BiDrac Industry 4.0 Framework: Application to an Automotive Paint Shop Process

Elma Sanz[a,b,c,1], *Joaquim Blesa*[b,c], *Vicenç Puig*[b,c]

[a] *Industrial Automation Maintenance, SEAT, S.A.,08760 Martorell, Barcelona, Spain*
[b] *Automatic Control Department Technical University of Catalonia (UPC) Barcelona, Spain*
[c] *Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Carrer Llorens Artigas, 4-6, 08028 Barcelona, Spain*

Abstract

To implement an Industry 4.0 Framework in an ongoing industrial manufacturing process to prepare it for complex Artificial Intelligence use cases is not an easy task. The automotive industry is undergoing a large transformation due to a variety of disruptive factors through the introduction of CASE (Connected, Autonomous, Shared, Electrified) technologies. Production improvements will enable future development of CASE capabilities. This paper presents an Industry 4.0 framework, named BiDrac, that integrates computing, communication and control under an Industrial Cyber-Physical System (ICPS) ecosystem. It combines Artificial Intelligence and Industrial Internet of Things (IIoT) inside the Industry 4.0 paradigm for the Predictive Maintenance of an Automotive Paint Shop Process inside an Industrial Cloud (IC) as an open platform that provides community of cloud-based solutions to enable the development of new digital solutions. Several real use cases of the proposed platform have been included in the results section to illustrate the potential of the proposed framework.

Keywords: Cyber-Physical System (ICPS), Artificial Intelligence, cloud computing, Industry 4.0, Predictive Maintenance, Automotive, Paint Shop

[1] This is to indicate the corresponding author.
Email address: elma.sanz@upc.edu (UPC), elma.sanz@seat.es (SEAT, S.A.), vicenc.puig@upc.edu (UPC), joaquim.blesa@upc.edu (UPC)
URL: https://www.iri.upc.edu/ (*Institute of Robotics and Industrial Informatics (CSIC-UPC)*), https://cs2ac.upc.edu/en (*Automatic Control Department Technical University of Catalonia (UPC)*), https://www.seat.com/ *(SEAT, S.A.)*

1. Introduction

Digital Transformation meets technology and people to improve the way Industrial Cyber-Physical System (ICPSs) is built or adapted, to understand industrial processes deeply with each intrinsic dimensionality and to add new analytics perspectives to solve problems (Ustundag and Cevikcan, 2018). So smart and automatic solutions should be included in industrial process using Artificial Intelligence (AI-driven framework) to be competitive (Porter and Heppelmann, 2014) inside the Industry 4.0 paradigm that basically influences the manufacturing industry (Stock and Seliger, 2016).

This paper presents the BiDrac framework, that aims to profit from the integration of computing, communication and control under an ICPS ecosystem (Colombo et al., 2017) combined with Artificial Intelligence (Russell and Norvig, 2016) and Industrial Internet of Things (IIoT) (Garcia et *al*., 2019) inside the Industry 4.0 paradigm for the Predictive Maintenance (Ran et al. 2019) of an Automotive Paint Shop Process (Streitberger and Dössel, 2008).

BiDrac is the Ecosystem where to integrate Equipments, Technical Locations, PLCs, sensors, communication protocols (OPC-UA (Pauker et *al*. 2016), MQTT (Mott, 2020)), production networks, industrial networks, corporate networks, Complex Infrastructure Systems (CIS), ETL tools, Data Bases, Datawarehouse, Data Lake, Digital Platform, Algorithms, Machine Learning models, Artificial Intelligence models, Infrastructure, MES, ERPs, among others. BiDrac has been created to allow predictive maintenance in the Paint Shop of a car manufacturer. This paper shows how the BiDrac framework allows applying Artificial Intelligence and Predictive Maintenance illustrating several use cases.

1.1 Literature review

The car coating process (Sreitberger and Dössel, 2008), considered as case study in this paper, has also been considered in other contributions. This process is a good case study for building an Ecosystem to integrate all the “Things” of the “Industrial Internet of Things (IIoT)” (Wollschlaeger et *al*. 2017; Jeschke et *al*., 2017) with data, advanced analytics and artificial intelligence in order to add value and knowledge with their

interactions (see Machine to Machine (M2M) described in (Wu et *al*., 2011)) and lead to a Smart Factory (Wang et *al*. 2016).

Automatization and Control with IT-Shopfloor are in the first stage, Edge Computing with data in the second one, Digital Production Platform with Artificial Intelligence engines with metrics of goodness are in the third one and automated or manual actions on the first stage or on the ERPs system (schedule actions) in the fourth.

A wide range of different Frameworks have been developed to improve advanced monitoring in industrial processes. Depending on the companies and the availability for developing new solutions, on-premise or cloud-based, the architecture proposed is going to be different (Kazemi et al., 2019).

Data is collected from the Technical Locations and Equipments (installation) and provided by different information systems and data sources that are not integrated. Information systems act as a concentrator to send (push) heterogeneous data packages in real-time from the SCADA (WinCC 7.5 Simatic S7 Project) to the MFS framework software and an Extract, Transform and Load tool (ETL), connects to the information systems Data Warehouse to extract in near real-time data to the BiDrac Staging Area (Sajid et *al*., 2016). ETL tools (Kimball, et *al*. 2011) pre-processes data, applies transformations and loads the pre-processed data to the edge computing BiDrac Data Lake (O'Leary, 2014; Miloslavskaya and Tolstoy, 2016). Clean Data is stored into the Data Lake, to detect any abnormality or fault in data, and to perform diagnosis and prognosis (Schwabacher, 2015; Reis and Gins, 2017).

Cloud Computing (Buyya et *al*., 2010) treats everything as a service. The services are defined under a layered system including Infrastructure as a Service (IaaS) (as e.g. virtual servers, networks or storage and where users can deploy and run software), Platform as a Service (PaaS) is where applications reside and run into the cloud infrastructure (as e.g. users develop and run scalable applications using high speed servers and storage using programming languages on the cloud infrastructures). Software as a Service (SaaS) and Function as a Service (FaaS) allows to eliminate the service applications and functions on local devices of individual users (Alcácer and Cruz-Machado, 2019), under the paradigm of Everything as a Service (XaaS).

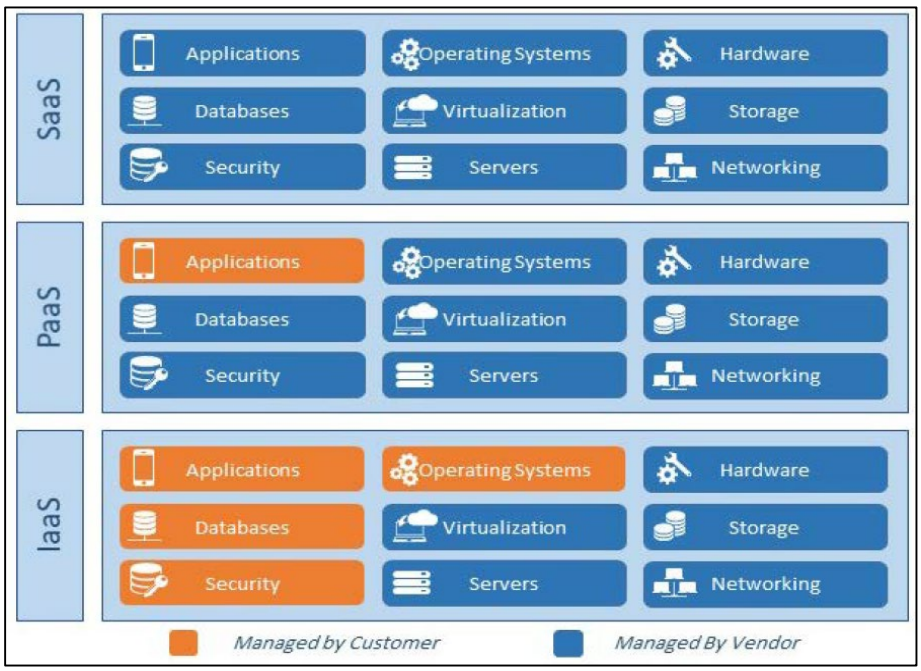


Fig. 1: Cloud Computing schema (Alqaryouti, 2018)

The largest cloud service provider in 2020 is Amazon with the scalable, high availability and dependability, cloud computing platform Amazon Web Services (AWS) (AWS, 2019). Amazon offers flexibility to the users to build their own applications in a secured environment.

Digital Production Platform (DPP) or Digital Manufacturing Platform (DMP) is a part of a layered architecture, commonly deployed in the cloud, that integrates operational state and ICPSs data provided by an industrial process, by their information control systems or directly from the PLCs or the sensors, using networks and communication protocols as interfaces and bring the data to a third party applications (EFFRA, 2019). DPP is a multi-tenant, secure and scalable platform consisting of data, device, and analytics services. DPP provides standardized device connectivity and management services that allow production facilities to manage and operate their plant floor machines and devices.

The Volkswagen Group and Amazon Web Services developed an Industrial Cloud (Fig. 2) that will integrate and combine data of all machines, plants and systems from

all the 124 production facilities of the Volkswagen Group (VW-AWS, 2019). In July 2020, up to 18 production plants have been connected to the Industrial Cloud and started developing functions, services and applications with their partners (VW-AWS Industrial Cloud Hub, 2020). IT at the production level of machinery, equipment and systems will be standardized and networked across the plants. Volkswagen has chosen the AWS portfolio of services including Internet of Things (IoT), machine learning analytics and compute services to construct the IC. It is entirely built on AWS native services, but other providers can join it. It uses the suite of AWS IoT Services, including AWS IoT Greengrass, AWS IoT Core, AWS IoT Analytics, and AWS IoT SiteWise, to detect, collect, organize, and run sophisticated analytics on data from the plant floor. The architecture is the new Digital Production Platform (DPP) used by Volkswagen. All plants and companies out of the Group are going to dock their system architectures into the platform. This platform will standardize and simplify data exchange between systems and plants. The aim is to achieve significant productivity improvements in the plants through the development of new use cases and solutions that will increase plant efficiency and uptime, improve production flexibility and increase vehicle quality for Volkswagen. Before DPP's arrival to the BiDrac project, BiDrac Data Lake had been constructed and prepared to be integrated in the DPP and several works on premise developed using Python environment have been tested.

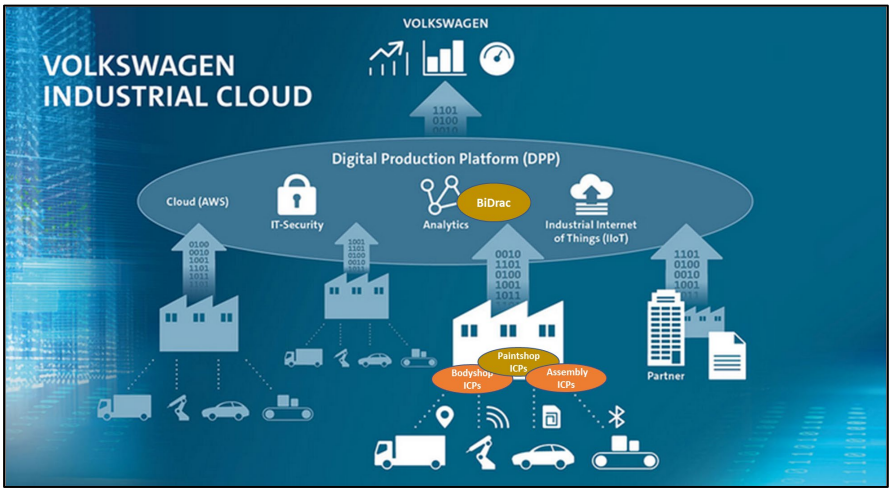


Fig. 2: BiDrac inside the Volkswagen Industrial Cloud with Amazon Web Services (based on VW-AWS, 2019)

The IC is designed as an open platform, community and marketplace of cloud-based solutions. It is an open architecture that connects industrial processes (production and supply chain) to the cloud and standardizes the sharing and exchange of data. An IC includes as partners diverse communities of part suppliers, logistics providers, technologies providers, systems integrators (SIs), independent software vendors (ISVs) and original equipment manufacturers (OEMs) (VW-AWS Industrial Cloud Hub, 2020). The IC facilitates the convergence of Operational Technology (OT) and Information Technology (IT) by providing methodologies and standards to integrate OT into the cloud platform. The Community or Industrial Partner Network is a large consortium of "builders" on the IC. The marketplace is a trusted and secure environment for sharing, selling, or acquiring proven applications. The IC helps to approach major challenges, such as to reduce the barrier of entry for new technologies, lower costs for production and innovation, stable and secure environment, enable the rapid deployment of solutions, integrate fragmented IT Landscape and IoT platforms (because of a lack of a harmonized, singular, open, industry-standard platform), include other companies, complexity of processes and supply chain, and competing customer demands. The IC also allows organizations focusing their resources on their challenges without needing to build an infrastructure on their own – it is a multiplier force. More benefits are: long-term partnerships, co-developed solutions, early access to the latest IoT technologies, 24/7 system, setting of new modern data-driven models, developing, launching, selling or acquiring services and cross-functional solutions via IC marketplace. The IC provides solutions to key manufacturing use-cases, including and not limited to: Digital Shop Floor Management, Smart Production processes, Predictive Maintenance, Smart and Predictive Quality, Smart Identification and Localization, Transparent N-Tier SCM and Material Flow Track/Trace. The objective is to create a continually growing worldwide industrial ecosystem. BiDrac Project aims to help focusing on Predictive Maintenance, Smart Production processes and Digital Shop Floor Management without losing the Smart and Predictive Quality inside the Paint Shop Process because of the surface defects due to malfunctions of the automated process.

The Reference Architecture Model Industry 4.0 (RAMI4.0) (Pauker, 2016) appears in Germany as the guidance for the implementation of Industry 4.0 technologies.

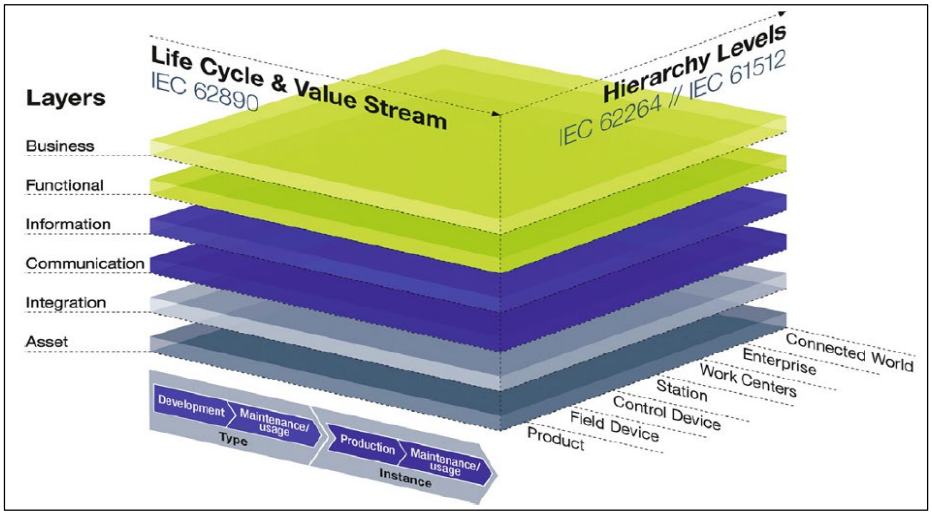


Fig. 3: Reference Architecture Model (RAMI4.0) (Pauker, 2018)

The 3D model of Figure 3 enables to identify the existing standards and among it allows the step-by-step migration from the current to the future manufacturing environments.

Advanced digital production (ADP) technologies change manufacturing production process by introducing digital production systems close to the physical processes. As discussed in the "*Industrial Development Report 2020: Industrializing in the digital age*" report from United Nations Industrial Development Organization (UNIDO, 2020), ADP technologies – combining hardware (advanced robotics and additive manufacturing), software (artificial intelligence, big data analytics and cloud computing), and connectivity (OPC UA, MQTT and Internet of Things among others) – used under the ethics paradigm can foster inclusive and sustainable industrial development (ISID) and the achievement of Sustainable Development Goals (SDGs). Technological paradigm aims to accelerate innovation and increase the value-added content of production in manufacturing industries. ADP increases efficiency and productivity of industrial production processes transforming them as a part of a smart factory. Digital transformation needs more skilled and knowledge-based sectors. The UNIDO report provides strategic areas that deserve particular attention: (i) developing framework conditions; (ii) fostering demand and leveraging ongoing initiatives; and (iii) strengthening required skills and research capabilities. The goal is to build

dynamic, sustainable, innovative and people-centred industrial processes with ADP technologies as a one driver more of our industrial development.

When ICPSs work in complex environments (e.g., HVACs, Coating Cabins, Ovens, Robots) with high-dependencies between them, it is mandatory to build an optimal maintenance strategy, based on accurate positioning of fault locations and prediction of fault conditions (Liu *et al.*, 2019). Improvement of maintenance scheduling tasks need to know, just in time, the health status of each ICPS. As in the methodologies for enabling Digital-Twin (DT), technical locations and equipment must be mapped under ICPSs' tree to cover the entire process. Each location will be influenced by the degradation of the health status of the mapped resources and as a result they are used in context of predictive maintenance models (Aivaliotis et al., 2019; Tao et al., 2018). The scope of the BiDrac project does not include Digital-Twins modelling and enabling but the concept allows to introduce the mapping of the ICPSs in it.

The ERP SAP Plant Maintenance (PM) module (Liebstuckel, 2011) includes in the master data equipment (PM-EQM-EQ), technical locations (all ICPSs in the process) for functional location (PM-EQM-FL) or a combination of them (equipment at functional locations), class (specifications for objects), and maintenance orders management and event register for each ICPSs. PM orders functional locations and equipment inside a 7-level tree map and classifies them into process, auxiliary and transport. This structure forms the basis for implementing data integration and, also, for implementing advanced and artificial intelligence modelling to introduce predictive maintenance.

Using the SAP PM multi-level tree map in the data model assures data integration inside the BiDrac Data Lake and allows the multi-attribute and multi-criterion analyses. In the literature, the concept of super-network (Liu et *al*., 2019) is used to describe a multilayer complex network. This network's features are multi-level, multi-attribute or multi-criterion and are being used to describe the interaction and influence between networks. The super-network consists of three sub-networks: data physical layer, data virtual layer and data service layer. Mapping between data physical layer sub-network is the SAP PM multi-level tree map.

Inside an automated industrial process, several relevant KPIs are used to describe the health of ICPSs in the Maintenance Department (e.g. MTTRi, MTBFi, OEE, and Availability) (Yan, 2014). Also, in Maintenance, different kinds of data (related to production, energy and environment) can be found. Thus, this is the suitable atmosphere to build and apply advanced monitoring and predictive maintenance techniques using signals, parameters, warnings, alarms, faults, etc. Predictive Maintenance algorithms are a tool which allow scheduling preventive maintenance work orders from predictive analytics results in order to act when and where is needed due to the health of the equipment or installation.

1.2 Motivation and Contribution

This paper presents a developed innovative industrial ecosystem. Most companies still do not analyse the process data of the ICPSs that, in the best case, are obtained from several shop floor systems with advanced data analytics algorithms in an integrated way. From these data can be obtained relevant information about the status and behaviour of the physical elements inside the factory. However, this information is not easy to analyse because of the large amount of data and sources in which they are stored and the data heterogeneity problem (Jirkovsky, 2016).

Monitored real-time CIS systems produce a huge amount of data, which generally includes the key issues of the actual status and behaviour of the monitored industrial and automated process. SCADA systems (as e.g., WinCC) are effective, but data should be conveniently collected and processed to enlighten this information appropriately. Big Data Analytics is an emerging and growing-up field which aims to extract valuable information from the huge amount of data available. From this information, it is possible to provide advice about the system behaviour. This process embraces from the raw data collection (gathered from signals of the sensors or PLCs) to the isolation of undesirable behaviours in the plant operation, including signal data validation/reconstruction and prediction because data abnormality detection, imputation and fault analysis for ICPS is the base of a good data analysis.

BiDrac Ecosystem aims to solve research questions as the ones in the following list:

Research question 1: What is the optimal Ecosystem architecture that fits the standard of SEAT, S.A. as a member of the VW-Konzern to go from the signals of sensors or PLCs to the artificial intelligence techniques applied to real process data?

Research question 2: Inside the Paint Shop process, considering the current status of installations, networks and communications protocols, which is the most suitable way to collect, store and distribute the raw and pre-processed data to the corporate network?

Research question 3: Which is the most suitable Data Lake technical design to integrate pre-processed data under one data model? Which is the data model that satisfies the needs and the key performance indicators (KPIs) related to the Paint Shop process and Maintenance area?

Research question 4: Which is the most suitable Data Model for this specific industrial process? How can a metadata layer (based on this data model) be built to label each physical element of our installations (Equipments, Technical Locations) with each signal from sensors or PLCs and label each signal with their characteristics, types, classifications and unities of measure? How can this heterogeneous data problem be dealt with?

Research question 5: Which is the most suitable digital production platform to use to apply artificial intelligence techniques into specific data sets to deal with different use cases?Which are the most suitable tools to bring the Artificial Intelligence algorithms results to the physical elements and to the installation? Artificial Intelligence results may come back as actions to do into the process. There are two ways to get the results: as an advert in a smart peripherical device or as an automated signal to a PLC.

Research question 6: How can inclusive and sustainable industrial KPIs be introduced into the Paint Shop process focused on data analytics from BiDrac Ecosystem? How can these digital skills and capabilities be transferred to the final users of BiDrac Ecosystem?

This paper illustrates the application of BiDrac framework presenting some use cases' results which answer some of the previous research questions in the automotive paint shop process of the car manufacturer SEAT (part of VW group) in Martorell (Spain).

1.3 Structure

The rest of this paper is organized as follows: Section 2 provides an overview of the proposed approach. Section 3 focuses on the description of the methodology designed to achieve the goals and the progress beyond existing solutions. Section 4 presents the BiDrac application results gained in some selected use cases. Finally, Section 6 draws the conclusions.

2. BiDrac approach

The wished output is to have complete descriptive and predictive input data sets with their results, inside the BiDrac Ecosystem (Fig. 4), to post-perform the corresponding action in the industrial process, giving an accurate architectural framework for distributed ICPS by the design, implementation and operation. The proposed infrastructure includes edge computing with BiDrac Data Lake, cloud computing with Digital Product Platform for artificial intelligence algorithms, smart gateway, industrial and corporate network, and assuring safety, privacy and security for ICPS. Several analytics use cases will be presented to prove the applicability and effectiveness of the proposed Ecosystem for a Paint Shop industrial process, with remarkable potential to optimize installations and reduce production volume loss to their final users.

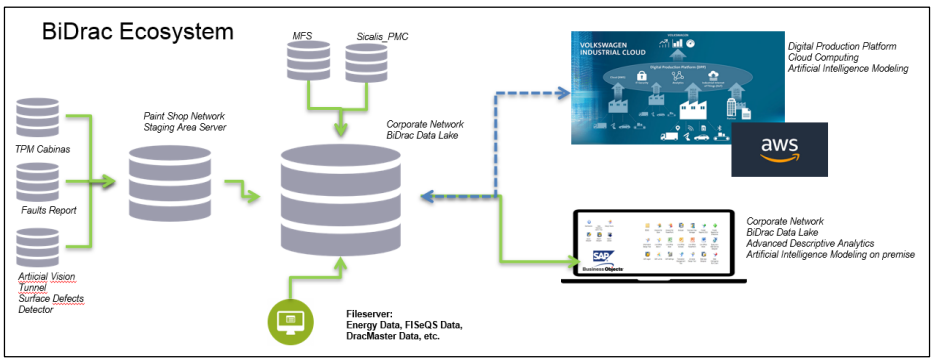


Fig. 4 BiDrac Ecosystem

Paint Shop industrial process systems are complex infrastructure systems facing new challenges in their real-time operations because of several issues (as e.g., limited

resources, intensive energy requirements, growing production, growing installations, costly and ageing infrastructure, increasingly stringent product quality criterions, and increased attention towards the environmental impact of automotive industry).

The number of bodies painted per day in the Paint Shop of the Martorell car manufacturer Seat, S.A. ranges from 2.200 to 2.400. This corresponds to almost one body coated per minute of production time. In an automotive factory the surface treatment is the process that consumes most energy, water and chemicals, and produces most waste and pollution. Roughly 50% of the energy is used in the paint shop with an average consumption of 800-1000 kWh per body. Within the paint shop, the dominating energy cost is the heating and ventilation of air (HVAC) in the booth (50%) followed by the ovens (25%).

3. Methodology description

The better way to introduce the proposed methodology is to start from its main premise that "*all that can be integrated is going to be integrated*". Actually, BiDrac methodology considers integrating and labelling as main keywords.

The different phases of BiDrac methodology are: (1) Ecosystem design and technical specification and requirements; (2) Research work to achieve project's objectives and (3) The development of the analytics use cases applying advanced analytics and artificial intelligence in order to validate and value the proposed approach. The methodology is applied from general to specific. This case focuses on the specific industrial process of coating bodies, but this methodology can be applied to all industrial process over different industry sectors.

The first phase consists in the physical process data analysis and the creation of a mapping tree of labelled signals to the Equipment or Technical Location that contains them, and in which information control system are located. Then, the pre-processing, transformations and data validation processes have been developed.

Then, a relational data model, where all the data are integrated, is developed in the Corporate Network. The BiDrac Data Lake contains all significant data from several data sources prepared for analytics. The mapping tree of all the ICPSs allow to add new sensors, signals, equipment, or technical locations, making easier the growth of the BiDrac Data Lake.

ETL processes assure the feed of the BiDrac Data Lake. The extraction is scheduled in order to not affect the Paint Shop Production Network and bring the data to the Corporate Network.

The applications and frameworks used to manage data from the BiDrac Data Lake are in the Corporate Network, so the queries run under this network.

The technical design of the BiDrac Data Lake includes several work packages to minimize or avoid design problems that could stop the project:

- Collect all the Technical Documentation available of the two main control systems in the Paint Shop: MFS and Sicalis-PMC. MFS for the product and processes tracking and Sicalis-PMC for installation signals and parameters (faults, warnings, sensor values, parameters, boundaries, etc.). The technical documentation was not updated nor digitalized at the beginning of this project. Digitalized version was distributed after the work package was done.
- Both control systems had obsolescence and presented a high-risk to stop the process. The application servers and old Datawarehouse Unix-based information systems with only 60 shifts of historical data were analysed. At the beginning of this project there was no migration project scheduled for them. Actually, both systems have been updated to the last versions available and located in a CPD under the IT maintenance service.
- Both control systems (Sicalis-PMC and MFS) were not integrated. Sicalis-PMC made groups of 10 variables; the name of the group informs about the technical location where the contained variables can be found. As technical locations have big amounts of signals (thousands), it is difficult to list all the signals related to one technical location. MFS has an ID per technical location, to do the body tracking and also an ID per body, each body has a unique PIN (VIN) and the Moby-I tag on the movement unit, SKID, contains all the parameters needed to coat the body. This information is stored in the MFS control system. The ERP is SAP and the maintenance department use the module PM that provides a 7-level mapping tree for Technical Locations and Equipments. To integrate all the data, a new semantic layer (master data and metadata) (Liu et al., 2019) has been constructed including all data into the mapping to prioritize the most relevant for advanced monitoring and artificial intelligence modelling.

– From BiDrac, it is possible to extract a specific data set with enough accuracy to create each required model on premise in a local server or, in the future, when the models are going to be developed and stored into the Digital Production Platform (DPP) where a net of algorithms is connected under the same industrial process ecosystem. Inside the DPP, algorithms are considered as gears; each algorithm could be part of the input of other algorithm as an infinite spider network with exponential growth.

4. Application Results

4.1 BiDrac Data Lake

A data model for the integration in the Paint Shop have been developed in order to make analysis with all the significant data where and when it was needed without taking into account the initial data source where it was stored.

First step was to introduce the SAP PM Fenix mapping tree (Fig. 5), as the methodology of digital twins proposed to target all the ICPSs and including the main control systems in the Paint Shop, Sicalis-PMC and MFS.

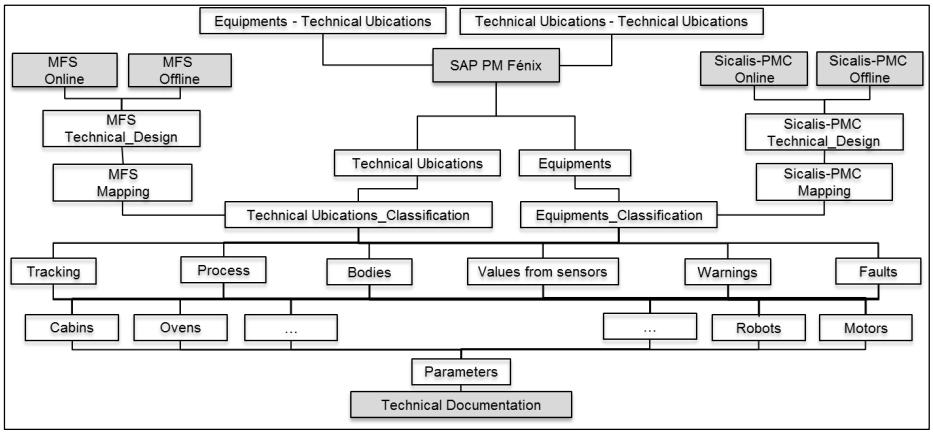


Fig. 5. SAP PM Fenix as integrator of Sicalis-PMC and MFS (online and offline)

Second step was to understand the process data intersections and correlations: time series models, dependencies between productive, auxiliary and transport Technical

Locations. How this gear is going to work? For instance, Fig. 6 shows how a fault influence the coating process and examples of analysis that could be done when the fault appears and is predicted by an AI model.

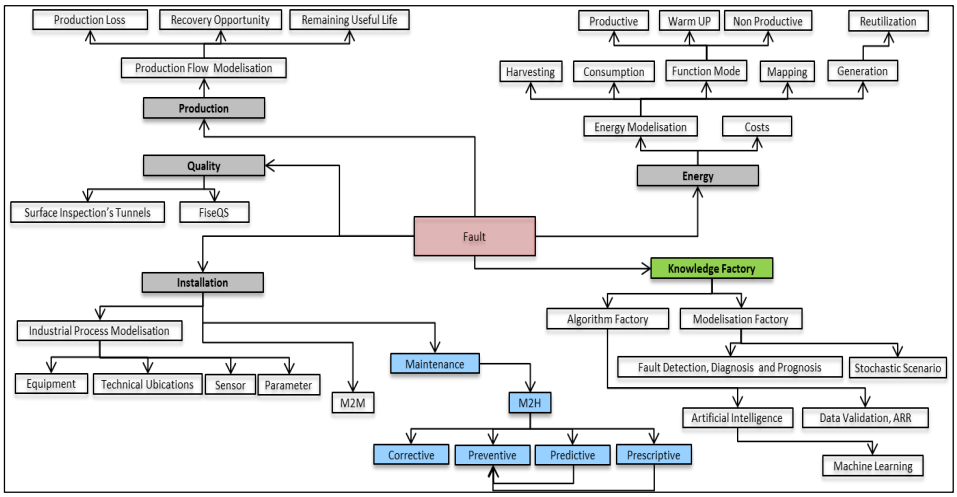


Fig. 6: Fault influence on the coating process.

This task is the basis of the data model orchestration allowing to discover how the tables are going to be constructed and which relations are needed. When a fault appears in one of the ICPs of the Paint Shop Process mapped in BiDrac (Fig. 6), the data model constructs aims to correlate the data from the faulty ICP with other ICPs (the production flow and scheduling, the nearest related ICPs, automated transport system, auxiliary ICPs (HVAC), energy consumptions and management, quality data of the coated bodies, maintenance and production KPIs) using advanced statistics, artificial intelligence and machine learning algorithms.

In this phase, more data are added from other data sources to the BiDrac Data Lake that contain suitable data for complex and bigger data models, descriptive or predictive models (Fig. 7).

A data model with 157 final tables and more than 700 FKs and relations is constructed as shown in Fig. 7.

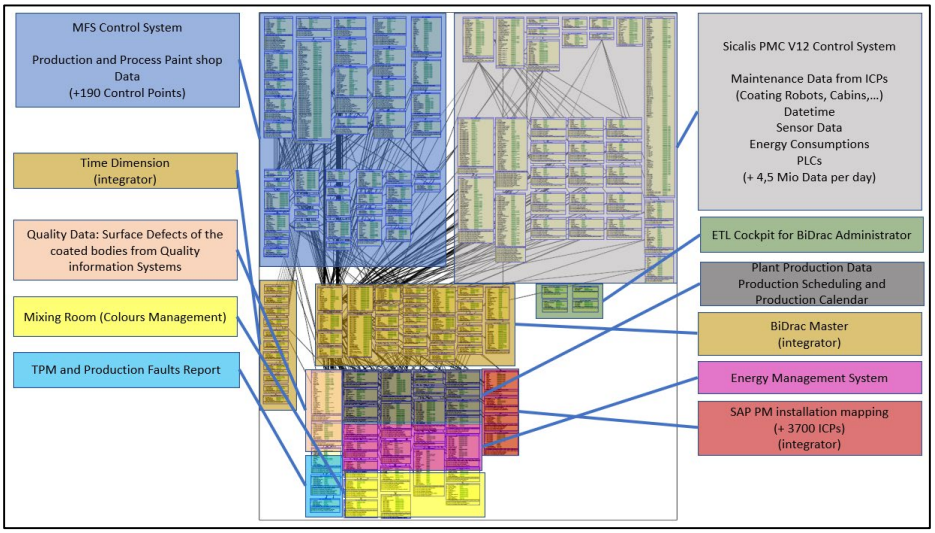


Fig. 7. Data model of the BiDrac Data Lake

BiDrac Data Lake included five years of historical Data prepared to be analysed.

Finally, a semantic layer has been included. The software chosen to develop reporting with advanced monitoring and descriptive analytics is the SAP Business Objects 4.2 (SAP BO) (see (SAP, 2020) for more details). The SAP BO adds a semantic layer to the data models known as universes and contain dimensions, metrics and attributes using business language.

BiDrac Data Lake has 11 linked universes with energy, faults, TPM, Surface Defects tunnels with automatic detection, Surface Defects from FISeQS quality system, plant calendar, program order, mixing room, and DracMaster data.

The end users of BiDrac Ecosystems use the web environment of SAP BO to make their own reporting or to consult the existing ones. Every user can analyse what, where and when he needs. For this reason, this was a big challenge at the beginning of the project.

4.2 Advanced Monitoring and Descriptive Analytics

The application of the proposed methodology is illustrated presenting several use cases in real scenarios. These use cases are developed on edge in a local server to show the goodness and robustness of the BiDrac Data Lake. Future work is to replicate them in the DPP.

A. Sensor validation and reconstruction use case

The BiDrac solution has been applied to the Paint Shop process in this paper for illustrative purposes extending the preliminary results presented in Sanz et al. (2017). The data validation/reconstruction module has been applied to the sensors of the Workshop 4 (T4) and 5 (T5) (see Fig. 8).

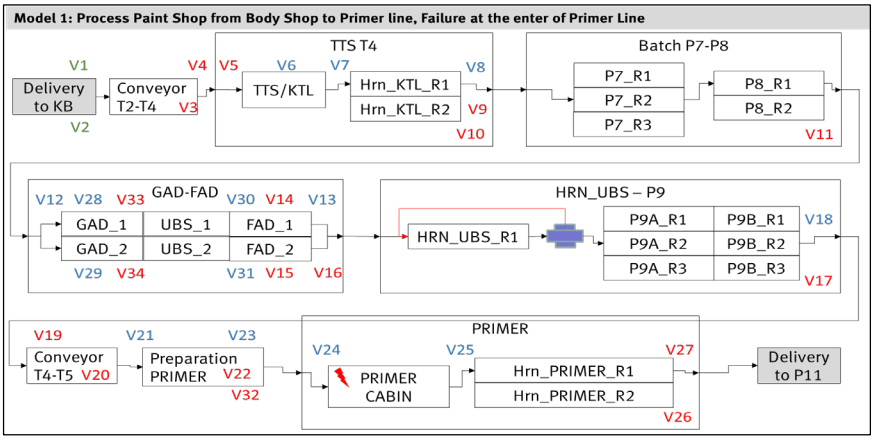


Fig. 8: Model for the Production Flow analyses from T4 to Primer line in T5

A model that relates every sensor with the physically related ones will be obtained using historical data. This model will be used to check the consistency of the sensor measurements to validate or invalidate the data following the methodology presented in Garcia *et al.* (2016).

Figure 8 presents the 34 sensors considered that belong to three control systems (the two-monitoring systems in the Paint Shop and the Fabrik Information Systeme (FIS) that is the production scheduling system used in Seat, S.A.). This set of sensors includes 18 variables from the Sicalis-PMC Control System, 14 variables come from the MFS Control System and 2 from the FIS System.

Using these 34 variables and basic mathematical flow relations by means of the structural analysis using the ranking algorithm described in Blanke *et al.*, (2016), 21 analytical redundancy relations (ARR) can be obtained. Every ARR allows generating a residual that will be used in the fault diagnosis module. Figure 9 details the complete list of variables and ARRs.

The structural method used to generate ARRs allows also to build the Fault Signature Matrix (FSM) where for each ARR (row) is indicated which fault (columns) in sensitive by means of a one in the crossing cells as presented in Fig. 10.Through FSM the residuals are detected as inconsistents (Escobet et al., 2012) and the fault can be detected and isolated. To enhance the robustness of the fault detection, an adaptive threshold for every residual is used through a set-membership approach as proposed (Puig, 2010). If the residual value is smaller than its threshold, the ARR is consistent; otherwise, if it is larger it is indicated as inconsistent.

ID_Rel	Form_Rel	System
R1	V1(k)=V2(k)	AS
R2	V3(k)=V1(k)-V4(k)+V3(1)	AS-PMC
R3	V6(k)=V5(k-T1)	MFS-PMC
R4	V7(k)=V6(k-T2)	MFS
R5	V8(k)=V7(k-T3)	MFS
R6	V10(k)=V5(k)-V9(k)+V10(1)	MFS-PMC
R7	V10(k)=V5(k)-V8(k)+V10(1)	MFS-PMC
R8	V11(k)=V8(k)-V12(k)+V11(1)	MFS-PMC
R9	V14(k)=V28(k)-V30(k)+V14(1)	MFS-PMC
R10	V15(k)=V29(k)-V31(k)+V15(1)	MFS-PMC
R11	V13(k)=V16(k)	MFS-PMC
R12	V17(k)=V138k)-V18(k)+V17(1)	MFS-PMC
R13	V19(k)=V18(k-T4)	MFS-PMC
R14	V20(k)=V19(k)-V21(k)+V20(1)	MFS-PMC
R15	V22(k)=V21(k)-V23(k)+V22(1)	MFS-PMC
R16	V23(k)=V21(k-T5)	MFS
R17	V24(k)=V23(k-T6)	MFS
R18	V25(k)=V24(k-T7)	MFS
R19	V27(k)=V25(k-T8)	MFS-PMC
R20	V26(k)=V24(k)-V27(k)+V26(1)	MFS-PMC
R21	V4(k)=V5(k)	PMC

Fig. 9: ARRs obtained from the production model presented in Fig. 8

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31
R1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R4	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R5	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R6	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R7	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R8	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
R10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
R11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
R13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
R14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
R15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
R16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
R17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
R18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
R19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
R20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
R21	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 10: Production model FSM. (own elaboration)

As an example of the diagnosis result obtained, Fig. 11 presents residuals R6 and R7 that are generated comparing the occupancy measured in the TTS_T4 process and the estimated one. From these residuals, residual R7 shows coherence

between the signals while residual R6 incoherence. Matching this situation with the FSM, the diagnosis result indicates a fault in sensor V9 that provides the measured output volume of the TTS T4 PMC.

This methodology is very useful to guarantee the consistency and quality of the sensor measurements collected from the two monitoring systems. In case that some sensor measurement is invalidated it can be reconstructed by means of its model.

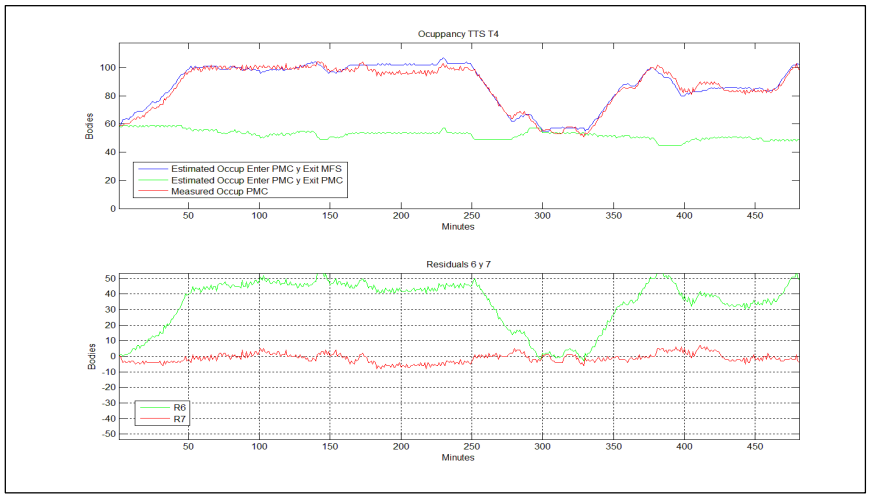


Fig. 11: Residuals R6 and R7 monitoring occupancy sensors of TTS T4

B. Prognosis use case

Once the application of sensor validation/reconstruction module has been presented, the prognosis module will be illustrated. After the sensor values are validated, conclusions can be extracted regarding the volumes and occupancies in the different phases of the painting process identifying potential problems (as e.g., body absences, blockages and breakdowns).

The prognosis module will be used to evaluate the Remaining Useful Life (RUL) based on a predetermined Failure Threshold (FT) based on the maximum occupancy as follows

$$RUL \in N \mid \hat{y}(t + RUL \mid t) = FT \tag{1}$$

where $\hat{y}(t+RUL\,|\,t)$ is the RUL-step ahead forecast at time *t* of the corresponding predictive model ($\hat{y}$).

To derive the predictive models from the data collected, the Brown's Double Exponential Smoothing has been used (Brown et al., 1963). Itis based on the following multi-step forecast formula

$$y_1(t) = \alpha y(t) + (1-\alpha) y_1(t-1) \tag{2}$$

$$y_2(t) = \alpha y_1(t) + (1-\alpha) y_2(t-1) \tag{3}$$

$$\hat{y}(t+h\,|\,t) = \left(2+\alpha\frac{h}{1-\alpha}\right) y_1(t) - \left(1+\alpha\frac{h}{1-\alpha}\right) y_2(t) \tag{4}$$

where *h* is the forecast horizon and α the smoothing parameter. The parameter α is obtained from historical data using parameter estimation by means of the least squares' algorithm.

Paint drying is a critical process while the time spent inside the oven by the body should not exceed the maximum allowed heating time (aprox. 25 min) and temperature. Figure 12 presents the effect in the Primer Cabin of a robot failure at time *t* = 250 minutes. As a result of this problem, an increase of occupation is observed in the Primer process and in the rest of subsequent processes upstream of the UBS oven. The occupation of the Preparation Primer shows no change because their occupancy was already saturated. The T4-T5 Bridge saturates very fast since it is very close to maximum occupancy while the P9 batch starts increasing its occupancy. Because of the drop in Primer occupancy and the observed rise of the P9 batch occupancy, an anomaly is detected at the entrance (block) of the Primer process. By means of the proposed prognosis module, the RUL (corresponding in this case to the time of reaching the maximum capacity of P9 batch) can be estimated, preserving the maximum capacity of the UBS Oven. This will allow ensuring the bodies evacuation of the UBS Oven and avoid to block the UBS exit with the vehicles inside the process.

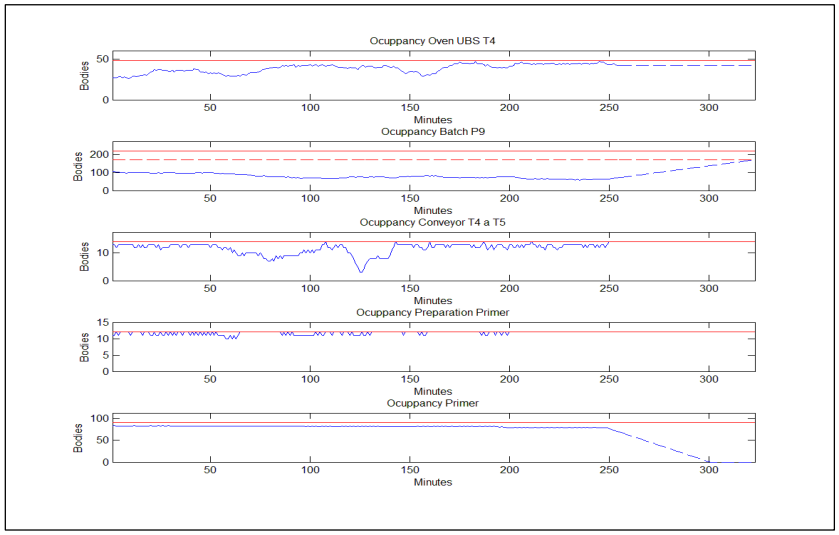


Fig. 12: Effects in the occupancy of the processes when a robot failure occurs

The prognosis module will be able to estimate when the drop in Primer occupancy will create a problem at the vehicle entrance of the painting process. After the problem detecting, the occupancy forecasting based on (2)-(4) with smoothing parameter $\alpha = 0.8$ is used to predict the occupancy in P9. This will allow determining the RUL by using (1) considering FT equal to the maximum occupancy of P9 batch minus the maximum occupancy of the UBS Oven to ensure the UBS Oven is not going to be blocked. The obtained RUL is equal to 68 minutes. Thus, if in this time the robot failure in the Primer Booth cannot be solved, the UBS Oven will be blocked.

4.3 Machine learning use case to predict HVAC energy consumption.

This use case has been developed in a local server to show the goodness and robustness of the BiDrac Data Lake. Future work will be devoted to replicate them in the DPP.

The issue of energy performance of industrial installations is of great concern to management nowadays as it translates to cost. Some industries have adopted energy savings targets for their industrial processes to reduce air pollution and climate change in urban areas as well as regionally and globally.

The plant purchases electricity and natural gas from the utility's companies. Electricity is used to power the equipment. Natural gas is mostly used for space heating and paint curing. Main energy conversion and transmission happens at the Energy Centre Purchased energy from the utility companies is converted to different energy forms (Technological Heating (steam), chilled water, compressed air, and so on) to cover the main production area needs.

The main goal of this use case is to use information about time and weather to predict energy demand of one specific process based on historical data and weather predictions. Easy-to-implement models with minimal input requirement and high accuracy have been considered. Such models will benefit managers of facilities, smart grid and industrial areas commissioning projects. For industry facilities, if they can predict the energy use of all their installations, they can make plans to optimize the operations of chillers, boilers and energy storage systems. The model will produce accurate energy demand forecasts that the Energy Centre can use to decide the optimal amount of electricity to purchase in the future and minimize cost. In industrial areas commissioning, engineers need to verify the energy savings after energy-saving measures have been implemented. However, it is difficult to have enough data points with the same conditions before and after the changes. Therefore, engineers need to interpolate and/or extrapolate the data (Cugueró-Escofet et al., 2016). This is also an important application of this research. Questions of the metering and data systems need to be made clear before modelling. Regression models correlating the energy consumption with the weather information and productivity or simple time series models with historical data are both good choices. Detailed physical models or statistical models can be built based on the data availability.

In this use case, different Machine Learning methods have been applied to the same dataset to:

1. Identify clusters (Unsupervised Learning), at this point expert knowledge is used in the dataset to discover the inherent groupings in the data,
2. Predict energy demand using the clustered inputs (Supervised Learning) and
3. Prepare, for furthers studies, the inputs needed to optimize the program of the controllers (PIDs) in different parts of an air supply group (Reinforcement Learning). Considering the set-points for the temperature and humidity inside the coating cabin and the State Meteorological Agency (AEMET) weather

predictions, a reward function that minimizes the energy consumption is added.

In brief, very different algorithms with very different goals working on the same dataset are concatenated to maximize the knowledge discovered to solve a common problem: energy saving based on time (working or non-working hours), weather (data from meteo-station and AEMET predictions) and sensor data involved.

Machine Learning is frequently applied to energy demand prediction and energy saving in industrial areas. Energy modelling and analyses have been widely studied to understand how and where the energy is used inside of the industrial processes (trends and patterns). The most sensitive variables affecting energy consumptions have been investigated identifying the different clusters and finding the best model that maximizes the accuracy for the energy predictions. While considering the current process metering status, the proposed approach has progressed in information sharing and improved suggestion determination.

There are many papers out there regarding the topics of energy demand prediction and energy saving. But it is more difficult to find studies that apply complementary Machine Learning techniques that solve the same problem from different perspective. This is one of the reasons why the scope of the use case includes the techniques described above applied in same season data (summer) from two consecutive years and with sensor data of one specific installation.

The industrial process environment is controlled through an HVAC system. For the most part, heating energy is provided through hot water from natural gas and cogeneration system, and cooling energy is provided through chilled water, mainly from electricity. One of the main causes of fluctuation in the monthly purchased energy is local seasonally weather changes. During the summer months, when the weather is hot, the heating energy (hot water/ technological heating (steam)) for processes is at the lowest point, but chilling energy (chilled water) for spacing cooling is at the peak. On the other hand, during the winter months, electricity used for generating chilled water is at the lowest point, but the natural gas for hot water is at the peak. This is one of the reasons why natural gas and electricity show a seasonal trend. It is also known that the cogeneration system runs at its full capacity year-round, all energy consumptions show a stable linear trend during all the seasons of the year.

Coating cabin, as painting spray booths, are small separate rooms isolated from the Paint Shop areas to prevent particle matters and gases like volatile organic compounds (VOCs) form paint to release into the working environment. Meanwhile, the painting spray processes required controlled temperature and humidity to provide a high-quality finish. It needs certain amount of air blowing from the roof of the booth to collect the sprayed paint and prevent residuals from affecting the next coming body. It is known that the energy used air conditioning to maintain the booth environment is huge.

In steady state, temperature, humidity and flow rate of the air inlet are controlled to be constant. The sensors in the unit will measure the temperature and relative humidity to the inlet air in the different parts inside an air supply group. The measured temperature and humidity will be used to compare with target parameter. Controllers (PIDs) will decide whether the air needs to be dehumidified, heated or cooled. Direct heating and cooling process is straightforward. The air goes through the heat exchanger (hot water heat exchanger for heating or chilled water cool exchanger for cooling) to reach the target temperature. Humidity is controlled through nozzles to increase water content. The studied case uses a cooling process for dehumidification. To include the weather information in the regression model is a good idea to make the model more informed and robust. The data studied is the summer one, so our electric consumption is going to be at the peak.

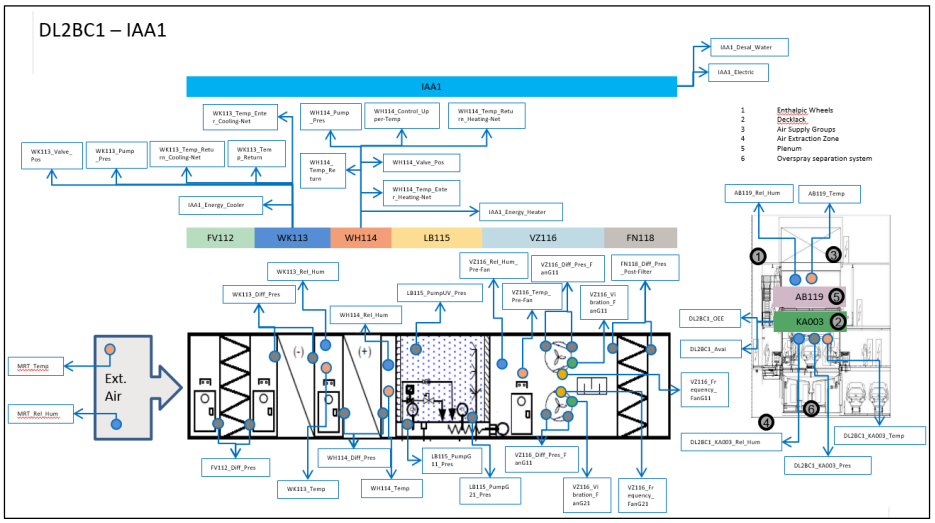


Fig. 13. Vertical Section of an HVAC of a Coating Cabin and sensors installed

The first step is to understand the data system. “Where are the meters that record the data located?" (see Fig. 13). Questions of the metering and data system need to be made clear before modelling. For plants that lack data systems, it is required either to install feasible meters for data collecting or use utility bill information instead.

Before describing each variable in the data set, it is necessary to relate the parts of a Coating Cabin, as appear in its vertical section:

- Enthalpic Wheels
- Air Supply Group
- Plenum
- Coating Cabin / Station
- Overspray Separation System
- Air Extraction

To predict the energy consumption of a coating cabin, the first step is to classify the different types of energy that can be found. Actually, for some large-scale coating cabins in industrial processes, there are seven types of energy consumption: electricity, chilled water, technological heat (steam), gas, water, demineralized water and compressed air. Chilled water is for cooling and steam is for heating.

In one of the most intensive energy processes, it is possible to find the relation between many different energy forms: hot and chilled water for process conditioned environment, electricity for power equipment and robots, compress air for coating process and robots.

The energy demand in every process can be calculated through enthalpy (before and after each phase of an air supply group, temperature and humidity are measured). In a scenario in which the air needs to be dehumidified, energy demand is the sum of enthalpy change in cooling process and enthalpy change in heating process. Dataset includes several meters for energy consumption, avoiding using enthalpy calculations unless the data collected are not significant or incorrect.

Here, it is considered one station of a coating cabin and it is analysed the energy consumption data of May, June and July in two consecutive years, 2018 and 2019, focusing on summer seasonality data. For each variable in our dataset, the control system provides us data hourly, that it, 24 data/day.

The temperature from our meteostation and the temperature inside our coating process are relevant variables for the use case (see Fig. 14). The temperature inside the

process should be maintained in the range 23+/-2 ºC with a PID control system to obtain good coating results.

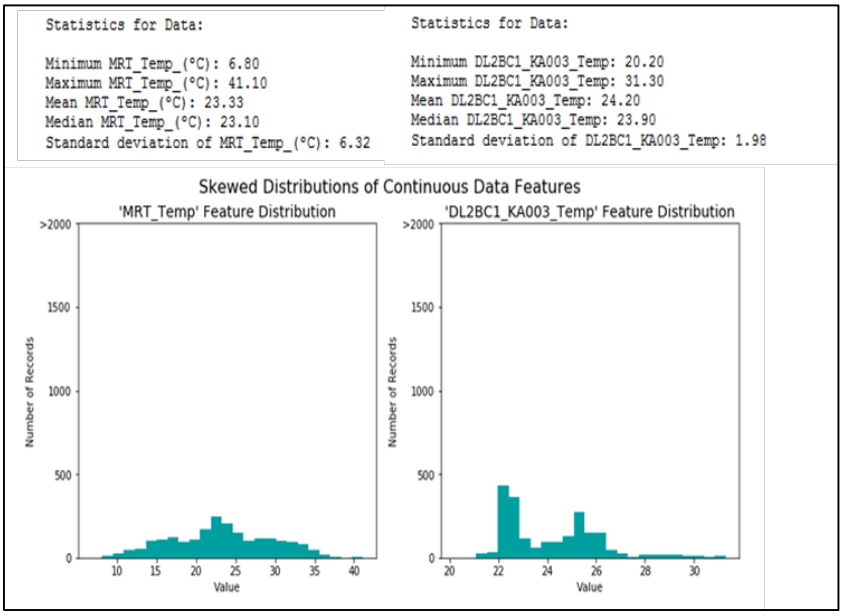


Fig. 14. Statistics of temperature outside in Martorell and inside the DL2 Coating Cabin.

Inside the coating cabin the standard deviation for temperature is lower than the outside temperature. This result is good because the air supply group is working well, a priori, and also because seasonality component is present inside our dataset (summer). In the considered scenarios, PID controllers are working under the summer program.

	IAA1_Electric	IAA1_Energy_Heater
count	2138.000000	2138.000000
mean	31.969130	6.616932
std	23.053808	7.430390
min	0.000000	0.000000
25%	13.800000	1.000000
50%	29.200000	4.000000
75%	48.700000	9.000000
max	180.300000	51.000000

Fig. 15. Statistics of energy consumptions for the entire HVAC and for the heating battery of the DL2 Coating Cabin

In summer, the bigger consumption is on the IAA1_Energy_Cooler variable (Cooler Battery), due to high temperatures in Martorell, but the data of consumption of the IAA1_Energy_Heater (Heating Battery) is not depreciable (because of the night lower temperatures) (see Fig. 15).

One of the methods to reduce energy consumption is the modification of the set point temperatures adapted to the external climate conditions but maintaining coating process specifications. In particular, as above commented, this process has to respect a temperature restriction, while it should be maintained in the range, 23+/-2 ºC. Out of this boundaries quality of coated bodies could be affected because of the quantity of surface defects arise. So, the coating environment is tightly controlled, temperature and humidity are sensitive variables. To introduce how the inferencing technology can be used in the energy management, a pattern-based energy consumption analysis by chaining Principal Component Analysis (PCA) and logistic regression is presented to correlate energy and operations and further use the power data to predict when operation events of interest (e.g. start up, idle, peak operation, etc.) occur, resulting in determining how current energy usage levels in manufacturing operations compares to the optimal usage patterns (Oh, 2016).

Although, the goal is to predict the IAA1_Energy_Cooler variable, so a PCA analysis of the Heating Battery variables is carried out.

Computing pairwise correlation of columns is also a good idea to reduce the number of variables in furthers analysis, and, also when the unsupervised analysis is applied, PCA with a subset of variables.

The final model created for this use case combines several different machine learning techniques. This problem is, from a macro view, an example of a machine learning regression and classification task because it requires predicting a continuous target variable (energy consumption of the air supply group) based on one or more explanatory variables (values from sensors installed). This problem is a supervised task because the targets for the training data are known ahead of time and the model will learn based on labelled data.

Going deeper in the macro view described above, a two-phase use case is proposed:

1. Unsupervised Learning approach: PCA, K-Means.
2. Supervised Learning approach: Xgboost, GradientBoostingClassifier, BaggingClassifier, scoring method using GridSearchCV() and a decision tree regressor to tune the model.

In order to clarify these methods, a description for each one and its parameters is given. Principal Component Analysis is a classification method that is often used to reduce dimensionality of large data sets but still contains most of the information in the large set. Principals components are eigenvectors of the data's covariance matrix. When a large dataset is to be clustered into a user specified number of clusters (k), which are represented by their centroids, K-Means will cluster the data by minimizing the squared error function (MSE), and often misclassifies some data due to outliers; also, the time complexity will be greater. To overcome these problems, principal components analysis (PCA) can be used to reduce the dataset to a lower dimension, while ensuring that the least information is lost, and providing a better centroid point for clustering. K-Means clustering partitions a dataset into different groups of similar objects. Clusters that are highly dissimilar from the others are regarded as outliers and discarded. Logistic regression is an efficient regression predictive analysis algorithm. Logistic regression is used in the description and analysis of data to explain the relationship between one dependent binary variable and one or more independent variables.

XGBoost (Chen and Guestrin, 2016; Nielsen, 2016) is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. It gives slightly better results than GradientBoostingClassifier. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting. The 'xgboost' is an open-source library that provides machine learning algorithms under the gradient boosting methods. The two reasons why use XGBoost are also the two goals of the project: Execution Speed and Model Performance. Generally, XGBoost is fast. Really fast compared to other implementations of gradient boosting. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems.

Gradient boosting is an approach where the new models created predict the residuals or errors of prior models and then add together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modeling problems.

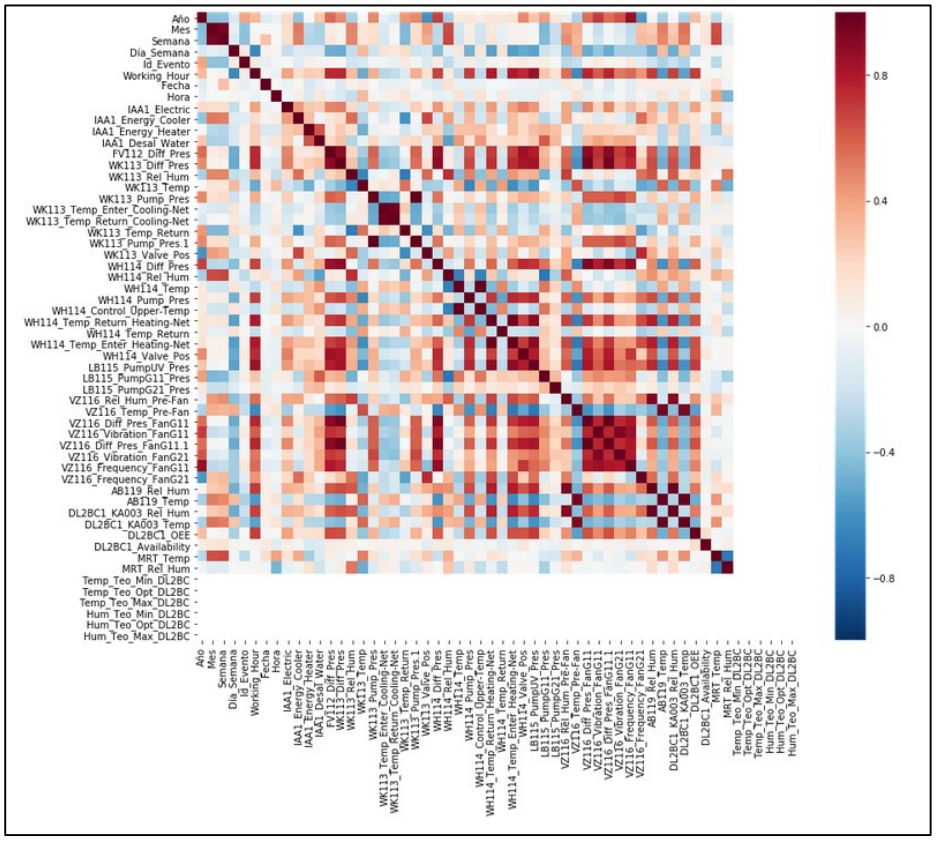
Bagging is an ensemble machine learning algorithm that combines the predictions from many decision trees. Bagging is provided via the BaggingRegressor and the BaggingClassifier classes. Both models operate the same way and take the same arguments that influence how the decision trees are created.

Machine learning algorithms have hyperparameters that allow you to tailor the behaviour of the algorithm to your specific dataset. Hyperparameters are different from parameters, which are the internal coefficients or weights for a model founded by the learning algorithm. Know which values use for the hyperparameters of a given algorithm on a given dataset is a challenge, therefore it is common to use random or grid search strategies for different hyperparameter values. The more hyperparameters of an algorithm are needed to tune, slower the tuning process will work. Therefore, it is desirable to select a minimum subset of model hyperparameters to search or tune. The most important parameter for bagged decision trees (BaggingClassifier) is the number of trees (*n_estimators*). In the case of Random Forest, the most important parameter is the number of random features to sample at each split point (*max_features*). The gradient boosting algorithm has many parameters to tune. There are some parameters pairings that are important to consider. The first ones are the learning rate, also called shrinkage or eta (learning_rate in [0.001, 0.01, 0.1]), and the number of trees in the model (n_estimators [10, 100, 1000]). Both could be considered on a log scale, although in different directions. Another important pairing is the number of rows or subset of the data to consider for each tree (subsample in [0.5, 0.7, 1.0]) and the depth of each tree (max_depth in [3, 7, 9]). These could be grid searched at a 0.1 and 1 interval respectively, although common values can be tested directly.

Grid search is the process of performing hyper parameter tuning to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified. GridSearchCV() does for each iteration, a test with all the possible combinations of hyperparameters, by fitting and scoring each combination separately.

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision trees regression normally use mean squared error (MSE) to decide to split a node in two or more sub-nodes. As a supervised machine learning model, a decision tree learns to map data from outputs in what is called the training phase of model building. During training, the model is fitted with any historical data that is relevant to the problem domain and the true value we want the model to learn to predict. The model learns any relationship between the data and the target variable. When we want to make a prediction the same data format should be provided to the model to make a prediction. The prediction will be an estimate based on the train data that it has been trained on. Decision trees regression normally use mean squared error (MSE) to decide to split a node in two or more sub-nodes.

Correlation matrix (Fig. 16 and Fig. 17) and PCA (Fig. 18) are applied to the data set. The first and second principal components (PCA(n_components=11)) explain 93.76% of the data variance, while the first four principal components explain 99.36% of the data variance. That is a good approach because considers the seasonality of data. Data from May to July months of two consecutive years, 2018 and 2019, are considered.



765 Fig. 16: Correlation matrix

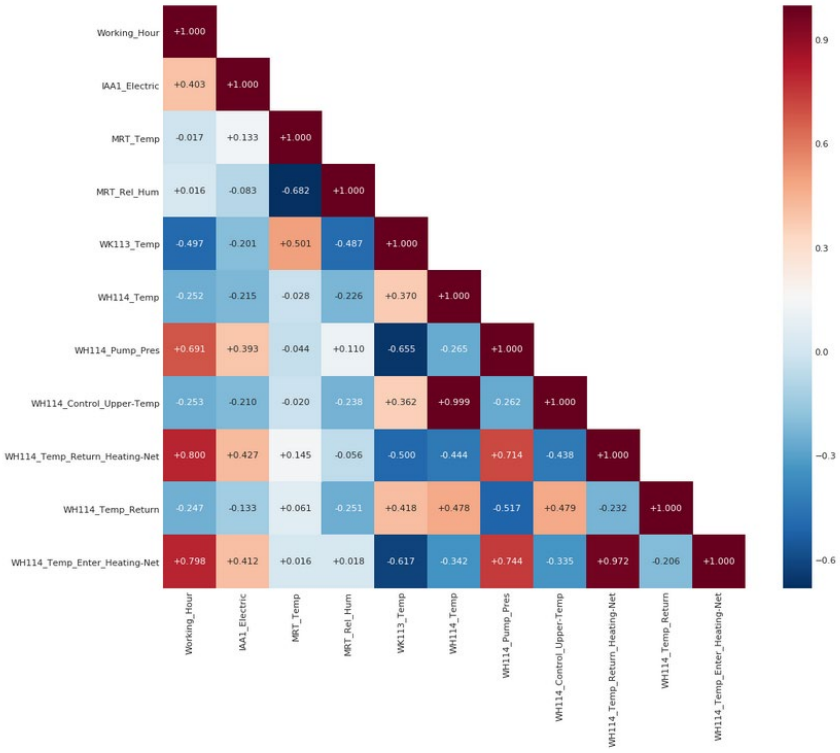


Fig. 17: Detail of the correlation matrix presented in Figure 16

From the heating map of the dataset, a data subset is extracted (see Figure 17) including heating battery signals and other needed to induce a good medialisation of our data (called semi-unsupervised learning because technical knowledge is added to the unsupervised techniques).

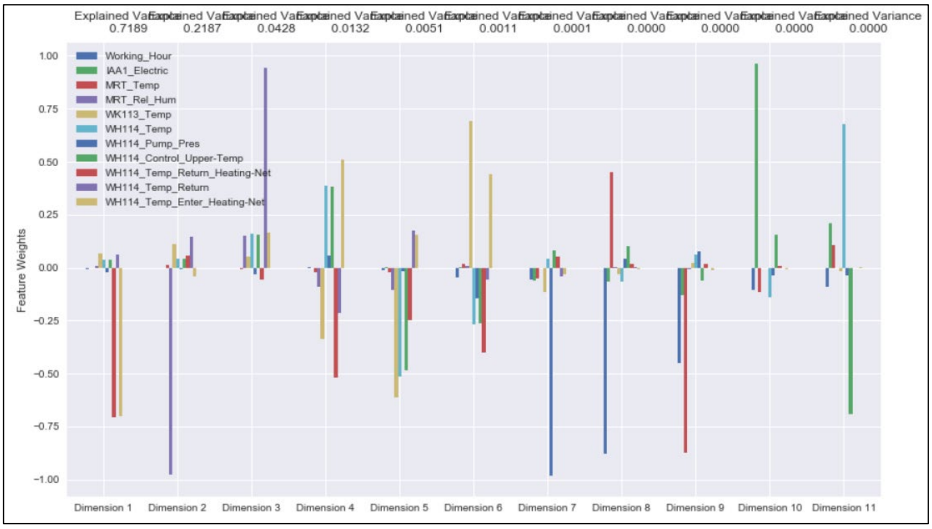


Fig. 18: PCA of the data set

The data exploration had underscored the need for dimensionality reduction. Luckily, one common technique for reducing the number of features, Principal Components Analysis (PCA), is an unsupervised technique that does not require an understanding of the physical representation of the features. PCA played a crucial role in reducing the input number of features into the algorithm and as discussed in the Data Pre-processing section. Figure 18 presents the result of the PCA analysis of the considered data set. Knowing that the first and second principal components explain 93.76% of the variance, considering a rough hypothesis of how many clusters to expect, a K-means clustering algorithm is used where the number of clusters is 2.

After our unsupervised Learning Approach, the output was a cluster prediction for each row in the data subset. So, at this point, the characteristics from our data, that are not obvious and stayed hidden inside the number of rows of data, have been identified.

To properly evaluate the performance of each model, a training and predicting pipeline is created to allow us to quickly and effectively train models using various sizes of training data and perform predictions on the testing data.

In the initial model evaluation, training the 1%, 10% and 100% of the data set (1710 samples) XGBClassifier has given the best results, followed by the GradientBoostingClassifier. Unless solve it considering time as a constraint, in this case the BaggingClassifier with few goodness penalization has been taken. In order to refine,

improve the results, from the XGBClassifier a grid search optimization is performed for the model over the entire training set (X_train and y_train). By tunning parameters to improve the model's F-score, the initial model used is: XGBClassifier(subsample=1.0, min_samples_split=3, random_state=0, verbose=3) and the initial parameters are: parameters = {'n_estimators': [100, 200, 500], 'max_depth': [8, 10], 'learning_rate': [0.05, 0.08, 0.1]}. After evaluating 18 combinations of parameters in 26.37 minutes, the final accuracy score and the F-score on the testing data is 0.5047. Optimized Model Parameters:

XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0, learning_rate=0.08,
max_delta_step=0, max_depth=10, min_child_weight=1,
min_samples_split=3, missing=None, n_estimators=200, n_jobs=1,
nthread=None, objective='multi:softprob', random_state=0,
reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
silent=None, subsample=1.0, verbose=3, verbosity=1)

Then, fine tuning the chosen model, GridSearchCV is applied. The model evaluation and validation are also applied. To validate the robustness of this model and its solution it might be applied sensitivity analysis. The final model gives a reasonable solution aligned with early expectations even though a need to improve final parameters appears. Small perturbations in training data affects the robust model because of the seasonality correlation between features and the need of a bigger dataset, which add more rows with the same columns.

Produce learning curves for varying training set sizes and maximum depths with *ModelLearning* and *ModelComplexity* functions. The final accuracy score and the F-score on the testing data after the fine tune is 0.4486. Parameter 'max_depth' is 9 for the optimal model.

The model final solution and its results are compared to the benchmark established before:

RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=10, max_features=3, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=3, min_samples_split=8,
min_weight_fraction_leaf=0.0, n_estimators=150, n_jobs=1,

oob_score=False, random_state=42, verbose=0, warm_start=False)

The accuracy of the logic by passing the data through a model that hast the feature importance method, the Ada Boost Classifier. The final model trained on reduced data final accuracy score and the F-score on the testing data is 0.4930.

The considered variable to be predicted is "IAA1_Energy_Cooler" since the summer season is analysed. The predictions can be compared with the historical data in the Figure 19 to show the obtained fitting.

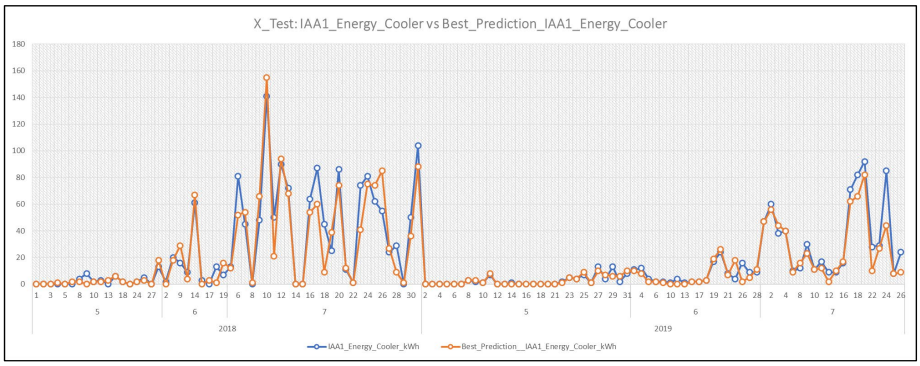


Fig. 19: Graph for the Performance of the Best Prediction vs real data on the variable IAA1_Energy_Cooler (Cooler Battery) on X_Test

XGBClassifier provides the best results as can be seen in Figs. 20 and 21, followed by the GradientBoostingClassifier, unless the time is a constraint. In this case, the BaggingClassifier with few goodness penalizations is considered.

Fig 20: Graph for the Performance Metrics for Three Supervised Learning Models

XGBClassifier

	1%	10%	100%
acc_test	0.331776	0.364486	0.478972
acc_train	0.440000	0.746667	0.973333
f_test	0.331776	0.364486	0.478972
f_train	0.440000	0.746667	0.973333
pred_time	0.015607	0.093751	0.296860
train_time	0.109395	1.296907	17.765822

GradientBoostingClassifier

	1%	10%	100%
acc_test	0.280374	0.364486	0.490654
acc_train	0.363333	0.746667	0.996667
f_test	0.280374	0.364486	0.490654
f_train	0.363333	0.746667	0.996667
pred_time	0.000000	0.015644	0.046875
train_time	0.312518	1.656254	19.000192

BaggingClassifier

	1%	10%	100%
acc_test	0.322430	0.362150	0.483645
acc_train	0.433333	0.716667	0.990000
f_test	0.322430	0.362150	0.483645
f_train	0.433333	0.716667	0.990000
pred_time	0.000000	0.015625	0.000000
train_time	0.031251	0.046875	0.515630

Fig. 21: Table for the Performance Metrics for three Supervised Learning Models

5. Conclusions

This paper has presented the BiDrac project, that aims to profit from the integration of computing, communication and control under an Industrial Cyber-Physical Systems (ICPSs) ecosystem combined with Artificial Intelligence and Industrial Internet of Things (IIoT) inside the Industry 4.0 paradigm on an Automotive Paint Shop Process.

BiDrac is the Ecosystem in which integrate Equipments, Technical Locations, PLCs, sensors, communication protocols, production networks, industrial networks, corporate networks, Complex Infrastructure Systems (CIS), ETL tools, Data Bases, Datawarehouse, Data Lake, Digital Platform, Algorithms, Machine Learning models, Artificial Intelligence models, Infrastructure, MES, ERPs and so on, and it is constantly growing to help solving problems in the Paint Shop. Focusing on the ICPSs, a Framework for applying Artificial Intelligence and Predictive Maintenance models have been presented and several use cases have been implemented for testing purposes.

Four different uses cases of BiDrac have been presented showing how three important problems can be addressed: sensor data validation/reconstruction using fault diagnosis techniques, prognosis of the fault effect in the Paint Shop process and finally energy consumption prediction.

As a future work, the proposed approach is being extended to other uses cases in the SEAT plant in Martorell profiting from the BiDrac framework already available.

Acknowledgements

Thanks to all the Departments in the Paint Shop and in the SEAT, S.A. who provided insight and expertises that greatly assisted the research. Special thanks to all the Maintenance Department for their contributions, collaborations and for the great work in the installations.

This work has been funded by SMART Project (ref. num. EFA153/16 Interreg Cooperation Program POCTEFA 2014-2020). Joaquim Blesa acknowledges the support from the Serra Húnter program.

References

Aivaliotis, P., Georgoulias, K., Arkouli, Z., & Makris, S. (2019). Methodology for enabling digital twin using advanced physics-based modelling in predictive maintenance. *Procedia CIRP*, *81*, 417-422.

Alcácer, V., Cruz-Machado, V. (2019). Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems, Engineering Science and Technology, an International Journal,Volume 22, Issue 3, 899-919.

Alqaryouti, O., & Siyam, N. (2018). Serverless Computing and Scheduling Tasks on Cloud: A Review. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *40*(1), 235-247.

AWS re:Invent 2019: Volkswagen takes production to the cloud (MFG204) (https://www.youtube.com/watch?reload=9&v=t9ED1BseanA).

Blanke, M., Kinnaert, M., Lunze, J. and Staroswiecki, M. (2016) "Diagnosis and fault-tolerant control". Springer-Verlag Berlin Heidelberg.

Box G.E.P., Jenkins G. M. (1970) "Time Series Analysis Forecasting and Control" Holden-Day

Broberg, J., & Goscinski, A. M. (Eds.). (2010). Cloud computing: Principles and paradigms (Vol. 87). John Wiley & Sons.

Brown, R.G. (1963) "Smoothing Forecasting and Prediction of Discrete Time Series". Englewood Cliffs, NJ: Prentice-Hall.

Brownlee, J. (2017). Arima for time series forecasting with python. Machine Learning Machinery (https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/).

Brownlee, J. (2017). Grid search ARIMA hyperparameters with python. Machine Learning Machinery (https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/)

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

Colombo, A. W., Karnouskos, S., Kaynak, O., Shi, Y., & Yin, S. (2017). Industrial cyberphysical systems: A backbone of the fourth industrial revolution. *IEEE Industrial Electronics Magazine*, *11*(1), 6-16.

Cugueró-Escofet, M.A., García, D., Quevedo, J., Puig, V., Espin, S., Roquet, J. A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network, Control Engineering Practice, Volume 49, Pages 159-172, 2016.

EFFRA (2019), The European Factories of the Future Research Association. Digital Manufacturing Platforms. Available online: https://www.era.eu/digital-manufacturing-platforms (accessed on 18 May 2019).

Escobet, T., Quevedo, J., and Puig, V. (2012) "A Fault / Anomaly System Prognosis using a Data- driven Approach considering Uncertainty," IEEE World Congress on Computational Intelligence, pp. 10–15.

Feng, L. & Mears, L. (2016). Energy Consumption Modeling and Analyses In Automotive Manufacturing Plant. Journal of Manufacturing Science and Engineering. 138. 10.1115/1.4034302.

García, C. G., Valdez, E. R. N., Díaz, V. G., Bustelo, B. C. P. G., & Lovelle, J. M. C. (2019). A review of artificial intelligence in the Internet of Things. IJIMAI, 5(4), 9-20.

Garcia, D., Creus, R., Minoves, M., Pardo, X., Quevedo, J. and Puig, V. (2016) "Prognosis of quality sensors in the Barcelona drinking water network". Conference on Control and Fault-Tolerant Systems, SysTol'2016, pp. 446-451.

Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., & Eschert, T. (2017). Industrial internet of things and cyber manufacturing systems. In Industrial Internet of Things (pp. 3-19). Springer, Cham.

Jirkovský, V., Obitko, M., & Mařík, V. (2016). Understanding data heterogeneity in the context of cyber-physical systems integration. *IEEE Transactions on Industrial Informatics*, *13*(2), 660-667.

Kazemi, Z., Safavi, A.A., Pouresmaeeli, S., Naseri, F. (2019). A practical framework for implementing multivariate monitoring techniques into distributed control system,

Control Engineering Practice, Volume 82, Pages 118-129, ISSN 0967-0661.

Kimball, R., & Caserta, J. (2011). *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. John Wiley & Sons.

Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. Procedia Computer Science, 88, 300-305.

Mqtt.org

Nielsen, D. (2016). *Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?* (Master's thesis, NTNU).

Liebstuckel, K. (2011). *Plant maintenance with SAP*. SAP PRESS.

Liu, Z., Chen, W., Zhang, C., Yang, C., & Chu, H. (2019). Data Super-Network Fault Prediction Model and Maintenance Strategy for Mechanical Product Based on Digital Twin. IEEE Access, 7, 177284-177296.

Oh, S., Hildreth, A. (2016). Analytics for Smart Energy Management. Tools and Applications for Sustainable Manufacturing. 10.1007/978-3-319-32729-7. ISSN: 1860-5168.

O'Leary, D. E. (2014) "Embedding AI and Crowdsourcing in the Big Data Lake," in IEEE Intelligent Systems, vol. 29, no. 5, pp. 70-73, Sept.-Oct. 2014. doi: 10.1109/MIS.2014.82

Pauker, F., Frühwirth, T., Kittl, B., & Kastner, W. (2016). A systematic approach to OPC UA information model design. Procedia CIRP, 57, 321-326.

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard business review*, *92*(11), 64-88.

Puig, V. (2010). Fault diagnosis and fault tolerant control using set-membership approaches: Application to real case studies. International Journal of Applied Mathematics and Computer Science. Vol. 20(4), pp. 619–635.

Ran, Y., Zhou, X., Lin, P., Wen, Y., & Deng, R. (2019). A Survey of Predictive Maintenance: Systems, Purposes and Approaches. *arXiv preprint arXiv:1912.07383*.

Reis, M., & Gins, G. (2017). Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. Processes, 5(3), 35.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.

Sajid, A., Abbas, H., & Saleem, K. (2016). Cloud-assisted IoT-based SCADA systems security: A review of the state of the art and future challenges. *IEEE Access*, *4*, 1375-1384.

SAP Business Objects (2020). https://www.sap.com/products/bi-platform.html

Sanz, E., Matey, J.L., Blesa, J., Puig, V. (2017). Advanced monitoring of an industrial process integrating several sources of information through a data warehouse. 2017 4th International Conference on Control, Decision and Information Technologies, CoDIT 2017, pp. 521-526.

Schwabacher, M. (2005). A survey of data-driven prognostics. In Infotech@ Aerospace (p. 7002).

Streitberger, H. J., & Dössel, K. F. (Eds.). (2008). Automotive paints and coatings (Vol. 1002). Weinheim: Wiley-Vch.

Stock, T., & Seliger, G. (2016). Opportunities of sustainable manufacturing in industry 4.0. *Procedia Cirp*, *40*, 536-541.

Tao, Fei & Cheng, Jiangfeng & Qi, Qinglin & Zhang, Meng & Zhang, He & Sui, Fangyuan. (2018). Digital twin-driven product design, manufacturing and service with big data. The International Journal of Advanced Manufacturing Technology. 94. 10.1007/s00170-017-0233-1.

Ustundag, A., Cevikcan, E. (2018). Industry 4.0: Managing The Digital Transformation. Springer.

Vogel-Heuser, B., Ocker, F. (2018) Maintainability and evolvability of control software in machine and plant manufacturing — An industrial survey, Control Engineering Practice, Volume 80, Pages 157-173.

VW-AWS, (https://www.volkswagenag.com/en/news/2019/03/volkswagen-and-amazon-web-services-to-develop-industrial-cloud.html) (2019).

VW-AWS Industrial Cloud Hub, (www.industrialcloudhub.com) (2020)

UNIDO (2020). Industrial Development Report 2020. https://www.unido.org/resources-publications-flagship-publications-industrial-development-report-series/idr2020

Wang, S., Wan, J., Li, D., & Zhang, C. (2016). Implementing smart factory of industrie 4.0: an outlook. International Journal of Distributed Sensor Networks, 12(1), 3159805.

Wollschlaeger, M., Sauter, T., & Jasperneite, J. (2017). The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0. IEEE industrial electronics magazine, 11(1), 17-27.

Wu, L., et al. (2012) "Improving efficiency and reliability of building systems using machine learning and automated online evaluation." Systems, Applications and Technology Conference (LISAT), 2012 IEEE Long Island. IEEE.

Wu, G., Talwar, S., Johnsson, K., Himayat, N., & Johnson, K. D. (2011). M2M: From mobile to embedded internet. IEEE Communications Magazine, 49(4), 36-43.

Yan, J. (2014). Machinery prognostics and prognosis-oriented maintenance management. John Wiley & Sons.