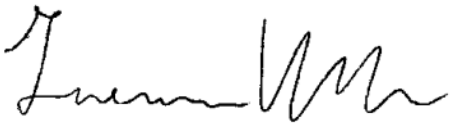
PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Francesco Vitale

Cycle: XXXVII

Training and Research Activities Report

Academic year: 2022-2023 - PhD Year: Second



Tutor: Prof. Nicola Mazzocca

Co-Tutor: Dr. Federico Papa

Date: October 23, 2023

1. Information:

- **PhD student: Francesco Vitale** **PhD Cycle: XXXVII**
- **DR number: DR995865**
- **Date of birth: 30/06/1997**
- **Master Science degree: Computer Engineering** **University: Università degli Studi di Napoli Federico II**
- **Scholarship type: Funding company**
- **Tutor: Prof. Nicola Mazzocca**
- **Co-tutor: Dr. Federico Papa**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
IoT Data Analysis	Course	12	4	09/01/2023 – 27/01/2023	Prof. Raffaele Della Corte	Y
Using Deep Learning properly	Course	10	4	10/01/2023 – 24/01/2023	Dr. Andrea Apicella	Y
Cambridge Advanced C1	Course	60	6	27/01/2023 – 23/05/2023	Ms. Janet Parker	Y
Cybercrime and Information Warfare: National and International Actors	Seminar	2	0.4	18/11/2022	Proff. Roberto Natella and Simon Pietro Romano	Y
Privacy and Data Protection	Seminar	2	0.4	22/11/2022	Proff. Roberto Natella and Simon Pietro Romano	Y
POWL: Partially Ordered Workflow Language	Seminar	2	0.4	16/05/2023	Alessandro Berti, M.Sc.	Y
WorkflowProcess discovery supporting a	Seminar	2	0.4	23/05/2023	Alessandro Berti, M.Sc.	Y

desirable event log while avoiding an undesirable event log						
Insights by Comparison	**Seminar**	**2**	**0.4**	**30/05/2023**	**Alessandro Berti, M.Sc.**	**Y**
Conformance Checking & Incremental Process Discovery Using Trace Fragments	**Seminar**	**2**	**0.4**	**06/06/2023**	**Alessandro Berti, M.Sc.**	**Y**
Held a lecture for the "Architetture dei Sistemi Digitali" course	**Tutorship**	**2**	**0.4**	**14/12/2022**	**Prof. Nicola Mazzocca**	**Y**
Study on the state-of-the-art on process mining for anomaly detection	**Research**	**30**	**6**	**-**	**-**	**N**
Study on the state-of-the-art on cyber-physical anomaly detection	**Research**	**20**	**4**	**-**	**-**	**N**
Study on the state-of-the-art on modeling and simulating cyber-physical systems	**Research**	**20**	**4**	**-**	**-**	**N**
Application of process mining for anomaly detection in railways from Hitachi case-studies	**Research**	**25**	**5**	**-**	**-**	**N**
Preparation of the paper "Evaluating Virtualization for Fog Monitoring of Real-time Applications in Mixed-Criticality Systems"	**Research**	**20**	**4**	**-**	**-**	**N**
Preparation of the paper "A Process Mining-based Unsupervised Anomaly Detection Technique for the	**Research**	**20**	**4**	**-**	**-**	**N**

Industrial Internet of Things"						
Publication of the paper "Digital Twins for Anomaly Detection in the Industrial Internet of Things: Conceptual Architecture and Proof-of-Concept"	**Research**	**20**	**4**	**22/02/2023**	**-**	**N**
Preparation of the paper "A Comparison Framework for Control-Flow Anomaly Detection in Event Logs of Information Systems"	**Research**	**20**	**4**	**-**	**-**	**N**
Preparation of the paper "A Two-Level Fusion Framework for Cyber-Physical Anomaly Detection"	**Research**	**20**	**4**	**-**	**-**	**N**
Preparation of the paper "Combining Process Mining and Unsupervised Machine Learning for Monitoring in Resilient Computer Systems"	**Research**	**20**	**4**	**-**	**-**	**N**
Held a seminar entitled "Combining Alignments and Dimensionality Reduction for Anomaly Detection in Event Logs of Information Systems"	**Research**	**2.5**	**0.5**	**20/06/2023**	**RWTH Aachen University**	**N**
Held a seminar entitled "Managing Complexity in	**Research**	**2.5**	**0.5**	**26/09/2023**	**Università Campus Bio-**	**N**

Modern Computer Systems with Process Mining"					**Medico di Roma**	

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**0.8**	**12**	**0.4**	**13.2**
Bimonth 2	**4**	**0**	**8**	**0**	**12**
Bimonth 3	**0**	**0**	**6**	**0**	**6**
Bimonth 4	**6**	**1.6**	**6**	**0**	**13.6**
Bimonth 5	**0**	**1.2**	**6**	**0**	**7.2**
Bimonth 6	**4**	**0**	**6**	**0**	**10**
Total (2nd year)	**14**	**3.6**	**44**	**0.4**	**62**
Total (1st + 2nd year)	**44**	**6.2**	**95**	**0.8**	**146**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

My research activity has delved deeply into the topics that I started during my first PhD year and focused on concretely exploring the opportunities opened by *combining Machine Learning (ML) and Process Mining (PM) for Anomaly Detection (AD) in Cyber-Physical Systems (CPSs).*

3.1. The context:

My research activity frames in the context of addressing the key concern of dependability issues in modern CPSs due to their architectural complexity, involving software, business processes, networking, and physical processes. CPSs generate huge amounts of data from the smart devices deployed within them. This fact allows for AD through data-driven techniques from ML and PM to discover faults in CPSs.

3.2. The state-of-the-art research on AD with ML and PM:

The scientific literature brims with techniques for AD with ML and PM. On one hand, ML is suited for numerical data and provides many paradigms (e.g., clustering, dimensionality reduction, and neural networks) for normative characterization and assessing deviations. On the other hand, PM handles event logs and provides native support for normative characterization through process discovery and assessing deviations by conformance checking.

Clearly, ML and PM have their strengths and weaknesses. The idea is to include ML and PM in an overarching **framework**, depicted in Fig. 1, that can be specialized for the domain in which the CPS is implemented and AD performed. The framework includes pre-processing with feature extraction,

normative characterization, and assessing deviations. All these steps are data-driven and may involve PM, ML, or both, depending on the reference CPS data and domain.

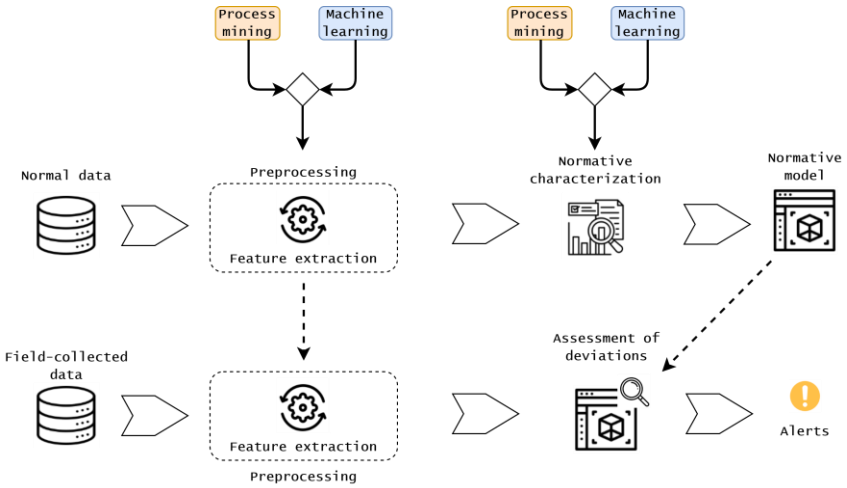


Fig. 1: The overarching framework that integrates PM and ML for AD in CPSs

Overall, the framework requires a thorough formalization that takes into account methodological and domain-related research gaps linked to AD with PM and ML. In the following, these research gaps and the related contributions will be briefly reported.

3.3. Methodological research gaps and contributions

Throughout my research, I have discovered the key methodological research gaps in applying PM for AD:

A. **Preprocessing and converting time series to event logs to apply PM techniques;**
B. **Fair comparison of PM approaches to AD in business processes of information systems;**
C. **Underfitting and overfitting problems related to PM for AD.**

These methodological gaps can be addressed by combining ML with PM. I will refer to this common theme as

- ♠ **The combination of PM with ML for AD.**

Regarding A., I have taken on the Hitachi project (Catenary Inspection System, CIS) that I had been working on throughout my first PhD year. CIS involves AD for train runs on railways by monitoring the behavior of the pantograph, which is an on-board component that connects the train roof with the overhead catenary. The pantograph's behavior is described by time series related to some of its characteristics, such as its height profile throughout the whole train run. Therefore, in order to apply PM, time series require pre-processing so these are converted into event logs.

Not only CIS but also another case study in the Industrial Internet of Things was targeted for AD by PM with time-series pre-processing. This case study comes from a scaled-down replica of a manufacturing plant that colleagues from University of Messina have developed and monitored. Likewise, this case study involves time series whose pre-processing is necessary for PM to be applied.

It is worth noting that, in both case studies, ♠ is involved. Indeed, ML has been integrated into the methods' pipelines for time series pre-processing.

While the collaboration with University of Messina led to the preparation and a couple of review rounds of a paper, "A Process Mining-based Unsupervised Anomaly Detection Technique for the Industrial Internet of Things", the Hitachi project has yet to result in a concrete paper. However, I have developed loads of material in the form of technical reports, presentations, prototypes, and experimental reports.

Regarding B., I have researched on the existing approaches to AD in business processes of information systems, discovering that PM has been proposed and used for AD in this domain. However, there are many other proposals that handle event logs to perform AD and involve ML. The scientific literature is rather unclear about how all techniques compare with each other, thus there is a lack of fair comparison among them.

Throughout my period abroad at the Process and Data Science (PADS) group, RWTH Aachen University, I have delved deeply into this topic and developed two contributions: a new technique for AD in event logs of business processes of information systems that combines alignment-based conformance checking with ML (♠); and a comparison framework to provide fair comparison among techniques for AD in business processes of information systems. These contributions have been developed in collaboration with PADS and resulted in a paper, "A Comparison Framework for Control-Flow Anomaly Detection in Event Logs of Information Systems".

Related to this research is the paper that I have published at IEEE, "Digital Twins for Anomaly Detection in the Industrial Internet of Things: Conceptual Architecture and Proof-of-Concept". Although AD is considered as a digital-twin service, I had the opportunity to explore ♠ in the experimental part of that paper, where I dealt with normal and faulty simulations of RBC/RBC Handover, a scenario from the ERTMS railways case study.

Regarding C., I have noticed that the state-of-the-art research on applying PM for AD in software applications focuses on thresholding the fitness metric, which is a metric evaluated by conformance checking when replaying event logs against normative process models. However, the performance of this approach is impaired by noisy event logs, which in turn lead to underfitting and/or overfitting process models, making the approach based on the fitness metric unreliable.

Thus, I have investigated this issue more in detail and proposed ♠ to enhance control-flow AD in software applications. These contributions resulted in a paper, "Combining Process Mining and Unsupervised Machine Learning for Monitoring in Resilient Computer Systems".

3.4. Domain-related research gaps and contributions

Whereas 3.1. relates to methodological research gaps, this subsection presents domain-related research gaps, shedding light on the domains where PM has been/can be applied for AD. The domains are the following.

I. **State of components in CPSs.**
II. **Control flow of business processes of information systems.**
III. **Control flow of computer systems' software.**

Regarding I., cyber-physical systems involve the monitoring of physical components by means of sensors that inform the cyber world of their state. Often, such monitoring leads to time series. Thus, AD with PM in domain I. requires pre-processing, which links to A. in subsection 3.1.

Regarding II., business processes of information systems involve collecting activities performed throughout the lifecycle of, e.g., manufacturing car pieces in automotive factories or the treatment of COVID-19 patients in hospitals. There is a plethora of approaches for AD in business processes of information systems, together with the lack of concrete proposals for combining PM with ML. These observations link to B. in subsection 3.1.

Regarding III., computer systems' software involve monitoring the software of applications running in, e.g., railways for correct and attentive supervision of train runs by on-board and on-track components or the correct handling of users' interactions with websites when shopping. The application of PM to AD in software links to C. in subsection 3.1.

Thus, my contributions not only improve the methodological aspects in applying PM to AD, mostly by combining its techniques with ML during pre- and post-processing, but also show its practical application in domains I., II., and III.

3.5. Results

The results of my research are categorized by the methodological and domain-related gaps that I have previously mentioned.

In "A Process Mining-based Unsupervised Anomaly Detection Technique for the Industrial Internet of Things", I have shown that pre-processing time series with ML by dimensionality reduction leads to better detection performance with PM (♠). This has been shown by evaluating performance metrics such as Accuracy, Recall, Precision, and F1-Score in the presence and absence of ML pre-processing by dimensionality reduction. Other than providing better performance, ML pre-processing also provides less variability in the results obtained, thus leading to a more reliable solution.

Within the CIS project, I have observed that analyzing PM results with ML enhances AD by allowing fault classification of different anomalies injected in time series related to the pantograph's height and stagger profiles (♠).

In "Digital Twins for Anomaly Detection in the Industrial Internet of Things: Conceptual Architecture and Proof-of-Concept" I showed how ♠ enhances the detection performance with respect to the ML-only counterpart, following the intuition that PM seamlessly handles event logs without the need of numerical feature extraction.

This aspect is explored in much more detail and related to the state-of-the-art on AD for business processes of information systems by the work "A Comparison Framework for Control-Flow Anomaly Detection in Event Logs of Information Systems". Although this work shows that ♠, implemented by my proposal regarding the combination of alignment-based conformance checking and ML, may enhance control-flow AD, the comparison also highlights that results are also influenced by the nature of the datasets and the type of pre-processing applied to event logs.

In "Combining Process Mining and Unsupervised Machine Learning for Monitoring in Resilient Computer Systems" I showed how ♠ enhances the reliability of detection performance when monitoring software control flow in spite of noisy data leading to underfitting and/or overfitting process models.

3.6. Side projects and related research gaps, contributions, and results

Although my main research deals with AD in the Industrial Internet of Things, business processes of information systems, and software, I've also carried out some research in other side projects:

a) **Decision fusion for cyber-physical AD.** The contributions led to the paper "A Two-Level Fusion Framework for Cyber-Physical Anomaly Detection".
b) **Exploiting virtualization techniques in multiprocessor system-on-chips to assure real-time constraints while monitoring the application behavior of critical tasks that enforce control laws in safety-critical applications.** The contributions led to the paper "Evaluating Virtualization for Fog Monitoring of Real-time Applications in Mixed-Criticality Systems".

Regarding a), the main research gap concerns the lack of flexible frameworks for cyber-physical AD that provide explainability of detection results. To address this gap, the referenced paper proposes a framework that allows flexible deployment of individual detectors in sections (i.e., segments) of industrial cyber-physical systems, whose collaborative effort in AD is combined with explainable decision fusion with time-varying dynamic Bayesian networks.The results show that our proposal enhances the detection results compared to other decision fusion techniques that combine decisions from multiple individual detectors (e.g., majority voting), while providing explainable results.

Regarding b), the main research gap concerns the lack of guidance for real-time systems developers in designing, deploying, and evaluating mixed-criticality systems on multi-processor system-on-chips, with reference to the safety-critical ITER case study. To address this gap, the referenced paper proposes a formalism for designing the deployment model of mixed-criticality systems on virtualized multi-processor system-on-chips and a system development process for evaluating whether real-time constrains related to the system specifications are met. The results show that the guidance provided by our formalism and system development process allows for evaluating real-time constrains by inspecting the impact of design choices on real-time metrics (e.g., the impact of the inter-VM communication mean on the latency experienced by the real-time control task).

4. Research products:

Journal papers:

J1. A. De Benedictis, F. Flammini, N. Mazzocca, A. Somma, F. Vitale, "*Digital Twins for Anomaly Detection in the Industrial Internet of Things: Conceptual Architecture and Proof-of-Concept*," IEEE Transactions on Industrial Informatics, vol. 19, no. 12, pp. 11553-11563, 2023, https://doi.org/10.1109/TII.2023.3246983.

J2. M. Cinque, L. De Simone, N. Mazzocca, D. Ottaviano e F. Vitale, "*Evaluating Virtualization for Fog Monitoring of Real-time Applications in Mixed-Criticality Systems*," Real-Time Systems, 2023. (**accepted for publication**).

J3. F. Vitale, F. De Vita, D. Bruneo e N. Mazzocca, "*A Process Mining-based Unsupervised Anomaly Detection Technique for the Industrial Internet of Things*." (submitted to Internet of Things (Netherlands), currently **under minor review**, second round).

J4. S. Guarino, F. Vitale, F. Flammini, L. Faramondi, N. Mazzocca e R. Setola, "A Two-Level Fusion Framework for Cyber-Physical Anomaly Detection." (submitted to IEEE Transactions on Industrial Cyber-Physical Systems, currently **under major review**, second round).

J5. F. Vitale, M. Pegoraro, W. M. P. Van der Aalst e N. Mazzocca, "A Comparison Framework for Control-Flow Anomaly Detection in Event Logs of Information Systems." (submitted to IEEE Transactions on Knowledge and Data Engineering, currently **waiting for first decision**).

J6. F. Vitale, F. Flammini, M. Caporuscio e N. Mazzocca, "Combining Process Mining and Machine Learning for Monitoring Resilient Computer Systems." (currently **in progress**)

Prototypes:

P1. Anomaly detection module for the Hitachi CIS project for diagnosing the train pantograph's behavior throughout its journey

5. Conferences and seminars attended

Seminars:

- *Combining Alignments and Dimensionality Reduction for Anomaly Detection in Event Logs of Information Systems*, RWTH Aachen University, 20/06/2023; I held the seminar.
- *Managing Complexity in Modern Computer Systems with Process Mining*, Università Campus Bio-Medico di Roma, 26/09/2023; I held the seminar.

5. Periods abroad and/or in international research institutions

I have chosen to spend my period abroad at the PADS group, RWTH Aachen University, led by Prof. Dr. Ir. Wil M. P. van der Aalst. This group specializes in several PM areas, such as the development of new algorithms for enhancing PM types and application of PM to business domains for various services, including AD.

I have carried out independent research within the group regarding the comparison of several approaches to AD that integrate PM and ML (see 3. Research activity). I have received feedback from the Prof. about my work throughout the whole experience with one-on-one meetings.

Following are the relevant dates and information regarding the experience.

- **Start date**: 01/02/2023.
- **Seminar on my research**: 20/06/2023.
- **End date**: 31/07/2023.
- **Period of stay**: 6 months.

7. Tutorship

I held a lecture for the "Architetture dei Sistemi Digitali" course (taught by Proff. N. Mazzocca and A. De Benedictis) on 14/12/2022, entitled "Progettazione di macchine complesse: Prodotto scalare".

8. Plan for year three

My plan for year three involves:

- Following the review process of the papers that I have submitted (J3, J4, J5);
- Finalizing the in-progress papers and submit them to suitable venues (J6);
- Provide a final prototype for the CIS project and lay out the results in a paper (P1);
- Extend my research to other related domains, such as leveraging PM for developing digital twins of cyber-physical systems and self-adaptation of anomaly detection;
- Start organizing my PhD thesis, whose draft title is "A Framework for Anomaly Detection with Process Mining and Machine Learning in Cyber-Physical Systems".