related characteristics are provided, sited in Appendix A1 A.2.

The results of the above studies generally show that PPM predictions can benefit significantly from the information encapsulated in OCELs. Consequently, these findings on how to exploit the information about the interactions between objects mark a notable milestone in the landscape of academic research in the field of PPM.

4.5.2 Classical Event Logs

In this section, we will show the approaches in the field of PPM that deal with classical event logs. Following the defined search methodology in Section 4.3, we will show the details of the most recent proposals, referring to the period in which there has been an increasing innovation in the considered field of research. The proposals are then classified according to the previously selected variables in Section 4.4, i.e. the prediction task, the method and the contribution area.

Di Francescomarino et al. [31] propose an online classification model for outcome-based prediction to identify normal or deviant process traces. The method extracts prefixes from historical traces, encodes them via frequency or sequence-based approaches, and clusters them using model-based clustering or DBSCAN, and the uses classifiers like Decision Trees (DT) or Random Forests (RF) to predict trace behavior. Tested on BPIC 2011, an hospital event logs about cancer treatment process, the framework shows competitive F1-scores, improvements in earliness, low failure rates, and reduced CT, demonstrating its effectiveness for fast online predictions in healthcare.

Marquez et al. [74] develop a run-time PPM approach using evolutionary algorithms to generate decision rules for predicting process values both of next events and until process completion. The method includes event-window encoding and strategies for imbalanced datasets, with performance evaluated using precision, recall, specificity, F-measure, and AUC. The approach, integrated into ProM and Camunda frameworks for training and real-time prediction, offers interpretable models. For the experimentation datasets from BPIC2013 about Volvo IT incident management and the Andalusian Health Service about IT healthcare are used, covering the IT service management domain.

Tax et al. [118] propose an approach using LSTM networks for next activity and timestamp prediction, without explicit process representation.

Activities are encoded as one-hot vectors, and time-based features are input to three LSTM architectures,i.e. single-task, shared multi-task, and hybrid. The efficacy of the proposed framework is evaluated using MAE for time predictions and accuracy for next-event predictions, with the analysis extending to suffix and remaining cycle time prediction. The efficacy of the proposed framework is demonstrated by its application to the datasets about helpdesk IT service management and BPIC12 on financial services, as evidenced by the utilization of Damerau-Levenshtein Similarity (DLS) and MAE. However, it is important to note the limitations imposed by repetitive sequences, which are identified by the evaluation process.

Polato et al. [97] propose a PPM approach for estimating remaining time in dynamic processes, combining control-flow and contextual data. The authors compare SVR, SVR with Transition Systems (TS), and Data-Aware Transition Systems (DATS), the latter of which employs Naïve Bayes classifiers and SVR regressors. The DATS model demonstrates efficacy in static scenarios, while the SVR model exhibits superior performance in dynamic scenarios. Experiments on helpdesk IT service management and road traffic fine management demonstrate that while remaining time predictions are accurate, suffix prediction results are comparable to random algorithms, based on Damerau-Levenshtein similarity and common prefix measures.

Wang et al. [138] apply Natural Language Processing (NLP) techniques for outcome prediction. Their method involves the transformation of multi-classification tasks into binary ones, by treating events as sentences. The proposed methodology employs an Att-Bi-LSTM network, which integrates attention mechanisms to capture contextual features and assign weights based on their impact on case outcomes. Categorical attributes are one-hot encoded, and numerical ones are normalized. The efficacy of the proposed methodology is evaluated on six diverse datasets, including BPIC 2012 and BPIC 2017 on financial services, sepsis cases, production environment, road traffic fines, and hospital billing. The experimental findings demonstrated that the Att-Bi-LSTM network consistently achieved higher levels of accuracy while exhibiting minimal efficiency trade-offs.

Park et al. [90] combine LSTM-based predictions for remaining time and next activity with optimisation-based resource allocation for BPM.

The architecture includes shared LSTM layers and task-specific layers optimized using the Adam algorithm. The offline training phase utilises one-hot encoded Event Logs, whereas the online phases are dedicated to resource allocation. Evaluations on both artificial and real-life BPIC12 logs, respectively about emergency department healthcare and financial services, demonstrate an improvement in terms of Total Weighted Completion Time (TWCT), though a corresponding rise in CT is observed as a consequence of the inherent complexity of the resource allocation process.

Camargo et al. [19] present a sequence prediction framework for next-event, suffix, and remaining cycle time predictions, with pre-processing, model architecture, and post-processing phases. The pre-processing phase involves converting categorical variables to embeddings and scaling continuous variables. Sequences are created using n-grams to manage temporal dimensions. The model incorporates stacked LSTM layers and dense output layers with specialized and shared variants. Post-processing addresses arg-max issues with random selection. Evaluations across Helpdesk in IT service management, BPIC 2012 on financial services, BPIC 2013 on IT service management, and BPIC 2015 in municipality demonstrate enhanced suffix prediction, substantial performance on extended sequences, and comparable outcomes to baselines for residual cycle time, employing Damerau-Levenshtein (DL) and MAE metrics.

Pasquadibisceglie et al. [92] propose a method for next activity prediction using Convolutional Neural Networks (CNNs). They transform process traces into prefixes, which are converted into 2D images with rows representing event timestamps and columns representing activities. Each pixel contains a vector for activity frequency and elapsed time. The model is evaluated on helpdesk Event Log in IT service management and BPIC 2012 on financial services using accuracy, precision, and recall metrics. Results show that the approach outperforms LSTM-based models in predicting the next activity but struggles with infrequent activities.

Marquez et al. [75] propose CAP3, a methodology for Process Performance Indicator (PPI) prediction that improves accuracy by incorporating contextual information. Using an extended version of the ORGANON methodology, the framework identifies relevant context through domain knowledge and expert feedback, enhancing predictive monitoring for both single-instance and aggregated PPIs. Predictions are made with Random

Forest (RF) and Extreme Gradient Boosting (XGB). The evaluation on IT incident management and IT incident management Healthcare Provider datasets, using MAE and RMSE, show that context inclusion outperforms traditional models.

Mehdiyev et al. [78] propose a novel approach to next event prediction that combines n-gram techniques, feature hashing, and extended feature matrices from control flow, data flow, resources, and organisational perspectives. The utilization of a DL framework with unsupervised pre-training and supervised fine-tuning for multi-class classification improve the efficacy of the proposed method, surpassing LSTM-based networks and traditional algorithms. To demonstrate the efficacy of the proposed method, experiments are performed on BPIC2012 on financial services, BPIC2013 on IT service management, and Helpdesk in IT service management. In particular, hyperparameter optimization and a Radial Basis Function (RBF) NN demonstrated to enhance accuracy, particularly in the context of rare but critical events.

Park et al. [91] propose a framework for predicting business process performance using traffic congestion prediction techniques. The proposed methodology involves the creation of annotated process models, the construction of spatial and temporal matrices, and the training of models such as CNN, LSTM, and LRCN. The efficacy of the proposed framework is validated through a comprehensive set of real-life event logs, including a healthcare dataset from a South Korean hospital emergency department, BPIC12 on financial services, and an IT service management dataset from an Italian help desk. The experimental results demonstrate the superior performance of the proposed framework in both short-term and long-term prediction tasks, particularly in scenarios where data availability is limited. The study emphasizes the significance of spatio-temporal information and network topology.

Pasquadibisceglie et al. [93] propose PREMIERE, a method for next activity prediction that converts event logs into images and uses a 2D CNN with Inception blocks. Trace prefixes are first mapped into feature vectors from multiple process perspectives, and then are converted into RGB images. Finally, the inception modules reduce network parameters and mitigate overfitting. Experiments on BPIC 2012 on financial services, receipt phase in public administration, and BPIC13 on IT service management, demonstrate that PREMIERE outperforms LSTMs and RNNs in accuracy, precision, recall, and F-score.

Taymouri et al. [119] present a GAN-based framework for next event label and timestamp prediction, employing a min-max game where the generator enhances predictions and the discriminator detects fake data. By integrating LSTM-based RNNs, the model effectively processes sequential temporal data. Experiments on Helpdesk in IT service management, and BPIC12 and BPIC17 on loan applications in financial services datasets, demonstrate superior accuracy and reduced MAE compared to baselines. The key innovation lies in adapting GANs for sequential data in process performance monitoring.

Bukhsh et al. [17] propose the ProcessTransformer, leveraging self-attention for next activity, event time, and remaining time prediction. The architecture captures long-range dependencies using a single attention block with trace embedding and positional encoding. Employing Scaled Dot-Product attention, it achieves over 80% accuracy for next activity prediction and lower MAE for time predictions across 9 event logs, including Helpdesk in IT service management, BPIC12 on loan applications, BPIC13 on incident management, BPIC20 on permit applications, hospital Event Log on healthcare workflows, and road traffic fines management.

Kratsch et al. [65] compare ML and DL techniques for outcome prediction, emphasizing the ability of DL methods to automate feature engineering. Using a two-level generalization framework, they evaluate LSTM, Deep Neural Network (DNN), RF, and SVM on five event logs including BPIC 2011 on cancer treatment, BPIC13 on IT incident management, road traffic fines, production log in manufacturing, and academic paper reviews, categorized by data and control-flow perspectives. The results evaluated with accuracy, F-Score, and ROC AUC, indicate that DL outperforms ML in scenarios with high variant-to-instance or activity-to-instance ratios and that is more robust with imbalanced targets.

Pasquadibisceglie et al. [94] propose a multi-view DL approach for predicting subsequent activities, combining multiple process perspectives. The proposed methodology is structured in two phases, starting from Model Extraction about the construction of a predictive model from Event Logs and ending with the Real-time Prediction of future process behaviours. Multi-input representations have been shown to embed categorical variables and pad numerical data. The efficacy of the proposed approach was evaluated through experiments on datasets spanning financial services (BPIC2012, BPIC2017, BPIC2020), BPIC2013 on IT service management, and receipt event log in public administration. The evaluation utilised various multi-class metrics, including overall accuracy, macro-precision, macro-recall, macro-Fscore, macro-AUROC and macro-AUCPR, demonstrating that RNNs with LSTM layers and a softmax output layer enhance accuracy, yet lack interpretability.

Kotsias et al. [64] focus on predicting the next activity by applying three discovery algorithms and using the inductive miner for model selection. The approach combines predictive and prescriptive PM with reinforcement learning (Q-learning), to handle incomplete traces and find optimal policies. The methodology involves five steps, including Event Log extraction, process discovery, and Reinforcement Learning (RL). The evaluation results on BPIC 2017 on financial services show that the model achieves stable results after approximately 500 training episodes.

Fahren et al. [39] introduced a prescriptive process monitoring framework that optimizes alarm generation to prevent undesired outcomes. Utilizing a parameterized cost model, the framework achieves a balance between costs and effectiveness, incorporating features such as delayed alarms and prefix-length-dependent thresholds. The experimentation on the BPIC 2017 on financial services, in road traffic fines, and social services datasets evaluated using metrics such as processing cost, mitigation effectiveness, and F-score, demonstrate the effectiveness of the framework in minimizing false alarms.

Chiorrini et al. [25] utilize Instance Graphs and Deep Graph Convolutional Neural Networks (DGCNN) for the prediction of subsequent activities. The BIG algorithm constructs instance graphs from Event Logs and process models, repairing anomalous traces prior to input to the DGCNN. Experiments on datasets such as Helpdesk on IT service management, BPIC12 on financial services, and BPIC20 on reimbursement processes, and administrative processes in university, demonstrate that the proposed method consistently outperforms LSTM, CNN, MLP, and GCNN in terms of accuracy and F1 score, by leveraging multi-perspective data.

Oberdorf et al. [88] propose an enterprise process network for predicting production disruptions. The framework integrates event logs with contextual data from various departments, using a multi-headed NN (MH-NN) for predictions. The evaluation on a dataset from a German manufacturing company, capturing production and logistics workflows with disruptions, shows that the MH-NN effectively combines data sources, achieving strong performance on multi-class accuracy, precision, recall, and F1-measure, although challenges like concept drift and prediction time persist.

Peeperkorn et al. [96] examined the generalisation of LSTM for next event prediction. Their proposed framework generates the full behaviour of a process model, followed by resampling at the variant level. By utilizing the leave-one-variant-out cross-validation (LOVOCV) approach, they assess the efficacy of LSTMs in generalising control-flow behaviour. The outcomes of this study indicate that LSTMs encounter difficulties in acquiring process structures, yet anti-overfitting measures prove beneficial. However, it is observed that LSTMs may generate variants that are not present in the original log, particularly in cases involving parallel behaviour.

Theis et al. [121] extended Decay Replay Mining (DRM) for next event prediction by combining NN with Petri Nets. The method involves concatenating time-state samples with event sequence vectors to enhance the Petri net, and then using this as input for the NN, to make the prediction. In particular, key extensions include the use of Reachability Graph exploration, Masking Vectors, and NN adaptations. The evaluation of the method on public datasets, including Helpdesk data on IT service management, BPIC 2012 on loan applications, and BPIC 2013 on incident management from Volvo IT, using multi-class mAUROC metrics, demonstrate an enhancement in prediction quality, particularly for events that occur infrequently.

Bousdekis et al. [16] propose using Reinforcement Learning (RL) for next activity prediction. The approach begins by applying a process discovery algorithm on the extracted event logs to select a process model, then computes transition probabilities based on activity and transition frequencies. The RL model is trained using Q-learning, a model-free algorithm, where the agent interacts with the environment and updates a Q-table based on rewards. A new “frozen” state is added to represent unfinished cases. The method is evaluated on BPIC 2017 Event log on loan application process in a financial institute, showing that Q-learning performs better for simpler problems, while Deep Q Networks (DQN), which use a

NN to replace the Q-table, are more suited for complex tasks, though they require more input data. The results indicate that DQN achieves better performance, especially in complex scenarios.

Razo et al. [103] propose the AXDP algorithm, which employs Adjacency Matrices and DL to circumvent overgeneralisation in next event prediction. The method retains the integrity of the event sequence by employing pairs of consecutive events to create adjacency matrices for input to a neural network, thereby outperforming 75% of state-of-the-art models. The efficacy of the algorithm is demonstrated through its evaluation using publicly accessible datasets, namely BPIC12 on loan applications, BPIC13 on IT incident management at Volvo, Receipt on building permit process, BPIC17 on loan applications, and BPIC20 on internal financial processes in education. This evaluation on the AUC, underscores the significance of preserving event sequences. However, challenges persist, including the long CT for large event logs.

Rama et al. [100] propose TACO, an innovative Recurrent Graph Convolutional Process Predictor that aims to facilitate next activity prediction. The integration of process models and event traces is facilitated by a directed Petri net graph, which provides a framework for spatial components, while event sequences are employed for temporal components. The method employs Split Miner to discover a Petri net, converts it into a place graph, and trains a stacked GRNN with feature matrices. An independent LSTM solves the loop overwriting problem. The efficacy of TACO is demonstrated by its application to 10 publicly available datasets, including Helpdesk, BPIC2012 on loan applications, BPIC2013 on IT incident management, building and environmental permits, demonstrates that TACO outperforms state-of-the-art methods. The evaluation on number of loops, edges, transitions, places, and loop lengths, i.e. the process characteristic metrics, highlight the importance of structural process information.

De Smedt et al. [28] propose a methodology for predicting process development by converting Event Logs into time series based on DFGs, utilizing equisize and equitemporal aggregation methods. The method involves the division of logs into intervals with equal numbers of DF relations, while equitemporal employs intervals of equal duration. The approach incorporates well-established forecasting techniques such as ARIMA and GARCH. The efficacy of the proposed method is evaluated through its application to six distinct datasets, namely BPIC12, BPIC17, BPIC18, Sepsis, Italian Help Desk, and road traffic fines. The prediction of process alterations is achieved through the utilization of the aforementioned method, assessing the performance through Mean Absolute Percentage Error (MAPE).

Fani et al [41] propose an Event Log sampling method for PPM with the dual objectives, i.e. enhancing training efficiency and preserving accuracy. The approach utilises LSTM and XGBoost with sampling strategies to achieve a balance between speed and accuracy for tasks such as next activity, remaining time, and outcome prediction. The efficacy of the proposed method was evaluated using real-life datasets, i.e. BPIC 2012 on financial services, road traffic management, and sepsis cases. The results demonstrate that the method can reduce training time while maintaining minimal accuracy loss, particularly for next activity and outcome prediction. However, it should be noted that the prediction of remaining time is more sensitive to the sampling strategy employed. The proposed method is well-suited for dynamic environments characterized by frequent retraining, with the potential for further optimization.

Galanti et al. in [47] introduce an explainable predictive process analytics framework using Shapley values to improve trust in PPM technologies. The framework utilizes CatBoost for accelerated training and superior accuracy in comparison to LSTMs, with the capacity to predict KPIs such as remaining time, activity occurrence, and customer satisfaction. The efficacy of the proposed framework is evaluated using several datasets, including Bank Account Closure, BPIC 2012, BPIC 2013, HelpDesk 2017, and road traffic fine management. The performance of the framework is measured by prediction accuracy, MAE, F1-score, task completion, and system usability. The key innovation lies in the integration of Shapley value explanations into IBM PM, thus making explainable analytics more accessible to business users.

Gunnarson et al. in [53] introduce the Complete Remaining Trace Prediction (CRTP) model, using LSTM for suffix and runtime prediction. In contrast to Single Event Prediction (SEP)-LSTM, CRTP-LSTM makes direct predictions of the entire remaining trace using only observed events. The incorporation of case and event data has been demonstrated to enhance the accuracy of the model. The evaluation of CRTP-LSTM on BPIC20 about university payment requests, BPIC19 on invoice handling, BPIC17 and BPIC12 on financial services, and on building permit applications, demonstrates the model's superiority over competing models in most instances, emphasizing the significance of utilizing all available event data.

Gunnarson et al. [52] introduce the LS-ICE framework, which enhances PPM by incorporating intercase features to represent the overall process state. The framework employs load points and the CRTP-LSTM model for the purpose of making remaining trace and runtime predictions. The efficacy of the LS-ICE framework is supported by empirical evaluations conducted on BPIC20 on payment requests, BPIC17 and BPIC12 on financial services, and on luggage handling in Brussels Airport. These evaluations demonstrate that the LS-ICE framework consistently enhances the prediction accuracy when compared to baseline models that utilize intracase features.

Yarlagadda et al. [141] propose a DL model for PPM, integrating predictive analysis into process models using a multi-layered architecture with fine-tuned hyperparameters. The model employs an LSTM network with input embeddings to predict the next activity in a process sequence. Evaluations on a consumer helpdesk and BPIC datasets demonstrate that the model outperforms existing benchmarks, achieving 76% accuracy on the helpdesk dataset and 78% on the BPIC dataset. This approach establishes a novel performance benchmark for predicting next activities in processes.

Rama et al. [101] focus on activity suffix prediction utilizing a Recurrent Graph Neural Network (RGNN)-based encoder-decoder architecture with an attention mechanism. The model predicts the entire sequence of future activities following a given prefix in an Event Log, employing beam search with length normalization during inference. The efficacy of the proposed model was evaluated on BPIC 2012 on financial services, BPIC 2013 on IT incident management, building permits application, Helpdesk in IT service management, NASA process management, and Sepsis cases, employing Damerau-Levenshtein normalized distance. Results demonstrate enhanced performance compared to prevailing methodologies.

Kaftantzis et al. [59] propose a predictive business process monitoring (PBPM) methodology for next activity prediction using automated ML (AutoML). The method employs the Tree-Based Pipeline Optimization Tool (TPOT) to automate the design and optimisation of ML pipelines, exploring algorithms such as Logistic Regression (LR), DT, RF, Gradient Boosting (GB), Support Vector Machines (SVM), and k-Nearest Neighbors (kNN). Evaluations on the BPIC 12 dataset related to a loan application process, demonstrate enhanced prediction accuracy and facilitate greater accessibility of PBPM to non-ML experts. The automated approach exhibited superior performance in comparison to previous methods, eliminating the need for manual intervention.

For further details on the characteristics of the analyzed studies, please refer to the more comprehensive extensive Table A.2 in which all the related characteristics are provided, sited in Appendix A1 A.2.

4.6 Discussion

The SLR conducted thus far yielded valuable insights into the research domain of PPM, thereby unveiling opportunities and limitations that necessitate further investigation. In the previous Section 4.5 we presented the analyzed proposals, providing an overview of their main characteristics and a detailed explanation of the phases involved in structuring their approach. However, in this section we will take a step further into the discussion, thereby providing the reader with a comprehensive set of observations on the analyzed studies. This will allow for the highlighting of shared aspects of these approaches, as well as their differences, and for the outlining of emerging trends and key lessons learned.

In this section we will identify patterns in the proposals being examined, according to the characteristics chosen for the classification in Section 4.4, i.e. the Event Log type, the Prediction task, the Method, Contribution Area and the Domain.

4.6.1 Event Log

The two main categories of Event Logs analyzed, i.e. OCELs and Classical Event Logs, serve as the basis for the classification of the analyzed approaches. Therefore, proposals are divided into two categories,i.e. PPM approaches that focus on OCELs and those that focus on Classical Event Logs. This differentiation signifies a significant advancement in the field, as the incorporation of OCELs introduces novel methodological challenges and opportunities that were not addressed in earlier reviews. The presence or absence of information on multi-object interactions has a significant impact on recent developments in PPM, particularly with the transition from Traditional Event Logs to graph-based representations, which influenced the methodologies employed across different approaches.

Focusing on PPM proposals dealing with OCELs, three main categories of methodologies emerge, i.e. approaches relying on flattening OCELs, those preserving the graph structure, and others employing alternative methods. These categories reflect how the multi-dimensional nature of OCELs is addressed, particularly in terms of incorporating data on object interactions.

The majority of the approaches fall into the first category, characterized by the flattening of OCELs to resemble Classical Event Logs, with a focus on the encoding phase through the enrichment of supplementary data detailing object interactions. This process simplifies the data structure, which makes it compatible with traditional techniques, while enabling more sophisticated predictive analyses through the integration of interaction data. Examples in this category include [50], which enriches flattened event logs with object-related attributes to capture their interactions. Similarly, [104, 48, 46, 1, 63, 87] rely on flattened OCELs where interactions are encoded to facilitate predictive tasks using well-established techniques. Notably, comparisons between flattened and graph-based approaches [1, 46] show that flattening combined with effective feature engineering can achieve competitive results. It is evident from these studies that while flattening has the effect of simplifying the data structure, it retains sufficient information for accurate predictions when object interactions are encoded appropriately. The methodology employed for the encoding of object interactions in predictive modeling remains a subject that requires further investigation.

The second category is concerned with methods that preserve the graph structure of OCELs, leveraging the relationships between objects and events directly. This approach is grounded in graph-based techniques, such as GNNs and Graph Embedding, which are designed to handle graph-structured data. For instance, [1] applies GNNs directly to OCELs preserving their graph structure and multi-dimensional relationships, while improving predictive accuracy. Similarly, [46] combines GNNs, graph embeddings, and classical PPM techniques, highlighting how graph-based representations can enhance prediction performance. [108, 109] introduce the graph-based encoding HOEG, enabling the direct use of OCEL data, integrating events and objects into a graph structure to preserve a nuanced representation of the data. Also [105] aims to work directly with OCEL graphs. Additionally, [12] employs a graph-based approach for feature extraction rather than flattening, further reinforcing the trend towards preserving OCELs structure.

A third category comprises studies that explore alternative methodologies for PPM involving OCELs. For instance, [23] employs Colored Petri Nets combined with image-based techniques. Meanwhile, [68] outlines a conceptual framework for predictions but does not specify whether it utilizes OCELs directly.

While graph-based approaches offer considerable potential for harnessing the full complexity of OCELs, flattening methods continue to prevail due to their simplicity and efficiency, a factor that is further reinforced by effective feature engineering.

In terms of Classical Event Logs, differences also arise based on the selected process perspective. As previously discussed in Chapter 2 and proposed by [32], the input data can vary depending on the chosen process perspective, which significantly influences the manner in which data are used. This selection is also driven by the more favorable results obtained by incorporating additional process information. Some approaches focus exclusively on a specific perspective, such as control-flow, while others enhance the log by integrating structured and unstructured data from new perspectives, such as contextual data. For instance, also approaches focused on predicting control-flow behavior can enjoy from other perspectives, such as contextual data, with improved accuracy of control-flow predictions, thereby influencing subsequent data treatment. Studies have demonstrated that the inclusion of additional information typically leads to improved prediction outcomes [97, 92, 93, 65, 94, 25]. Therefore, integrating diverse perspectives is crucial for improving predictive accuracy.

In conclusion, while flattening OCEL data enables the application of established PM techniques, encoding methods designed for multi-perspective classical event logs offer valuable insights for OCEL encoding. This suggests promising avenues for developing novel strategies that directly integrate object interactions into OCEL models, ultimately yielding more comprehensive and accurate predictive approaches.

4.6.2 Prediction Task

The Prediction Task concerns the target variable within a process that is the object of prediction. As outlined in Chapter 2, this variable may be the outcome of the process, a numeric value, or the next event. Moreover, a distinction can be made between predictions made for the completion of the entire case or for specific points in the future. This distinction impacts both the selection of the prediction algorithm and the metrics employed. In the context of prediction, OCELs and Classical Event Log approaches address distinct challenges, reflecting their different approaches to input data and datasets.

In the context of applying PPM to OCELs, a range of categories of prediction tasks becomes evident, reflecting the multiple approaches adopted by researchers in addressing diverse aspects of process forecasting.

The predominant prediction task in OCEL-based PPM is the estimation of remaining time and throughput time, with numerous studies concentrating on the prediction of the time until process completion [50, 1, 108, 12, 23, 109]. Methods such as LSTM, Flattened, GNN, GNNML, and HOEG/GNN are used, and this task is crucial for process optimization and resource management. Another significant area of investigation is next activity/event prediction, a field that has seen significant contributions from the research community, as evidenced by [50, 104, 1, 63, 68], employing a range of techniques, including LSTM, GANs, GNN, Flattened, Petri Nets, Statistical Learning, CatBoost, and counterfactual approaches. The primary objective of these techniques is to forecast the sequence of future events along with their timestamps. Other emerging tasks include the prediction of KPIs and process outcomes [48, 46, 105, 68], using methods such as CatBoost, GCN, GGNN, LSTM, and counterfactual techniques. Anomaly detection [87, 12] is also explored, though it is not a traditional prediction task; rather, it represents a key objective in the utilization of PPM with OCELs in these studies. Notably, authors such as [50, 1, 12, 68] address multiple prediction tasks, demonstrating the broad scope of their contributions to the field by employing techniques such as GNN, GCNAE, Petri Nets, and LSTM.

The complexity of the information related to the graph in OCELs indicates that the choice of prediction task is not directly influenced by the structure of the event log. Instead, the methods are more focused on managing the structure of OCELs than on the specific prediction task itself. Furthermore, the diverse range of combined prediction tasks highlights the potential for advanced PPM, with ongoing research dedicated to identifying the most effective approaches for leveraging these capabilities.

In Classical Event Log-based approaches, research is focused on two key areas. Firstly, the prediction of the next event is typically investigated, involving activity labels and timestamps. Secondly, the forecasting of the entire case progression until its completion is studied, which includes suffix prediction and remaining time estimation.

The next event prediction task, the most extensively studied of the tasks under consideration, aims to forecast the next activity and its timestamp [118, 19, 17, 92, 119, 25]. Some works extend this to include next event time or KPIs [74, 103, 141]. However, the majority of these works rely on a sequential prediction approach rather than directly modeling the entire case trajectory.

A more comprehensive perspective is provided by remaining time prediction, which estimates the time until process completion [97, 90, 17]. Some studies integrate this with next activity or suffix prediction to enhance process forecasting [19, 53, 52]. In distinction to next event prediction, suffix prediction aims to forecast the entirety of process traces without the requirement for iterative next-step predictions [28, 53, 101]. This alternative approach serves to mitigate the cumulative errors of step-by-step forecasting.

Outcome prediction, another key area, typically follows a classification approach, focusing on binary outcomes [31, 65] or decomposing multi-class problems into multiple binary tasks [138]. Some studies target KPI prediction [75, 91], while others extend outcome forecasting to broader predictive tasks [39, 88].

The techniques employed in these tasks are not inherently constrained to a specific prediction category. Conversely, the adaptation process occurs in response to the problem formulation, encompassing classification for prediction of outcomes, regression for the estimation of residual duration, and sequence modeling for the prediction of subsequent events and suffixes. The incorporation of these varied prediction types, along with the leveraging of their interdependencies, has the potential to further enhance PM insights.

When considering outcome prediction, several studies focus directly on binary outcome prediction, such as [31]. Alternatively, other approaches address the issue of multiple outcome prediction by transforming it into multiple binary classification tasks, as outlined in [138].

While the study of such tasks traditionally occurred in isolation, recent research exhibited an increasing tendency to adopt a more holistic approach, integrating multiple predictions and leveraging the interdependency between different aspects of process execution. Future research could benefit from more integrated approaches, bridging next activity, suffix, remaining time, and outcome predictions to enhance decision-making in PM.

4.6.3 Method

The Method feature addresses the selection of the prediction model. As outlined in Chapter 2, the approaches employed within PPM can be categorized into two distinct groups: model-based or process-aware techniques, and non-process-aware approaches. The information contained in the event log is encoded by both categories and subsequently fed to a prediction model. The objective of this section is to analyze these prediction models, with a focus on identifying which methods are most commonly used, the benefits they offer in terms of performance, explainability, interpretability, and compilation time, regardless of whether they are process-aware or not. Firstly, a distinction is made between OCELs and classical event logs, which are then further classified into solutions based on ML and DL.

A review of the methods employed in PPM for OCELs reveals several distinct methodologies, thereby emphasizing the varying approaches researchers have adopted in order to leverage the unique features of OCELs. A categorization of these methods is possible based on their underlying models, which include GNNs, LSTMs, gradient-boosted decision trees, and other specialized techniques.

GNNs are the most commonly used method for OCELs, as they leverage the graph-based structure of OCELs to capture relationships between objects and events. Studies such as [1, 46] demonstrated the potential of

GNNs for next event and outcome prediction. Similarly, [108, 109] explore GNN-based methods, introducing HOEG encoding to integrate events and objects into graph structures. Also [87] employs GCNAE for anomaly detection, and [105] aims to utilise GNNs for KPI and outcome predictions. [12] employs GNNs indirectly for feature extraction in isolation forest-based anomaly detection. GNNs effectively handle OCEL complexity, but their long training times [46] motivate the exploration of alternatives like CatBoost and LSTMs. Indeed, with effective feature engineering, these methods can achieve competitive performance at a lower computational cost.

LSTM models represent the second most prevalent method within OCELs, employed on flattened event logs. [50] employs LSTMs using enriched and filtered event logs. Similarly, [46] uses LSTMs in a comparative analysis with other techniques, showing their effectiveness. [23] also incorporates LSTMs alongside Colored Petri Nets for specific time-related predictions, particularly for time-series and sequence prediction tasks highlighting the potential for hybrid approaches.

The study in [63] applies Petri nets and SL methods in order to predict subsequent activities. [104] represents a distinctive contribution in the application of GANs for PPM with the aim of generating realistic sequences in order to enhance the performance of models.

[12] uses ML techniques like Isolation Forests for anomaly detection and Regression Models for prediction tasks. Meanwhile, Gradient-boosted decision trees, particularly CatBoost, are another popular approach for OCELs, [68] introduces Catboost along with counterfactual reasoning to provide insights into the effects of changes in the process.

PPM methodologies for OCELs range from GNNs, due to their efficacy in handling graph structures, to simpler, faster alternatives such as LSTMs and gradient-boosted decision trees, the adoption of which is driven by the high computational cost of GNNs.

For Conventional Event Logs, LSTMs are recognized as the best performing DL technique due to their ability to handle long dependencies. LSTMs are applied across a variety of prediction tasks, from next activity prediction [118, 19, 119, 94, 96] to outcome prediction [138, 90, 65] and remaining time prediction [53]. In the domain of classical event logs, LSTM networks are employed in several different ways. Firstly, the standard architecture, which is employed in numerous studies including [118, 90, 19, 93, 96]. Secondly, with significant architectural modifications or in combination with other techniques [138, 119, 100, 53], or in comparison with other methods [91, 65]. Other approaches focus on the use of widely known methods such as Convolutional Neural Networks (CNN) [92, 93], and GNN [25], which apply these architectures to consider multi-perspective enriched instance graphs.

[17] employ transformers, which demonstrated to yield superior prediction results; however, they necessitate longer training times and yield less explainable results.

ML approaches involve using DT, RF, SVM for both regression and classification, and Q-Learning [64, 16]. These models provide good performance and explainable results.

The analysis of these proposals reveals that certain methods, including Transformers and DL-based approaches, demonstrate superior performance in comparison to traditional ML methods. However, experimental findings suggest that these methods also present certain challenges. The primary distinction among the proposals pertains to the requisite training time and the capacity of the selected method to provide explainable predictions. Specifically, the initial concern pertains to the training time, which is predominantly associated with networks such as GNN or Multi-Headed DNN, which necessitate a greater investment in training time compared to other DL and ML methods. Secondly, there is the over-generalization of LSTMs, as highlighted in [96, 103] where LSTMs have been observed to generate process behaviours not observed in the log. A further issue is the lack of explainability in the results, as DL-based models and Transformers are unable to provide clear explanations for their predictions.

4.6.4 Contribution Area

The analysis and identification of the scientific contribution within the same research field proves to be more intricate than the previous discussions carried out so far. The objective of this section is to emphasise the distinctions in the contributions of each approach by mapping them to the phases of the PPM framework. This task is particularly challenging, as it cannot be directly linked to either the type of prediction task or the methodology employed. While this is anticipated for newer approaches, such as those utilizing OCELs, which introduce a variety of contributions to address the encountered challenges, Classical Event Log approaches necessitate a deeper investigation to better understand these distinctions.

Within the diverse approaches, different types of contributions can be identified, which can be categorized according to the phases of the PPM framework.

PPM approaches that focus on OCELs primarily target interventions in Event logs. The primary objective of these proposals is to leverage the information derived from object interactions, either through modifications in data encoding or by utilizing the graph structure of the logs [50, 48, 1, 46].

PPM approaches that focus on Classical Event Logs can be further subdivided into various categories. A considerable number of studies have focused on the steps preceding model training, thereby acting on input data and making necessary adjustments to align with already established methods, as seen in [138, 92, 91, 93, 28]. Another set of approaches addresses the representation of input data, as exemplified in [118, 19, 78, 94, 25, 88, 121, 103, 100, 53]. These two categories are interconnected, as transforming the input data appropriately is essential for applying known algorithms.

Additionally, the previously discussed dimensions focusing on input data play a role in the selection of the prediction model. Some studies emphasize the choice or adaptation of the prediction model based on the assumptions made about the input data. For instance, [31, 118, 97, 93, 119, 17, 64, 16]. Some of the proposed modifications are improvements to existing models that are widely recognized in the research community for their effectiveness, such as modifications to LSTM networks. In contrast, some studies propose the use of specific networks that have not been previously applied in this domain.

A further category of studies focuses on comparing the most effective solutions proposed in the literature, with a view to identifying the factors contributing to observed differences in metric values. The present studies do not introduce new innovations; rather, they conduct a comparative analysis of existing methods to determine which performs best[65, 96].

In conclusion, it is possible to draw some general conclusions from the analyzed approaches. While categories have been identified, it is evident that each approach addresses a specific sub-problem. For instance, in the context of Event Log treatment, some studies focus on process dynamics, others prioritize the representation of particular constructs (e.g., loops), and some concentrate on the characteristics of the log (i.e., completeness or volume of available data). In the context of prediction models, the proposals are varied and include studies that emphasize offline construction of models, those that focus on the interpretability, accuracy, or timeliness of predictions, and finally those that address the problem of overgeneralization. This diversity of focus areas affects benchmarking, which is challenging to establish due to the varying assumptions and contributions of different approaches. This study reveals that the current direction of research is oriented toward different aspects, making the comparison challenging.

4.6.5 Domain

The connection between domain characteristics and PPM approaches is a key feature that needs to be explored. From the analysis above in Section 4.5, it is clear that OCEL-based approaches excel in domains involving intricate processes with multiple interacting objects, while Classical Event Log are employed in simpler process structures having single case IDs.

First of all, the domains explored in the context of OCELs are predominantly centered on order management processes. These processes are particularly well-suited to the OCEL format, as they naturally involve multi-object interactions, such as those between orders, items, and customers. Spanning various scenarios—from customer order management to order processing, order-to-cash, and procure-to-pay workflows—Order Management Event Logs have emerged as a primary use case for OCELs, as demonstrated by studies such as [50, 104, 63, 87, 108, 12, 109].

Beyond order management, OCELs have also been applied to other domains, including utility providers [48, 46] and financial services [1, 87, 108, 109]. In healthcare, OCELs have been used to integrate images and time series data for process analysis [23], while in the public sector, they have been explored for modeling public tenders [105].

The datasets analyzed exhibit significant variability in the number of object types involved, ranging from two to six. This variability raises important considerations regarding how the number of object types influences both methodological choices and model performance, highlighting a crucial factor in the effectiveness of OCEL-based predictive models. While order management processes remain the most prominent use case, production and logistics processes also benefit from OCELs due to their inherently multi-object nature and the interconnected components within these domains.

On the other hand, Classical Event Logs rely heavily on widely available public datasets, promoting comparability and reproducibility across studies. These approaches span a broader range of methodologies and prediction tasks, reflecting the maturity and versatility of research in this area.

The only common experimented domain between OCELs and Classical Event Logs is financial services, making it an interesting area for evaluating methodologies across different data structures. Comparing these approaches within the same domain could provide valuable insights into managing OCELs. On one hand, OCELs could benefit from the extensive experimentation already conducted in financial services using Classical Event Logs. On the other hand, this joint exploration could provide lessons for applying PPM to OCELs, leveraging insights from traditional event log analyses and adapting them to the OCEL format.

By comparing methodologies applied within the same domain but using different Event Log types, it becomes evident that OCEL-based approaches are more specialized and constrained by the complexity of the dataset complexity, whereas Classical Event Logs allow for a wider range of tasks and techniques, fostering broader applicability.

This analysis highlights the need to tailor PPM frameworks to the specific demands of the domain while acknowledging the potential of OCELs to expand the scope of PM into more complex and dynamic contexts.

4.7 Conclusion and future work

In this chapter, we proposed a SLR with the aim of analyzing and synthesizing the techniques employed in the PPM domain. Indeed, since the adoption of advanced Enterprise Resource Planning (ERP), CRM systems and big data analytics has driven DT in organizations, process-oriented PPM, focused on the prediction of future process behavior, received increasing attention.

In this SLR, the differences between OCELs and Classical Event Logs is discussed w.r.t. the Prediction Task, the Method, the Contribution Area and the Application Domain. The findings of this SLR can be summarized as follows.

Firstly, regarding the *Event Log type*, OCELs are distinguished by an inherent complexity resulting from the existence of multiple object types and the interactions between these object types. In contrast, Classical Event Logs are associated to a single case process. The inherent structural complexity of OCELs provides more comprehensive process information, but also increases the difficulty of effectively modeling data by including information about object interactions and predicting process behavior. An analysis of existing approaches revealed two main distinct research areas, i.e., studies focusing on OCELs flattening and enrichment and studies that focus on maintaining the graph structure of OCELs.

The distinction between OCELs and Classical Event Logs influences also the *Prediction Task*. Approaches based on OCELs address a wide range of predictive tasks, leveraging their richer information to integrate multiple predictions and explore broader PM applications. Their recent adoption has prompted experimentation that goes beyond classical prediction task, as researchers investigate the effective integration of OCELs into several PM techniques. In contrast, Classical Event Log-based approaches primarily focus on next event prediction, frequently leveraging iterative methods that are subject to cumulative errors.

As a consequence, the OCELs data structure influences also the utilized *Method*, which has been identified as being subject to challenges in achieving a trade-off between prediction accuracy and computational efficiency. In fact, working on the complex data structures of OCELs, such as graphs, influences the utilized method, which is often identified in GNN. However, this trade-off between prediction accuracy and computational efficiency is a common challenge in GNNs. On the other hand, methods based on Classical Event Logs primarily focus on improving prediction accuracy through ML and DL models such as LSTM, transformers, and CNNs. While these models enhance predictive performance, they also face limitations such as over-fitting and over-generalization, particularly in LSTMs, which may generate unrealistic process behaviors. Hybrid approaches, integrating both ML and DL techniques, such as LSTM-CNN or transformer-based models, are increasingly prevalent, offering enhanced accuracy, but giving rise to concerns regarding interpret ability.

As regards the contribution of the analyzed methods in a specific *Contribution Area* of the implementation framework of PPM, OCEL-based methods focus on enhancing event log representations through object interactions, while Classical Event Log approaches address data preprocessing, encoding techniques, and model selection. These methodologies exhibit several assumptions and problem scopes making a direct comparisons quite difficult. While some studies introduce novel predictive models, others refine existing techniques or benchmark competing approaches. The broad spectrum of research directions underscores the challenge of establishing standardized evaluation criteria, emphasizing the need for further investigations to enhance comparability and methodological consistency.

Moreover, several studies have focused on specific *Domain* revealing that the choice of a method often depends on data characteristics. OCEL-based approaches are particularly suited for domains with multiple interacting objects, such as order management, production, and logistics, where their capacity to capture relational structures enhances predictive performance. In addition, domain chosen for OCELs differ from each other also for the number of object types being included in data structure, feature that could be further investigated in order to understand if it influences the performances of the methodology. In contrast, Classical Event Logs remain prevalent in simpler process structures, benefiting from standardization and methodological diversity. The growing adoption of OCELs highlights their potential for modeling complex event interactions, yet their applicability is constrained by dataset intricacies. Future research should further explore domain-specific adaptations to refine predictive methodologies and enhance comparability across different log types.

Whilst OCEL-based methodologies show great potential in the capture of object interactions, their incorporation into broader PPM techniques continues to be in an emergent phase. This integration remains underdeveloped in comparison to that of Classical Event Logs and there is an acknowledged necessity to overcome the challenges presented by computational complexity and feature engineering. The challenges associated with OCELs, including their complexity and extended processing times, led to their limited adoption. Consequently, there is a necessity to concentrate on the enhancement of PPM methods by means of the optimization of encoding, the development of enhanced feature engineering techniques, and the reduction of complexity. The creation of standardized datasets and the development of benchmarking tools are imperative for the comparison of diverse methods and the enhancement of overall performance. Further research is necessary to refine prediction models and facilitate their application in real-world scenarios.

In summary, this SLR provide a support for improving the structure and methodology of the research field, creating a valuable resource for new researchers. The previous discussion can provide researchers with a starting point for a more focused research structuring.

Chapter 5

Process Mining and Predictive Process Monitoring in a new context

Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth.

Jules Verne

5.1 Motivation

The phenomenon of climate change represents a significant challenge for society, the urgency of which is intensifying with the rapid pace of global environmental deterioration. The imperative to address climate change has prompted a global search for alternative, low-carbon sources of energy and fuel. This has led to a notable increase in the demand for bio-fuel, which is projected to grow by 38 billion litres between 2023 and 2028, representing a near 30% expansion from the previous five-year period [3]. Furthermore, the rising population is exacerbating the problem of increased demand for food contributing to climate change. This is leading to an anticipated overall global food increase of between 35 and 56 percent between 2010 and 2050 [132]. As a consequence of the increase in demand for food and biofuel, land cultivation is being pushed towards larger areas, thereby driving the phenomenon of Large Scale Land Acquisitions (LSLAs) [24]. This phenomenon raised a number of questions pertaining to production models, the rights of individuals, and how resources are governed, particularly in relation to local populations. In response, in 2009 the Land Matrix Initiative was launched [1], with the aim to provide transparent information about the process of LSLAs. To this aim, the Land Matrix Initiative collects data from a range of heterogeneous information sources pertaining to LSLAs, encompassing several information on land acquisitions, such as the negotiation and implementation status of the acquisition, the countries involved, the intention of the investment (e.g. agriculture, forestry, mining), the size of the land and other details [44]. However, the status of land deals is subject to frequent alteration demonstrating high variability, with negotiations and implementations progressing from one status to another [79], often resulting in outcomes hard to predict.

The dynamic nature of the land acquisition process raises questions about whether the visited statuses during the land acquisition process influence the final outcome of the deal. By leveraging PM techniques, which use event logs (i.e. records of activities captured within IS) to reconstruct business processes [124], it is possible to model the entire land acquisition process, from negotiation to production, and to identify process behaviors. We believe that this approach can effectively model land acquisition and production processes, using event log derived through data processing from publicly available sources such as the Land Matrix Initiative.

In this work, we analyze LSLAs using PM, following a structured pipeline [133] involving planning, data extraction, preprocessing, and exploratory analysis. We aim to assess the applicability of PM to LSLAs and investigate how ML can support outcome prediction and process analysis, with the ultimate goal of understanding LSLAs behavior and dynamics.

The contribution of this chapter is threefold. First, we investigate how PM techniques can be applied to model the deal acquisition process of LSLAs. For this purpose, we use both process discovery [126] and trace clustering techniques [110]. The process extracted via PM presents complexity and high variability, resulting in a spaghetti process i.e., an highly

[1] https://landmatrix.org/

unstructured graph. To improve the interpretability of process behavior, we apply log decomposition and clustering to partition the event log into coherent subsets, each allowing for simpler and more readable process models [15] aligned with outcome classes.

Second, given the high process variability and class imbalance in the event log, we evaluated the impact of incorporating process-related information into outcome prediction using a Random Forest (RF) classifier. The results indicate that adding process control-flow data yields only marginal improvements—or, in some cases, slight decreases—in predictive accuracy, largely due to the complexity of the underlying processes. Nonetheless, the analysis shows that cases with richer historical traces tend to produce more accurate predictions, highlighting the potential value of process-aware features in supporting outcome classification.

Finally, to better capture this value, we evaluate the impact of different encoding strategies on predictive performance. We compare various encoding techniques, including advanced methods such as Bidirectional encoder representations from transformers (BERT) and Multilayer Perceptron (MLP), to generate enriched feature representations and assess their influence on classification results. The obtained results demonstrate that combining PM and ML approaches can uncover process-aware patterns and improve understanding of LSLAs deal trajectories, with implications for more informed policy analysis on land governance.

5.2 Background and Related Work

5.2.1 Background

In this section, we provide an overview of fundamental concepts related to Trace Clustering techniques, and the selected Trace Encoding method. For PM and PPM we refer to the definitions provided in Chapter 2.

Complexity and Trace Clustering The effectiveness of PM techniques is highly dependent on the quality and structure of the available event log data. In real-world domains (i.e., PA or Healthcare) [84], event logs often exhibit low structure and high variability, characterized by numerous process variants that complicate process discovery [110]. The discovered process models frequently contain a large number of activities and complex interconnections, bringing to *spaghetti models*, i.e. process models with tangled and difficult to interpret structures [40, 4]. To address this issue, the *trace clustering* technique has been widely adopted as a pre-processing step, in order to partition the event log into subsets of similar traces, enabling the discovery of more comprehensible process models within each group [111, 56]. A typical pipeline involves two phases: (i) **profiling**, where traces are encoded as feature vectors (e.g., control-flow or performance-based), and (ii) **clustering**, where these representations are grouped using distance-based ML techniques [40, 31]. Over the years, various clustering strategies have been proposed. A recent systematic review [142] classifies these methods according to the chosen profiling strategy, similarity metric, and clustering algorithm. Broadly, trace clustering methods can be categorized into three families [71]: (i) Feature-vector-based, where traces are encoded as vectors using activity frequency or transition patterns, and clustered using algorithms such as K-means or Agglomerative Hierarchical Clustering (AHC) [111, 85], (ii) Sequence-based, where are compared based on sequence similarity using edit distances or alignment techniques [15], which can struggle with structural variability, (iii) Model-based Clustering, driven by the quality of the models induced from trace groups (e.g., Petri nets, Markov chains) [29], which can be resource-intensive and difficult to scale. Early methods include Self-Organizing Maps (SOM) [73] and frequent sequence pattern mining [71], while more advanced approaches such as *Active Trace Clustering* (ActiTraC) use active learning to iteratively improve cluster quality based on model fitness [29].

Despite recent advancements, achieving a balance between scalability, interpretability, and accuracy remains challenging—particularly in complex domains. For these reasons, feature-vector-based clustering methods, such as K-means and AHC, continue to be widely applied. While their limitations are recognized, these algorithms provide a practical framework for exploratory process analysis in data-intensive settings.

Trace Encoding In order to apply clustering techniques as outlined in Paragraph 5.2.1, the event log need to be transformed into a suitable format, making the choice of encoding a fundamental step in complexity reduction. Many approaches convert event logs into vector-space models, followed by distance-based clustering algorithms, where obviously the trace encoding method plays a key role.

Among various encoding methods, we selected index-based encoding due to its proven predictive performance [70] and its lossless nature, allowing complete reconstruction of the original trace [112]. This encoding preserves the sequence order of activities, which is crucial to maintain the behavioral and temporal aspects of the process. Furthermore, index-based encoding is compatible with both clustering and classification techniques, supporting our dual-phase approach.

Also previous studies in PPM confirm the effectiveness of this encoding method [117, 61], and its use alongside clustering algorithms like DBSCAN and K-means [45]. In fact, this choice guarantees a trade-off between information fidelity and practical applicability in analyzing LSLAs processes.

In addition, considering the inherently sequential nature of event logs, also more sophisticated approaches, such as those DL-based (i.e., LSTM, Transformer-based embeddings), can provide valuable representations. In fact, for their nature, these models can capture complex temporal dependencies and richer behavioral patterns.

Therefore, their impact, particularly in PPM settings, will be explored more in details in the experimental section.

5.2.2 Related Work

In this section, we discuss previous literature related to the phenomenon of LSLAs.

Overview on LSLAs analysis

While LSLAs are a major driver of global economy since many decades, few works can be found in literature that addressed this complex phenomenon through computational approaches. A first approach is the one proposed by Seaquist et al. [107], consisting of a structural analysis of a network extracted from a dataset built combining information from Land Matrix and GRAIN (`www.grain.org`), another LSLAs database compiling information on deals from press articles and scientific publications. However, the data used in this work dates back to 2012, making most analyses obsolete. Later, with the aim to identify general trends characterizing the process of land acquisition, Mechiche-Alami et al. [77] proposed a study that focuses on the identification of different phases of activity in the land trade market. An advanced approach exploiting complex network analysis techniques to analyze and characterize LSLAs phenomena at global scale is presented in [58]. The authors perform an extensive quantitative and qualitative analysis about the topic, also defining a score (namely *LSLAs-score*) that allows to rank the countries based on their investing/target role in the land trade network. The paper also includes a correlation analysis between network features and several country development indicators and an analysis centered on the discovering and analysis of network motifs that provides an insight into statistically significant land trade schemes. A major finding of this work was that, other than the well known Global North-Global South one, other important dynamics exist in the global land trade market, such as a South-South one (e.g., Latin American countries investing in Africa), and the one involving emerging economies, such as the so-called BRICS countries (Brazil, Russia, India, China and South Africa). Interestingly, while several of these countries tend to have an *investor* profile, they are still often involved as target countries in several deals. With the aim to provide a deeper investigation on this phenomenon, that can be considered an anomaly in the land trade market, the same authors proposed a dedicated analysis about the countries showing a double investor/target profile [57], based on eigenvector centrality methods.

It can be noted that while previous works mainly focused on global-scale analyses and structural patterns of land deals, none of them have explored the internal dynamics of each individual LSLAs case. To the best of our knowledge, our work is the first who addresses this gap by leveraging PM techniques to analyze and predict the outcomes of individual deals, enabling a finer-grained understanding of the LSLAs lifecycle and offering potential explanation tools for the reason of reaching the final outcome of each deal and to predict such outcome.

5.3 Dataset Description

5.3.1 Raw Dataset Description

The data used in this study was sourced from the Land Matrix Initiative open access database [98]. The data were extracted on 01/07/2024 and include information about lands leased or sold prior to that date provided, shown in Appendix A.2, Table A.3. We focused on the deals, involvements and investors tables included in the database. For the purposes of this analysis, according to previous approaches [58, 57], only the transnational deals have been extracted (i.e., deals where at least one of the investors is located in a country different from the target one). The dataset includes 4,058 transnational deals, encompassing both acquisitions and sales across international borders (in this work, when we refer to *sold* land we may also imply contracts of different nature such as leasing or rent). We selected the following features for each deal:

- Deal ID: represents the unique identifier of the deal;
- Negotiation status: reports on the different statuses of the negotiation related to the deal contract;
- Implementation status: reports on the different statuses of implementation of the activity planned on the acquired land;
- Target country: represents the country in which the land is acquired;
- Top parent companies: lists the company (or companies) involved in the acquisition as investors, along with a Company ID;
- Intention of investment: represents the main activity for which the acquired land should be used (e.g., Agriculture, Forestry, Mining, and so on);
- Deal size: reports the size (in hectares) of the acquired land, as stated on the deal contract;
- Nature of the deal: reflects the nature of the deal contract, i.e., if it is a lease, a concession etc.;

Additional features about the companies involved in the deals were extracted from the Investors table, more precisely:

- Investor ID: the ID of the company;
- Name: the name of the company;
- Country of registration/origin: the country of the company;
- Classification: contains the information about the type of the operating company, if it is governmental, public, private etc.;

5.3.2 Data Preprocessing and Event Log Preparation

In the following, we will describe the preprocessing steps performed to obtain a proper event log describing the land acquisition process, starting from the data about land trade deals extracted from the Land Matrix database.

Event Log In order to organize data in the format of an *event log* (i.e., where each deal is associated to a sequence of activities and their timestamps) we exploited information from the *Negotiation status* and *Implementation status* attributes of the Deals table. The selected columns contain sequential information about the deal history, alternating statuses and associated timestamps (more specifically, a single cell can contain sequences of information separated by a given delimiting character). The final result of this process is a sequence reporting all the activities associated to each deal (i.e., both negotiation and implementation ones), from the oldest to the most recent one, with associated timestamps. In our event log dataset, we associated these sequences to their corresponding Deal ID through a couple of columns called *Activity* and *time* (where the order of the sequence gives the association between the two). Then, all the Deal IDs having at least one missing value in the *Activity* column or in the *time* column were removed. Finally, the *time* column was rearranged into a timestamp in order to be suitable to PM techniques. In the end, the event log presented in Appendix A.2, Table A.4, contains the following attributes: Deal ID, Activity and Timestamp, more in detail :

- **Deal ID**: the unique identifier of the deal, corresponding to the Case ID of the event log;
- **Activity**: representing the alternating status of negotiation and implementation. These statuses can be sequential or concurrent, as multiple statuses can alternate for each deal;
- **Timestamp**: represents the year, month and day in which the negotiation status or implementation status was recorded. The negotiation status and the implementation status may have the same timestamp, indicating that they follow a partial sequence. This is not a problem as these events could occur simultaneously. In such cases, the sequential ordering of activities ensures that the process can still be properly represented. Therefore, if both negotiation and implementation statuses have the same timestamp, the partial order can be maintained by alternating negotiation statuses with implementation statuses, ensuring a logical sequence in the process.

Attributes The attributes table contains additional information about each deal, extracted from the *Deals* and *Investors* tables of the original Land Matrix database. It contains the following columns:

- Deal ID;
- Target Country: it represents the country in which the land is acquired;
- Parent Countries: represents the countries of the companies involved in the acquisition (one acquisition can be done at the same time from companies located in one or more countries);
- Category: represents the goal of the investment. It was obtained by grouping the 'Intention of investment' attribute into four categories (according to the hierarchy reported on the Land Matrix website): 'Agriculture', 'Forestry', 'Renewable energy power plants', 'Other';
- Deal size: reports the size in hectares of the deal;
- Nature of the deal: reports on the contractual nature of the deal (e.g., leasing, rental, etc);

- Classification: represents the classification status of the companies involved in the acquisition (e.g., private company, stock-exchange listed company, etc.).

5.4 Event log Analysis

In this section we will explore the ground truth of the Land Acquisition Process, and then we will go through the event log inspection and the modeling of DFG [128].

The Land Acquisition Process is conceptually modeled as a structured, imperative process, meaning it follows a clearly defined sequence of phases. This structure enables its representation using BPMN, which serves as the theoretical ground truth of the process. However, the actual event log is far from imperative. In fact, it exhibits a high degree of flexibility, including partial traces, skipped steps, and non-linear transitions. As a result, for the purpose of discovering a process model from the event log, we opted for a DFG. DFGs offer a simpler and more robust abstraction, emphasizing the most frequent and direct activity relations rather than attempting to reconstruct a fully structured process.

Although more sophisticated discovery techniques—such as Inductive Miner and Split Miner—were also considered, the use of a DFG was deemed sufficient for this initial exploratory phase. In such contexts, simpler models like DFGs often prove more effective, as they highlight dominant behavioral patterns without being overwhelmed by noise or infrequent behaviors [124].

Moreover, this approach leaves space for future refinements through trace clustering and log filtering, which can support the development of more structured and accurate process models in subsequent stages.

5.4.1 Land Acquisition Process

In this section we describe the ideal land acquisition process, which serves as the ground truth against which the actual behavior extracted from the event log. The Land Acquisition process is theoretically represented by a simplified BPMN diagram (see Appendix A.3, Figure A.1), illustrating a structured sequence of macro-phases like Negotiation and Implementation.

This process represents how the process is expected to be structured and executed, based on official documentation.

The land acquisition process consists of various stages, each reflected by a specific deal status. As the project starts and progresses, different negotiation and implementation statuses may apply. Typically, the process starts with an *'Expression of Interest'* from the company, after which the deal enters in an *'Under Negotiation'* phase, followed by a *'Memorandum of Understanding'* which represents an initial agreement on the terms of the deal. The negotiation can lead to one of two outcomes: either the deal is concluded or it fails. If concluded, the deal can move to an *'Oral Agreement'* or *'Contract Signed'* status, with the latter marking formal closure. In the event of failure, the deal assumes a *'Negotiations Failed'* status. However, even concluded deals may revert to a *'Contract Cancelled'* status due to unforeseen issues. Moreover, external factors may trigger status transitions. For example, a contract may become obsolete, leading to a *'Contract Expired'* status, or the ownership of the deal may change, resulting in a *'Change of Ownership'* status.

Once the negotiation is settled, the implementation process begins. Depending on the stage of implementation, different statuses are recorded, which reflect the progress of the project. Initially, a *'Project Not Started'* status indicates that the project associated with the land acquisition has not actively started yet. If the project is in the early setup phase but not yet producing output, it is categorized as *'Start-up Phase (No Production)'*. As soon as the project is fully operational, it moves into the *'In Operation (Production)'* status, meaning that production has started and the project is in active execution. There are also cases where projects, although classified as 'Concluded', are not progressing as expected. For instance, a project might be categorized under *'Project Abandoned'* when operations on the ground cease, even if the concession contracts are still in place. These projects, while legally concluded, are effectively unsuccessful.

BPMN is used mainly to provide a clear conceptual model, while the actual execution captured in the event log is far more flexible and variable.

5.4.2 Event log Inspection

Each row in the event log represents an instance of the deal history, identified by a precise Case ID, an activity, and a timestamp.

In this section we present the results obtained from log inspection and variant analysis. The analysis and the charts are obtained by using the `PM4Py` Python library, which provides process mining capabilities including visualization of variants and frequency-based behavior. PM4Py was selected for its flexibility and seamless integration with Python-based data analysis workflows.

Event log Description The inspection of the subsumed event log revealed that there are 2198 different deals ID. In total it records 5138 events, with 13 different activities: ‘Intended (Under negotiation)’, ‘Project not started’, ‘Concluded (Contract signed)’, ‘Failed (Contract cancelled)’, ‘Startup phase (no production)’, ‘In operation (production)’, ‘Project abandoned’, ‘Intended (Expression of interest)’, ‘Contract expired’, ‘Concluded (Oral Agreement)’, ‘Intended (Memorandum of understanding)’, ‘Failed (Negotiations failed)’, ‘Concluded (Change of ownership)’.

The cases can be of variable length. The maximum trace length is determined by 11 events, while the minimum length is 1, which corresponds to an update in either the Negotiation or Implementation status. This implies that some traces do not reflect the full progression of the Deal ID, and are thus considered ‘incomplete’. In addition, the average number of events per Deal ID is 2.33, suggesting that most cases include only updates on the Negotiation and Implementation statuses, without additional updates on their progression, which would have resulted in a higher number of events per Deal ID. Preliminary insights into temporal execution patterns were obtained through a dotted chart analysis (see Appendix A.3, Figure A.2). This chart displays individual activities as dots, with timestamps on the x-axis and case IDs on the y-axis, sorted by ascending activity time within each case and by earliest timestamp across cases.

One recurring pattern involves a transition from ‘Concluded’, typically the initial activity in the process and represented in orange, to ‘In Operation’, shown in sky blue. In certain cases, a high frequency of activity updates is observed within individual cases, resulting in a denser event distribution. This behavior may correspond to multiple iterative phases of negotiation/implementation between the involved parties. Conversely, some traces exhibit noticeable temporal gaps between activities, suggesting delays in the progression to subsequent states. Such gaps may indicate a latency period following contract finalization, prior to the transition into the operational or production phase of the project.

Variants Analysis The analysis of the unique paths in the event log showed that there are several variants, presenting different distributions and frequencies. More precisely, the variant analysis revealed that there are 321 different variants in the event log. However, the majority of cases are covered by variants having 1 or 2 activities. The aforementioned variants are not included in the process analysis due to their length, they are not able to represent the history of the acquisition process. This means that the other almost 300 rare and different variants are different from each other, reflecting how the land acquisition process does not follow a general standard process, and each deal may represent a different case per se. This significantly increases the visual complexity of the process, thereby reducing the interpretability of the process model visualization.

As the above analysis shows, this process presents very high variability and heterogeneity of variants, therefore reflecting the intrinsic flexibility and noise in the LSLAs process. This behavior limits the applicability of standard process discovery algorithms. Indeed, the lack of a dominant, standardized process flow motivates our choice to focus on frequency-based behavioral representations (e.g., DFGs) rather than on detailed structural process models.

5.4.3 Directly Follows Graph

To model the process of deals transitioning between different statuses, a Directly Follows Graph (DFG) was generated. The rationale behind this decision is that our focus is not on the structure of the process itself. That is to say, we are not interested in how much the model is able to represent the event log, nor are we interested in decision points and concurrency constructs. Our interest is in the exploration of the behavior, that is the directly following relations between activities. This goal can be achieved by modeling the process through DFGs. This kind of graph captures the directly follows relations between activities, illustrating both the frequency of each activity and the transitions between them through the edges connecting the activities. The intensity of the color assigned to each activity node reflects its frequency, while the thickness of the edges represents the

Table 5.1. Statistical description of the DFG. event log measures refer to the characteristics of the input data, while DFG measures describe properties of the resulting graph used for process modeling.

	event log						DFG		
	# traces	# variants	variants / traces	# events	min. trace length	max. trace length	# vertices	# edges	|E| / |V|
event log	2198	321	0.15	5138	1	11	13	110	8.46

frequency of the directly follows relations between the activities. Usually, this visualization graph provides a clear representation of the event log structure by illustrating process dynamics, highlighting the most common transitions and key patterns in the process flow. Indeed, in this case, the DFG representation reveals the high complexity of the process (see Appendix A.3 Figure A.3). The graph is difficult to interpret due to the large number of variants, making the process opaque. The abundance of different paths and transitions results in a classical *'spaghetti chart'*, as can be observed from the structural characteristics shown in Table 5.1, highlighting the need for further simplification or filtering of variants to better understand the underlying process.

The high edge-to-vertex ratio reported in Table 5.1, $|E|/|V| = 8.46$, further confirms the need for simplification. The next sections introduce trace clustering and log decomposition as strategies to reduce this complexity.

5.5 Experimental Setup for Complexity Reduction and Predictive Analysis

This section outlines the framework and the methodologies adopted to address the problem of complexity reduction in the DFG modeled from the event log, and then to predict the outcome of the land acquisition process.

5.5.1 Experimental Plan

The analysis conducted in this paper are designed with three main objectives: (i) exploring process model behavior and reducing process complexity, (ii) evaluating the contribution of process-related information to outcome prediction, and (iii) assessing the impact of different encoding techniques on prediction accuracy.

Following the preprocessing phase, which includes Log Decomposition and event log encoding, the data is clustered using both K-means and AHC, with the goal of grouping similar traces and simplifying the process representation (i). Then, to assess the predictive value of control-flow information, features are extracted from two different datasets (i.e., one with event log information and one without) and used to train a RF classifier (ii). To clarify, the dataset with event log information includes the full sequence of recorded activities along with static case attributes. In contrast, the dataset with only attributes excludes the activity sequence entirely, retaining only static case-level information, which are the attributes. Finally, to analyze the impact of different encoding strategies, various techniques, such as embedding and DL-based methods, are applied to categorical attributes prior to classification with RF model (iii).

5.5.2 Log Decomposition

Log Decomposition is employed in the context of PM with the objective of reducing the complexity of process models. It consists in splitting the event log into sub-logs based on domain knowledge and end activities, thus abstracting the complexity of the entire event log into more manageable subsets. The objective is to ensure the preservation of the sequence of activities within the business process, and subsequently to group traces into clusters based on their sequence behaviors. In the context of LSLAss, the application of the Log Decomposition involves the extraction of different sub-logs based on the most recent status of the negotiation or implementation, which corresponds to the end activity of the process. The resulting sub-logs are then modeled as DFGs and evaluated through both visual inspection and the complexity metrics introduced in subsection 5.5.6.

The variant analysis presented in Paragraph 5.4.2 played a crucial role in guiding the choice of clustering. Initially, we considered applying filtering techniques to simplify the process model and enhance its readability. However, the analysis revealed that most variants are infrequent and structurally heterogeneous. As such, filtering them out would have led to the loss of valuable behavioral information, resulting in a biased and incomplete representation of the process. In other words, filtering would have oversimplified the process landscape, compromising our objective of capturing its full behavioral diversity. For this reason, we opted to preserve such variability by employing clustering and visualization strategies that allow us to manage complexity without discarding rare but potentially meaningful cases.

We chose to explore index-based encoding combined with two widely used clustering algorithms: K-means and AHC. The rationale behind this choice, already discussed at the end of section 5.2, specifically in the Paragraph 5.2.1, is primarily driven by the characteristics of the resulting feature vectors, which are high-dimensional and sparse. These properties make the application of alternative techniques, such as density-based methods like DBSCAN, which are less suitable, due to their sensitivity to sparsity and their reliance on precise parameter tuning.

5.5.3 Index-Based Encoding and Categorial Variable Encoding

In order to represent the event log as feature vectors, it is essential to perform an appropriate encoding procedure, which aims to store process data prior to the deployment of the algorithm. The encoding may be performed based on the process characteristic in order to store several pieces of information. In general, the objective is to store information about the control-flow perspective of the process, that provides information about the sequential process. To perform both clustering and classification, each trace in the process is represented as a point in a vector space using *index-based encoding*, which is a lossless encoding technique that enables the original sequence of activities to be preserved. This approach enables process analysis from a control-flow perspective by capturing the order of activities within each trace.

Using index-based encoding, the trace is transformed into a format that allows for the differentiation between static and dynamic information, where static information refers to information that remains constant across different events with the same case ID, and dynamic information refers to information that varies across different events with the same case ID. For a comprehensive definition of index-based encoding, the reader can refer to literature in [70, 32]. In the scenario under analysis, the *dynamic information* is represented by the activities of the case, and the *static* *information* is represented by the attributes of the case, which remain constant throughout the process. Subsequently, the features derived from index-based encoding for dynamic and static information, are processed using different encoding techniques to assess whether the choice of encoding impacts classification performance.

In this study, three different combinations of encoding techniques for dynamic and static information are employed to represent the features for classification. Hereafter, we refer to dynamic information as the sequence of activities and static information as the case attributes. For clustering, Ordinal Encoding was deemed sufficient for both dynamic and static information, as the goal of clustering is to identify inherent patterns and structure in the data, where the precise encoding of the sequential nature of activities is less critical. In contrast, for classification we decided to explore three encoding combinations, as we believe that the encoding of sequential information directly influences the predictive accuracy of the model, especially when dynamic features are key to the outcome. Specifically, the following encoding approaches are used:

- Ordinal Encoding for dynamic and static information;

- BERT Embedding for dynamic information + Ordinal Encoding for static information

- BERT Embedding for dynamic information + MLP-Transformed Ordinal Encoding for static information

Ordinal Encoding A prevalent encoding technique in the domain of categorical encoding is *ordinal encoding*, which implies the existence of an order among the dimensions. This approach is particularly effective for dynamic information because preserving the inherent sequence is crucial for analyzing control-flow behavior. While one-hot encoding is typically preferred for static information, it becomes impractical when dealing with a large number of categories. Therefore, in this study, we use ordinal encoding for static attributes too. Despite the potential bias induced by an artificial order, this method remains a practical and commonly-adopted choice for handling high-cardinality categorical variables.

BERT Embedding + Ordinal Encoding BERT is a transformer-based model designed specifically to handle sequential data. Within this experiment, we used BERT to process dynamic information in index-based encoded traces by utilizing its tokenization and embedding layers. During the tokenization phase, textual data is converted into tokens, which are then transformed into embeddings through BERT's deep layers. This allows BERT to capture the contextual relationships within the sequence. The resulting embedded vectors representing dynamic information are then concatenated with ordinal-encoded static information to form a complete vector representation of each process instance.

BERT Embedding + MLP-Transformed Ordinal Encoding In this experiment, the categorical attributes encoded with ordinal encoding are further transformed using an MLP to obtain a representation of the static information. The MLP maps these attributes into a dense vector space, enabling a more expressive representation that facilitates the integration of static and dynamic data. The resulting encoded attributes are then standardized and merged with the dynamic features extracted using BERT embeddings. The final combined feature vector, which incorporates both the transformed static attributes and the dynamically derived BERT representations, aims to provide a richer representation of the underlying process.

5.5.4 Clustering Algorithms

In this study, we focus on K-means and AHC applied to index-based and ordinal encoded traces, in order to identify homogeneous cases exhibiting consistent behavioral patterns. The selection of these algorithms is motivated both by their *theoretical properties* and by the *characteristics of the dataset* under analysis.

Due to the limited number of activities contained in each trace, the encoded data result in sparse and relatively low-density feature vectors. Under these conditions, density-based algorithms such as DBSCAN are known to be less effective, as the notion of density becomes less informative in sparse spaces and requires careful parameter tuning. Preliminary experiments confirmed these limitations, showing unstable cluster structures and strong sensitivity to parameter choices.

In contrast, K-means is well suited for partitioning data into clusters with comparable size and compactness, producing interpretable sub-logs that facilitate the subsequent analysis of DFGs. In addition, AHC offers a complementary hierarchical view, capturing subtle similarities between traces that may be overlooked by flat clustering methods and enabling multiple levels of granularity.

The use of these two algorithms allows us to balance scalability, interpretability, and sensitivity to data sparsity, while providing both global and local insights into process behavior. This methodological choice directly addresses the challenges outlined in section 5.2 and ensures that the resulting cluster structures are both robust and theoretically grounded.

K-means K means clustering algorithm is a partition-based clustering algorithm, considered one of the most used clustering algorithms in data analysis [140], which constructs k clusters by dividing the data into a predefined number of k groups. By using K-means clustering methods it is possible to divide event log Data into different K clusters of sub-logs. The main objective of K-means is to measure the difference between vectors based on the euclidean distance measure.

Agglomerative Hierarchical Clustering The AHC is a bottom-up clustering algorithm that is widely employed in the field of data mining. The AHC treats each data point as a cluster and then performs a recursive merging of the two closest clusters based on a chosen similarity measure, such as the euclidean distance [102]. In order to represent the data, a hierarchical tree structure (dendrogram) is provided, which allows users to visualize cluster structures at varying levels of granularity.

5.5.5 Classification

In this study, we decided to use RF for classification of the output of the cases. RF was selected due to its strong generalization ability, resistance to overfitting, and interpretability. Initially, RF is first applied to ordinal-encoded data, comparing two datasets: one with only attribute information and the other with both activities and attributes. Then, we extend the comparison to include the complete event log (activities and attributes) with different encoding techniques. The algorithm, used for both classification and regression tasks, works by generating a large number of decision trees during the training phase, with each tree making its own prediction. The final output is determined by aggregating the results of all the trees, based on a voting process for classification tasks and on average for regression tasks.

5.5.6 Evaluation Metric

With the aim to assess the quality of the obtained results, we adopt different evaluation metrics: for trace clustering, for the process model complexity, and for classification task. More in details, clustering results will be assessed by means of silhouette score. Additionally, to evaluate the process models complexity of the obtained graphs we follow [56]. Specifically, we opted for *structural complexity metrics*, in particular we choose the following metrics based also on the metrics used in [22], in which the complexity of the DFG is evaluated:

- Number of *traces* and number of trace *variants*;
- *Variability*, which is the ratio of the number of variants over the number traces;
- Number of *events*;
- Minimum (Min.) and Maximum (Max.) *trace length*;
- Number of arcs and nodes which are respectively the number of *edges* (directly follow relations) and number of *vertices* (activities);
- *Density*, which is the ratio of the number of vertices to the number of edges;
- Just visual analysis;

Since we employed DFGs for process discovery, we relied on structural complexity metrics for evaluation. Unlike executable models such as BPMN or Petri nets, DFGs lack formal execution semantics, making traditional semantic quality metrics, such as fitness and precision, which are inapplicable in this context.

Then, the classification results will be evaluated utilizing accuracy, balanced accuracy and weighted avg precision, recall, and F-1 score.

5.6 Results

In order to evaluate the proposed framework, the event log obtained by the preprocessing LSLAs data (shown in Appendix A.2, Table A.4) was used as the input .

5.6.1 Log Decomposition

The event log Analysis in section 5.4 illustrates that there is a considerable degree of complexity, as evidenced by both the event log inspection and the visual analysis of the resulting DFG. In the event log the number of activities does not contribute to the complexity; however, the high number of different variants is translated into a large number of paths, which results in a high complexity level of the log. Accordingly, given the event log's variability, we implement log decomposition and extract three discrete sub-logs, as delineated by the final activity, as shown in subsection 5.5.2. Based on domain expert knowledge, we propose to divide the log information into three sub-datasets, based on the final status of the trace (i.e., corresponding to the current status of the deal):

- **Work In Progress**: the dataset includes all cases corresponding to deals that are still in progress, therefore ending with 'Intended (Under negotiation)', 'Intended (Expression of interest)', 'Intended (Memorandum of understanding)', 'Concluded (Contract signed)', 'Concluded (Oral Agreement)', 'Concluded (Change of ownership)', 'Startup phase (no production)', 'Project not started' activities;

- **Production**: the dataset comprises all events related to deal processes currently in production, therefore ending with 'In operation (production)' activity;

- **Failed**: the dataset comprises all instances of failed deal processes, therefore ending with 'Project abandoned','Failed (Contract cancelled)','Failed (Negotiations failed)','Contract expired' activities.

Subsequently, the obtained aforementioned sub-logs were modeled using DFGs, and then evaluated using process model complexity measures. The process models obtained by dividing the event log based on the outcome type were evaluated using the process model complexity measures described in subsection 5.5.6. As shown in the first rows of Table 5.2, Table 5.3, and Table 5.4, this evaluation demonstrated that the models still exhibited high complexity, with a high density measure and evident spaghetti behavior. These results are also reflected in the DFGs shown in Appendix A.4, Figures A.4, A.5, and A.6.

5.6.2 Complexity Reduction

Subsequently, in order to further reduce the complexity of the model, trace encoding and trace clustering were adopted on such sub-logs as shown in subsection 5.5.3 and subsection 5.5.4. Consequently, index-based encoding which includes feature extraction and trace encoding, was employed, resulting in the identification of seven features pertaining to the control-flow behavior of the process, i.e. the activities of the process. These features were successively encoded via ordinal encoding.

After that, trace clustering was used on the encoded event log. This process was performed on each sub-log, with clustering results evaluated using the silhouette score as a performance metric. The optimal number of clusters was determined by combining the elbow method with domain knowledge, particularly by analyzing the trace lengths. Specifically, the number of clusters was influenced by the variable prefix length of the traces, which ranged from 2 to 7 activities. To mitigate the risk of over fitting, the number of clusters was limited, even in cases where a higher silhouette score was achieved with more clusters. In this decision, the number of observed trace variants played a crucial role. Specifically, avoiding a number of clusters too close to the number of variants helped prevent trivial separations with limited analytical value, which would have simply grouped individual variants without allowing for meaningful generalization. Hence, the final number of clusters was selected to strike a balance between data complexity, variant density, and the need for interpretable, semantically coherent process models. This approach ensured a balance between capturing variations in trace complexity and maintaining meaningful distinctions within the data. Subsequently, a DFG was retrieved from each cluster derived from the sub-logs and then was subjected to visual inspection and comparison using process model complexity measures subsection 5.5.6. The effectiveness of the employed complexity reduction technique is evaluated on each event log taken into consideration. To this

Table 5.2. Failed Sub-Log description before and after clustering

	event log						DFG		
	# traces	# variants	variants / traces	# events	min. trace length	max. trace length	#vertices	#edges	|E| / |V|
Failed Sub-Log	262	86	0.33	760	1	8	13	72	5.54
C0_AHC	98	18	0.18	199	2	3	10	18	1.8
C1_AHC	13	13	1	76	5	7	12	34	2.83
C2_AHC	93	25	0.27	279	3	3	13	32	2.46
C3_AHC	47	27	0.57	188	4	4	13	42	3.23
C0_KM	59	39	0.66	260	4	7	13	55	4.23
C1_KM	96	28	0.29	288	3	3	13	35	2.69
C2_KM	96	16	0.17	194	2	4	11	17	1.55

Table 5.3. Work In Progress Sub-Log description before and after clustering

	event log						DFG		
	# traces	# variants	variants / traces	# events	min. trace length	max. trace length	#vertices	#edges	|E| / |V|
WIP Sub-Log	867	140	0.16	1795	1	10	12	78	6.50
C0_AHC	17	17	1	99	5	7	11	33	3.00
C1_AHC	55	20	0.36	166	3	4	11	26	2.36
C2_AHC	50	38	0.76	215	4	7	11	42	3.82
C3_AHC	19	13	0.68	64	3	4	9	15	1.67
C0_KM	230	43	0.19	46	4	7	11	51	4.64
C1_KM	177	36	0.20	55	3	4	12	38	3.17
C2_KM	237	28	0.12	73	3	4	11	31	2.81

Table 5.4. In Production Sub-Log DFG description before and after clustering

	event log						DFG		
	# traces	# variants	variants / traces	# events	min. trace length	max. trace length	# vertices	# edges	|E| / |V|
In Production Sub-Log	1069	95	0.09	2583	1	11	12	55	4.58
C0_AHC	37	33	0.89	201	5	7	11	38	3.45
C1_AHC	51	9	0.18	153	3	3	7	12	1.71
C2_AHC	65	36	0.55	260	4	4	12	39	3.25
C3_AHC	132	4	0.03	396	3	3	5	7	1.40
C0_KM	183	13	0.07	549	3	3	8	18	2.25
C1_KM	65	36	0.55	260	4	4	12	39	3.25
C2_KM	65	33	0.51	201	5	7	11	38	3.45

aim, it is evaluated by considering both event log characteristics which describe the event log structure and DFG characteristics by utilizing the graph complexity measures Table 5.2, Table 5.3, Table 5.4.

The obtained results suggest that AHC tends to produce a larger number of smaller clusters, often characterized by lower density and a higher risk of overfitting. While this fine-grained segmentation may capture subtle behavioral differences, it can reduce the analytical value by limiting the potential for generalization.

On the other hand, K-means clustering allows for more controlled grouping, typically resulting in fewer clusters, each representing a broader set of behaviors. This leads to denser and more informative DFGs, where the process models capture recurring patterns more effectively, without being overly fragmented.

5.6.3 Outcome Prediction

For the outcome prediction task we desing two different experiments on the Event Log without applying any complexity reduction: one regarding the evaluation of the inclusion of process control-flow related information, and one considering the different encoding methodologies, all compared and evaluated using RF. We utilized the original event log dataset for outcome prediction in order to preserve the full behavioral variability of the process, and to ensure a valid and informative predictive modeling setting.

Event Log Manipulation The classification task was performed using RF, as outlined in section 5.5, and the results were evaluated through the classification metrics in subsection 5.5.6. In particular it was assessed the capability of the ML algorithm of classifying the outcome of the process in the three labels:

- Failed: 0
- Production: 1
- Work In Progress: 2

Subsequently, the data underwent manipulation in order to facilitate the generation of a prediction. In particular, traces comprising only two activities were removed, and the final activity (which is the designated outcome) was excluded. Subsequently, the data underwent through the encoding phase, using the encoding outlined in subsection 5.5.3. From the data manipulation stage, it resulted that the event log is characterized by an imbalanced distribution of outcome labels. Specifically, *Production* (label 1) is the most represented, with 313 instances, followed by *Work In Progress* (label 2) with 214 instances, and by *Failed* (label 0) with 180 instances. This imbalance in class distribution influences the results.

Classification and Evaluation First, to assess impact of incorporating activity transitions history (i.e. the event log) in the acquisition process, we evaluated the classification task on different encoded features. To this end, a comparison between two scenarios has been evaluated: one including both the sequence of activities and attribute information (i.e., event log + metadata), and one including only static attribute information (i.e., metadata alone).

A comparison of the results using the classification metrics is provided in Table 5.5. These results show a slight decrease in classification performance when both activities and attributes are used. This performance drop can be attributed to the class imbalance in the dataset, which results in the under-representation of certain classes. In addition, as demonstrated in section 5.4, the presence of several process variants results in a higher degree of variability in process behaviors, which complicates the learning process due to a lack of consistent patterns in the data. However, from the confusion matrix obtained with the experiments, we understand that when sufficient historical data is available, it enhances the prediction of the final outcome of the deal process, as the model has more information to base its predictions on. Based on these findings, we conducted further experiments with different encoding techniques to better capture process activity information.

Comparison of Encoding Techniques We explored alternative encoding strategies to address the aforementioned issues and better capture the underlying patterns in the highly variable processes, aiming to improve overall predictive performance.

Specifically, we employed two additional encoding combinations and used RF for classification, as outlined in subsection 5.5.3. For BERT, we followed a two-step process: tokenization and embedding, followed by a classification layer. The BERT-based uncased variant [30] was utilized for

Table 5.5. Classification results for the attributes and activities dataset encoded using three encoding techniques and evaluated across five metrics

Model	Accuracy	Balanced Accuracy[1]	Precision	Recall	F1 Score[2]
O_attr	0.58	0.55	0.58	0.58	0.57
O_act_attr[3]	0.55	0.51	0.57	0.55	0.54
BERT+O_act_attr[4]	0.62	0.59	0.62	0.62	0.61
BERT+MLP_act_attr[5]	**0.63**	**0.61**	**0.63**	**0.63**	**0.63**

embedding. During tokenization, indexed data were converted into tokens, which were then transformed into embeddings through BERT's deep layers.

Next, we combined the BERT embeddings with MLP-transformed ordinal-encoded attributes to capture both sequential patterns from the activity traces and the case attributes. The MLP was used to transform the ordinal-encoded features into a high-dimensional space, enabling the model to better capture the patterns in the data. The results, presented in Table 5.5, show that incorporating BERT embeddings for process traces improves model performance. Additionally, the concatenation of BERT embeddings with MLP-encoded features yields better results compared to using BERT alone. While the overall accuracy remains similar, the improvement is particularly notable in predicting underrepresented outcome categories in the original dataset.

The implementation of the framework necessitated the use of the PM4Py library, written in Python[14].

5.7 Conclusion

In this Chapter we explore the application of PM techniques in the context of LSLAs, with the goal of making process behaviors more transparent and interpretable to domain experts, ultimately supporting more informed land policy analysis.

We started by constructing an event log through data preprocessing, which we then used to model the LSLAs process. The resulting process model, represented by DFGs, revealed significant behavioral variability and complexity, characteristic of spaghetti process models. To address this complexity, we applied trace clustering, combining domain knowledge with index-based encoding and both K-means and AHC algorithms. The results demonstrated that clustering effectively reduces model complexity and enhances interpretability, with K-means maintining a balance between process model density and representation.

We further assessed the predictive power of process-related information by comparing different datasets, one including control-flow features and another based solely on attributes, using a RF classifier. The experiment results indicated limited gains from incorporating control-flow data, which is due in part to data sparsity and variability.

Therefore, to further assess the impact of capturing process control flow-related information on outcome prediction, we investigated alternative encoding strategies using DL models, including BERT embeddings and MLPs. These techniques produced more balanced classifications across outcome categories, suggesting that advanced encoding can capture richer behavioral signals and reduce bias toward dominant classes.

Overall, our findings highlight the relevance of incorporating process-related information in the analysis of LSLAs, which is an aspect largely overlooked in prior research. They reinforce the value of PM techniques and DL models for enhancing outcome prediction, and emphasize the critical role of encoding strategies when dealing with complex, irregular processes. This lays the foundation for more accurate predictive modeling and more informed policy evaluation in land governance.

Chapter 6

Conclusions

6.1 Conclusion

This dissertation addressed the persistent gap between the operational reality of DT within PAs and the need for more efficient and effective digital services, as outlined by the objectives set by the EU. Despite EU promoted several and extensive investments in DT initiatives, many administrative organizations still struggle to keep pace with their infrastructures and organizational structures. In fact, often PAs operate in contexts with fragmented ISs, heterogeneous data sources and lack of interoperability across departments. These conditions constrain end-to-end process visibility and timely adoption of data-driven decision-making.

This dissertation proposes a *process-aware* perspective that integrates BPM, PM, and PPM to address the aforementioned challenges, reinforcing the alignment between strategic objectives, process execution, and technological implementation.

Following an exploratory approach that progresses from general concepts to specific applications, three research lines were pursued.

First, the empirical analysis on-site of the judicial offices, conducted through the application of BPM, highlighted the structural issues affecting PAs, namely, non-integrated IS, manual data handling, and the absence of shared, process-oriented platforms. This analysis allowed the identification of process redesign solutions, which were operationalized by proposing PoC functionalities. Then, with a *Process Evaluation* phase, it was possible to derive functional requirements for subsequent IS. Finally, the study also shed light on the object-centric nature of PA processes.

Second, the SLR on PPM clarified the current research landscape and identified key challenges in extending predictive analytics to real-life contexts, such as the ones in PA. The study highlighted the transition from classical case-centric event logs to OCELs, which allow the modeling of multiple interacting entities. While OCELs enhance representational richness and predictive accuracy, the review also identified unresolved issues, such as computational scalability, encoding heterogeneity, and the absence of standard benchmarks, therefore pointing to the need for more mature methodological foundations.

Third, the experimental application of PM and PPM techniques to the Land Matrix dataset demonstrated that process-oriented analysis can be successfully applied even to semi-structured, irregular, and "spaghetti-like" event data. In fact, despite the noisy nature of the data, PM uncovered meaningful behavioral patterns, while predictive models achieved improved performance through clustering and contextual encoding (e.g., BERT embeddings).

These results confirm that PM and PPM are not limited to structured enterprise environments, but can also generate valuable insights in unstructured and complex domains.

By combining the top-down modeling capabilities of BPM with the bottom-up evidence derived from PM and PPM, this dissertation demonstrates how process-oriented methods can effectively bridge the gap between strategic planning and operational execution. This integrated perspective promotes transparency, interoperability, and continuous improvement, enabling PAs to evolve from fragmented, document-centric practices toward intelligent, data-driven, and adaptive process governance.

Conceptual Framework The conceptual framework proposed in this dissertation aligns with the principles of BPM frameworks, integrating top-down strategic guidance with bottom-up empirical evidence to ensure that DT initiatives in PAs are both goal-driven and informed by real process behavior.

At the *strategic level*, the organizational objectives guide process analysis and redesign.

Then, at the *process level* through BPM it is provided a structured representation of workflows and their inefficiencies.

The *knowledge level*, informed by the SLR, consolidates insights on PPM, identifying approaches that can be generalized across contexts.

Finally, the *application level* employs PM and PPM on empirical event log data to generate actionable insights, detect operational issues, and support continuous improvement.

6.1.1 Limitations and Future Directions

In this dissertation, each research line presents both contributions and specific limitations which are discussed below.

The first research line offered valuable insights, highlighting opportunities for process redesign. However, organizational constraints prevented the implementation of the proposed functional requirements in a real IS, restricting the empirical evaluation of the full integration between BPM and PM.

The second research line provided important insights into OCEL process analysis, nevertheless, the relatively small number of object-centric studies in the literature limits the ability to derive generalizable recommendations for this area.

The third research line explored the application of PM and PPM to the Land Matrix log. Since this log represents a novel dataset, benchmarking against similar datasets is not yet possible, complicating the validation of the proposed methods. Moreover, the datasets used in the experimental evaluation are limited in both scale and scope.

Future work will focus on extending the application of PM and PPM methods to larger and more heterogeneous event logs across diverse administrative settings with object-centric characteristics.

Appendix A

Appendix

A.1 Predictive Process Monitoring

Table A.1. Classification of Object-Centric Event Logs (Ordered by year and author) based on **Study, Prediction Task, Method, Metrics, Contribution, Domain**, where the number of object types (ot) of the dataset is reported in brackets.

Study	**Prediction Task**	**Method**	**Metrics**	**Contribution**	**Domain**
Gherissi et al. [50]	next activity, next event time, remaining time	LSTM	accuracy, MAE	enriched event logs	customer order management (3 ot)
Rohrer et al. [104]	next activity sequences, next timestamp	GANs	sequence similarity, MAE	rich object-attribute information	order management (4 ot)

Galanti et al. [48]	KPI	CatBoost	F1-Score, MAE	flattened logs with aggregation attributes	utility provider (5 ot)
Adams et al. [1]	next activity, next timestamp, remaining time	GNN, Graph Embeddings, Flattened Event Logs Approach	accuracy, MAE	graph structure of event logs	financial services (2 ot)
Galanti et al. [46]	KPI	GCN, GGNN, LSTM, CatBoost	F1 Score, MAE	flattened logs with object interaction	technology (3 ot), utility provider (5 ot)
Knopp et al. [63]	next activity	Petri Nets, Statistical Learning	EMD, Behavioral Conformance, Timing Behavior	-	order management (3 ot), procure-to-pay (3 ot)
Niro et al. [87]	anomaly detection	GNN, GCNAE	F1-score, PR-AUC, ROC-AUC, Recall	-	financial services (2 ot), order management (3 ot)
Ruffini et al. [105]	outcome, KPI	GNN	AUC, F1-score	rich inter-object relationships	public tenders

Smit et al. [108]	remaining time	GNN	MAE, MSE, MAPE, training and prediction times	-	financial services (3 ot), order management (3 ot), operational system (3 ot)
Van del Aalst et al. [130]	-	-	multi-object interactions	-	OC-PM (6 ot)
Berti et al. [12]	anomaly detection, throughput time	ML (isolation forests, regression models)	Isolation Forest Scores, MAPE, RMSPE	interaction graphs	order management (3 ot), order-to-cash (all), procure-to-pay (4 ot), recruiting (3 ot)
Chifeng et al. [23]	activity duration	Petri Nets, LSTM	Precision, Recall, mAP, MSE, RMSE, MAE, R-squared	multimodal data	healthcare
Le et al. [68]	KPI, next activity	counterfactual reasoning, CatBoost	-	dynamic process	-

Smit et al. [109]	remaining time	HOEG, GNN	MAE, MSE	graph encoding	financial services (2 ot), order management, financial institution

The analyzed proposals are summarized in Table A.2 below.

Table A.2. Classification of Classic Event Log (Ordered by year and author) based on **Study, Prediction Task, Method, Metrics, Contribution, Domain**

Study	**Prediction Task**	**Method**	**Metrics**	**Contribution**	**Domain**
Di Francesco-marino et al. [31]	outcome	DT, RF, DBSCAN	F1-score, earliness, failure rate, CT	offline training	healthcare
Marquez et al. [74]	next event, remaining event	Evolutionary Algorithms	precision recall specificity F-measure AUC	event-window encoding	IT service management
Tax et al. [118]	next activity, next timestamp, remaining event, remaining time	LSTM	accuracy, MAE, DLS	multi-task learning	financial services, IT service management

Study	**Prediction Task**	**Method**	**Metrics**	**Contribution**	**Domain**
Polato et al. [97]	remaining time	DATS, SVR, SVR+TS	MAPE, RMSPE	process dynamism	IT service management, municipality
Wang et al. [138]	outcome	Att-Bi-LSTM	accuracy, earliness, execution time	NLP	administrative process, financial services, healthcare, municipality, production
Park et al. [90]	remaining time, next activity	LSTM	TWCT, CT	resource scheduling	financial services, healthcare
Camargo et al. [19]	next activity, suffix, remaining time	LSTM	Damerau–Levenshtein (DL), MAE	embedding	financial services, healthcare
Pasquadibisceglie et al. [92]	next activity	CNN	accuracy, precision, recall	temporal data into spatial data	IT service management, financial services
Marquez et al. [75]	single KPI aggregated KPI	RF, XGB	MAE RMSE	incorporation of contextual information	IT service management

Study	Prediction Task	Method	Metrics	Contribution	Domain
Mehdiyev et al. [78]	next event	DL multi-stage	avg. accuracy, avg. precision, avg. recall, avg. F-measure, MCC, AOC	feature hashing	financial services, IT service management
Park et al. [91]	KPI	CNN, LSTM, LRCM	MAPE, MAE	traffic congestion	financial services, healthcare, IT service management
Pasquadibisceglie et al. [93]	next activity	2D CNN	Inception Architecture	traces into images	financial services, IT service management, municipality
Taymouri et al. [119]	next event label, timestamp	GANs (LSTM–FNN)	accuracy, MAE	RNNs to GANs	financial services, IT service management
Bukhsh et al. [17]	next activity, next event time, remaining time	Transformers	accuracy, F-score, MAE	self-attention mechanism	financial services, health-care, IT service management, municipality

Study	Prediction Task	Method	Metrics	Contribution	Domain
Kratsch et al. [65]	outcome	DL (LSTM–DNN), ML (RF–SVM)	accuracy, F-score, ROC AUC	event log structures	academy, healthcare, IT incident management, manufacturing, municipality
Pasquadibisceglie et al. [94]	next activity	LSTM	overall accuracy, macro-precision, macro-recall, macro-Fscore, macro-AUROC, macro-AUCPR	multi-process perspectives	administrative process, financial services, IT service management, municipality
Kotsias et al. [64]	next activity	Q-Learning	training score	RL no environment	financial services
Fahrenkrog et al. [39]	outcome	probability model	processing cost, mitigation effectiveness, F-score	parameterized cost model	financial services, municipality, social services
Chiorrini et al. [25]	next activity	BIG-DGCNN	accuracy, F1-score, AR, SRR, ranking R	BIG	administrative services, financial services, IT service management

Study	Prediction Task	Method	Metrics	Contribution	Domain
Oberdorf et al. [88]	multi-class	MLP, three-headed DNN	balanced multi-class accuracy, precision, recall, F1-measure	event log contextual information	manufacturing
Peeperkorn et al. [96]	remaining trace	LSTM	designed metrics	LSTM overgeneralization (LOVOCV approach)	artificial log
Theis et al. [121]	next event	Decay Replay Mining	mAUROC	reachability graph, masking vectors, FNN	financial services, IT service management
Bousdekis et al. [16]	next activity	RL, Q-Learning, DQN	training score, accuracy	RL (no environment)	financial services
Razo et al. [103]	next event	AXDP	AUC	LSTM generalization	administrative process, financial services, IT service management, municipality
Rama et al. [100]	next activity	TACO (GRNN–LSTM)	process metrics	spatio-temporal information	financial services, IT service management, municipality

Study	Prediction Task	Method	Metrics	Contribution	Domain
De Smedt et al. [28]	process model	time-series techniques	MAPE	time-series	financial services, healthcare, IT service management, municipality
Fani et al. [41]	next activity, remaining time, outcome	LSTM, XGBoost	accuracy	event log sampling	financial services, healthcare, municipality
Galanti et al. [47]	remaining time, activity occurrence, KPI	CatBoost, LSTM	MAE, F1-score, Shapley value	business users	financial services, IT service management, municipality
Gunnarson et al. [53]	suffix, remaining time	CRTP-LSTM	MAE, RMSE	no-hallucinated logs	administrative processes, financial services, municipality
Gunnarson et al. [52]	remaining trace, runtime	CRTP-LSTM	LS-ICE framework, intercase context	5 real-life datasets	administrative processes, financial services, logistics

Study	Prediction Task	Method	Metrics	Contribution	Domain
Yarlagadda et al. [141]	next activity	LSTM	accuracy, precision, recall, F1-score, AUC, sensitivity, specificity, FDR, MCC	novel DL model	customer helpdesk
Rama et al. [101]	activity suffix	Recurrent Graph Neural Network (RGNN)	Damerau–Levenshtein normalized distance	RGNN encoding	financial services, healthcare, IT service management, process management, municipality
Kaftantzis et al. [59]	next activity	TPOT, ML (LR, DT, RF, GB, SVM, kNN)	accuracy	ML pipeline automation	financial services

A.2 Dataset Overview

In this appendix, the datasets utilized in the study are presented.

Table A.3. Excerpt of selected columns from the LSLA raw dataset, extracted from publicly available data. Each row corresponds to a Deal ID. The columns Negotiation status and Implementation status describe the chronological evolution of each deal. The Intention of Investment column is shown as an example of other deal attributes, providing information about the type of intended investment.

Deal ID	Intention of investment	Negotiation status	Implementation status
10362	2022#current#7818#For carbon sequestration/REDD	2016#current#Concluded (Contract signed)	2022#current#In operation (production)
10363	2021-06-01#current#2863#For carbon sequestration/REDD	2014#current#Concluded (Contract signed)	2021-09#current#In operation (production)
10364	2014#current#165000#For carbon sequestration/REDD	2014#current#Concluded (Contract signed)	2014#current#In operation (production)
10391	#current##Solar park	2017#current#Concluded (Contract signed)	2017##Startup phase (no production) 2018-12#current#In operation (production)
10399	2002#current#12182#For carbon sequestration/REDD, Forestry unspecified	2022#current#Concluded (Contract signed)	#current#In operation (production)

A.3 LSLA Process Overview

This appendix provides additional process-related information, including the BPMN diagram of the ground truth process, the dotted chart derived from the event log analysis, and the Directly-Follows Graph (DFG) generated from the event log.

Table A.4. Excerpt of selected columns of the LSLA event log. Each row represents an event belonging to a case, which is identified by the Deal ID.

Index	Deal ID	Activity	Timestamp
5133	9958	Failed (Negotiations failed)	2023-01-01
5134	9958	Project abandoned	2023-01-01
5135	9962	Intended (Memorandum of understanding)	2019-05-22
5136	9962	Project not started	2020-01-01
5137	9990	Failed (Contract cancelled)	2022-01-01

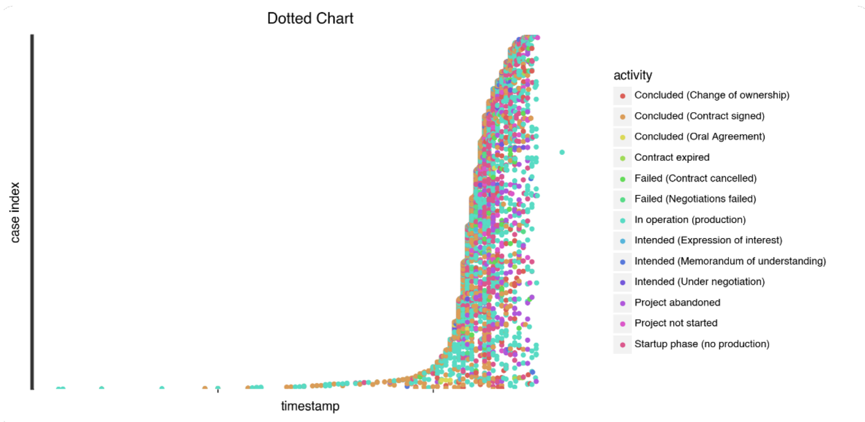

Figure A.2. Dotted chart of the event log. The x-axis represents timestamps; the y-axis indicates case indices. The color legend on the right maps activities to dots.

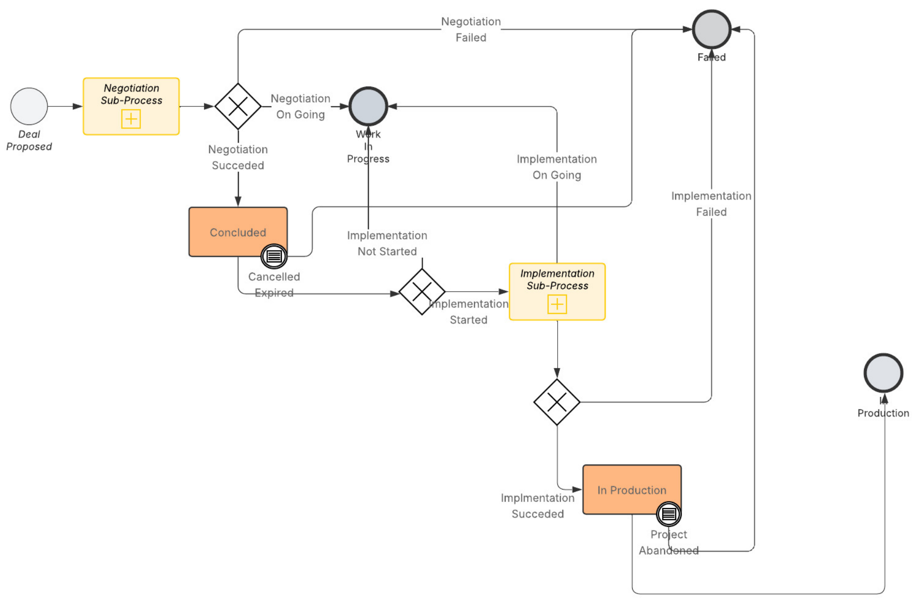

Figure A.1. BPMN representation of Land Acquisition Process

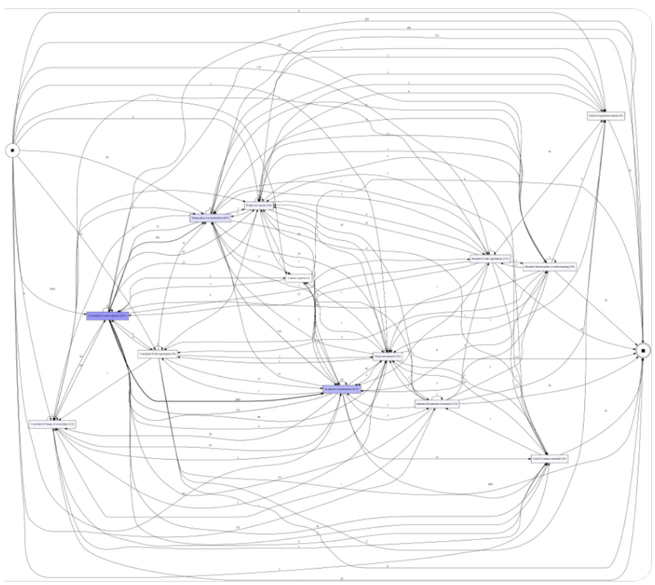

Figure A.3. Directly Follow Graph of the event log showing high complexity and low interpretability.

A.4 Pre-Clustering DFGs of Sub-Logs

In this appendix, we present the DFGs obtained for the failed, in-production, and work-in-progress sub-logs before applying clustering, showing that they are still characterized by high complexity.

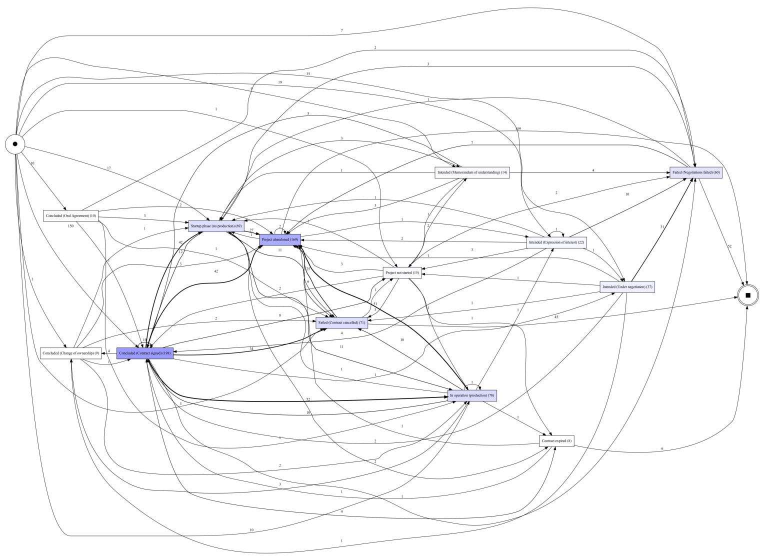

Figure A.4. Directly Follow Graph of the failed sub-log showing high complexity and low interpretability.

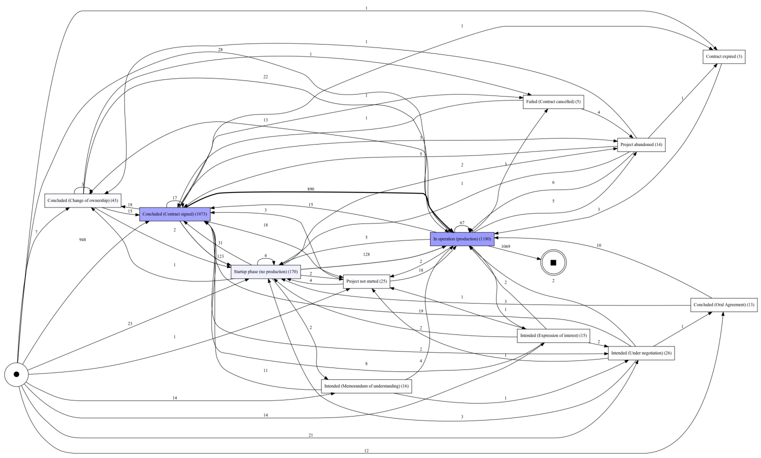

Figure A.5. Directly Follow Graph of the in production sub-log showing high complexity and low interpretability.

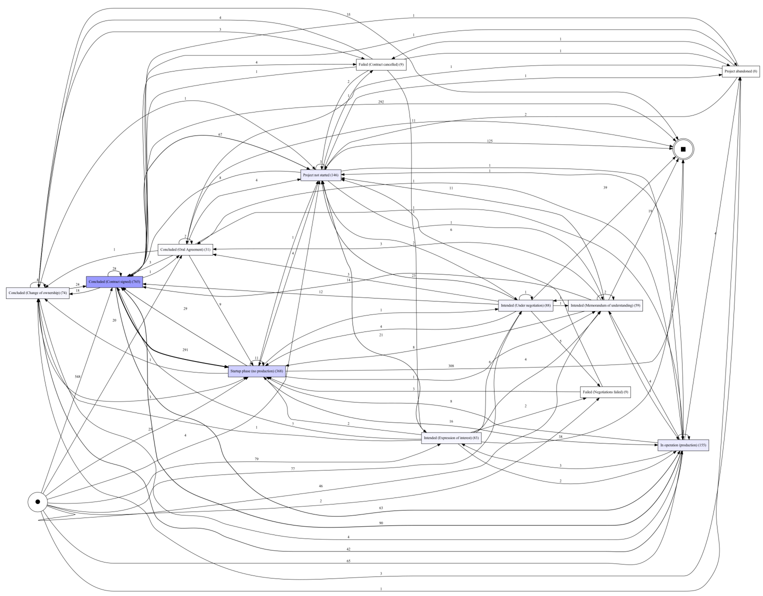

Figure A.6. Directly Follow Graph of the work in progress sub-log showing high complexity and low interpretability.

Bibliography

[1] Jan Niklas Adams, Gyunam Park, and Wil MP van der Aalst. Preserving complex object-centric graph structures to improve machine learning tasks in process mining. *Engineering Applications of Artificial Intelligence*, 125:106764, 2023.

[2] Jan Niklas Adams and Wil M. P. van der Aalst. Addressing convergence, divergence, and¬†deficiency issues. In Jochen De Weerdt and Luise Pufahl, editors, *Business Process Management Workshops*, pages 496–507, Cham, 2024. Springer Nature Switzerland.

[3] International Energy Agency. Renewables 2023, IEA, Paris. https://www.iea.org/reports/renewables-2023, 2024. [Accessed 05-09-2024].

[4] Razi Ahmed, Muhammad Faizan, and Anwer Irshad Burney. Process mining in data science: A literature review. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, pages 1–9. IEEE, 2019.

[5] Flora Amato, Simona Fioretto, Eugenio Forgillo, Elio Masciari, Nicola Mazzocca, Sabrina Merola, and Enea Vincenzo Napolitano. Evolving justice sector: An innovative proposal for introducing ai-based techniques in court offices. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 75–88. Springer, 2023.

[6] Flora Amato, Simona Fioretto, Eugenio Forgillo, Elio Masciari, Nicola Mazzocca, Sabrina Merola, and Enea Vincenzo Napolitano. Introducing ai-based techniques in the justice sector: A proposal for digital transformation of court offices. In *SEBD*, pages 497–504, 2023.

[7] Adriano Augusto, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, Andrea Marrella, Massimo Mecella, and Allar Soo.

Automated discovery of process models from event logs: Review and benchmark. *IEEE transactions on knowledge and data engineering*, 31(4):686–705, 2018.

[8] Egidijus Barcevičius, Guoda Cibaitė, Cristiano Codagnone, Vaida Gineikytė, Luka Klimavičiūtė, Giovanni Liva, Loreta Matulevič, Gianluca Misuraca, Irene Vanini, et al. Exploring digital government transformation in the eu. *Publications Office of the European Union, Luxembourg*, 2019.

[9] OV Belyakova et al. Digital transformation of public administration: achievements and problems. *European Proceedings of Social and Behavioural Sciences*, 2021.

[10] Alessandro Berti. Filtering and sampling object-centric event logs. *arXiv preprint arXiv:2205.01428*, 2022.

[11] Alessandro Berti, Johannes Herforth, Mahnaz Sadat Qafari, and Wil MP van der Aalst. Graph-based feature extraction on object-centric event logs. *International Journal of Data Science and Analytics*, pages 1–17, 2023.

[12] Alessandro Berti, Johannes Herforth, Mahnaz Sadat Qafari, and Wil MP van der Aalst. Graph-based feature extraction on object-centric event logs. *International Journal of Data Science and Analytics*, 18(2):139–155, 2024.

[13] Alessandro Berti and Wil MP van der Aalst. Oc-pm: analyzing object-centric event logs and process models. *International Journal on Software Tools for Technology Transfer*, 25(1):1–17, 2023.

[14] Alessandro Berti, Sebastiaan van Zelst, and Daniel Schuster. Pm4py: A process mining library for python. *Software Impacts*, 17:100556, 2023.

[15] RP Jagadeesh Chandra Bose and Wil MP van der Aalst. Trace clustering based on conserved patterns: Towards achieving better process models. In *Business Process Management Workshops: BPM 2009 International Workshops, Ulm, Germany, September 7, 2009. Revised Papers 7*, pages 170–181. Springer, 2010.

[16] Alexandros Bousdekis, Athanasios Kerasiotis, Silvester Kotsias, Georgia Theodoropoulou, Georgios Miaoulis, and Djamchid Ghazanfarpour. Modelling and predictive monitoring of business processes under uncertainty with reinforcement learning. *Sensors*, 23(15):6931, 2023.

[17] Zaharah A Bukhsh, Aaqib Saeed, and Remco M Dijkman. Processtransformer: Predictive business process monitoring with transformer network. *arXiv preprint arXiv:2104.00721*, 2021.

[18] Andrea Burattin. Process mining techniques in business environments. *Lecture Notes in Business Information Processing*, 207(1):220, 2015.

[19] Manuel Camargo, Marlon Dumas, and Oscar González-Rojas. Learning accurate lstm models of business processes. In *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*, pages 286–302. Springer, 2019.

[20] Josep Carmona, Boudewijn van Dongen, Andreas Solti, and Matthias Weidlich. Conformance checking. *Switzerland: Springer.[Google Scholar]*, 56:12, 2018.

[21] Paolo Ceravolo, Marco Comuzzi, Jochen De Weerdt, Chiara Di Francescomarino, and Fabrizio Maria Maggi. Predictive process monitoring: concepts, challenges, and future research directions. *Process Science*, 1(1):1–22, 2024.

[22] David Chapela-Campa, Marlon Dumas, Manuel Mucientes, and Manuel Lama. Efficient edge filtering of directly-follows graphs for process mining. *Information Sciences*, 610:830–846, 2022.

[23] Shao Chifeng and Wang Qianqian. Object-centered petri net process prediction: A case study of multi-system intelligent healthcare. *IEEE Access*, 12:170661–170676, 2024.

[24] Andrew Chilombo, Janet A Fisher, and Dan van Der Horst. A conceptual framework for improving the understanding of large scale land acquisitions. *Land Use Policy*, 88:104184, 2019.

[25] Andrea Chiorrini, Claudia Diamantini, Laura Genga, and Domenico Potena. Multi-perspective enriched instance graphs for next activity prediction through graph neural network. *Journal of Intelligent Information Systems*, pages 1–21, 2023.

[26] Antonio Cordella, Francesco Gualdi, and Mindel van de Laar. Digital skills within the public sector: A missing link to achieve the sustainable development goals (sdgs). *Information Polity*, 29(1):13–33, 2024.

[27] Tímea Czvetkó, Alex Kummer, Tamás Ruppert, and János Abonyi. Data-driven business process management-based development of industry 4.0 solutions. *CIRP journal of manufacturing science and technology*, 36:117–132, 2022.

[28] Johannes De Smedt, Anton Yeshchenko, Artem Polyvyanyy, Jochen De Weerdt, and Jan Mendling. Process model forecasting and change exploration using time series analysis of event sequence data. *Data & Knowledge Engineering*, 145:102145, 2023.

[29] Jochen De Weerdt, Seppe vanden Broucke, Jan Vanthienen, and Bart Baesens. Active trace clustering for improved process discovery. *IEEE Transactions on Knowledge and Data Engineering*, 25(12):2708–2720, 2013.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[31] Chiara Di Francescomarino, Marlon Dumas, Fabrizio Maria Maggi, and Irene Teinemaa. Clustering-based predictive process monitoring. *IEEE transactions on services computing*, 12(6):896–909, 2016.

[32] Chiara Di Francescomarino and Chiara Ghidini. *Predictive Process Monitoring*, pages 320–346. Springer International Publishing, Cham, 2022.

[33] Chiara Di Francescomarino, Chiara Ghidini, Fabrizio Maria Maggi, and Fredrik Milani. Predictive process monitoring methods: Which one suits me best? In *International conference on business process management*, pages 462–479. Springer, 2018.

[34] Beniamino Di Martino, Antonio Esposito, and Luigi Colucci Cante. Multi agents simulation of justice trials to support control management and reduction of civil trials duration. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13, 2021.

[35] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A Reijers. Introduction to business process management. In *Fundamentals of business process management*, pages 1–33. Springer, 2018.

[36] Marlon Dumas, Marcello La Rosa, Jan Mendling, Hajo A Reijers, et al. *Fundamentals of business process management*, volume 2. Springer, 2018.

[37] Gianluca Elia, Gianluca Solazzo, Antonio Lerro, Federico Pigni, and Christopher L Tucci. The digital transformation canvas: A conceptual framework for leading the digital transformation process. *Business Horizons*, 67(4):381–398, 2024.

[38] European Commission: Directorate-General for Communications Networks, Content and Technology, Capgemini, IDC, Politecnico di Milano, and Sogeti. egovernment benchmark 2025 – building online, user-friendly, and interoperable services – background report. Technical report, Publications Office of the European Union, Luxembourg, 2025. Written by Capgemini, IDC, Politecnico di Milano, and Sogeti for the European Commission Directorate-General for Communications Networks, Content and Technology.

[39] Stephan A Fahrenkrog-Petersen, Niek Tax, Irene Teinemaa, Marlon Dumas, Massimiliano de Leoni, Fabrizio Maria Maggi, and Matthias Weidlich. Fire now, fire later: alarm-based systems for prescriptive process monitoring. *Knowledge and Information Systems*, 64(2):559–587, 2022.

[40] Muhammad Faizan, Megat F Zuhairi, and Shahrinaz Ismail. Process discovery enhancement with trace clustering and profiling. *Annals of Emerging Technologies in Computing (AETiC)*, 5(4):1–13, 2021.

[41] Mohammadreza Fani Sani, Mozhgan Vazifehdoostirani, Gyunam Park, Marco Pegoraro, Sebastiaan J van Zelst, and Wil MP van der Aalst. Performance-preserving event log sampling for predictive monitoring. *Journal of Intelligent Information Systems*, 61(1):53–82, 2023.

[42] Diogo R. Ferreira. *Control-Flow Perspective*, pages 15–29. Springer International Publishing, Cham, 2020.

[43] Marcus Fischer, Florian Imgrund, Christian Janiesch, and Axel Winkelmann. Strategy archetypes for digital transformation: Defining meta objectives using business process management. *Information & management*, 57(5):103262, 2020.

[44] Land Portal Foundation. Land Matrix | Land Portal — landportal.info. https://landportal.info/organization/land-matrix. [Accessed 05-09-2024].

[45] Chiara Di Francescomarino, Marlon Dumas, Fabrizio Maria Maggi, and Irene Teinemaa. Clustering-based predictive process monitoring. *IEEE Transactions on Services Computing*, 12(6):896–909, 2019.

[46] Riccardo Galanti and Massimiliano de Leoni. Predictive analytics for object-centric processes: Do graph neural networks really help? In *International Conference on Business Process Management*, pages 521–533. Springer, 2023.

[47] Riccardo Galanti, Massimiliano de Leoni, Merylin Monaro, Nicolò Navarin, Alan Marazzi, Brigida Di Stasi, and Stéphanie Maldera. An explainable decision support system for predictive process analytics. *Engineering Applications of Artificial Intelligence*, 120:105904, 2023.

[48] Riccardo Galanti, Massimiliano De Leoni, Nicolò Navarin, and Alan Marazzi. Object-centric process predictive analytics. *Expert Systems with Applications*, 213:119173, 2023.

[49] Victor Galaz. Pandemic 2.0: can information technology help save the planet? *Environment: Science and Policy for Sustainable Development*, 51(6):20–28, 2009.

[50] Wissam Gherissi, Joyce El Haddad, and Daniela Grigori. Object-centric predictive process monitoring. In *International Conference on Service-Oriented Computing*, pages 27–39. Springer, 2022.

[51] Massimiliano Giacalone, Eugenia Nissi, and Carlo Cusatelli. Dynamic efficiency evaluation of italian judicial system using dea based malmquist productivity indexes. *Socio-Economic Planning Sciences*, 72:100952, 2020.

[52] Björn Rafn Gunnarsson, Seppe vanden Broucke, and Jochen De Weerdt. Ls-ice: A load state intercase encoding framework for improved predictive monitoring of business processes. *Information Systems*, 125:102432, 2024.

[53] Björn Rafn Gunnarsson, Seppe vanden Broucke, and Jochen De Weerdt. A direct data aware lstm neural network architecture for complete remaining trace and runtime prediction. *IEEE Transactions on Services Computing*, 16(4):2330–2342, 2023.

[54] Deepak Gupta and Rinkle Rani. A study of big data evolution and research challenges. *Journal of information science*, 45(3):322–340, 2019.

[55] Nitin Harane and Sheetal Rathi. *Comprehensive Survey on Deep Learning Approaches in Predictive Business Process Monitoring*, pages 115–128. Springer International Publishing, Cham, 2020.

[56] Mohammad Imran, Maizatul Akmar Ismail, Suraya Hamid, and Mohammad Hairul Nizam Md Nasir. Complex process modeling in process mining: A systematic review. *IEEE Access*, 10:101515–101536, 2022.

[57] R. Interdonato, J. Bourgoin, Q. Grislain, and A. Tagarelli. The new abnormal: Identifying and ranking anomalies in the land trade market. *PLOS ONE*, 2022 (to appear on December 1st, 2022).

[58] R. Interdonato, J. Bourgoin, Q. Grislain, M. Zignani, S. Gaito, and M. Giger. The parable of arable land: Characterizing large scale land acquisitions through network analysis. *PLOS ONE*, 15(10):1–31, 10 2020.

[59] Savvas Kaftantzis, Alexandros Bousdekis, Georgia Theodoropoulou, and Georgios Miaoulis. Predictive business process monitoring with automl for next activity prediction. *Intelligent Decision Technologies*, 18(3):1965–1980, 2024.

[60] Seunghyun Kim, Byungchul Choi, and Yong Kyu Lew. Where is the age of digitalization heading? the meaning, characteristics, and implications of contemporary digital transformation. *Sustainability*, 13(16):8909, 2021.

[61] Sungkyu Kim, Marco Comuzzi, and Chiara Di Francescomarino. Understanding the impact of design choices on the performance of predictive process monitoring. In *International Conference on Process Mining*, pages 153–164. Springer, 2023.

[62] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[63] Benedikt Knopp, Mahsa Pourbafrani, and Wil M. P. van der Aalst. Discovering object-centric process simulation models. In *2023 5th International Conference on Process Mining (ICPM)*, pages 81–88, 2023.

[64] Silvester Kotsias, Athanasios Kerasiotis, Alexandros Bousdekis, Georgia Theodoropoulou, and Georgios Miaoulis. Predictive and prescriptive business process monitoring with reinforcement learning. In *Novel & Intelligent Digital Systems Conferences*, pages 245–254. Springer, 2022.

[65] Wolfgang Kratsch, Jonas Manderscheid, Maximilian Röglinger, and Johannes Seyfried. Machine learning in business process monitoring: a comparison of deep learning and classical approaches used for outcome prediction. *Business & Information Systems Engineering*, 63:261–276, 2021.

[66] Sascha Kraus, Paul Jones, Norbert Kailer, Alexandra Weinmann, Nuria Chaparro-Banegas, and Norat Roig-Tierno. Digital transformation: An overview of the current state of the art of research. *Sage Open*, 11(3):21582440211047576, 2021.

[67] Yulia Kumar, Jose Marchena, Ardalan H Awlla, J Jenny Li, and Hemn Barzan Abdalla. The ai-powered evolution of big data. *Applied Sciences*, 14(22):10176, 2024.

[68] Ngoc-Diem Le. Concept-drift-aware prescriptive analytics for object-centric processes. 2024.

[69] Volodymyr Leno, Abel Armas-Cervantes, Marlon Dumas, Marcello La Rosa, and Fabrizio M Maggi. Discovering process maps from event streams. In *Proceedings of the 2018 international conference on software and system process*, pages 86–95, 2018.

[70] Anna Leontjeva, Raffaele Conforti, Chiara Di Francescomarino, Marlon Dumas, and Fabrizio Maria Maggi. Complex symbolic sequence encodings for predictive monitoring of business processes. In *Business Process Management: 13th International Conference, BPM 2015, Innsbruck, Austria, August 31–September 3, 2015, Proceedings 13*, pages 297–313. Springer, 2015.

[71] Xixi Lu, Seyed Amin Tabatabaei, Mark Hoogendoorn, and Hajo A Reijers. Trace clustering on very large event data in healthcare using frequent sequence patterns. In *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*, pages 198–215. Springer, 2019.

[72] Fabrizio Maria Maggi, Chiara Di Francescomarino, Marlon Dumas, and Chiara Ghidini. Predictive monitoring of business processes. In *Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings 26*, pages 457–472. Springer, 2014.

[73] Ronny S Mans, MH Schonenberg, Minseok Song, Wil MP van der Aalst, and Piet JM Bakker. Application of process mining in healthcare–a case study in a dutch hospital. In *Biomedical Engineering Systems and Technologies: International Joint Conference, BIOSTEC 2008 Funchal, Madeira, Portugal, January 28-31, 2008 Revised Selected Papers 1*, pages 425–438. Springer, 2009.

[74] Alfonso E Márquez-Chamorro, Manuel Resinas, Antonio Ruiz-Cortés, and Miguel Toro. Run-time prediction of business process indicators using evolutionary decision rules. *Expert Systems with Applications*, 87:1–14, 2017.

[75] Alfonso E Márquez-Chamorro, Kate Revoredo, Manuel Resinas, Adela Del-Rio-Ortega, Flávia M Santoro, and Antonio Ruiz-Cortes. Context-aware process performance indicator prediction. *IEEE Access*, 8:222050–222063, 2020.

[76] Alfonso Eduardo Márquez-Chamorro, Manuel Resinas, and Antonio Ruiz-Cortés. Predictive monitoring of business processes: a survey. *IEEE Transactions on Services Computing*, 11(6):962–977, 2017.

[77] Altaaf Mechiche-alami, Carlo Piccardi, Kimberly A Nicholas, and Jonathan W Seaquist. Transnational land acquisitions beyond the food and financial crises. *Environmental Research Letters*, 14(8), 2019.

[78] Nijat Mehdiyev, Joerg Evermann, and Peter Fettke. A novel business process prediction model using a deep learning method. *Business & information systems engineering*, 62:143–157, 2020.

[79] Peter Messerli, Markus Giger, Michael B Dwyer, Thomas Breu, and Sandra Eckert. The geography of large-scale land acquisitions: Analysing socio-ecological patterns of target contexts in the global south. *Applied Geography*, 53:449–459, 2014.

[80] Flavia Michelotto and Luiz Antonio Joia. Organizational digital transformation readiness: An exploratory investigation. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4):3283–3304, 2024.

[81] Gianluca Misuraca and Colin Van Noordt. Ai watch-artificial intelligence in public services: Overview of the use and impact of ai in public services in the eu. *JRC Research Reports*, (JRC120399), 2020.

[82] Flavia Monti, Jerin George Mathew, Francesco Leotta, Agnes Koschmider, and Massimo Mecella. On the application of process management and process mining to industry 4.0. *Software and Systems Modeling*, 23(6):1407–1419, 2024.

[83] Birgit Moser-Plautz and Lisa Schmidthuber. Digital government transformation as an organizational response to the covid-19 pandemic. *Government Information Quarterly*, 40(3):101815, 2023.

[84] Jorge Munoz-Gama, Niels Martin, Carlos Fernandez-Llatas, Owen A Johnson, Marcos Sepúlveda, Emmanuel Helm, Victor Galvez-Yanjari, Eric Rojas, Antonio Martinez-Millana, Davide Aloini, et al. Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*, 127:103994, 2022.

[85] Lakshmi Narayana N and Jagadishwari V. Trace clustering techniques for process mining. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–6, 2023.

[86] Dominic A Neu, Johannes Lahann, and Peter Fettke. A systematic literature review on state-of-the-art deep learning methods for process prediction. *Artificial Intelligence Review*, pages 1–27, 2022.

[87] Alessandro Niro and Michael Werner. Detecting anomalous events in object-centric business processes via graph neural networks. In *International Conference on Process Mining*, pages 179–190. Springer, 2023.

[88] Felix Oberdorf, Myriam Schaschek, Sven Weinzierl, Nikolai Stein, Martin Matzner, and Christoph M Flath. Predictive end-to-end enterprise process network monitoring. *Business & Information Systems Engineering*, 65(1):49–64, 2023.

[89] Daniel E O'Leary. Digitization, digitalization, and digital transformation in accounting, electronic commerce, and supply chains, 2023.

[90] Gyunam Park and Minseok Song. Prediction-based resource allocation using lstm and minimum cost and maximum flow algorithm. In *2019 international conference on process mining (ICPM)*, pages 121–128. IEEE, 2019.

[91] Gyunam Park and Minseok Song. Predicting performances in business processes using deep neural networks. *Decision Support Systems*, 129:113191, 2020.

[92] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. Using convolutional neural networks for predictive process analytics. In *2019 international conference on process mining (ICPM)*, pages 129–136. IEEE, 2019.

[93] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. Predictive process mining meets computer vision. In *Business Process Management Forum: BPM Forum 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, pages 176–192. Springer, 2020.

[94] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. A multi-view deep learning approach for predictive business process monitoring. *IEEE Transactions on Services Computing*, 15(4):2382–2395, 2021.

[95] Justin Paul, Akiko Ueno, Charles Dennis, Eleftherios Alamanos, Lucill Curtis, Pantea Foroudi, Agnieszka Kacprzak, Werner H Kunz, Jonathan Liu, Reza Marvi, et al. Digital transformation: A multidisciplinary perspective and future research agenda. *International Journal of Consumer Studies*, 48(2):e13015, 2024.

[96] Jari Peeperkorn, Seppe vanden Broucke, and Jochen De Weerdt. Can recurrent neural networks learn process model structure? *Journal of Intelligent Information Systems*, 61(1):27–51, 2023.

[97] Mirko Polato, Alessandro Sperduti, Andrea Burattin, and Massimiliano de Leoni. Time and activity sequence prediction of business process instances. *Computing*, 100:1005–1031, 2018.

[98] Land Portal. Land matrix data. https://landmatrix.org/list/deals/, 2024. [Accessed 01-07-2024].

[99] Efrén Rama-Maneiro, Juan Vidal, and Manuel Lama. Deep learning for predictive business process monitoring: Review and benchmark. *IEEE Transactions on Services Computing*, 2021.

[100] Efrén Rama-Maneiro, Juan C Vidal, and Manuel Lama. Embedding graph convolutional networks in recurrent neural networks for predictive monitoring. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[101] Efrén Rama-Maneiro, Juan C Vidal, Manuel Lama, and Pablo Monteagudo-Lago. Exploiting recurrent graph neural networks for suffix prediction in predictive monitoring. *Computing*, 106(9):3085–3111, 2024.

[102] Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8):8219–8264, 2023.

[103] Martha Razo and Houshang Darabi. Adjacency matrix deep learning prediction model for prognosis of the next event in a process. *IEEE Access*, 11:11947–11955, 2023.

[104] Timo Rohrer, Anahita Farhang Ghahfarokhi, Mohamed Behery, Gerhard Lakemeyer, and Wil MP van der Aalst. Predictive object-centric process monitoring. *arXiv preprint arXiv:2207.10017*, 2022.

[105] Giulia Ruffini. Object-centric process mining for process analysis and operational support. In Jan Martijn E. M. van der Werf, Cristina Cabanillas, Francesco Leotta, and Laura Genga, editors, *Doctoral Consortium and Demo Track 2023 at the International Conference on Process Mining 2023 co-located with the 5th International Conference on Process Mining (ICPM 2023), Rome, Italy, October 27, 2023*, volume 3648 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[106] Colin Rule. Online dispute resolution and the future of justice. *Annual Review of Law and Social Science*, 16:277–292, 2020.

[107] J W Seaquist, Emma Li Johansson, and Kimberly A Nicholas. Architecture of the global land acquisition system: applying the tools of network science to identify key vulnerabilities. *Environmental Research Letters*, 9(11):114006, nov 2014.

[108] Tim Smit. How object-centric is object-centric predictive process monitoring? introducing objects into object-centric predictive process monitoring. Master's thesis, 2023.

[109] Tim K Smit, Hajo A Reijers, and Xixi Lu. Hoeg: A new approach for object-centric predictive process monitoring. In *International Conference on Advanced Information Systems Engineering*, pages 231–247. Springer, 2024.

[110] Minseok Song, Christian W Günther, and Wil MP Van der Aalst. Trace clustering in process mining. In *Business Process Management Workshops: BPM 2008 International Workshops, Milano, Italy, September 1-4, 2008. Revised Papers 6*, pages 109–120. Springer, 2009.

[111] Minseok Song, H Yang, Seyed Hossein Siadat, and Mykola Pechenizkiy. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. *Expert Systems with Applications*, 40(9):3722–3737, 2013.

[112] Hadi Sotudeh. *"Predicting undesired business process behavior using supervised machine learning.* PhD thesis, MA thesis. KTH, School of Electrical Engineering and Computer Science (EECS), 2018.

[113] Matthias Stierle, Jens Brunk, Sven Weinzierl, Sandra Zilker, Martin Matzner, and Jörg Becker. Bringing light into the darkness - a systematic literature review on explainable predictive business process monitoring techniques. In *European Conference on Information Systems*, 2021.

[114] Ana-Marija Stjepić, Lucija Ivančić, and Dalia Suša Vugec. Mastering digital transformation through business process management: Investigating alignments, goals, orchestration, and roles. *Journal of entrepreneurship, management and innovation*, 16(1):41–74, 2020.

[115] Bayu Adhi Tama and Marco Comuzzi. An empirical comparison of classification techniques for next event prediction using business process event logs. *Expert Systems with Applications*, 129:233–245, 2019.

[116] Luca Tangi, Marijn Janssen, Michele Benedetti, and Giuliano Noci. Barriers and drivers of digital transformation in public organizations: Results from a survey in the netherlands. In *International Conference on Electronic Government*, pages 42–56. Springer, 2020.

[117] Gabriel M Tavares, Rafael S Oyamada, Sylvio Barbon Junior, and Paolo Ceravolo. Trace encoding in process mining: A survey and benchmarking. *Engineering Applications of Artificial Intelligence*, 126:107028, 2023.

[118] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with lstm neural networks. In *Advanced Information Systems Engineering: 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Proceedings 29*, pages 477–492. Springer, 2017.

[119] Farbod Taymouri, Marcello La Rosa, Sarah Erfani, Zahra Dasht Bozorgi, and Ilya Verenich. Predictive business process monitoring via generative adversarial nets: the case of next event prediction. In *Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, pages 237–256. Springer, 2020.

[120] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–57, 2019.

[121] Julian Theis and Houshang Darabi. Improving predictive process monitoring through reachability graph-based masking of neural networks. *IEEE Transactions on Computational Social Systems*, 10(4):1927–1938, 2023.

[122] Bui Quang Truong, Anh Nguyen-Duc, and Nguyen Thi Cam Van. A quantitative review of the research on business process management in digital transformation: a bibliometric approach. *Software*, 2(3):377–399, 2023.

[123] Wil Van Der Aalst. Process mining: Overview and opportunities. *ACM Transactions on Management Information Systems (TMIS)*, 3(2):1–17, 2012.

[124] Wil van der Aalst. *Process Mining: The Missing Link*, pages 25–52. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[125] Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

[126] Wil M. P. van der Aalst. *Foundations of Process Discovery*, pages 37–75. Springer International Publishing, Cham, 2022.

[127] Wil MP van der Aalst. Object-centric process mining: Dealing with divergence and convergence in event data. In *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*, pages 3–25. Springer, 2019.

[128] Wil MP Van Der Aalst. A practitioner's guide to process mining: Limitations of the directly-follows graph, 2019.

[129] Wil MP van der Aalst. Object-centric process mining: An introduction. In *International Colloquium on Theoretical Aspects of Computing*, pages 73–105. Springer, 2021.

[130] Wil MP van der Aalst. Object-centric process mining: unraveling the fabric of real processes. *Mathematics*, 11(12):2691, 2023.

[131] Wil MP van der Aalst and Josep Carmona. *Process mining handbook*. Springer Nature, 2022.

[132] Michiel van Dijk, Tom Morley, Marie Luise Rau, and Yashar Saghai. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nature Food*, 2(7):494–501, July 2021.

[133] Maikel L. van Eck, Xixi Lu, Sander J. J. Leemans, and Wil M. P. van der Aalst. PM^2: A process mining project methodology. In Jelena Zdravkovic, Marite Kirikova, and Paul Johannesson, editors, *Advanced Information Systems Engineering*, pages 297–313, Cham, 2015. Springer International Publishing.

[134] Ilya Verenich, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Irene Teinemaa. Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4):1–34, 2019.

[135] Peter C Verhoef, Thijs Broekhuizen, Yakov Bart, Abhi Bhattacharya, John Qi Dong, Nicolai Fabian, and Michael Haenlein. Digital transformation: A multidisciplinary reflection and research agenda. *Journal of business research*, 122:889–901, 2021.

[136] Mark Von Rosing, Stephen White, Fred Cummins, and Henk De Man. Business process model and notation-bpmn., 2015.

[137] Johannes Vrana and Ripudaman Singh. Digitization, digitalization, and digital transformation. In *Handbook of nondestructive evaluation 4.0*, pages 107–123. Springer, 2022.

[138] Jiaojiao Wang, Dongjin Yu, Chengfei Liu, and Xiaoxiao Sun. Outcome-oriented predictive process monitoring with attention-based bidirectional lstm neural networks. In *2019 IEEE International Conference on Web Services (ICWS)*, pages 360–367. IEEE, 2019.

[139] Mathias Weske. Business process management architectures. In *Business process management: Concepts, languages, architectures*, pages 333–371. Springer, 2012.

[140] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Philip S Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14:1–37, 2008.

[141] Vasanth Yarlagadda, Abishi Chowdhury, Amrit Pal, Shruti Mishra, Sandeep Kumar Satapathy, Sung Bae Cho, Sachi Nandan Mohanty, and Ashit Kumar Dutta. Predictive monitoring in process mining using deep learning for better consumer service. *IEEE Transactions on Consumer Electronics*, 2024.

[142] Fareed Zandkarimi, Jana-Rebecca Rehse, Pouya Soudmand, and Hartmut Hoehle. A generic framework for trace clustering in process mining. In *2020 2nd International Conference on Process Mining (ICPM)*, pages 177–184, 2020.

Author's publications

1. A. Anniciello, S. Fioretto, E. Masciari, and E.V. Napolitano. Covid-19 impact on health information technology: the rapid rise of e-Health and Big Data driven innovation of healthcare processes, Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2759–2764, 2022. DOI: 10.1109/BIBM55620.2022.9995515.

2. F. Amato, S. Fioretto, E. Forgillo, E. Masciari, N. Mazzocca, S. Merola, and E.V. Napolitano. Introducing AI-Based Techniques in the Justice Sector: A Proposal for Digital Transformation of Court Offices, Proceedings of the SEBD 2023 Conference, 497–504, 2023

3. S. Fioretto. Business Process Extraction from Documents with AI, Proceedings of the SEBD 2023 Conference, 697–702, 2023.

4. S. Fioretto, E. Masciari, and E.V. Napolitano. Can the Study of Trajectories Help to Extract Information from Business Processes?, Proceedings of the PMAI Workshop at IJCAI 2023, 13–24, 2023.

5. A. Anniciello, S. Fioretto, E. Masciari, and E.V. Napolitano. Digital Twins for Traffic Congestion in Smart Cities: a novel solution using Data Mining techniques, Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023), vol. 3, pp. 241–248, 2023. DOI: 10.5220/0012208100003598.

6. A. Anniciello, S. Fioretto, E. Masciari, E.V. Napolitano. How Pandemic Affected the Adoption of e-Health Systems, IDEAS '23: Proceedings of the 27th International Database Engineered Applications Symposium, pp. 94-98, DOI: 10.1145/3589462.3589496

7. S. Fioretto, E. Masciari, and E.V. Napolitano. Dossier classification to support workflow management optimization, Proceedings of the 4th Inter-

national Conference and Summer School "Numerical Computations: Theory and Algorithms" (NUMTA 2023), Lecture Notes in Computer Science, vol. 14478, pp. 286–292, 2023. DOI: 10.1007/978-3-031-81247-7_25.

8. F. Amato, S. Fioretto, E. Forgillo, E. Masciari, N. Mazzocca, S. Merola, E. V. Napolitano. Evolving Justice Sector: An Innovative Proposal for Introducing AI-Based Techniques in Court Offices, 12th International Conference on Electronic Government and the Information Systems Perspective (EGOVIS 2023), pp. 75-88, DOI: 10.1007/978-3-031-39841-4_6

9. S. Fioretto. Process Mining Solutions for Public Administration, Proceedings of the European Conference on Advances in Databases and Information Systems (ADBIS 2023), in New Trends in Database and Information Systems, Lecture Notes in Computer Science, pp. 668–675, 2023. DOI: 10.1007/978-3-031-42941-5_60.

10. S. Fioretto, E. Masciari, E.V. Napolitano. A Joint Analysis of Trajectory Mining and Process Mining for Smartphone User Behaviour, Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023, vol. CCIS, 2135, Springer, Cham, DOI: 10.1007/978-3-031-74633-8_43

11. S. Fioretto, E. Masciari, E.V. Napolitano Machine Learning for KPI Development in Public Administration, Proceedings of the 13th International Conference on Data Science, Technology and Applications, DATA 2024, pp. 522-527, SciTePress, DOI: 10.5220/0012820300003756

12. S. Fioretto, D. Ienco, R. Interdonato, E. Masciari. Integrating predictive process monitoring techniques in smart agriculture, 2024 International Symposium on Methodologies for Intelligent Systems (ISMIS 2024), LNAI,vol. 14670 pp. 306-313, Springer, Cham, DOI: 10.1007/978-3-031-62700-2_27

13. S. Fioretto, E. Masciari. A Conceptual Framework for Predictive Process Monitoring in Public Administration, 2024 International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2024), LNDECT,vol. 87, pp. 423-431, Springer, Cham, DOI: 10.1007/978-3-031-70011-8_40

14. M.L. Conza, S. Fioretto, E. Masciari, E.V. Napolitano Enhancing Employee Health Through an Experimental Diet: Insights from Machine Learning Analysis, 33rd Euromicro international Conference on Parallel, Distributed, and Network-Based Processing (PDP), pp. 283-290, IEEE, DOI: 10.1109/PDP66500.2025.00046

15. S. Fioretto, E. Masciari. A comparative analysis of predictive process monitoring: object-centric versus classical event logs, International Journal of Knowledge and Information Systems, vol. 67, pp. 7355-7398, 2025, DOI: 10.1007/s10115-025-02461-y.