Industrial data management strategy towards an SME-oriented PHM

N. Omri[1, 2], Z. Al Masry[1], N. Mairot[2], S. Giampiccolo[2], N. Zerhouni[1]

[1] FEMTO-ST institute, Univ. Bourgogne Franche-Comté, CNRS, ENSMM,
24 rue Alain Savary, Besançon cedex, 25000, France
[2] SCODER
1 rue de la Forêt Z.A. l'Orée du Bois, Pirey 25480, France

Abstract

The fourth industrial revolution is derived from advances in digitization and prognostic and health management (PHM) disciplines to make plants smarter and more efficient. However, an adapted approach for data-driven PHM process implementation in small and medium-sized enterprises (SMEs) has not been yet discussed. This research gap is due to the specificities of SMEs and the lack of documentation. In this paper, we examine existing standards for implementing PHM in the industrial field and discuss the limitations within SMEs. Based on that, a novel strategy to implement a data-driven PHM approach in SMEs is proposed. Accordingly, the data management process and the impact of data quality are reviewed to address some critical data problems in SMEs (e.g., data volume and data accuracy). A first set of simulations was carried out to study the impact of the data volume and percentage of missing data on classification problems in PHM. A general model of the evolution of the results accuracy in function of data volume and missing data is then generated, and an economic data volume notion is proposed for data infrastructure resizing. The proposed strategy and the developed models are then applied to the Scoder enterprise, which is a French SME. The feedback on the first results of this application is reported and discussed.

Keywords: Small and medium-sized enterprises, Data-driven PHM, Industrial data management, Data quality metrics, PHM implementation strategy

1. Introduction

The industrial domain has evolved rapidly over time through the various industrial revolutions, from steam engines to cloud computing technology [1]. Throughout these changes, small and medium-sized enterprises (SMEs) have always played an essential role in global business. SMEs are defined as an organization with less than 250 workers, and in terms of turnover, they do not exceed 40 millions euros [2]. Thus, SMEs represent about 90% of all companies and actively contribute to job creation and global economic development [3]. In [4], the authors argue that the sustainability of SMEs depends on their ability to satisfy their customers effectively. However, they still cannot benefit from the industry 4.0 technologies because they suffer from many problems that limit their adoption of such technologies. These problems are mainly related to the fact that the paradigm industry 4.0 is created by big groups to be deployed in their production process. In [5], Sahal et al. define the concept of industry 4.0 as a general framework that enables industries with new elements of tactical intelligence using new technologies such as the internet of things and big data. Moreover, industry 4.0 can be defined as "a new approach for controlling production processes by providing real-time synchronization of flows and by enabling the unitary and customized fabrication of products" [6]. This industrial revolution is based on data collection and valorization not only for maintenance perspective but for a more general objective which covers performance evaluation, equipment (production) memory creation, process deviation prediction, and decision support [7]. From these new needs of industrial companies, the industrial maintenance paradigm has evolved to give birth to the Prognostics and Health Management (PHM) discipline.

PHM is a cutting-edge science that brings together different technologies such as diagnosis, prognosis, and decision support. In the case of data-based approaches, PHM presents solutions for data management and valorization. The main idea of this approach consists of studying the health state of an equipment and predicting its future evolution. This concept allows us to control the production process and to implement suitable maintenance strategies

[8]. Thus, data-driven PHM approach takes advantage of industry 4.0 techniques and technologies to make the plant factory smarter and improve their performance [1]. Nevertheless, SMEs are not, for instance, involved in the deployment of advanced manufacturing technologies [9]. They are expected to be in the third industrial revolution, which makes it more complex to carry out industry 4.0 technologies in their production processes. The complexity can be summarized in two main aspects: (i) the lack of dedicated documentation for such projects and (ii) the absence of the required technologies infrastructure.

In [2], Mittal et al. review the characteristics of SMEs and identify the issues that need to be addressed to ensure successful digital transformation within SMEs. In [10], the authors argue that SMEs are interested in including new technologies in their manufacturing process. However, they consider that standards belonging to these technologies can make their know-how accessible to their competitors. It is clear that SMEs feel the importance of digital transformation and the need to include new technologies such as PHM in their workshops, but they still fear the consequences of this transformation and its success. Up to our knowledge, there still no PHM guideline or strategy implementation for SMEs in the literature. In this paper, we here are interested in developing a first strategy for PHM implementation in SMEs. To do, the limitations of the SMEs are discussed to propose a set of best practices that facilitate the success of a PHM strategy in small companies. Moreover, the impact of the data quality (DQ) on the results of this strategy is addressed. To be more precise, DQ dimensions and metrics are reviewed, and the most critical data issues in SMEs are studied. These issues concern data volume and data completeness essentially. Then, a general model of the impact of these problems on PHM classification problems is proposed. This model is developed using a set of simulations conducted from several public datasets. Finally, the proposed strategy is deployed in a real case study, and the general evolution model is used to resize the data infrastructure.

The rest of this paper is organized as follows. The evolution of maintenance strategies and standards in the industrial domain as well as the gap between the PHM documentation and technologies are presented in Section 2. Then, industrial PHM concept and applications are described in Section 3 while discussing the factors that limits its applicability in SMEs. To fill this gap, a new strategy to implement data-driven PHM within SMEs is proposed in Section 4. Section 5 and 6 deal with the data management process and the impact DQ aspect on the PHM results. A real case study for Scoder factory is provided in Section 7. Finally, conclusions and perspectives are drawn in Section 8.

2. Historical perspectives and problem statement

As shown in Figure 1, data are always collected in the industrial domain. However, nature and utility of these data have evolved over time with the evolution of the manufacturing process. In fact, the new needs in terms of product quality and process optimization have forced companies to collect more data to increase the reliability and capability of their manufacturing processes [7]. This numerical transformation has been supported by the advances in the Information Technologies (IT) [11]. These technologies have enabled enterprises to have more accurate information about their production process. Thus, the valorization of the industrial data has been evolved from absenteeism rate calculation to a more global PHM framework which allows to optimize the maintenance function, to improve the production process, and to reduce the operational cost. In this section, we propose to review the evolution of the industrial data collection throughout the different industrial revolutions. Then, we propose to study the impact of this evolution on the evolution of the maintenance function within industrial companies. Moreover, the evolution of documentation that standardizes the data management process for maintenance applications is addressed (See Figure 1).

2.1. The evolution of manufacturing data

For a long time, data is generated throughout the manufacturing process. The complexity of these data depends on the complexity of the industry and its degree of development. Usually, data is recorded with different technologies, from paper to big data hubs. The utility of these data depends on the needs of the companies. The role of data has evolved in line with the evolution of the industrial activities, which are characterized by four revolutions (from industry 1.0 to industry 4.0). For that purpose, we propose to study the manufacturing data in line with these revolutions.

The first industrial revolution (industry 1.0). This revolution was triggered by the introduction of steam engines into the production process [12]. According to Tao et al. [11], this revolution has no impact on the collection, storage, and analysis of data. Only a few data are managed manually by the workers through papers, but the rest is stored in human memory as know-how. These data usually relate to workers and serve very limited purposes. Thus, information such as assistance, productivity, and performance are collected to evaluate worker performance.

The second industrial revolution (industry 2.0). The second revolution is characterized by mass production. In fact, electric machines are used with more sophisticated management principles (e.g., the Bessemer process, the Taylorism, etc.) [12]. Unlike the first revolution, Industry 2.0 has imported many changes in data management. As a matter of fact, more data are recorded in more formal documents such as charts and logbooks [11]. As a result, the utility of these data has been expanded to reach other applications such as operation planning, quality control, and failure rate.

The third industrial revolution (industry 3.0). The third revolution was introduced thanks to the development introduced in the domain of computers and semiconductors in the 1960s [13]. This revolution is characterized by the use of Logic Controllers, Computer Numerical Control (CNC) robotics, which gives birth to the concept of fully automated factories [12]. As a result, data are collected automatically, saved in computers, and managed by information systems. From these data revolution was born a set of information systems (eg. ERP, MES,etc.) to manage the maintenance, production, supply chain, and financial data [11].

The fourth industrial revolution (industry 4.0). The fourth industrial revolution was triggered thanks to the emergence of technology of Internet of Things (IoT), Artificial Intelligence (AI), and Big Data analytics, which are integrated into the manufacturing process. All these technologies make it possible to collect and manage a huge quantity of data from several heterogeneous sources [11]. These data describe the product throughout its life-cycle [14].

To sum up, the evolution of data collection, technologies and utility are evolved in line with the development of industries. Also, the data analysis techniques are evolved increasingly to valorize these data and extract knowledge from it. These techniques can be regrouped in a general framework such as the PHM [15]. The evolution of this concept is discussed in the next section.

2.2. Towards predictive manufacturing: The evolution of PHM

As mentioned above, manufacturing systems have evolved throughout four industrial revolutions. This industrial evolution has been accompanied by a change in the manufacturing functions such as the maintenance one [16]. This function was evolved from ”repair work” to PHM (See Figure 1). For a long time, maintenance is considered as repair work where machines are not maintained except in case of breakdown [17]. Since there is no available data for the machines, the prediction of the breakdowns was impossible. A little more time later, developments in the maintenance paradigm give birth to a new concept that is Time Based Maintenance (TBM). TBM is based on the breakdowns historic of the machine to introduce preventive maintenance planning. However, this strategy ignores many information (since they do not exist) that can affect machine degradation. Because of these drawbacks and the evolution of the data technologies, the TBM was replaced by a more sophisticated maintenance method which is the Condition-Based Maintenance (CBM). The CBM is based on preventive actions taken after the apparition of failure symptoms [18]. Thus, CBM allows analyzing data coming from machines to avoid failure [19, 20]. Despite this significant evolution in the maintenance paradigm, many disadvantages have been entrusted to the CBM. The drawbacks concern the cost-effectiveness specially and the scope; in fact, this method becomes useless and expensive in the case of non-critical components. Moreover, CBM ignores many aspects such as safety, reliability and the economic aspect. As a result, the CBM concept has evolved to give birth to the PHM discipline [7]. PHM goes beyond CBM scope and integrate many aspects such as logistics, security, reliability, mission criticality and cost-effectiveness [21].

2.3. The evolution of data-driven industrial PHM Standards

PHM process involves different domains and applications, making it challenging to develop a general approach for PHM installation in enterprises. In this context, Vogl et al. [22], assert that until now, there was no consistent guide for conducting a PHM study, and even the standardization of specific methods for PHM was ineffective since each application had its own requirements. As the PHM paradigm has evolved from a maintenance function to a more global framework, documentation in this field has also evolved from a general maintenance context (FD X60250 standard) to a closer PHM context (ISO 13374). Many international institutes and organizations have worked in this domain and proposed standards and guidelines for PHM process implementation. These organizations include the Air Transport Association (ATA), the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), the Society of Automotive Engineers (SAE), and the United States Army (US Army) [15].

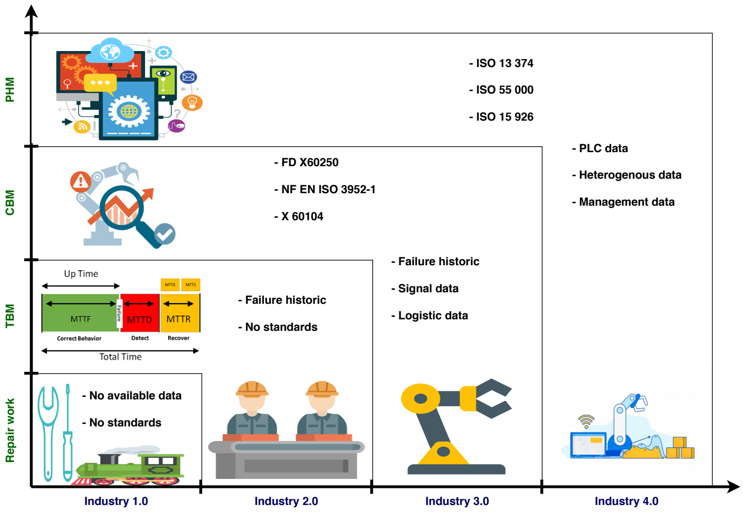


Figure 1: The evolution of maintenance paradigm within the industrial revolutions.

During the first and second industrial revolutions, no standards were identified in the context of asset maintenance and management, and even the few existing documents were only applied in the military domain. One of the first standards in this area is $X60104$ (1981), which defined the requirements of maintenance contracts. Then, the ISO 3952-1 was published to introduce the symbols system and to facilitate the documentation of machine components by modeling their interaction. In addition, there is the $FDX60250$ standard (1983), which presents technical maintenance documentation for users and recommendations for its implementation. Later in the 2000s and with the emergence of the data technologies, the maintenance documentation has evolved in order to follow the technological revolutions. In this context, the $ISO55000$ standard has introduced the concept of asset management and widen the maintenance scope to integrate related activities such as the planning, design, implementation, and review of asset management activities. $ISO15926$ was published for the representation of process plant life-cycle information via a data model with a consistent context for data definitions. In the same context of data management standards, the ISO 13374 series (Condition monitoring and diagnostics of machines) was introduced. This standard is presented in 4 parts:

- ISO 13374-1 aims to provide the basic requirements for open software specifications to facilitate the transfer of data among various condition monitoring software, regardless of platform or hardware protocols [23].
- The ISO 13374-2 goes beyond the data transformation and provides requirements for a reference information model and a reference processing model for an open Condition Monitoring and Diagnostics (CM&D) architecture [24].
- The third part of the ISO 13374 standard (ISO 13374-3 [25]) defines the communication requirements for any open CM&D systems to aid the interoperability of such systems [25].
- The ISO 13374-4 [26] details the requirements for the presentation of information for technical analysis and decision support in an open architecture for condition monitoring and diagnostics.

To facilitate the uses of these standards, the Machinery Information Management Open Systems Alliance (MIMOSA) publishes an open CMD information specification known as the MIMOSA Open Systems Architecture for Enterprise Application Integration (OSA-EAI) [27], which is free for download and compliant with the requirements outlined in ISO 13374-1 and ISO 13374-2. Based in the feedback of this standard, MIMOSA has elaborated another documentation (MIMOSA Open Systems Architecture for Condition Based Maintenance (OSA-CBM)), which is the most used in both the research and industrial domains [28].

Generally, the development and deployment of PHM within an industrial organization is a very complex task. In [15], Guillen et al. affirm that there is a gap in terms of PHM documentation. Establishing general methodological approaches to guide the design and implementation of PHM process is hence needed. Moreover, up to our knowledge, there are no documentations dedicated to SMEs. In [9], the authors concluded that SMEs are not involved in the deployment of advanced manufacturing technologies. Thus, SME's constraints are studied and discussed in the next section in order to propose a generic PHM implementation strategy.

3. Industrial PHM applications and their limits within SMEs

The increased amount of data in the industrial field requires appropriate treatment to meet the challenge of zero defect manufacturing [29]. In this context, data-driven PHM of industrial systems has attracted the attention of researchers and industrialists during the last decade [30]. Their works concern many fields such as the manufacturing, energy and transportation industries [31]. In this section, we review these studies while presenting the conditions for implementing a data-driven PHM in the industrial domain.

3.1. The PHM paradigm

Prognostics and Health Management is a science that study the health state of an equipment and predict its future evolution [7]. This concept allows to better control the systems and to implement suitable maintenance strategies [8]. In [7], the authors define PHM as "a set of tools that can be used in cascade or separately to monitor the health state of a system, predict its future evolution and/or optimize decisions". In [32], the authors affirm that PHM can be implemented using model-based or data-driven approaches. The first approach consists of building analytical models that are directly related to the physical processes which influence the health state of systems. Thus, a good comprehension of the physical process of components degradation and interaction is required. The second approach consists in using historical monitoring data to model the evolution of the system until a failure occurs. In this case, the understanding of the physical process of the system could not be necessary, but results only depend on the quality of historical data.

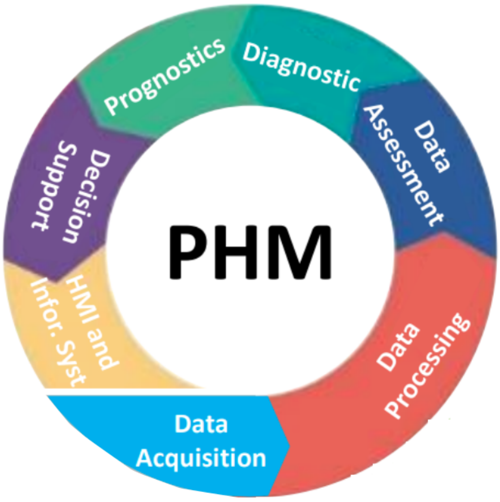


Figure 2: The traditional PHM cycle.

Traditionally, the PHM is decomposed in three main steps (observation, analysis and decision) which are detailed in 7 steps from data acquisition to the Human Machine Interface (HMI) as shown in Figure 2. This configuration of the PHM process assumes that the studied system is already defined and that the necessary detection and analysis resources are available. Most of the existing PHM works does not cover how to define the PHM system and assumes that the PHM will improve its performance [33]. However, in the industrial domain, there are many systems to be studied with different impacts on the production system. Moreover, in the case of SMEs, available resources are generally limited and their allocation need to be optimized to maximize the benefit of the PHM study. In [33], the authors state that choosing the most suitable PHM project is a difficult activity. Thus, a clear estimation methodology is therefore needed. This methodology can be based on PHM indicators and their expected evolution after the application of the PHM. Moreover, the financial impact of the PHM depends on the available resources and data. In fact, unless useful data (or the related resources) are available, a PHM project may be insignificant.

3.2. Industrial PHM applications

The PHM concept has been widely applied in the industrial domain [30]. In this context, Toshiba collaborate with NEC to develop an IoT-based PHM system. Thus, data are collected from the Toshiba's devices and saved in data centers managed by NEC, then Toshiba maintenance team analyzes the data to respond to the costumer's requests [31]. A similar collaboration have been developed between Nidec and IBM. In fact, PHM services are proposed by IBMs based on the collected data from Nidec's machine [31]. As for the automotive industry, cars from General Motors, Tesla, BMW, and other manufacturers are equipped by an application programming interfaces (APIs). The APIs allow applications built by third parties to use the collected data. This enables the development of applications for IoT-based PHM that add value by increasing connectivity, availability, and safety [31]. Concerning the construction and mining industries, they benefit from the development of the connectivity technologies in order to control their equipment, since working sites are generally in isolated locations. In this context, Komatsu [31] developed a data-driven PHM module in order to monitor and diagnosis their construction equipment via satellite communications. In [34], the authors assumes that PHM is a multidisciplinary activity that require significant time and effort which make it very expensive. Feldman et al. [35] applies the PHM concept to an electronics Line-Replaceable Unit (LRU) in the Boeing 737 aircraft. The results are obtained for 300 flights and shown that the PHM cost for each LRU is 700 $ (value is in 2008 U.S. dollars). These studies prove that PHM implementation requires an important investment which seems to be very expensive for an SME.

3.3. Factors that limit the implementation of an SME-oriented PHM approach

PHM is considered as a cutting edge technology that requires specific resources which make its implementation within SMEs a difficult task. In [7], the authors review the SMEs domain and propose to classify the constraints that limit the development of such approach into two main classes: Resources-based constraints and Organization-based constraints. Moreover, Müller et al. [36] conduct a study on German SMEs, and they affirm that standardization, personnel resources, financial resources, and a belief on digitization are unique constraints for SMEs to integrate advanced technologies. We here propose to study the constraints that limit the development of an SME-oriented PHM process: Human, Organizational factors, and Resources factors.

The human aspect is a crucial factor in the success or failure of the implementation of a data-driven PHM process. In this context, Lee et al. [37] affirm that data become useless unless its analysis by the right expert in the right context. However, the PHM process is generally complex and includes a huge volume of data that exceeds the user's capacity [38]. This disadvantage creates a kind of mistrust between the human and the PHM process, particularly in the case of incorrectly predicted events. Thus, a misunderstanding is created between the user and the PHM system. The PHM process requires human expertise to improve its performance, but the user feels that these technologies will replace him. So, the workers rely on false predictions to justify the uselessness of the PHM. In addition, the required resources and time for a PHM process are unknown. All these problems make the PHM a doubtful project for the SMEs managers. SMEs are characterized by central management, where the owners are involved in all the decision-making process from the strategic level to the operational level [39]. Thus, the first challenge in implementing a PHM process in an SME is to convince the manager of the effectiveness of these technologies. In addition, communication in SMEs is generally informal and very close between workers [40]. Hence, information that concerns the manufacturing process is not documented since they are kept in the mind of the manager and key workers [7]. Nevertheless, a reliable data valorization process inside companies requires a certain level of documentation that guarantees no loss of knowledge [41]. Usually, activities related to documentation and corporate memory creation are supposed to be useless since their added value is not yet explored in SMEs. This calls into question the limited financial resources in an SMEs, which limits the adoption of new technologies that seems to be expensive [42]. In [2], Mittal et al. conclude that SMEs are financially constrained, which affect their ability to adopt advanced manufacturing technologies. Moreover, SME's limited resources have affected the research and development field among small enterprises [43]. After all, it can be concluded that the development of SME-oriented PHM is limited by many constraints. Thus, an implementation strategy is required to guarantee the success of data-driven PHM approaches within SMEs.

4. Strategy to conduct data-driven PHM projects inside SMEs

Taking into account the previously detailed constraints, the best appropriate way to implement a data-driven PHM process within SMEs is to start by existing data. Before collecting new data, it is necessary to digitize the existing data. Since SMEs do not have a lot of resources (financial resources in particular) to install sophisticated data acquisition devices. It is recommended to use simple acquisition systems and use non-expensive storage solutions. Thus, smartphones and tablets are the simplest solutions when they are coupled with existing and free mobile applications. Free cloud solutions could be used to ensure real-time acquisition, but this solution calls into question the confidentiality and security of the collected data. Once the company's ordinary data is digitized, potential valorization applications are discussed. A matrix that links each data group to the valorization application with its associated benefit can help to set priorities for the data analysis phase. One should note that the SME world is not accustomed to sophisticated processes such as data-driven PHM, so it is better to think about working quickly with existing data to quickly provide useful results and prove the feasibility of the project. This step can convince the managers, implicate workers, and introduce PHM culture into the company at the same time. Workers are in central of any PHM strategy; thus, a user-friendly PHM framework should be developed where the user can communicate with it and integrate his expertise on it. Another important issue that can be started in parallel is the standardization of the communication process within the enterprise in order to guarantee the no loss of information. We here propose a set of best practices that should be followed to successfully implement the PHM in SMEs (See Figure 3). In particularly, we focused on three main elements which are detailed in the following: data inventory, scope identification and PHM metrics.

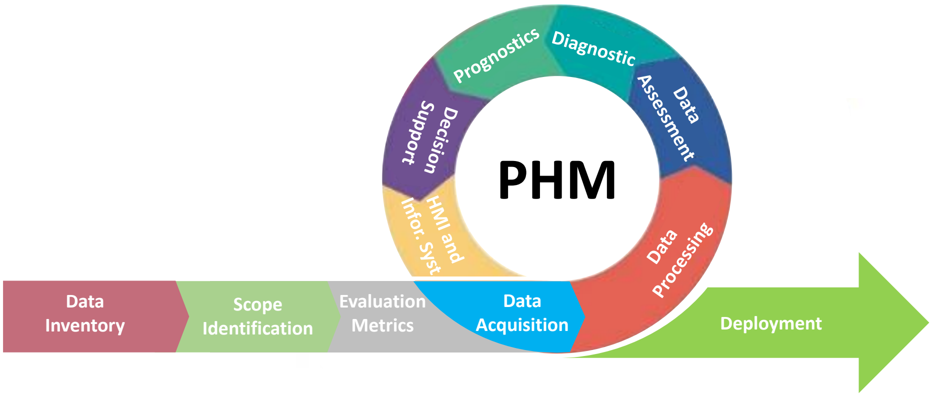


Figure 3: The extended PHM cycle.

4.1. *Data inventory*

Data inventory is a deep analysis of the circulating data around the manufacturing process. More particularly, data inventory is a quest to collect information about the existing data. These information concern different elements present in Table 1.

In addition to the process of collecting information about the data, data inventory also aims to regroup data around potential PHM projects and identify the missed data to accomplish these projects. To do, a list of necessary attributes for each system can be defined. This list depend on the nature of the studied system. In [44], the author propose a set of variables that can be collected to make PHM study on 8 different domains (see [44] for more details). This list can help to define the needed variables for each system and then to compare them with the existing ones in order to identify the missed variables.

4.2. Scope identification

Since available resource in SMEs are limited, it is mandatory to limit the scope of a PHM study and focus only in the relevant projects. Thus the potential projects identified in the data inventory step should be ranked and select the most relevant ones. Projects relevance depend on the following points:

- The available digital data;
- The data to be collected or digitized;
- PHM profitability.

To do, some techniques are proposed in the literature. As example of projects raking techniques we can cite the Analytical Hierarchical Process (AHP) [45] and the Data Envelopment Analysis (DEA) [46]. In the PHM case, some works are done to optimize the sensors selection [47]. The selection process is generally converted in an optimization problem that can be solved by the traditional techniques such as linear programming. The common point between these techniques is the assignment of an importance order to each characteristic in order to prioritize them and calculate a global score for each system. This score is used to rank the available equipment, to select the relevant PHM equipment applicability and to guide the investment plan to maximize the benefits of the study.

Information	**Description**
Title	Dataset name
Features	Dataset attributes
Purpose	Data creation aim
Type	Text, images, numbers
The owner of the data	Production team, Marketing team
Location of the data	ERP, Server
The volume of the data	1 GB, 1TB
The format of the data	Papers, CSV files
Data transfer	Data usgae per team
Update frequency	1 hours, 1 day, 1 week
Restrictions	Confidentiality, Accessibility

Table 1: Attributes of a Data Inventory Quest.

4.3. PHM metrics

One of the most important steps in a PHM study is the evaluation of the whole approach. To do, a set of metrics are needed to assess the objectives of the project either for big or small companies. Also, metrics are needed to better describe the performance of the used technologies to satisfy these objectives. In [48], the author proposes a set of PHM metrics in relation with the different PHM themes and theirs benefits. We here propose to classify the PHM metrics in relation with the characteristics of the company, the objectives, the collected data and the used PHM techniques.

Table 2 shows an example of PHM metrics that can be used in the different steps of a PHM study. Accordingly to the fixed objectives, a set of performance of the used techniques can be identified. Moreover and based on these performance, the data required data quality can be fixed. In real case study, it is difficult some times to ensures the

Company	Objectives	PHM tools	Data quality
R & D budget	Reduce maintenance frequency	Accuracy	Completeness
%of skilled workers	Improve products quality	False alarm	Time to value
Documentation level	Optimize resources allocation	Prediction lead time	Volume

Table 2: Example of PHM metrics in SME.

required data quality which necessitates a modification in the initial objectives [7]. PHM metrics are common to all sizes of companies, but one should think about specific metrics for SMEs in relation with their limits. In this context, the percentage of skilled workers, the research and development (R & D) budget and the documentation level can be used as PHM metrics in the case of SMEs.

To sum up, the proposed PHM strategy deals extends the PHM cycle (See Figure 3). This process is indeed very long, but in our opinion, it is the most adapted one for SMEs with few resources, as mentioned previously. The performances of data acquisition devices are limited in the case of SMEs which impact the data quality and consequently the PHM results. In the sequel, the data management phase is detailed. Moreover, this phase allows us to asses and improve the data quality, to define the data requirements and to resize the PHM infrastructure.

5. Industrial data management

Data are real-world objects with storage, retrieval, and development capabilities, and it can communicate over a network [49]. The process that deals with their valorization are called Data management. In [50], the authors start from the definition of ”management” which means ”the process of dealing with or controlling ... [and] having responsibility for ...”, and they claim that data management is more than ” deal with ”since it includes many tasks such as identifying resources to meet objectives, organizing these resources, defining and implementing a data management strategy to achieve a fixed set of objectives. For the Data Management Association (DAMA) [51], data management is ”the business function that develops and executes plans, policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data”. In the same context, another definition is proposed by Fisher in [52], which considers data management as “a consistent methodology that ensures the deployment of timely and trusted data across the organization”. In [22], authors affirm that data management concerns data collection, processing, visualization, and storage. Hereafter, we are interested in the first two steps (data collection and processing) by detailing the industrial data sources and studying the impact of the quality of these data on the processing phase.

5.1. Industrial data sources

As mentioned above, the implementation of the PHM must take into account the existing data in the enterprise and propose adequate solutions to deal with the problems that characterize these data. Here, we propose to clarify the data architecture in the manufacturing organizations and to detail the different data sources that can be used as input in a data-driven PHM process. In this context, data can be extracted from different levels and sectors of production. These levels are usually represented in a pyramidal architecture where information flows from the bottom to the top of the pyramid, unlike the control flows that flow from top to bottom [53]. Figure 4 represents the different data levels into an industrial company. From the top to the down, this pyramid is decomposed into Enterprise Resource Planning (ERP), Manufacturing Execution Systems (MES), Control Level, and Device Level.

ERP systems are created as a solution to a classic problem in the industrial field. This problem involves treating the activities and transactions separately, without any link between them [54]. For this, ERP systems provide a common systems platform that enhances data visibility. In [55], ERP is defined as a method for planning and controlling the required resources to respond to customer orders. These tasks are satisfied using a software package that manages the data flowing in the enterprise. These data concern a global representation of the companies, including the different

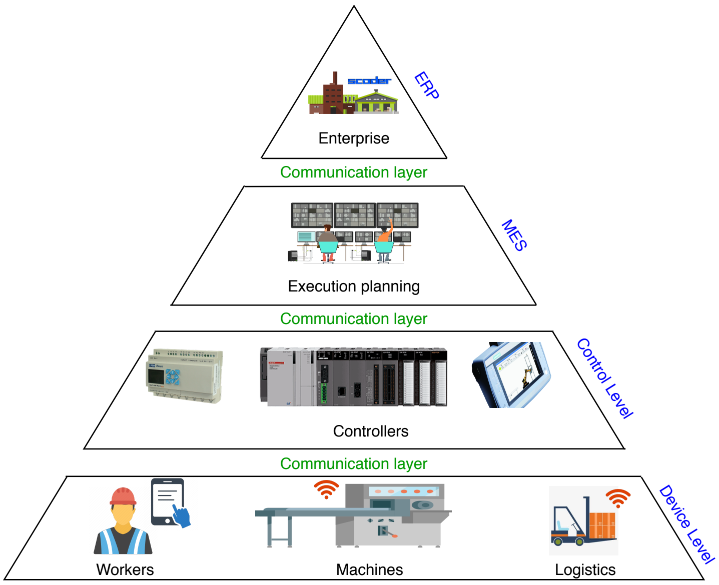


Figure 4: Automation pyramid in the industrial domain.

resources (human, machines, and raw materials). Also, the ERP database includes information about sales, historical production data, accounting, and production range [55]. ERP systems are used for long term activities planning without focusing on the shop scheduling to accomplish these tasks [53]. Unlike ERP systems, the MES software focuses on the digitization of the production process to enable real-time control of the different activities [56]. The MES provides information to optimize activities throughout the production process. Using current and accurate data, an MES system guides, initiates, response, and reports on workshop activities as they occur. [57]. In other words, the MES manages finer data in terms of granularity. These data concerns manufacturing instructions, design engineering data, the status of resources, the progress of activities, and all events that occurred during the production activities. The control level is composed of all forms of computer or programmable cards that control the evolution of the system state during its operational mode to detect and avoid failures. In addition, these digital computers are responsible for controlling the industrial environment to reflect the atmosphere of the workshop. The device-level is characterized by low-level devices that are represented by the machines or sensors. These devices generate data that is needed to perform process optimization or to detect problems in the production flow.

From these different sources, many types of data can be collected, such as tabular, image and time series, etc. These data generally present some quality problems. However, the results of a PHM process depend heavily on the quality of the input. In this context, Hyunseok [58] describes this phenomenon using a well-known proverb ”Garbage in, garbage out” which means that if the used data are of low quality, the poor results are unavoidable. In [7] and concerning the data problems in the SMEs, the authors report that the well-known problems are the missing data, the manually recorded data, the small volume of data, and the irrelevant data. Missing data refers to incomplete elements in a database. This missing data are due to a problem in the acquisition system or difference in terms of the acquisition frequency. For manually recorded data, this problem applies to all businesses regardless of size, but it should be noted that this problem is more related to SMEs because they do not have the technology to digitize their data. The next sections are dedicated to detail the data quality issues while focusing on those that affect the SMEs’ data.

5.2. Data quality dimensions

In the literature, we find many definitions of high-quality data. Generally, these definitions link quality to a set of requirements to satisfy. The ISO/IEC 25012 standard [59] defines high-quality data as ”the degree to which a set of characteristics of data fulfills requirements”. In the same context, authors in [60] define poor DQ as ”the degree to which a set of characteristics of data does not fulfill the requirements”. High-quality data can also be defined as ”data that is fit for use by data consumers” [50]. From a quality management point of view, high-quality data is ”appropriate for use or to meet user needs, or it is quality of data to meet customer needs" [61]. As we can see, DQ is evaluated with a set of requirements. These requirements define the constraints, fixed by the user [62], that should be satisfied for the resolution of a problem. Moreover, these requirements represent the goals behind a data analysis task. Thus, we define high-quality data as all data with a minimum level of quality that guarantees the satisfaction of objectives set by the owner. Wang and Strong in [63] have presented one of the first approaches for DQ dimensions identification and present many DQ dimensions which are reduced in a first step in 20 dimensions and then in only 15 dimensions regrouped in 4 classes:

1. The intrinsic category: It includes dimensions that express the natural quality of the data such as accuracy, believability, objectivity, and reputation;
2. The contextual category: It expresses the fact that the quality of the data must be considered in a specific context. These dimensions include the amount of data, its completeness, relevancy, value-added and timeliness;
3. The representational category: It refers to dimensions related to the format and meaning of the data such as the ease of understanding, the interpretability, and the consistency;
4. The accessibility category: It refers to dimensions that express how data is accessible to users [60], including the ease of access and its security.

Many DQ dimensions were proposed and discussed in the literature, but until now, there is no consensus on the essential DQ dimensions for DQ evaluation [50]. However, there is a shortlist of DQ dimensions, which are the most cited and discussed in the literature. According to Redman [64], this list contains:

- Accuracy: Degree to which data is correctly recorded and represents the real word.
- Completeness: It evaluate the ratio of missing values for a variable.
- Timeliness: Evaluate whether data is up to date.
- Consistency: Evaluate if the data respect all the constraints imposed by the data context.

Only quality problems that characterize the SME's data are here studied. As mentioned above, SMEs suffer from a lack of the required data infrastructures. As a result, data that comes from SME's production process are usually incomplete and with a small volume. For that purpose, only these two DQ problems will be studied in the rest of this paper. However, to evaluate the DQ and to propose strategies to improve it, the quantification of these qualities is a crucial activity [60]. In this context, a DQ metric can be defined as a function that transforms a quality dimension to a numerical value [65]. Below, a set of metrics is proposed to measure the previously mentioned DQ dimensions (Volume and completeness).

- Volume is one of the most critical DQ dimensions. It refers to the available amount of data for the construction of a PHM model. In this paper, we consider data volume as the number of instances (observations) in a dataset.
- Completeness is the data characteristics that deal with the problem of missing data. In the literature, completeness is often defined as the "breadth, depth, and scope of information contained in the data" [66]. In [63], divide completeness into Schema completeness, column completeness, and population completeness (See section 2 for more details about data modeling). In the literature, many works [66, 67] use the following formula to calculate completeness:

$$Completeness = 1 - \frac{Number\ of\ incomplete\ elements}{Total\ number\ of\ elements}. \tag{1}$$

In the case of SMEs, we are dealing with a particular type of missing data. Usually, one or more features are completely missing due to the absence of sensors or because they are not digitized. Thus, completeness is no longer the percentage of available values, but it is represented as the percentage of available features in relation to the total number of features that describe the problem. For that, in this paper, the following metric is used to measure the data completeness:

$$Completeness = \frac{Number\ of\ available\ features}{Total\ number\ of\ features}. \tag{2}$$

6. Modelling the data quality impact: Numerical simulations

The idea in this section consists of testing the most used classification tools using different datasets to model the DQ impact on a data-driven PHM approach. Thus, we suppose that data are only collected from the main sensors, and the rest of the sensors are supposed to have one by one, an ISO impact on the results.

Data-driven PHM is usually based on data mining techniques to perform advanced tasks in relation to diagnosis and prediction activities. Since the available data in the industrial domain are generally with poor quality, it is hence needed to identify the impact of DQ on the PHM results. In this context, classification techniques are well applied in the data-driven PHM since they are able to perform different tasks such as fault detection and system diagnosis [68]. A set of simple classification techniques is tested to define their behavior regarding the data problems in the SMEs world. These techniques are *Artificial Neural Network*, *Decision Tree*, *Support Vector Machine*, *K-Nearest Neighbors* and *Gaussian Naive Bayes*. We first recall the basic concept of these tools, and we then follow three strategies of simulation:

- Artificial neural network: An Artificial Neural Network (ANN) is a supervised machine learning tool that aims to define a function that links m input variables to n output variables. Given a set of features and a target, it can learn a non-linear function that can be used for classification or regression. ANN is different from logistic regression by the fact that between the input layer and the output layer, it can be one or more non-linear layers, called hidden layers [69].

- Decision tree: The decision tree (DT) is a supervised, non-parametric machine learning tool that can be used both in classification and regression. The main idea of the DT algorithm is to learn from the data to create simple inferred rules that will be used to segment the data and make predictions [70]. In the literature, many tree techniques are proposed, such as Chi-squared Automatic Interaction Detection (CHAID), Classification And Regression Trees (CART), and C4.5 [71]. The main advantage of the DT method is the possibility to explain the developed model. However, the performance of this model is generally lower than that of other machine learning techniques.

- Support Vector Machine: The Support Vector Machine (SVM) aims to find a separating hyperplane that separates the different classes. The hyperplane that reduces the number of wrongly classified samples in the training phase is called Optimal Separating Hyperplane (OSH). To find it, the SVM technique focuses only on the training instances that are in the border of each class distribution [72]. To do this, the OSH is defined as the hyperplane situated between the classes that maximize the margin between them [73].

- K-Nearest Neighbors: The nearest neighbor technique is one of the biggest approaches of learning. It can deserve both supervised (classification and regression) and unsupervised (clustering) learning. The nearest neighbors approach is based on finding a fixed number k of samples from a training dataset, which are the closest ones, in terms of distance, to the new instance to predict its label. The KNN classifier is a particular application of this approach; it consists in predicting the class of a new point based on the classes of the k closest instances to this later [74]. Many techniques to measure the distances between samples are used, and the Euclidean distance remains the most used in KNN algorithm implementation.

- Naive Bayes: The Naive Bayes (NB) method is a simple, probabilistic, and supervised classifier. This later is based on coupling the *Bayes theorem* with the *Naive* hypothesis of conditional independence between every pair of features given the value of the class variable. This coupling support a well known and very efficient classifier [75]. More details about this technique are presented in this work [76].

Based on the selected tools, we now come to deal with the DQ problems tested with different datasets obtained from the Machine Learning Repository [77]. The idea consists of using many public datasets with different quality levels in relation with two quality aspects that we consider them the most important in the SMEs domain. These aspects concern the data completeness and volume. The used datasets concerns the *Letter Recognition Data Set* [78], *Magic Gamma Telescope Data Set* [79] and *Covertype Data Set* [80] where their characteristics are presented in Table 3. The first dataset is for English alphabet recognition; the second one is for predicting the forest cover type while the third dataset is for the Gamma signal detection. The different datasets are modified to test the impact of DQ problems on the results of the used classifiers. To do, three simulation strategies are proposed:

Dataset	**# of instances**	**# of features**	**# of classes**
Letter	20000	16	26
Forest cover	581012	54	7
MAGIC Gamma	19020	11	2

Table 3: Characteristics of the used datasets.

- Strategy 1 (volume test strategy): To study the impact of the data volume on the results of the different classifiers. The used datasets are divided into 10% for test and the remaining 90% for the train. Then, different train iterations are performed; the first one consists of using only 1% of the training dataset to train the model. Then, in the second iteration, 2% of the training subset is used. These iterations are repeated until reaching 100% of the training subset. The objective of this strategy is to imitate the case when no historic data are available to conduct a data-driven PHM study within SMEs. To more imitate the reality of SMEs, where generally we observe one class more than the others, an intelligent sampling is used to guarantee that a class is present more than the other ones in the training subset.

- Strategy 2 (missing test strategy): This strategy aims to study the impact of the missing data on the classification task. For this purpose, missing data are processed from the point of view of SMEs where missing data may relate to one or more sensors, which means that one or more variables are completely missing. To imitate this reality, different training iterations are performed; the first consists of deleting a characteristic (a column of the training subset). This procedure continues in subsequent iterations until only one feature remains in the data set. At each iteration, the developed model is tested in a test subset with the same features like the training one.

- Strategy 3 (volume and missing test strategy): This strategy aims to summarize the previous strategies. The objective is to understand the impact of DQ issues (volume, missing data) on the classification results.

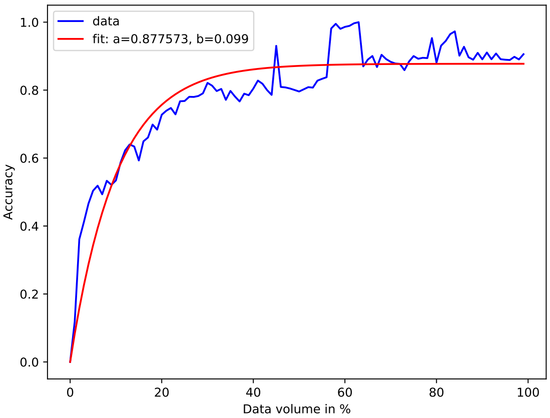


Figure 5: Accuracy evolution of the training data volume. The blue curve represents the mean of the obtained results from the different tested classifiers on the different datasets. The red curve refers to the fitting one.

We then come to study the impact of the previously discussed problems on the results are studied and discussed. The results of this study will be used to satisfy two objectives: the first one is to determine the best techniques to deal with each DQ problem, and the second is to evaluate the impact of the DQ on the pertinence of the results. Noted that these results are normalized to be between 0 and 1 where 1 refers to the case when the predictions are 100% correct. Figure 5 shows the results of the first strategy of simulations. These evolutions prove that the larger the data, the greater the probability of having all the classes of the output variable in the training phase. Thus, the developed model will be more general since it takes into account all the possible cases. The mean accuracy obtained from the different classification tools tested on several datasets is used to fit an empirical function that models the evolution of the accuracy in function of the data volume. The fitting results are shown in Figure 5, and we can conclude that accuracy increases exponentially with increasing data volume. This evolution can be modeled using the following equation:

$$Accuracy = Acc_0 \times (1 - e^{-b \times V}) \tag{3}$$

where Acc_0 is the maximum precision that can be achieved by the model, b is a constant that characterizes the speed of the model to reach a stable level of accuracy, c is a regularization constant, and V is the data volume (in percentage).

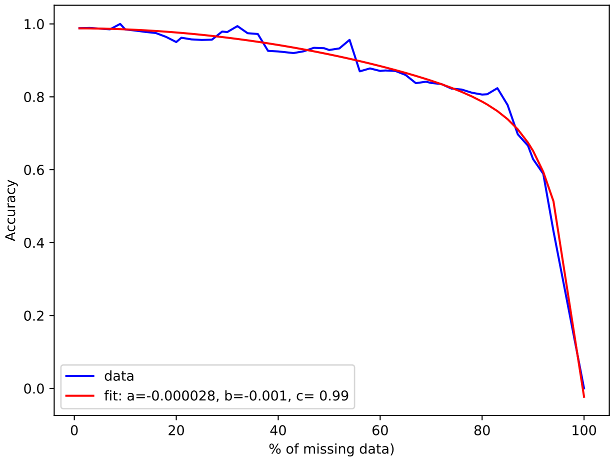


Figure 6: Accuracy evolution of the missing data percentage. The blue curve represents the mean of the obtained results from the different tested classifiers on the different datasets. The red curve refers to the fitting one.

The figure 6 shows the obtained results from the second simulation strategy. The results show that accuracy decreases with increasing percentage of missing data. This conclusion is explained by the fact that the missing data partly represents the reality which affects the quality of the forecast model. Again, an adjustment algorithm is used to extract an empirical equation that models the change in accuracy as a function of the percentage of missing data. This evolution can be modeled using the following equation:

$$Accuracy = a \times (M^2 + e^{-b \times M^2}) + c \tag{4}$$

where a, b and c are constant that can be determined empirically while M is the percentage of missed data.

Figure 7 represents the mapping of the accuracy function of the data volume (%) and the percentage of missing data. This figure shows that with a small number of missing data, we can obtain satisfactory results even if the data volume is small. However, when the percentage of missing data is high, the accuracy is low, even with a big volume of data. This result can be explained by the fact that knowledge is represented by a set of features, and if a set of them is missing, the information is incomplete, and we cannot achieve the expected result.

The results of this DQ impact study represent a first element of the solution to implement data-driven PHM within SMEs. For that, the next section is dedicated to detail the used case study while presenting the obtained results.

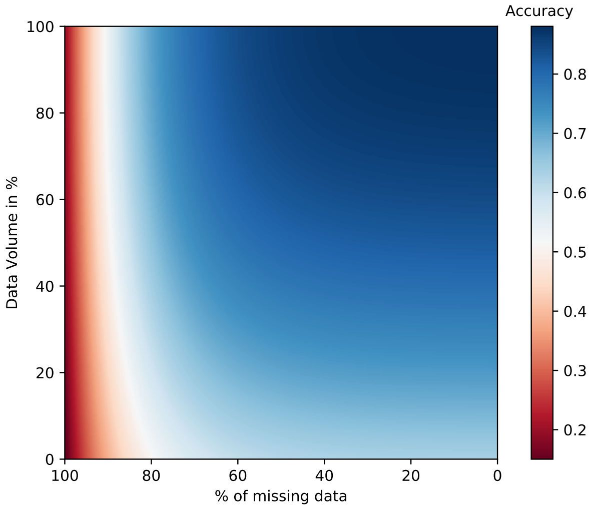


Figure 7: Accuracy map function of the data volume and the percentage of missing data.

7. Case study

We here consider the Scoder case study as a real application of the proposed approach. For that, several assumptions are here considered:

- Systems are monitored;
- Systems are assumed to operate in a nominal situation;
- Features are extracted following an order of importance.

Scoder is a French SME specialized in ultra-precise stamping for automotive applications. This case study consist of sheet metal forming lines. Three machines need to be studied but we have the resources to study only one. The machines are equipped by an integrated acquisition systems to record the breakdowns historic. More data about the metal proprieties, the quality rates and the historic of maintenance are collected by different teams. The complexity of these data and their degree of digitization differs from one machine to another. For that purpose, the proposed methodology is applied in the Scoder factory in order to conduct a data-driven PHM study.

7.1. Application of the proposed strategy in the Scoder factory

Different data which comes from heterogeneous sources are shared in the factory. These data are generally recorded in papers that make it difficult for their deployment in this study. As proposed in the strategy, the first step to conduct a PHM study within SMEs is to start with the data inventory. Figure 8 describes the physical and informational flows in the factory.

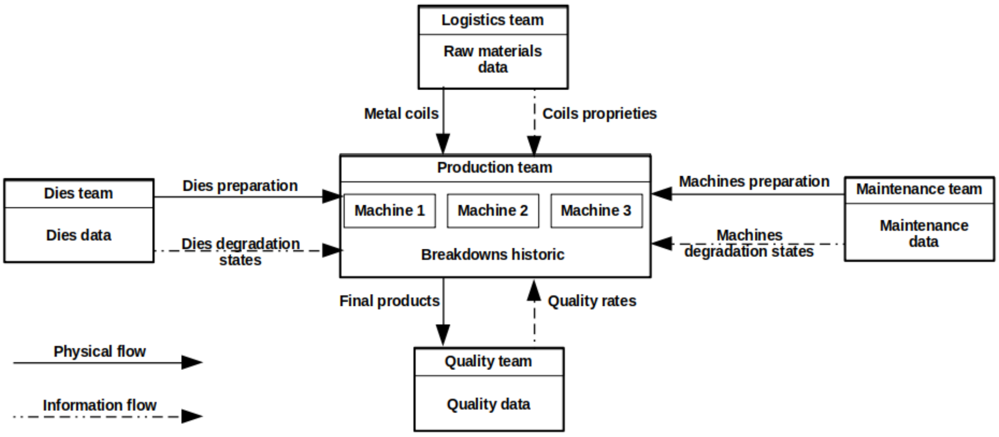
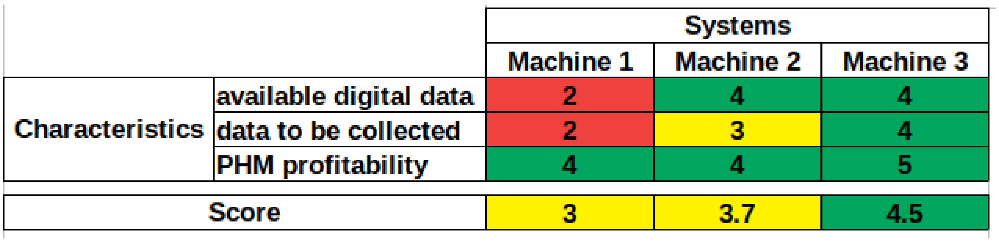


Figure 8: Data inventory in the Scoder case study.

The objective is to ensure a stable production but with the available resources, we can only study one machine. A scope selection is first done according to the characteristics mentioned in Subsection 4.2. Then, these characteristics are used to affect empirically a score to each machine and rank them to select the most suitable one for a PHM study (See Figure 9). One should note that it could be possible to formulate this score mathematically based on the feedbacks of the conducted PHM projects. To take into account the importance level of each characteristic, a constant coefficient is affected to them. Then the final score is calculated as below:

$$Score = 0.2 \times available\ digital\ data + 0.3 \times data\ to\ be\ collected + 0.5 \times PHM\ profitability. \tag{5}$$



		Systems		
		Machine 1	Machine 2	Machine 3
Characteristics	available digital data	2	4	4
	data to be collected	2	3	4
	PHM profitability	4	4	5
Score		3	3.7	4.5

Figure 9: Machines ranking for potential data-driven PHM projects. Red color refers to a poor score, yellow color represents an acceptable score and green color indicates a good score.

Figure 9 shows that the third machine is the most suitable to conduct a data-driven PHM study. The PHM project was initiated with the objective of ensuring a stable production by reducing machine failures and improving productivity. The production performance is affected by the used metal, the die, and the mechanical press. However, a study was conducted inside the factory showed that the used metal coil characteristics have the most important impact on production. From this existing data, a PHM study is conducted to determine an ”Id card” for each sheet metal coil. This Id card represents the characteristics of the coil, the caused press breakdowns, and the quality rate of the products fabricated from it. However, only the sheet metal characteristics are available without any indication about the quality rate, date, and time of the use of each metal coil. To overcome this lack of data, a very simple data acquisition system is installed on the Scoder's machine. This system consists of using a tablet at the beginning of the production line to scan the bar code of each coil to save its date of use, and another tablet at the end of the production line is used to collect the quality data (See Figure 10).

These data are coupled with the machine breakdowns data to describe the production process. Data are structured and saved to be analyzed later using intelligent algorithms and extract knowledge from them. The characteristics of the Scoder dataset are presented in Table 4.

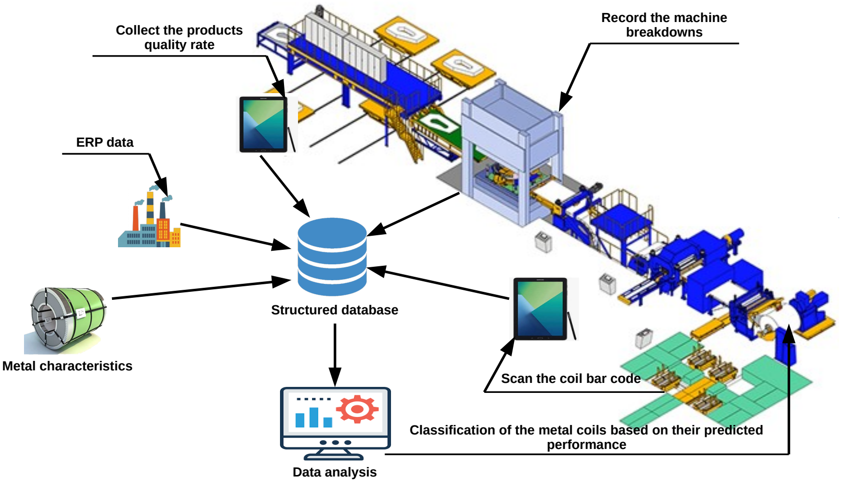

Collect the products quality rate
Record the machine breakdowns
ERP data
Structured database
Metal characteristics
Scan the coil bar code
Classification of the metal coils based on their predicted performance
Data analysis

Figure 10: Details of the SCODER case study.

Dataset	# of instances	# of features	# of classes
Scoder dataset	90	16	2

Table 4: Characteristics of the Scoder dataset.

However, existing data concern only the characteristics of the sheet metal, which means a 60% of missing data (other data from the production process exist but are not digitized). Based on this information and the evolution curve of the accuracy function of the percentage of missing data, we can conclude that at best, the used PHM algorithm can reach an accuracy of 88% (See Figure 11). Moreover, the volume of the created database is very small (only 90 instances), which means that the accuracy result will be less than this level.

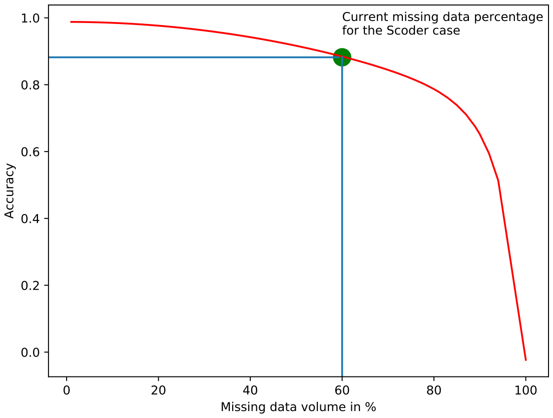


Figure 11: Expected accuracy for the Scoder case study regarding the missing data factor.

In this case study, the DT algorithm is used to classify the metal coils in different categories regarding their expected performance. The DT method is chosen because it is an explainable machine learning tool, which means that workers have an idea about the built classification rules. Figure 12 shows the evolution of the accuracy rate in function of the volume of training data (since the real output of the n^{th} training iteration added to the training subset of the $(n+1)^{th}$ iteration). One can point out that after a few iterations, we can reach about 50% of accuracy. This low accuracy could be explained by the fact that the existing data doesn't contain information about the health state of the die and the press. Moreover, the data volume is not sufficient.

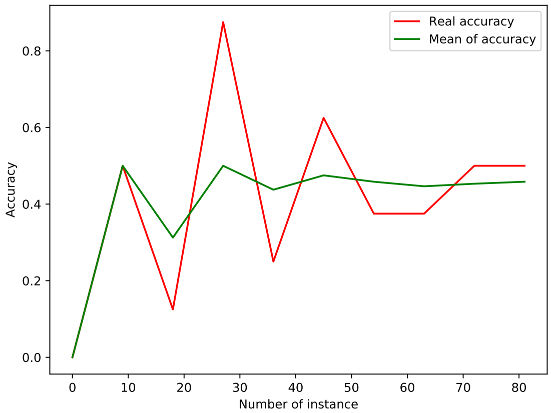


Figure 12: Evolution of the accuracy function of the number of instance used in the training phase. These results are obtained using the DT algorithm.

Based on the previous DQ study, it is possible to reach a stable level of accuracy when 50% or more of the needed data volume is used. In addition, Figure 12 shows that the accuracy of the prediction is about 50%, which means that the existing data represents about 10% of the needed data volume (See Figure 13). Since the accuracy will stabilize from the threshold of 50% we can conclude that we need four times more data (360 instances) to reach the maximal accuracy that can be obtained with the existing features. Thus, we need to collect more data for one year since the existing data are collected in three months. This conclusion affirms that a stable prediction model can be obtained in one year, which will be proved in the next months. Therefore, an economical (optimal) volume of data is defined (450 instances), which means that, with the existing features, we need only 450 instances to reach the maximal level of accuracy and build a stable prediction model. The data collect frequency can be reduced after this period, and only a few data will be collected that will serve to adjust the classification model over time.

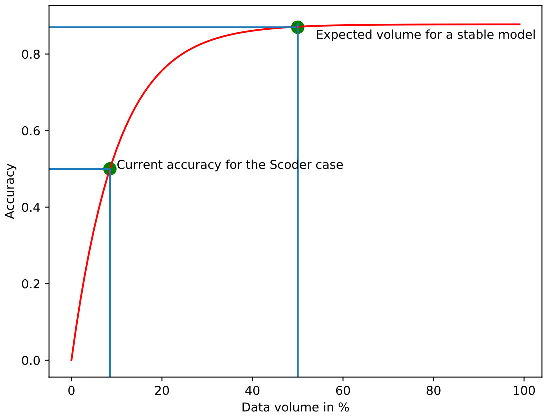


Figure 13: Expected results for the Scoder case study regarding the data volume factor.

Despite the unsatisfied results of the PHM project, the performance of the studied machine has been improved. The considered horizon for parts production is 115 days. The PHM study have been started on day 50 of the production. Figure 14 shows the evolution of the produced quantity of parts between two breakdowns during 115 production days. This indicator allows to quantify the occurrence of breakdowns by taking into account the produced quantity. The results shows that the productivity have been improved with more than 80% after PHM deployment.

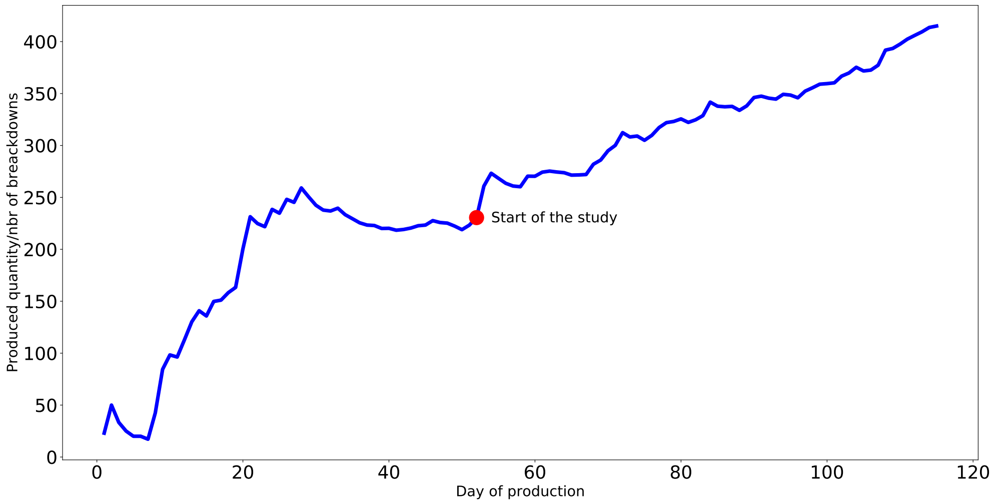


Figure 14: Impact of the PHM study on the evolution of the number of produced parts between breakdowns.

7.2. Discussion

Recall that this paper addresses the two most critical issues that limit the implementation of data-driven PHM in industry and particularly in SMEs. These problems are the lack of documentation and the impact of DQ. On the basis of this study, a generic strategy is proposed to implement a data-driven PHM process in SMEs while focusing on the impact of DQ. Based on the DQ study, general models of the evolution of the accuracy function of the data volume and/or the missing data percentage are identified. Thus, it is possible to define a required data volume and a percentage of missing data that allows reaching an expected accuracy rate. This result means that only with a little amount of data we can do a PHM study, and based on the obtained results, we can make a projection over time to determine the time of data requirements and the expected results.The objective is to resize the PHM infrastructure and give a clear idea about the data requirements that should be satisfied. These requirements include the data acquisition system, storage hub, analysis tools and devices, frequency of collect, volume of data, etc. Thus, a temporal and technological boundary can be affected to each PHM project, and in this way, the cost of the PHM strategy can be calculated. In fact, one of the non-announced PHM limitations is its unknown cost, which still an understudied topic.

8. Conclusion and perspectives

PHM discipline is widely used in the industrial domain to make plants smarter and more efficient. For a long time, PHM was applied only in the big companies since they have the required human and technological resources. However, the current PHM tool is inapplicable in the case of SMEs due to many characteristics that differ them from big corporations. In this paper, we pointed out the issues that limit the integration of new technologies within a small organization. These problems concern the lack of resources and human factors essentially. Based on these limits, an adapted PHM has been proposed. Finally, a study was conducted to characterize the needed data in terms of volume and completeness to satisfy the fixed objectives. Table 5 summarizes the main differences between big companies and SMEs for PHM implementation as well as the proposed solutions for each challenge.

The proposed strategy is a first framework for an semi-developed domain, which can be improved and completed by other aspects such as the business impact. Thus, one should think about developing a global PHM cost model. Moreover, the human factor in the PHM implementation is not yet considered despite its key impact on the success of the strategy. In this context, an interactive and user-friendly PHM framework must be developed. Thus, the use of explainable data analysis techniques may help in the integration of the users in the PHM strategy.

Factor	Attributes	Big companies	SMEs	Proposed solutions
Technology	Infrastructure	Available	Limited	Scope identification
	Data quality	Medium	Medium	DQ assessment and improvement
Human	Skilled workers	High	Medium	User friendly PHM framework
	Resistance to change	Low	Medium	Automate the existing data projects
Organization	Documentation	Developed	Semi-developed	Knowledge capitalization process
	Objectives metrics	Detailed	Global	Standards PHM metrics

Table 5: The SME-oriented PHM challenges and solutions.

References

[1] J. Wang, Y. Ma, L. Zhang, R. X. Gao, D. Wu, Deep learning for smart manufacturing: Methods and applications, Journal of Manufacturing Systems 48 (2018) 144–156.

[2] S. Mittal, M. A. Khan, D. Romero, T. Wuest, A critical review of smart manufacturing & industry 4.0 maturity models: Implications for small and medium-sized enterprises (smes), Journal of manufacturing systems 49 (2018) 194–214.

[3] C. Zheng, X. Qin, B. Eynard, J. Bai, J. Li, Y. Zhang, Sme-oriented flexible design approach for robotic manufacturing systems, Journal of Manufacturing Systems 53 (2019) 62–74.

[4] W. Li, K. Liu, M. Belitski, A. Ghobadian, N. O'Regan, e-leadership through strategic alignment: An empirical study of small-and medium-sized enterprises in the digital age, Journal of Information Technology 31 (2016) 185–206.

[5] R. Sahal, J. G. Breslin, M. I. Ali, Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case, Journal of Manufacturing Systems 54 (2020) 138–151.

[6] D. Kohler, J.-D. Weisz, Industrie 4.0: les défis de la transformation numérique du modèle industriel allemand, La Documentation française, 2016.

[7] N. Omri, Z. Al Masry, S. Giampiccolo, N. Mairot, N. Zerhouni, Data management requirements for phm implementation in smes, in: 2019 Prognostics and System Health Management Conference (PHM-Paris), IEEE, 2019, pp. 232–238.

[8] M. Pecht, Prognostics and health management of electronics, Encyclopedia of Structural Health Monitoring (2009).

[9] J. Kennedy, P. Hyland, et al., A comparison of manufacturing technology adoption in smes and large companies, in: Proceedings of 16th Annual Conference of Small Enterprise Association of Australia and New Zealand, 2003, pp. 1–10.

[10] K. Blind, A. Mangelsdorf, Alliance formation of smes: empirical evidence from standardization committees, IEEE Transactions on Engineering Management 60 (2012) 148–156.

[11] F. Tao, Q. Qi, A. Liu, A. Kusiak, Data-driven smart manufacturing, Journal of Manufacturing Systems 48 (2018) 157–169.

[12] E. S. Madsen, A. Bilberg, D. G. Hansen, Industry 4.0 and digitalization call for vocational skills, applied industrial engineering, and less for pure academics, in: Proceedings of the 5th P&OM World Conference, Production and Operations Management, P&OM, 2016.

[13] K. Schwab, The fourth industrial revolution, Currency, 2017.

[14] J. Li, F. Tao, Y. Cheng, L. Zhao, Big data in product lifecycle management, The International Journal of Advanced Manufacturing Technology 81 (2015) 667–684.

[15] A. J. Guillén, V. González-Prida, J. F. Gómez, A. Crespo, Standards as reference to build a phm-based solution, in: Proceedings of the 10th World Congress on Engineering Asset Management (WCEAM 2015), Springer, 2016, pp. 207–214.

[16] J. Lee, M. Holgado, H.-A. Kao, M. Macchi, New thinking paradigm for maintenance innovation design, IFAC Proceedings Volumes 47 (2014) 7104–7109.

[17] S. Takata, F. Kirnura, F. J. van Houten, E. Westkamper, M. Shpitalni, D. Ceglarek, J. Lee, Maintenance: changing role in life cycle management, CIRP annals 53 (2004) 643–655.

[18] M. Xu, X. Jin, S. Kamarthi, M. Noor-E-Alam, A failure-dependency modeling and state discretization approach for condition-based maintenance optimization of multi-component systems, Journal of manufacturing systems 47 (2018) 141–152.

[19] A. K. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance, Mechanical systems and signal processing 20 (2006) 1483–1510.

[20] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, J. Lin, Machinery health prognostics: A systematic review from data acquisition to rul prediction, Mechanical Systems and Signal Processing 104 (2018) 799–834.

[21] G. W. Vogl, B. A. Weiss, M. Helu, A review of diagnostic and prognostic capabilities and best practices for manufacturing, Journal of Intelligent Manufacturing 30 (2019) 79–95.

[22] G. W. Vogl, B. A. Weiss, M. A. Donmez, Standards for prognostics and health management (PHM) techniques within manufacturing operations, Technical Report, National Institute of Standards and Technology Gaithersburg United States, 2014.

[23] ISO, SO 13374-1:2003 - Condition monitoring and diagnostics of machines Data processing, communication and presentation Part 1: General guidelines, 2003. URL: `http://www.iso.ch/cate/d27688.html`.

[24] ISO, ISO 13374-2:2007 - Condition monitoring and diagnostics of machines Data processing, communicaon and presentaon Part 2: Data processing, 2007. URL: `http://www.iso.ch/cate/d27688.html`.

[25] ISO, ISO 13374-3:2012 - Condition monitoring and diagnostics of machines Data processing, communicaon and presentaon Part 3: Communication, 2012. URL: `http://www.iso.ch/cate/d27688.html`.

[26] ISO, ISO 13374-4:2015 Condition monitoring and diagnostics of machine systems – Data processing, communication and presentation – Part 4: Presentation, 2015. URL: `http://www.iso.ch/cate/d27688.html`.

[27] A. Mathew, L. Zhang, S. Zhang, L. Ma, A review of the mimosa osa-eai database for condition monitoring systems, in: Engineering Asset Management, Springer, 2006, pp. 837–846.

[28] M. Thurston, M. Lebold, Standards developments for condition-based maintenance systems, Technical Report, PENNSYLVANIA STATE UNIV UNIVERSITY PARK APPLIED RESEARCH LAB, 2001.

[29] K.-S. Wang, Towards zero-defect manufacturing (zdm)—a data mining approach, Advances in Manufacturing 1 (2013) 62–74.

[30] M.-A. Koulali, S. Koulali, H. Tembine, A. Kobbane, Industrial internet of things-based prognostic health management: a mean-field stochastic game approach, IEEE Access 6 (2018) 54388–54395.

[31] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani, M. G. Pecht, Iot-based prognostics and systems health management for industrial applications, IEEE Access 4 (2016) 3659–3670.

[32] M. G. Pecht, A prognostics and health management roadmap for information and electronics-rich systems, IEICE ESS Fundamentals Review 3 (2010) 4_25–4_32.

[33] S. Adams, M. Malinowski, G. Heddy, B. Choo, P. A. Beling, The wear methodology for prognostics and health management implementation in manufacturing, Journal of Manufacturing Systems 45 (2017) 82–96.

[34] X. Gao, O. Niculita, B. Alkali, D. McGlinchey, Cost benefit analysis of applying phm for subsea applications, in: Proceedings of the European Conference of the PHM Society, 2018.

[35] K. Feldman, P. Sandborn, T. Jazouli, The analysis of return on investment for phm applied to electronic systems, in: 2008 International Conference on Prognostics and Health Management, IEEE, 2008, pp. 1–9.

[36] J. Müller, K. Voigt, Industry 4.0—integration strategies for small and medium-sized enterprises, in: Proceedings of the 26th International Association for Management of Technology (IAMOT) Conference, Vienna, Austria, 2017, pp. 14–18.

[37] J. Lee, E. Lapira, B. Bagheri, H.-a. Kao, Recent advances and trends in predictive manufacturing systems in big data environment, Manufacturing letters 1 (2013) 38–41.

[38] R. Kothamasu, S. H. Huang, W. H. VerDuin, System health monitoring and prognostics—a review of current paradigms and practices, The International Journal of Advanced Manufacturing Technology 28 (2006) 1012–1024.

[39] S. Bridge, K. O'Neill, Understanding enterprise: Entrepreneurship and small business, Macmillan International Higher Education, 2012.

[40] S. Durst, I. Runar Edvardsson, Knowledge management in smes: a literature review, Journal of Knowledge Management 16 (2012) 879–903.

[41] K. Yew Wong, E. Aspinwall, Characterizing knowledge management in the small business environment, Journal of Knowledge management 8 (2004) 44–61.

[42] R. S. Wadhwa, Flexibility in manufacturing automation: A living lab case study of norwegian metalcasting smes, Journal of Manufacturing Systems 31 (2012) 444–454.

[43] P. Julien, C. Ramangalahy, Competitive strategy and performance of exporting smes: An empirical investigation of the impact of their export information search and competencies, Entrepreneurship Theory and Practice 27 (2003) 227–245. URL: `https://doi.org/10.1111/1540-8520.t01-1-00002`. doi:10.1111/1540-8520.t01-1-00002. `arXiv:https://doi.org/10.1111/1540-8520.t01-1-00002`.

[44] S. Cheng, M. H. Azarian, M. G. Pecht, Sensor systems for prognostics and health management, Sensors 10 (2010) 5774–5797.

[45] T. Saaty, Planning, priority setting, resource allocation, The analytic hierarchy process (1980).

[46] W. W. Cooper, L. M. Seiford, J. Zhu, Data envelopment analysis, in: Handbook on data envelopment analysis, Springer, 2004, pp. 1–39.

[47] J. Xu, Y. Wang, L. Xu, Phm-oriented sensor optimization selection based on multiobjective model for aircraft engines, IEEE Sensors Journal 15 (2015) 4836–4844.

[48] J. J. Luna, Metrics, models, and scenarios for evaluating phm effects on logistics support, in: Proceedings of Annual Conference of the Prognostics and Health Management Society, 2009.

[49] T. Kong, T. Hu, T. Zhou, Y. Ye, Data construction method for the applications of workshop digital twin system, Journal of Manufacturing Systems (2020).

[50] L. Sebastian-Coleman, Measuring data quality for ongoing improvement: a data quality assessment framework, Newnes, 2012.

[51] Dama-dmbok (2nd edition): Data management body of knowledge, DAMA International - 2017 (2017) i–. doi:9781634622349.

[52] T. Fisher, The data asset: How smart companies govern their data for business success, volume 24, John Wiley & Sons, 2009.

[53] M. Hoffmann, C. Büscher, T. Meisen, S. Jeschke, Continuous integration of field level production data into top-level information systems using the opc interface standard, Procedia CIRP 41 (2016) 496–501.

[54] J. W. Ross, M. R. Vitale, The erp revolution: surviving vs. thriving, Information systems frontiers 2 (2000) 233–241.

[55] I. Madanhire, C. Mbohwa, Enterprise resource planning (erp) in improving operational efficiency: Case study, Procedia CIRP 40 (2016) 225–229.

[56] P. D. U. Coronado, R. Lynn, W. Louhichi, M. Parto, E. Wescoat, T. Kurfess, Part data integration in the shop floor digital twin: Mobile and cloud technologies to enable a manufacturing execution system, Journal of manufacturing systems 48 (2018) 25–33.

[57] B. Saenz de Ugarte, A. Artiba, R. Pellerin, Manufacturing execution system–a literature review, Production planning and control 20 (2009) 525–539.

[58] H. Oh, M. H. Azarian, S. Cheng, M. G. Pecht, Sensor systems for phm, Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things (2018) 39–60.

[59] ISO/IEC, Software engineering software product quality requirements and evaluation (square) data quality model, in: ISO/IEC, Tech. Rep.

ISO/IEC 25012, 2008, 2008.

[60] N. Laranjeiro, S. N. Soydemir, J. Bernardino, A survey on data quality: classifying poor data, in: 2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC), IEEE, 2015, pp. 179–188.

[61] F. G. Alizamini, M. M. Pedram, M. Alishahi, K. Badie, Data quality improvement using fuzzy association rules, in: 2010 International Conference on Electronics and Information Engineering, volume 1, IEEE, 2010, pp. V1–468.

[62] B. Becerik-Gerber, F. Jazizadeh, N. Li, G. Calis, Application areas and data requirements for bim-enabled facilities management, Journal of construction engineering and management 138 (2011) 431–442.

[63] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (1996) 5–33.

[64] T. C. Redman, Data Quality for the Information Age, 1st ed., Artech House, Inc., Norwood, MA, USA, 1997.

[65] Ieee standard for a software quality metrics methodology, IEEE Std 1061-1998 (1998) i–. doi:10.1109/IEEESTD.1998.243394.

[66] C. Batini, M. Scannapieco, Data and information quality: Concepts, methodologies and techniques, 2016.

[67] Y. Wand, R. Y. Wang, Anchoring data quality dimensions in ontological foundations, Communications of the ACM 39 (1996) 86–95.

[68] M. Sharp, R. Ak, T. Hedberg Jr, A survey of the advancing use and development of machine learning in smart manufacturing, Journal of manufacturing systems 48 (2018) 170–179.

[69] R. Zemouri, N. Omri, F. Fnaiech, N. Zerhouni, N. Fnaiech, A new growing pruning deep learning neural network algorithm (gp-dlnn), Neural Computing and Applications (2019) 1–17.

[70] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (2007) 1761–1768.

[71] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.

[72] A. Mathur, G. M. Foody, Multiclass and binary svm classification: Implications for training and classification users, IEEE Geoscience and remote sensing letters 5 (2008) 241–245.

[73] J. Wang, S. Liu, R. X. Gao, R. Yan, Current envelope analysis for defect identification and diagnosis in induction motors, Journal of Manufacturing Systems 31 (2012) 380–387.

[74] M. Khanzadeh, S. Chowdhury, M. Marufuzzaman, M. A. Tschopp, L. Bian, Porosity prediction: Supervised-learning of thermal history for direct laser deposition, Journal of manufacturing systems 47 (2018) 69–82.

[75] G. H. John, P. Langley, Estimating continuous distributions in bayesian classifiers, in: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[76] I. Rish, et al., An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, volume 3, 2001, pp. 41–46.

[77] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[78] P. W. Frey, D. J. Slate, Letter recognition using holland-style adaptive classifiers, Machine learning 6 (1991) 161–182.

[79] R. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savickỳ, S. Towers, et al., Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 516 (2004) 511–528.

[80] J. A. Blackard, D. J. Dean, Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables, Computers and electronics in agriculture 24 (1999) 131–151.