2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3503-0080-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/METROXRAINE58569.2023.10405727

The MICS Project: A Data Science Pipeline for Industry 4.0 Applications

Loredana Cristaldi
Politecnico di Milano
Milano, Italy
loredana.cristaldi@polimi.it

Parisa Esmaili
Politecnico di Milano
Milano, Italy
parisa.esmaili@polimi.it

Giambattista Gruosso
Politecnico di Milano
Milano, Italy
giambattista.gruosso @polimi.it

Alessio La Bella
Politecnico di Milano
Milano, Italy
alessio.labella@polimi.it

Massimo Mecella
Sapienza Università di Roma
Roma, Italy
massimo.mecella@uniroma1.it

Riccardo Scattolini
Politecnico di Milano
Milano, Italy
riccardo.scattolini@polimi.it

Ala Arman
Sapienza Università di Roma
Roma, Italy
ala.arman@uniroma1.it

Gian Antonio Susto
Università di Padova
Padova, Italy
gianantonio.susto@dei.unipd.it

Letizia Tanca
Politecnico di Milano
Milano, Italy
letizia.tanca@polimi.it

***Abstract*— The goal of the MICS (Made in Italy, Circular and Sustainable) project is to develop new models and techniques to support the entire pipeline for applying Data Science algorithms to data from industrial processes. Although many libraries and tools are already available to aid the analysis of data, we believe that each different application domain requires individual study to propose appropriate methods and tools that accommodate the specific peculiarities of its data. In this position paper, we discuss the following points by also outlining our case studies.**

Keywords—digital factories, data management, machine learning pipeline, case study

I. Introduction

The fourth industrial revolution, known as Industry 4.0, implies an industrial infrastructure based on cyber-physical systems with sensor networks connected to the physical world through the Industrial Internet of Things (IIOT) [1]. This definition provides a framework for a smart industry where all the processes are digitalized and characterized by real-time analysis, shifting towards increasing sustainability in manufacturing [2]. Such an automated solution offers several advantages, including optimizing energy consumption in intensive industries as well as small but critical productive assets. By introducing Industry 4.0, a transition toward the in-process use of longer-range sensors has received considerable attention not only in the field of monitoring and condition tracking but also as a key component to establishing a virtual representation of complex physical systems known as Digital Twin (DT) [3]. Undoubtedly, DT is one of the pillars of Industry 4.0 towards achieving comprehensive control and communication between labor, machines, and management. A working definition of DT is a representation of a fabrication process or service in the digital world where its main complexity lies within both the reliable sensor networks for environment realization and accurate functional representation of the physical system in a consequential virtual representation. This results in digital transformation and description in a cyber-world context of real-world processes/items along with their surroundings where the Energy Management Information System (EMIS) and Prognostic Health Management System (PHMS) can be applied directly. Consequently, a complement EMIS/PHMS following the Digital Twin (DT) paradigm, which is the aim of MICS project.

As highlighted in the 2021 Annual Report of the World Manufacturing Forum, the development of a country must be directed towards enhancing sustainable and circular performances to optimize resource consumption and consequently minimize waste and emissions. To maintain competitiveness, the Italian industrial ecosystem is called to develop best practices on research and technology transfer to be adopted by small-medium enterprises. To do this, the Italian government has launched an economic recovery plan called the National Recovery and Resilience Plan (PNRR), which fundamentally relies on the MICS. Mainly, the MICS focuses on promoting digital transformation, innovation, and administrative simplification in various sectors of the Italian economy. As shown in Figure 1, it proposes eight thematic areas that need to be focused on to address the challenges that currently confront existing models of design, production, consumption, the End-of-Life of materials, products, production technologies, and processes necessary for moving towards greener and circular pathways and patterns.

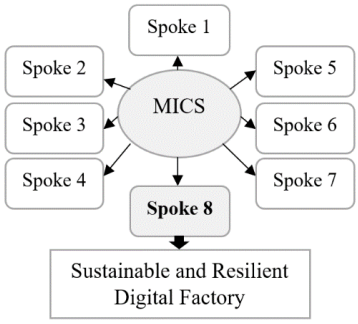

Fig. 1. MICS: Thematic areas

979-8-3503-0080-2/23/$31.00 ©2023 IEEE

One thematic area, so-called Spoke 8, has the ambition to create a new concept for a sustainable and resilient digital factory in which Artificial Intelligence (AI), digital technologies, and collaborative robotics will establish a trustworthy human-machine coevolution relationship and lead to high-performance, inclusive, sustainable human-machine working systems.

To this end, we aim to devise new models and techniques for the support the entire pipeline of application of Data Science algorithms to data from industrial process [4]. There are indeed various libraries and tools available to support the analysis of data, possibly coming from different sources. However, we claim that each specific application domain needs to be studied, to the end of proposing appropriate methods and tools to support the peculiarities of its data. Data Science integrates research tools and methods from statistics and computer science, applicable to research in various application domains, such as social science, digital humanities and, of course, industry. The 2015 National Science Foundation (NSF) report, summarizing the NSF-sponsored workshop on data science education, introduced a definition of data science that reflects the perspective of data science as a workflow: "Data science is a process, including all aspects of gathering, cleaning, organizing, analysing, interpreting, and visualizing the facts represented by the raw data." Data science is indeed commonly presented as an iterative workflow for generating value and data-driven actions from data (see Figure 2).

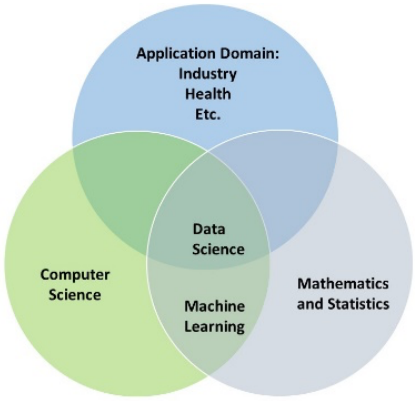

Fig. 2: Data Science Venn Diagram

In this position paper, we propose a specific pipeline for the MICS project, and we outline possible industrial case studies to be adopted for experimenting and validating our outcomes.

II. Data Science Pipeline

Figure 3 reports a typical Data Science pipeline.

Data Acquisition consists in selecting the data needed for the analysis on the basis of the *analysis objective*. This is a critical step in the data science pipeline. It involves collecting data coming from different sources and collecting it accurately and comprehensively is vital to obtain a good result.

Different strategies must be adopted considering the nature and origin of the data to assess the quality of the collected data. For instance, data coming from sensors must be qualified from a metrological point of view.

Data cleaning is motivated by the consideration that real-world data is often inaccurate, noisy, uncertain, and incomplete, leading to low-quality, possibly wrong, and even dangerous results of the Data Science algorithms. Manual intervention can often solve these problems, but it is time-consuming, and thus not scalable with respect to the amount of data nowadays available. The activities of *data cleaning* (possibly to be placed or repeated after data integration) include data transformation, data reduction, deduplication, error detection, missing value imputation, and space transformations in the case of multimedia data [5][6][7]. The overall task is complicated also by the fact that optimizing one dimension of data quality might cause a quality loss for another dimension. In this respect, the preliminary study of the specific analysis objective will also be useful to support the reconciliation of conflicts that exist among the different data quality dimensions.

As for data annotation, we propose associating appropriate descriptions called Metadata to the acquired data. These descriptions are partly provided by the data owners and partly obtained as output from the data cleaning activity, and they are used to retain information that will be useful in various processing phases [8]. Metadata has different classifications; this paper presents an idea of general categories that must be refined and specialized during the project to create a classification appropriate for industrial process data.

- *Governance Metadata*: These cover all security and privacy policies, access rights, ownership and responsibility roles, acquisition information, data quality, data authenticity, and other legal requirements.
- *Data Life Cycle Metadata*: These are related to data provenance, including the source of the data, all transformations performed and existing versions, usage tracking, and information required to preserve and use the data, including technical specifications. This group is approximately comparable to what is also referred to as use metadata.
- *Descriptive Metadata*: These describe data for purposes of identification and discovery purposes. Since this is a very comprehensive definition, they are further divided into subcategories (*Business-specific*, describing the meaning of the dataset or domain-specific metadata; *Intrinsic*, characteristics of the dataset and its value, and *Inter-Relationship*, describing possible relationships among different datasets).

Data integration is the problem of combining data residing at different sources and providing the user with a unified view of this data [9][10][11][12][13][14]. This entails detecting correspondences between similar objects that come from different sources, solving conflicts, and performing final data fusion. In this project, we will study the need for a data integration process and explore specific tools that can be applied to industrial process datasets in various circumstances.

What are the main problems related to integrating data from an industrial environment? There are three main questions to address:

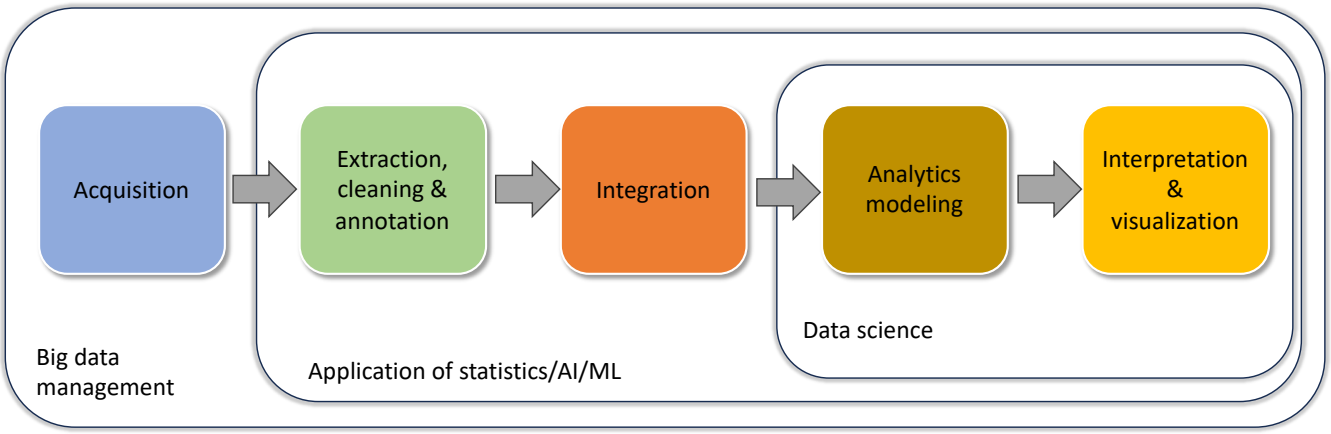

Fig. 3: Typical Data Science pipeline

- Is it necessary to 'put together' all the data related to a specific industrial process? Is this «putting together» somehow similar to the traditional idea typical of integration of business data?
- Do we need also to 'put together' the data related to different processes as well? Again, how is this similar to the traditional idea of data integration?
- Probably, we need to compare data of any kind (e.g., images, time series, etc...) to the end of deriving their similarities and differences. Data comparison is a typical step of data integration.

Data lake technologies are promising solutions for enhancing data management and supporting data integration for data science capabilities: *(i)* managing big-data volume and variety, as those typically found in industrial settings, and *(ii)* providing data analysts with a self-service environment in which advanced analytics can be applied. Such technologies should support:

- Connectivity;
- Integration of heterogeneous data;
- Data-storage capacity;
- Appropriate solutions to ingest, store, and process the enormous amount of heterogeneous information coming from all the industrial devices;
- Possibility to support data-analysis capabilities (Statistics, Data Mining, Machine Learning).

Data modeling and analysis consists of *(i)* parameters tuning, *(ii)* model training and deployment, *(iii)* analysis, and *(iv)* visualization and interpretation. Analysis is strongly dependent on the machine learning task considered in the application at hand. In the context of Industry 4.0, several applications have been developed in recent years to improve sustainability, quality and operations associated with manufacturing processes and related control and maintenance operations. In the following, a list of the most important data-driven industrial applications and the related data analysis task are summarized:

- Predictive maintenance (PdM) is a proactive maintenance strategy that uses data analysis and machine learning techniques to predict when equipment or machinery is likely to fail or require maintenance. By analyzing data from sensors, monitoring equipment, and other sources, predictive maintenance can help reduce downtime, optimize maintenance schedules, and increase efficiency and reliability. The PdM task is typically formalized as a Remaining Useful Life (RUL) prediction problem [15], making it a supervised regression task from a data analytics perspective. Other approaches in the literature also include supervised classification, leveraging the fact that PdM users are typically not interested in high granularity in the prediction but more in obtaining a good balance between unexpected breakdowns and unexploited lifetime [16].
- Anomaly Detection (AD) solutions are diagnostic tools capable of identifying unusual or unexpected patterns or events in processes. It involves analyzing data to identify outliers or anomalies that do not fit the expected or normal behavior of the system or process being monitored. AD is highly applicable in many scenarios, since it doesn't require labelled data [17]. AD solutions resort to unsupervised multivariate approaches capable of handling heterogeneous and complex data [18].
- Fault Detection (FD) is the supervised equivalent of anomaly detection (AD). If labeled data is available, diagnostic tools can not only provide an indication of abnormal conditions, but also they can point out which failure is currently happening [19]. In this case, supervised classification approaches are typically in place.
- Soft Sensing (SS) or Virtual Metrology/Virtual Sensing are software technologies that estimate quantities that are costly or impossible to be measured by leveraging historical data and, typically, supervised regression approaches. SS approaches are used to improve quality monitoring, control, and sampling in production [20];
- Dynamic Sampling (DS) strategies are approaches able to optimize measurement and sampling in complex manufacturing environments [21]. DS approaches typically rely on regression analysis and optimization approaches [22];

- Computer Vision-based control and monitoring is widely adopted in the industry thanks to the diffusion of low-cost effective industrial cameras and the incredible advancements of Deep Learning approaches in recent years. In this context, supervised classification and segmentation approaches to recognize defects [23] are some of the most popular techniques adopted, but unsupervised methodologies have been used in the industry [24].
- Optimization and dynamic model predictive control design [25] are popular methods nowadays for enhancing production capacity and energy consumption in the digital factory. According to a data-driven approach, and by means of Machine Learning methods, data collected from the plant are used to estimate a dynamic model of the system. This model is then used on-line to optimize the plant's behavior by minimizing a suitable cost function while adhering to constraints on the process variables. Adaptive versions can be employed to modify, in a largely autonomous way, the control algorithms to adapt to production scenarios characterized by strongly varying working conditions.

Data visualization and interpretation are of fundamental importance when dealing with Machine Learning (ML)-based approaches in the industry. One of the main reasons is that many of the aforementioned applications are consumed by users through decision support systems [26]. While ML tools provide predictions and elaborations, it is up to the users to take actions and make the final decision on many monitoring and control tasks. Users may need to have additional information beside the model prediction/outcome in order to act on the process/equipment. In light of this, users may benefit from eXplainable Artificial Intelligence (XAI) methods in order to get valuable additional knowledge [27]. XAI methods may provide for example:

- feature importance, making users and operators know which are the relevant variables for a ML model (global importance) or for its outcomes (local importance) [28];
- the relationship between output and input in complex black box models [29];
- in computer vision tasks, the pixels of an image that are relevant for the task at hand [30].

XAI approaches and, more in general, visualizations and statistics are fundamental tools also for increasing the trust of process/machine experts without a background in Machine Learning, fostering the adoption of such technologies in the industry. Nevertheless, XAI suggestions must be tailored for the industrial audience and must be provided in pseudo real-time in order to be used in the manufacturing environment.

III. Case studies

Two engineering fields closely linked to energy efficiency are upgradeable EMIS and PHMS. The EMIS focuses on monitoring and control systems that can transform the collected data into correlated and usable information, while the PHMS manages anomaly detections, identifies fault causes, and predicts remaining useful life to support effective decision-making driven by condition-based techniques. Both EMIS and PHMS combine sensor networks, software, and data to provide support in managing energy at the process, system, facility, and enterprise levels. To achieve a complement EMIS/PHMS following the Digital Twin (DT) paradigm, two industrial cases studies have been defined as follows:

A. Monitoring of apparatus energy consumption

In this case study, energy consumption monitoring is focused on industrial equipment such as high-speed and precision spindles used in small-scale machinery tools or highly modular machining centers developed for the large-scale woodworking industry. To perform such an analysis, the first step is to create a digital twin of the apparatus. Later, the model will be integrated with data acquired from physical assets for validation.

B. Energy forecasting and modeling in industrial building and process

The second case study is dedicated to energy consumption monitoring in industrial plants, especially focused on the woodworking industry. Therefore, the fundamental task is to create a scalable digital twin of the process or the plant. Once completed, the validation will be performed based on acquired data in the field.

Following the cases studies and using the digital twin created earlier, it is possible to generate synthetic data that can simulate inefficiencies, failures, or other critical aspects that cannot be easily replicated in the real world. This synthetic data can then be used to create and validate new machine-learning algorithms in terms of EMIS/PHMS.

IV. Conclusion

In this position paper, we introduce a specific data science pipeline aimed at enacting in the MICS project – spoke 8, with the specific aim of supporting Industry 4.0 applications in the Italian industrial landscape. We argue that the results of the activities will be applicable in similar scenarios, especially in Europe, where the features of industrial players of large vs. medium enterprises are quite like the Italian ones. Therefore, they will constitute an advancement over the current state of the art, which has been surveyed in this paper. Future activities will address all the specific challenges outlined in this paper.

Acknowledgment

This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-Generation EU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). CUP: C93C220052800, B53C22004130001, C93C220052800.

References

[1] Qin, Jian, et al. "A Categorical Framework of Manufacturing for Industry 4.0 and Beyond." Procedia CIRP, vol. 52, 2016, pp. 173–178.

[2] Barreto, L., et al. "Industry 4.0 Implications in Logistics: An Overview." Procedia Manufacturing, vol. 13, 2017, pp. 1245–1252.

[3] Javaid, Mohd, et al. "Significance of Sensors for Industry 4.0: Roles, Capabilities, and Applications." Sensors International, vol. 2, 2021, p. 100110.

[4] M. Koby; H. Orit, "What is Data Science?". Communications of the ACM. 66 (2): 12–13. doi:10.1145/3575663. ISSN 0001-0782.

[5] M. Calautti, S. Greco, C. Molinaro, I. Trubitsyna. Preference-based inconsistency-tolerant query answering under existential rules. In KR, 2020.

[6] L. Caruccio, V. Deufemia , G. Polese. Relaxed functional dependencies - a survey of approaches. TKDE, 28(1), 2016.

[7] M. Mazuran, E. Quintarelli, L. Tanca, S. Ugolini. Semi-automatic support for evolving functional dependencies. In EDBT, 2016.

[8] A. Chapman, P. Missier, G. Simonelli, R. Torlone. Capturing and querying fine-grained provenance of preprocessing pipelines in data science. PVLDB, 14(4), 2021.

[9] W.C. Lin, C.F. Tsai. Missing value imputation: A review and analysis of the literature (2006-2017). AI Rev., 53(2), 2020.

[10] S. Song et al. Enriching data imputation with extensive similarity neighbors. PVLDB, 8(11), 2015.

[11] S. Staworko, J. Chomicki, J. Marcinkowski. Prioritized repairing and consistent query answering in relational databases. AMAI, 64(2-3), 2012.

[12] D. Calvanese et al. Enriching ontology-based data access with provenance. In IJCAI, 2019.

[13] Fabio Azzalini, Davide Piantella, Emanuele Rabosio, Letizia Tanca: Enhancing domain-aware multi-truth data fusion using copy-based source authority and value similarity. VLDB J. 32(3): 475-500 (2023)

[14] Fabio Azzalini, Songle Jin, Marco Renzi, Letizia Tanca: Blocking Techniques for Entity Linkage: A Semantics-Based Approach. Data Sci. Eng. 6(1): 20-38 (2021)

[15] Ren, L., Dong, J., Wang, X., Meng, Z., Zhao, L., & Deen, M. J. (2020). A data-driven auto-CNN-LSTM prediction model for lithium-ion battery remaining useful life. IEEE Transactions on Industrial Informatics, 17(5), 3478-3487.

[16] Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. IEEE transactions on industrial informatics, 11(3), 812-820.

[17] Dandolo, D., Masiero, C., Carletti, M., Dalle Pezze, D., & Susto, G. A. (2023). AcME—Accelerated model-agnostic explanations: Fast whitening of the machine-learning black box. Expert Systems with Applications, 214, 119115.

[18] Carletti, M., Masiero, C., Beghi, A., & Susto, G. A. (2019, October). Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In 2019 IEEE international conference on systems, man and cybernetics (SMC) (pp. 21-26). IEEE.

[19] Ding, Y., Zhuang, J., Ding, P., & Jia, M. (2022). Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. Reliability Engineering & System Safety, 218, 108126.

[20] Jiang, Y., Yin, S., Dong, J., & Kaynak, O. (2020). A review on soft sensors for monitoring, control, and optimization of industrial processes. IEEE Sensors Journal, 21(11), 12868-12881.

[21] Dauzere-Péres, S., Rouveyrol, J. L., Yugma, C., & Vialletelle, P. (2010, July). A smart sampling algorithm to minimize risk dynamically. In 2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC) (pp. 307-310). IEEE.

[22] Susto, G. A., Maggipinto, M., Zocco, F., & McLoone, S. (2019). Induced start dynamic sampling for wafer metrology optimization. IEEE Transactions on Automation Science and Engineering, 17(1), 418-432.

[23] Tabernik, D., Šela, S., Skvarč, J., & Skočaj, D. (2020). Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing, 31(3), 759-776.

[24] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. (2022). Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14318-14328).

[25] Rawlings, J.B., D.Q. Mayne, and M. Diehl, Model predictive control: theory, computation, and design, 2017, Nob Hill Publishing Madison, WI.

[26] Kasie, F. M., Bright, G., & Walker, A. (2017). Decision support systems in manufacturing: a survey and future trends. Journal of Modelling in Management, 12(3), 432-454.

[27] Molnar, C. (2020). Interpretable machine learning. Lulu. com

[28] Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest. Engineering Applications of Artificial Intelligence, 119, 105730

[29] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(4), 1059-1086

[30] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that?. arXiv preprint arXiv:1611.07450.