Towards an adapted PHM approach: Data quality requirements methodology for fault detection applications

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

Increasingly, extracting knowledge from data has become an important task in organizations for performance improvements. To accomplish this task, data-driven Prognostics and Health Management (PHM) is introduced as an asset performance management framework for data management and knowledge extraction. However, acquired data come generally with quality issues that affect the PHM process. In this context, data quality problems in the PHM context still an understudied domain. Indeed, the quality of the used data, their quantification, their improvement techniques and their adequacy to the desired PHM tasks are marginalized in the majority of studies. Moreover, many PHM applications are based on the development of very sophisticated data analysis algorithms without taking into account the adaptability of the used data to the fixed objectives. This paper aims to propose a set of data quality requirements for PHM applications and in particular for the fault detection task. The conducted developments in this study are applied to Scoder enterprise, which is a French SME. The feedback on the first results is reported and discussed.

Keywords: Data quality metrics, Data quality assessment, Data-driven PHM, Data management, Impact of data quality on PHM results, data detectability.

Notation

$\sum$:	The studied system
Det:	Detectability state of the system $\sum$
O:	Observability state of the system $\sum$
Q:	Data quality of the dataset
P:	Performance of the used detectability algorithm
GQ:	Global data quality
LQ:	Local data quality
X_i:	Features that describe the system $\sum$ for $i = 1, \ldots, n$
Q_i:	Data quality of a features X_i
q_{Im}:	Imbalanced data ratio
q_{i1}:	Missing data ratio for a feature X_i
q_{i2}:	Noisy data ratio for a feature X_i
w_i:	Importance weight of the feature X_i
CD:	Cost of a negative detection
CI:	Needed cost to assess an imbalance ratio level
CM_i:	Required cost to assess a missing data ratio level
CN_i:	Required cost to assess a noisy data ratio level
$\lvert . \rvert$:	Cardinality of the data space

1. Introduction

Prognostics and Health Management (PHM) is a science that studies the health state of a part of equipment and predicts its future evolution [1]. This concept allows to better control systems and to implement suitable maintenance strategies [2, 3]. In [1], 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 [4], 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 component degradation 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 the results only depend on the quality of historical data. Recently, a new approach for implementing PHM solutions has emerged, which is the hybrid approach. Hybrid approach merges the advantage of data-driven and model-based techniques to implement an efficient PHM process.

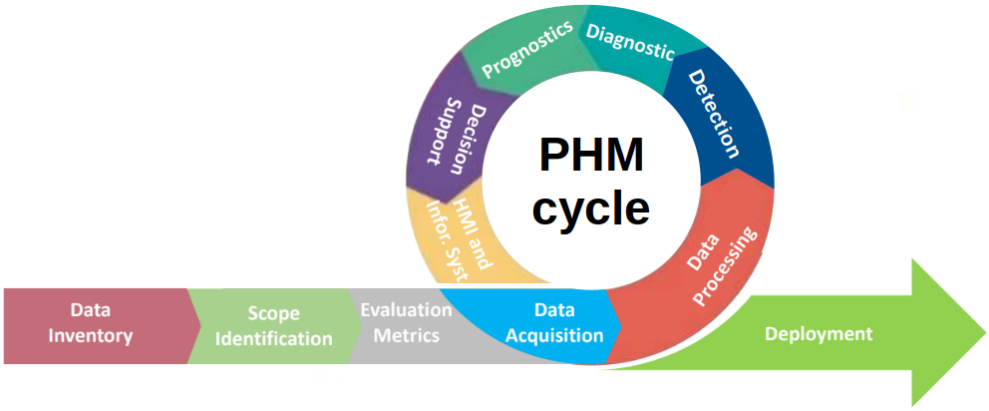

Figure 1: Extended PHM cycle [5].

Driven by the emergence of digitization technologies, data-driven approach for product life cycle management has attracted more attention in recent years [6]. Thanks to the huge volume of collected data, scalable, re-configurable and low cost PHM solutions can be easily implemented. Thus, data become the fuel for the PHM locomotive by facilitating its implementation in several fields, notably the industrial one. The industrial application of the PHM discipline generally concerns decision support for a more efficient and intelligent operation of machines [7]. To satisfy this objective, three main PHM tasks are identified [8]: (*i*) Fault detection, (*ii*) Fault diagnosis and (*iii*) Degradation prediction. Fault detection concerns the detection of abnormal system behaviors and their separation from normal ones. As for fault diagnosis, it is the separation of the various failure modes of the system and their classification into known classes. Finally, degradation prediction deals with the prediction of the evolution of the system health state to give an accurate information about its Remaining Useful Life (RUL) [9]. As shown in Fig. 1, these tasks are performed consecutively starting with fault detection and ending with decision support. Thus, the success of the fault detection task is a necessary condition for the success of the entire PHM process.

Many sophisticated algorithms have been proposed to deal with the problem of fault detection with impressive performances [10]. These algorithms become useless in the case when the data are not suitable for this task [11]. Data quality has a major impact on the success of the PHM implementation [5]. Nevertheless, the suitability of the existing data quality for the different PHM tasks still an understudied problem [11]. Despite the development of data modeling techniques for PHM tasks, only few methods exist to assess the suitability of data for these techniques [12, 11]. These works analyze the adequacy of an existing data set to the fixed objectives. This implies that the data acquisition step is carried out in advance. Moreover, these works are based on visualization techniques for data quality assessment without defining a generic metric to quantify data quality and its impact on PHM results.

To address this issue, this work aims to propose a new metric to assess the suitability of data to the fault detection task. For that purpose, a formalization of the problem is proposed which leads to understand the impact of data quality on the fault detection task. The proposed metric is firstly validated using some benchmarks and then applied to the Scoder application [5], which is a real case study. Throughout the paper, the terms ”data”, ”variable” and ”feature” are used interchangeably, as is often done in academic literature.

The remainder of this paper is organized as follows. Section 2 presents a brief review of related works that concern data quality impact on the PHM tasks. The problem statement is then illustrated in Section 3. In Sections 4 and 5, a formalization of the data detectability problem and an associated empirical metric are presented, respectively. These developments are applied in a real case study in Section 6. Finally, discussions and conclusions are displayed respectively in Sections 7 and 8.

2. Overview

The objective of this work is to position the data quality problem in a PHM context and to propose a new metric to assess the suitability of data to the fault detection task. Such a complex and multi-factorial problem brings together many disciplines such as *data quality*, *data analysis* and *PHM*. A clear understanding of these disciplines is required to satisfy the paper's objective. For that purpose, this section presents an overview of the data quality notions in the PHM context.

2.1. Data quality studies

Data quality (DQ) has been the subject of many research works where several definitions were proposed to characterize this concept. The ISO 8000-8:2015 standard [13] describes fundamental concepts of information and data quality, and how these concepts apply to quality management processes and quality management systems. In [14], Zaveri et al. assume that data quality problem refers to a set of issues that can affect the potentiality of the applications that use the data. The authors in [5] affirm that the majority of these definitions link data quality to a set of requirements to satisfy. The ISO/IEC 25012 standard [15] definition assumes that high data quality is ”the degree to which a set of characteristics of data fulfill requirements”. Indeed, authors in [16] define it as ”data that is fit for use by data consumers”. Data quality is usually defined according to a set of requirements that should be accomplished. We here adopt the data quality definition proposed in [5] and which assumes ”high quality data as all data with a minimum level of quality that guarantees the satisfaction of objectives set by the owner”.

	Category			
	Intrinsic	**Contextual**	**Representational**	**Accessibility**
Data quality	Accuracy Believability Objectivity Reputation	Completeness Relevancy Value-added Timeliness Data volume	Ease of understanding Interpretability Consistency	Ease of access Security

Table 1: Categories of data quality dimensions [5].

As previously stated, data quality is a multidimensional issue that is widely studied in the literature. Thus, a set of Data Quality Dimensions (DQD) is defined to characterize the data requirements [17, 18]. As the Table 1 shows, these dimensions are classified into four main categories: intrinsic, contextual, representational and accessibility dimensions. Each category has a set of data quality dimensions that describes the data. However, some dimensions are studied more than others. Redman [19] offers a short list of the most studied ones which includes accuracy, completeness, consistency, timeliness and consistency. Additionally, the most studied data quality dimensions are often reduced to three main dimensions in the context of industrial application:

- Data volume: Evaluate whether the data volume is sufficient for the study.

- Data accuracy: Represent the degree of representativeness of the correctly recorded data to the real world.
- Data completeness: Evaluate the ratio of missing values for a variable.

Despite the huge volume of studies that deal with the data quality problem, only few of them introduce this issue in a PHM context. The next paragraph reviews these works [1, 5, 11, 12].

2.2. Data quality in the PHM context

Data quality metrics differ from one application to another. Three data quality metrics are defined in [11] to characterize data for PHM applications. These metrics concern the aspects of detectability, diagnosability and trend-ability [11]. Detectability refers to the fault detection task in the PHM framework and it represents the ability of system abnormal behavior data to be detected and separated from the normal ones. Diagnosability fits the fault diagnosis within the PHM approach and it means that data allow a good separation of the different system failure modes. As for trend-ability, it concerns the degradation prediction and it describes the ability of data to estimate an accurate information about the RUL of the system [9]. To measure these data qualities, Jia et al. [11] proposed to use a statistic test based on the Maximum Mean Discrepancy (MMD) method which used to evaluate the difference between two data distributions. Despite the obtained results, this method strongly depends on the data modeling algorithm used which ignores the impact of the data quality. In coherence with this work, the authors in [12] define the cluster-ability as a metric that can be used in the PHM context. Cluster-ability measures the predisposition of data to be clustered in natural groups without referring to the real data labels [20]. Similarly, this method strongly depends on the used clustering algorithm. The proposed metrics describe the data quality in an aggregated way that is unrelated to the basic data quality issues (i.e. missing data, noisy data, incomplete data, etc.). Moreover, the authors in [1] propose a set of data quality requirements to be suitable with PHM applications. The data issues evoked in the mentioned paper concern mainly: data volume, data accuracy and completeness. Consequently, we propose in this paper to classify the data problems according to these characteristics.

2.2.1. Data volume

One of the most critical factors that lead to failure of a PHM project is the data unavailability. Data volume is the most important data quality dimensions and it concerns different aspects such as:

- **The number of instances**: It refers to the existing volume of data (number of observations) that can be used to build a PHM model. This data quality is measured by:

$$q_v = |R| \tag{1}$$

  where $|.|$ refers the cardinality of the data space and R is the ensemble of objects that make up the database.

- **The imbalanced data**: It is a form of between-class imbalance that arises when one data class dominate over another class. It causes the machine learning model to be more biased towards majority class. The following metric is used to quantify this aspect of data problem:

$$q_{Im} = 1 - \frac{|o \in S|}{|R|} \tag{2}$$

  where o is an observation and S is the objects ensemble of the subsampled class.

2.2.2. Data accuracy

Data accuracy is one of the most frequently cited dimensions of DQ in the literature. In [21], authors define accuracy as the ”distance” between the data or information and the world reality they describe. Data accuracy could control:

- **The outlier data:** It is one of the well known data quality problem that refers to data objects that do not correspond to expected behaviors [22]. It is defined by:

$$q_o = \frac{|\, o \in A \,|}{|\, R \,|} \tag{3}$$

where A is the ensemble of outlier observations in the dataset.

- **The noisy data:** It concerns data which are recorded with an error compared to the world reality they describe. From a logical point of view, a value is considered as noisy only if it impacts the detection result. We propose here to quantify the accuracy ratio as follows:

$$q_n = \frac{|\, o \in N \,|}{|\, R \,|} \tag{4}$$

where N is the ensemble of noisy observations for a specific feature $X_i, i = 1, \ldots, n$.

2.2.3. Data completeness

Completeness is the data dimension that deals with the problem of missing data. We here differ between two types of missing data:

- **Partially missing data:** It evaluates the ratio of missing values for a variable. Thus, completeness is explained in this case as the percentage of available values for a variable. The completeness ratio is calculated using the following metric:

$$q_m = \frac{|\, o \in M \,|}{|\, R \,|} \tag{5}$$

where M is the ensemble of missing observations for a specific feature $X_i, i = 1, \ldots, n$.

- **Completely missing data (Insufficient features):** This case is discussed in [5] and it concerns the case where one or more features are completely missing due to the absence of sensors or the fact that they are not measurable. Insufficient features ratio is quantified using:

$$q_{ins} = \frac{n}{d} \tag{6}$$

where n is the number of saved features $X_i, i = 1, \ldots, n$ and d refers to the number of identified variables during the data inventory step and that describe the system Σ.

3. Problem statement

As shown in Fig. 2, data quality management in the PHM context can be seen from two sides: (1) a direct process where data quality is assessed and its suitability for the PHM application is evaluated, and (2) a reverse process where a set of data quality requirements are defined to meet the fixed objectives. In the PHM context, there is little literature that addresses the data quality issue. These works analyze the adequacy of an existing data set to the fixed objectives. This implies that the data acquisition step is carried out in advance. Moreover, these works are based on visualization techniques for data quality assessment without defining a generic metric to quantify data quality and its impact on PHM results. We are here interested in the definition of a generic metric allowing the understanding and the quantification of the data quality impact on PHM tasks in relation to the expected performance for each task before the installation of the data acquisition system. Recall that the main PHM tasks are fault detection, diagnosis and degradation prediction. The fault detection task is the first one on the PHM process [11] and is considered in the rest of this study.

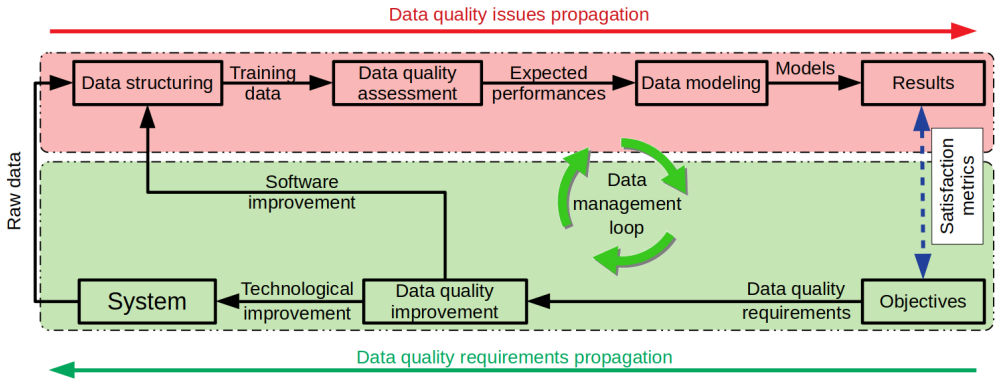

Figure 2: Data management process. In red, the direct process which consists in evaluating the suitability of the used data to the fixed objectives. While, the inverse process (in green) aims to set data quality requirement that should be respected in order satisfy the objectives. For that, data quality improvement actions are proposed in the system level and the data level.

In [23], the authors specify that fault detectability can be divided into two notions: *(i) Intrinsic detectability* and *(ii) Performance based fault detectability*. The intrinsic notion refers to the system's anomalies signature without any dependence on the used fault detection technique. This fits with the system's intrinsic propriety such as controllability and observability [24]. On the other side, performance based fault detectability is defined according to the fault detection algorithm used and it refers to the ability of this algorithm to detect anomalies [23]. As shown in Fig. 3, many factors can affect the fault detection task. These factors can be related to the performance of the used detection algorithm, the data quality issues or the system observability. For the first possibility, many sophisticated algorithms have been proposed to deal with the problem of defect detection with impressive performances. However, if the used data do not describe the studied system, it is not necessary to develop a sophisticated algorithm to solve the problem because it is impossible to meet the objectives due to the inadequacy of the data.

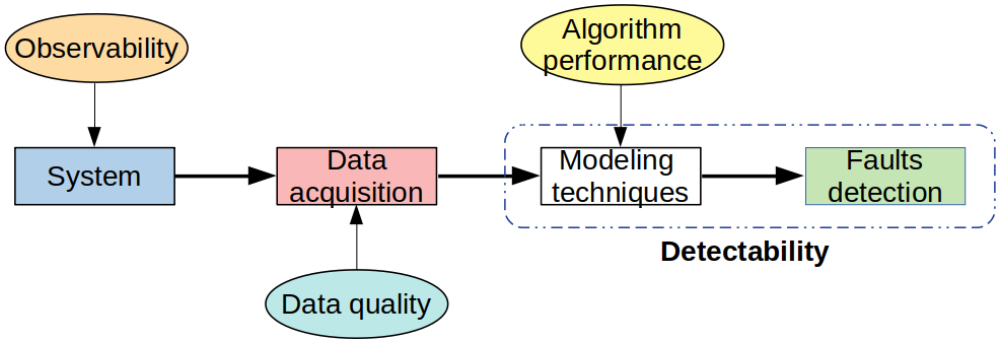

Figure 3: Factors that impact the detectability accuracy.

In this paper, we propose to formalize the data quality impact on the fault detection task. Some assumptions are made for our study:

(A1) The identified variables in the data inventory step provide a complete description of the system $Sigma$

(A2) $\sum$ is is observed during a sufficient horizon of time to collect the needed data.

(A3) The used detection algorithms are all able to perform equal results.

(A4) The detctability task is done in a supervised mode.

To sum up, this article aims to quantify detectability for fully observable systems and define data quality requirements in relation with the expected detection results.

4. Formulation of the data quality problem

Intrinsic detectability refers to the system's anomalies signature without any dependence on the used fault detection technique. This fit with the observability O as a system's intrinsic propriety. On the other side, the performance based fault detectability is defined according to the used fault detection algorithm and it refers to the ability of this algorithm to detect anomalies. However, the ability of an algorithm to detect anomalies can be a result of its intrinsic performance P and the quality of the used dataset Q. Thus, the detectability of a system $\sum$ can be expressed as a function of the observability, the data quality and the performance of the used detection algorithm:

$$Det = f(O, Q, P). \tag{7}$$

where $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ describes the link that exists between O, P, Q and the detectability.

As detailed above, we are only interested to study the data quality impact by considering (A1-A4). Thus, the detectability can be expressed as a function of the data quality:

$$Det = f(Q). \tag{8}$$

Data quality stands out as one of the most important criterion since it impacts the performance of the used detectability algorithm. We have to point out that there are global data quality issues that belong to the dataset (i.e. imbalanced data) and other local issues that pertain to variables (i.e. missing data or noisy data). Thus, each type of data quality acts differently on the detection task. The global quality issues (GQ) have an iso-impact on each feature X_i regardless its local quality problems (LQ_i) as shown below:

$$Q_i = GQ \times LQ_i \ , \ \forall \ i \in \mathbb{N}. \tag{9}$$

The GQ is the quality issues that concern the whole dataset which is described by

$$GQ = \prod_{j=1}^{m} GQ_j. \tag{10}$$

where m is the number of the considered global data quality problems GQ_i.

As for the local quality issues, their impacts differ from a variable X_i to another. The link between the local quality of a feature X_i with the different l quality problem that concern this feature is described by

$$LQ_i = \sum_{k=1}^{l} LQ_{ik}, \forall i \in \mathbb{N} \tag{11}$$

where LQ_{ik} is a local quality that depends on the quality characteristic q_{ik} and the feature importance weight w_i. Thus, LQ_{ik} is a complex function that connects these variables given by

$$LQ_{ik} = g(w_i, q_{ik}). \tag{12}$$

where $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ that describes the local quality of a feature X_i regarding a data problem k.

According to (9), (10) and (11), the quality of a feature X_i can be written by

$$Q_i = \prod_{j=1}^{m} GQ_j \times \sum_{k=1}^{l} LQ_{ik}. \tag{13}$$

Using (13) and (12), the quality of the recorded dataset can be explained in relation with the feature importance weights by

$$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} \Big[\prod_{j=1}^{m} GQ_j \times \sum_{k=1}^{l} g(w_i, q_{ik}) \Big]. \tag{14}$$

Hence, referring to the development detailed in (13) and (14), the detectability metric can be written as follows:

$$Det = \prod_{j=1}^{m} GQ_j \times \sum_{i=1}^{n} \sum_{k=1}^{l} g(w_i, q_{ik}). \tag{15}$$

5. The empirical data quality model

This section presents an empirical development of the detectability metric using the previously detailed problem formulation in Section 4. Thus, we propose to estimate the global and local data quality as proposed in (15). To do, it seems to be logical to estimate the parameters w_i from the ability of features to detect the abnormal mode of the system. However, it may be more difficult to estimate the global and local quality functions. Thus, we propose to estimate these elements empirically.

The features importance w_i are important parameters for any data analysis task that cannot be overlooked or marginalized. In this context, two solutions arise to define features importance: (i) based on the human expertise or (ii) based on manually collected data. The first solution seems to be easier, faster, less expensive, but imprecise. In fact, human expertise is limited in the case of complex problem. Since this work is the results of a practical approach, the second solution is adopted due to its precision. To do, data samples are collected carefully and manually and they are used to preliminary analyze the data and to quantify the importance of each feature. We refer here to the feature importance conducted implicitly by the *Random Forest* classifier based on the "Gini importance" [25]. According to this method, the importance of a feature X_i is computed by the sum of all impurity decrease measures of all nodes in the forest at which a split on X_i has been conducted and normalized by the number of trees [26]. The impurity for a tree t is usually computed by the Gini impurity given below

$$G^t(X_i) = \sum_{K=1}^{Category(X_i)} p_a(K) \times G(K). \tag{16}$$

where X_i is the feature, $p_a(K)$ is the fraction of category K in a feature X_i and $G(K) = \sum_{a=1}^{C} p_a(K) \times (1 - p_a(K))$ is the gini index of a category K.

Then, the feature importance is obtained as follow.

$$w_i = \frac{1}{n_{tree}} [1 - \sum_{t=1}^{n_{tree}} G^t(X_i)] \tag{17}$$

where n_{tree} is the number of trees.

We then turn to estimate the local and global quality functions. For that purpose, we considered 10 datasets (real and simulated datasets) and we tested the most used fault detection techniques in order to study their behavior regarding the data problems. Table 2 presents the training datasets which are used to study the behavior of the most used fault detection algorithms regarding data quality problems. In this study, the used algorithms include:

- Artificial neural network (ANN): Given a set of features and a target, an ANN 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 [33].

Dataset	Number of features	Number of instances	Application domain	Reference
Credit card	24	30000	Credit card default	[27]
DBWorld e-mails	4702	64	Announces detection	[28]
BCWD	10	699	Breast cancer detection	[29]
Car Evaluation	6	1728	Car safety detection	[30]
Balloons	4	16	Cognitive psychology	[31]
Audit	18	777	Fraudulent firm detection	[32]
Dataset 1	5	10000	Artificial data	-
Dataset 2	10	10000	Artificial data	-
Dataset 3	15	10000	Artificial data	-
Dataset 4	20	10000	Artificial data	-

Table 2: Details of the training datasets.

- Decision tree (DT): 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 [34].
- Support vector machine (SVM): The 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).
- K-nearest neighbors (KNN): The KNN classifier consists in predicting the class of a new point based on the classes of the k closest instances to this later [35].
- Naive Bayes (NB): The NB algorithm 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. More details about this technique are presented in this work [36].

Before detailing the obtained data quality models, this paragraph describe the injection of data quality problem in the training datasets. For the missing data problem, original values are replaced by the value 0. As mentioned above, a value is considered as noisy only if it impacts the detection result. However, variables are dependent which means that a variable X_i can be considered noisy or not regarding the accuracy of other features. For this, we randomly add noises ϵ_i to each feature X_i (such as $-mean(X_i) \leq \epsilon_i \leq mean(X_i)$) and we evaluate if these noises affect the detection result. Then we define the noise threshold for each feature X_i as $mean(\epsilon_i)$. Thus, added noises is superior than these thresholds. For the imbalanced data, the instances number of the faulty class is modified in order to create a between-class imbalance. More than 10^5 simulations have been carried out with different quality configurations. For each configuration, the data detectability is assessed. The overall mean of these simulation results is then used to develop a global detectability model taking into account each data quality issue (i.e. Imbalanced, missing and noisy issues). The obtained models are detailed below.

- **Imbalanced data model:** Numerical simulations performed on the different datasets have shown that detectability increases exponentially in function of the imbalanced data ratio. This evolution is illustrated in Fig. 4 and shows that the imbalanced data quality issue has no impact on the detectability result if its ratio is greater than 50%. The global quality, defined in (10) for $m = 1$ (since we only consider the imbalanced data as a global quality issue), is then given by

$$GQ(q_{Im}) = 1 - 0.52 \times e^{-0.07 \times q_{Im}} \tag{18}$$

  where q_{Im} is defined in (2).
- **Missing data model:** Fig. 5 displays the detectability evolution as a function of the missing data ratio per feature. These results show that the detectability decreases as the the missing data ratio increases. The impact of this problem is clearer when the missing ratio exceed 40%.

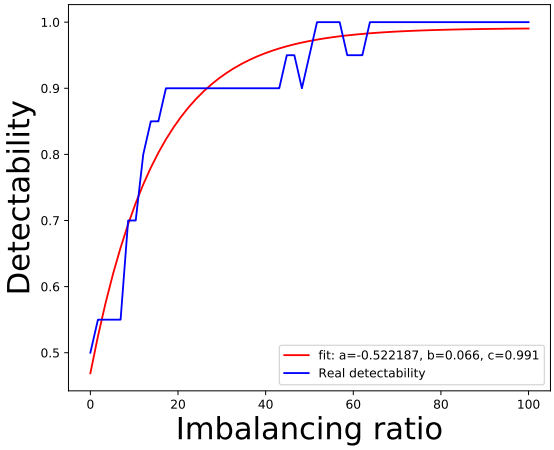

Figure 4: Detectability evolution as a function of the imbalanced data ratio.

The local quality, related to the missing data ratio q_{i1} of a feature X_i, depends on the evolution function of Fig. 5 and multiplied by the term $\frac{w_i}{w_{i_{min}}}$ to describe the detectability evolution function of

$$LQ_{i1} = g(w_i, q_{i1}) = \frac{w_i}{w_{i_{min}}}[1 - 2.10^{-6} \times (q_{i1}^2 + e^{0.11 \times q_{i1}})] \tag{19}$$

where q_{i1} is defined in (5).

- **Noisy data model:** Fig. 6 displays the detectability evolution function of the noisy data ratio per feature. These results show that detectability decreases when the noisy data ratio increases. Similarly to the missing data issue, the impact of this problem is more clearer when the noisy ratio exceed 20%.

  Therefore, the local quality related to the noisy data is given by

$$LQ_{i2} = g(w_i, q_{i2}) = \frac{w_i}{w_{i_{min}}}[1 - 10^{-6} \times (q_{i2}^2 + e^{0.07 \times q_{i2}})]. \tag{20}$$

  where q_{i2} is defined in (4).

- **Final detectability model:** The previously detailed models details one global data quality issue (imbalanced data) and two local data quality problem (noisy and missing data). By substituting them into the detectability model proposed in (15), the final detectability model can be defined as follows:

$$\begin{aligned} Det &= GQ(q_{Im}) \times \sum_{i=1}^{n} [g(wi, q_{i1}) + g(wi, q_{i2})] \\ &= (1 - 0.52 \times e^{-0.07 \times q_{Im}}) \times \sum_{i=1}^{n} \frac{w_i}{w_{i_{min}}} [2 - 2.10^{-6} \times (q_{i1}^2 + e^{0.11 \times q_{i1}}) - 10^{-6} \times (q_{i2}^2 + e^{0.07 \times q_{i2}})]. \end{aligned} \tag{21}$$

To better understand the impact of the studied data quality issues on the fault detection task, Fig. 7 displays the detectability map in function of the basic data quality issues. In fact, it is shown that the imbalanced data ratio have a fatal impact on the detectability when it is less than 20%. Moreover, the missing data ratio have an important impact when it is greater than 80%. One should note that the proposed detectability metric is derived from a real understanding of the behavior of the data and as well as the related data quality problems.

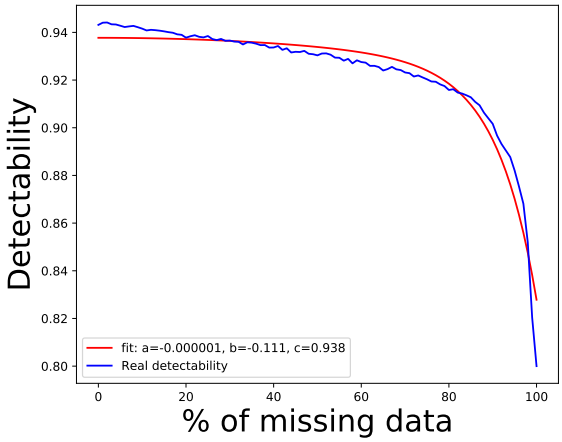

Figure 5: Detectability evolution function of the missing data ratio per feature.

From a technological point of view, it is expensive to guarantee these data quality levels for some variables. Also, it is costly for the company to waste time until reaching the required imbalanced data ratio. The goal is to determine the missing and noisy data ratios q_{i1} and q_{i2} for each variable X_i, so as to minimize total cost of negative detection (false alarm) and investment in the data acquisition technologies while maintaining a fixed detection level. Hence, we can formulate the problem as follows

$$\begin{cases} \min\ [CI \times q_{Im} + CD \times (100 - GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi, q_{i1}) + g(wi, q_{i2})]) + \sum_{i=1}^{n} CM_j \times (100 - q_{i1}) \\ \quad + \sum_{j=1}^{n} CN_j \times (100 - q_{j2})] \\ subject\ to: \\ \qquad 80 \leq GQ(q_{Im}) \times \sum_{i=1}^{n}[g(wi, q_{i1}) + g(wi, q_{i2})] \leq 100 \\ \qquad 0 \leq q_{i1} \leq 100,\ for\ i = 1, \ldots, n \\ \qquad 0 \leq q_{i2} \leq 100,\ for\ i = 1, 2, ..., n \end{cases} \tag{22}$$

where CD is the cost of a negative detection (false alarm), CM_j and CN_j are respectively the costs to guarantee a missing data ratio q_{j1} and a noisy data ratio q_{j2}, CI is the cost of wasting time until the data balancing and n is the number of variables.

Finally, we come to assess the accuracy of the developed model. A set of numerical simulations is used to validate the detectability model. Thus, the previously used fault detection algorithms are tested to define their behavior regarding to the data problems. Table 3 shows the results of the validation steps. For each dataset, 500 data quality configurations are tested. Results show that the developed model is able to predict the general evolution of detectability as a function of the quality of the used data. Detectability is predicted with an RMSE less than 0.1. We can remark that the difference between actual and expected detectability does not only belong to the developed model, but also to other data quality issues (i.e. outliers) that are not taken into account in this work.

Next section presents a practical application of this work in a real case study.

6. Case study and results

We here consider the Scoder case study as a real application of the proposed approach. Scoder is a French SME specialized in ultra-precise stamping for automotive applications. This case study consists of sheet metal forming

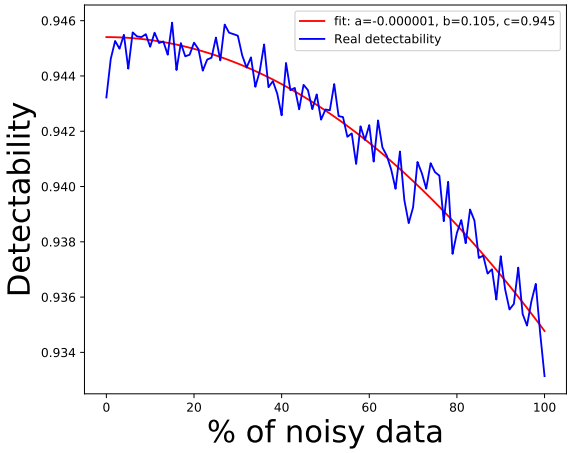

Figure 6: Detectability evolution function of the noisy data ratio per feature.

Dataset	Reference	# of features	# of instances	Application domain	RMSE
Diagnosis	[37]	6	120	Inflammation detection	0.05
Spam	[38]	57	4601	Spam detection	0.09
Blood Transfusion	[39]	4	748	Blood Transfusion	0.08
Caesarian	[40]	5	80	Caesarian Section detection	0.04
Cryotherapy	[41]	6	90	Wart treatment	0.09

Table 3: Results of the model validation step.

lines. The objective of the PHM project is to ensure a stable production by reducing machine failures and improving the productivity and the quality. The production performance is affected by the used metal coil characteristics. For that purpose, a PHM study is conducted to determine if a sheet metal is suitable for production or not according to the quantity of non-conform parts produced. This study is based on the characteristics of the coil, the caused press breakdowns, and the quality rate of the products fabricated from the sheet metal coil. The aim here is to identify the suitability of each metal coil with 80% as minimum rate of performance. Algorithm .1 presents the steps to be followed.

Algorithm .1 Data quality management algorithm

Step 1: Identify the problem and understand it.
Step 2: Collect some samples that can describe the problem.
Step 3: Compute the features importance and identify the most important ones.
Step 4: Apply the data quality models and identify the requirements according to objective.
Step 5: Install the data acquisition system.
Step 6: Control and improve the results.

A data inventory is first conducted to collect all the data that can be useful for project. Then, 24 variables are identified which consist in 12 metal proprieties, 6 types of machine's breakdown and 6 kinds of product's non conformity. Samples of these data are collected carefully and manually and they are used to preliminary analyze the data and to quantify the importance of each feature. Only 5 features are identified as pertinent for the study. Thus, the rest of features are eliminated and the rest of the study is based on these five variables. Table 4 shows the features importance from the used data subset. In fact, it is proven that the 5^{th} variable is the most important one to identify the capacity of the used coil to produce good quality parts.

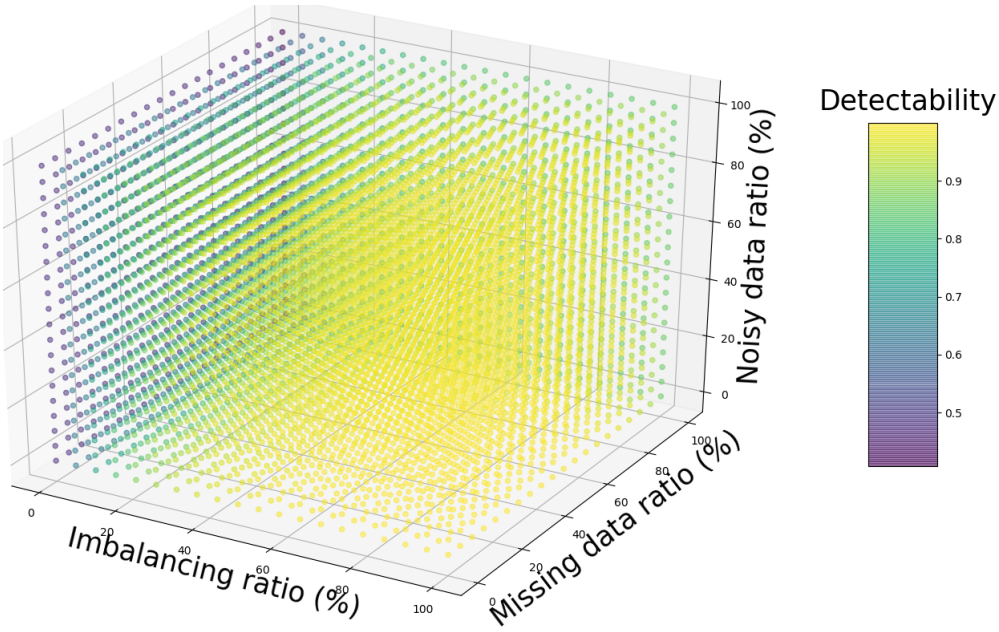

Figure 7: Detectability map in function of the basic data quality issues.

Variable	var_1	var_2	var_3	var_4	var_5
w_i	0.11	0.09	0.14	0.09	0.57

Table 4: Features importance in the Scoder case study.

Once the features are identified, the data quality issues for the Scoder dataset are analyzed (see Fig. 8). Results show that its authorized to have an imbalanced data ratio greater than 50%. Moreover, a percentage less than 20% and 30% of noisy data and missing data, respectively, has no impact on the detectability results.

A simple technique to set data quality requirements to satisfy the fixed objectives is to use these thresholds ($q_{Im} \geq 50\%$, $q_{i1} \leq 30\%$ and $q_{i2} \leq 20\%$) to guide the PHM implementation for the Scoder case study. However, this solution don't take into account the cost and the time to guarantee these data quality levels.

For the Scoder case study, advanced sensing technologies are required to ensure high data quality level for the fifth variable. As for the other variables, it can be done easily. In addition, it takes a lot of time to have a balanced dataset. For that reason, a significant cost is allocated to the imbalanced data ratio without forgetting the important cost of a negative detection. Table 5 shows the magnitude of these costs.

CD	CI	CM_1	CM_2	CM_3	CM_4	CM_5	CN_1	CN_2	CN_3	CN_4	CN_5
10	8	1	1	1	1	3	1	1	1	1	3

Table 5: Magnitude of different costs for the Scoder application.

Newton's optimization technique [42] is used to minimize the cost function given in (22) and identify the requirements according to Scoder objective.

The results of this problem is given in Table 6.

As a matter of fact, it is allowed to have 29% of missing data for var_5 and up to 60% for some other variables. For the noisy data, the percentages are between 36% and 77%. According to the developed data quality model, this configuration results in a detectability of 90% which satisfies the fixed objective. These requirements are respected during the installation of the data acquisition system. The expected installation cost is 11.05 MU which is optimized to be suitable for SMEs with limited resources. Figure 9 shows the Scoder data acquisition system which is based on

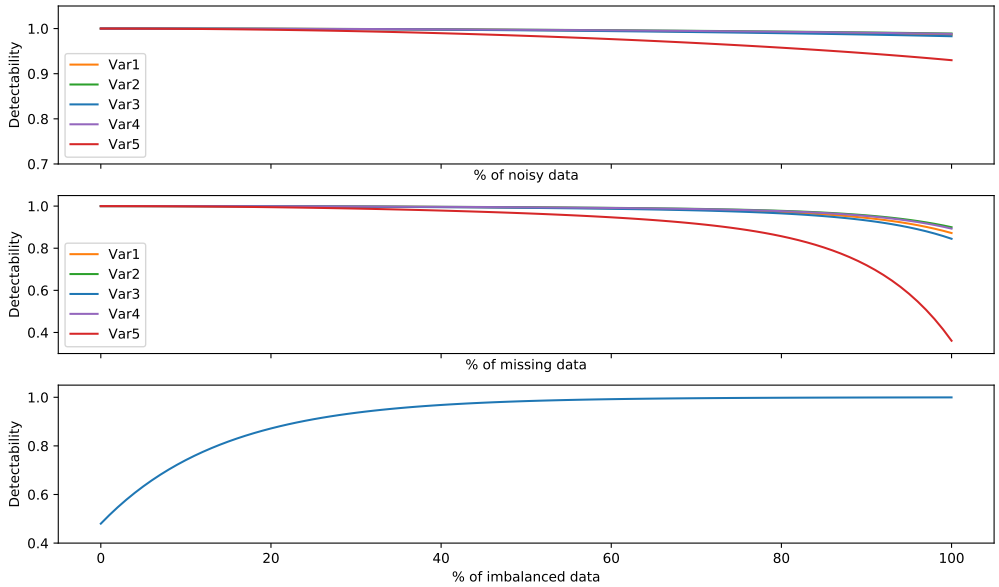

Figure 8: Synthesis in the Scoder case study.

Det	q_{Im}	q_{11}	q_{21}	q_{31}	q_{41}	q_{51}	q_{12}	q_{22}	q_{32}	q_{42}	q_{52}
90%	30%	25%	60%	6%	28%	29%	53%	36%	77%	71%	67%

Table 6: Data quality requirements for the Scoder application.

simple tablets to collect data throughout the production chain.

The properties of the metal coils are tested in a specialized test station where the previously defined requirements are met. The dataset was collected in more than 6 months during which the data quality has evolved to meet the defined requirements. Table 7 displays the evolution of the expected detectability versus the real one using the previously detailed detection algorithms. The results show that the developed model is able to predict the general evolution of the detectability as a function of the quality of the used data.

7. Discussion

This work proposes a new model to assess data quality and quantify their impact on the fault detection task of the PHM process. This model allows to define a set of data quality requirements to satisfy a fixed objective regarding the

Month	**Requirements**							**Real detectability (%)**					**Expected detectability (%)**
	Imbalanced (%)		**Var1**	**Var2**	**Var3**	**Var4**	**Var5**	**DT**	**SVM**	**ANN**	**KNN**	**NB**	
M1	5	**Missing (%)**	94	79	48	83	66	45	**50**	**50**	**50**	**50**	50
		Noisy (%)	0	3	9	0	5						
M2	11	**Missing (%)**	55	84	65	45	61	80	50	50	**65**	80	67
		Noisy (%)	5	1	4	10	8						
M3	18	**Missing (%)**	29	53	49	5	55	95	55	**80**	65	50	79
		Noisy (%)	18	18	8	27	21						
M4	25	**Missing (%)**	31	48	15	47	4	**90**	55	75	65	60	87
		Noisy (%)	13	1	10	9	23						
M5	33	**Missing (%)**	35	45	17	54	32	95	65	**90**	75	50	92
		Noisy (%)	21	23	24	19	18						
M6	40	**Missing (%)**	32	41	2	48	28	**95**	70	90	85	90	94
		Noisy (%)	20	10	51	14	19						

Table 7: Evolution of the expected detectability versus the real one.

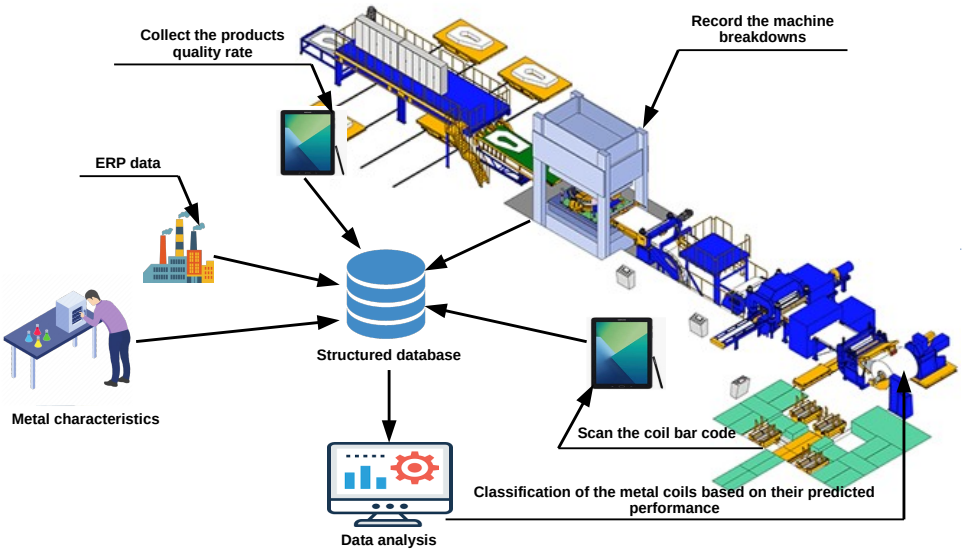

Figure 9: Details of the SCODER case study.

fault detection task. The algorithm .1 details the different steps to assess the suitability of data to the detection task. It should be noted that estimating the importance of features is a difficult task that has a large impact on the accuracy of the developed data quality model. In this work, we adopted a solution based on data samples collected manually and carefully. However, human expertise can be used to accomplish this task. In both cases, the task remains difficult, but from the point of view of the authors, the data quality cannot be represented independently of these parameters. Thus, the limitations resulting from the estimation of features importance should be considered further in future works.

As shown in Fig. 10, it is possible to formulate a clear idea of the data requirements to be satisfied from the set objectives and the available project budget. These requirements represent the needed data quality ratios. They can be extended to cover the data acquisition system, storage hub, analysis tools and devices and the frequency of collect.

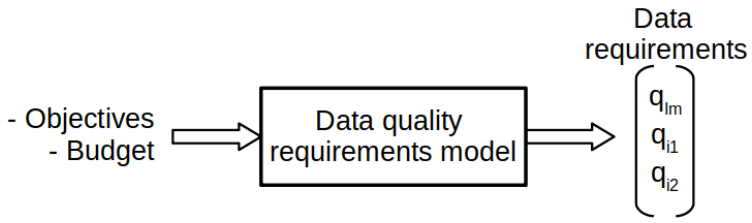

Figure 10: Identification of data quality requirements based on set objectives and available budget.

The proposed methodology for detectability assessment consists in starting from small dataset to quantify the features importance. Then, the developed models are used to set data quality requirements in line with the objective. The results are then used to install the needed data management infrastructure. Using the same methodology, other models for data diagnosability and trendability can be developed which allows to cover all the PHM process. 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 which is an understudied topic [5]. In the context of data quality management, it is necessary to talk about the data quality improvement techniques. Before investing in advanced sensing technologies it is more efficient to apply some algorithms to deal with data quality issues. Table 8 shows some examples of data quality improvement techniques that deal with the previously studied data quality issues.

Many works have been done to present and study the effectiveness of these techniques. However, to the best of

Imbalanced data	Missing data	Noisy data
- SMOTE [43]	- Mean imputation [44]	- AENN [45]
- CBO [46]	- SVM imputation [47]	- ModeFilter [48]
- GAN [49]	- Listwise [50]	- C45robustFilter [51]

Table 8: Data quality improvement techniques.

our knowledge, none of this work has studied them in a PHM context. In this context, a comparison between them in terms of complexity and cost of implementation, suitability for different types of data and their impact on the different PHM tasks is needed.

8. Conclusion

PHM discipline is widely used as a framework for data management and knowledge extraction. For a long time, PHM users consider that available data are suitable for PHM analysis without assessing its adequacy with the fixed objectives. However, the impact of the used data quality on the results of the PHM results still an understudied domain. As a summary, this paper proposed to study the data quality issues in the PHM context while evaluating their impact on the detection task. The conducted developments were applied to a real case study and the obtained results were reported and discussed. It should be noted that a PHM approach is not a straightforward tool and must combine staff knowledge with advanced analysis models. Thus, a good understanding of the problem through human expertise is a mandatory step for a PHM application.

To conclude, this study provided a first empirical model for data quality requirements identification for PHM applications. Only the detection task of the PHM process was considered. One should think about studying the impact of data quality issues on the rest of the PHM tasks (diagnosability and predictability, for example). In addition, this work can be considered as a first step to evaluate the data suitability for a defined PHM. Further work should be developed to define a technical protocol for data quality evaluation and improvements in a PHM context. This may allow to reduce the time and the cost of data processing and to improve the decision accuracy. Three research gaps arise and they could be considered as future works:

- Further theoretical formalization of the proposed model;
- Extension of the proposed model to cover the fault diagnosis and degradation prediction tasks;
- Development of a PHM cost model to optimize the PHM process implementation.

References

[1] 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.

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

[3] N. Julka, A. Thirunavukkarasu, P. Lendermann, B. P. Gan, A. Schirrmann, H. Fromm, E. Wong, Making use of prognostics health management information for aerospace spare components logistics network optimisation, Computers in industry 62 (2011) 613–622.

[4] R. Gouriveau, K. Medjaher, N. Zerhouni, From prognostics and health systems management to predictive maintenance 1: Monitoring and prognostics, John Wiley & Sons, 2016.

[5] N. Omri, Z. Al Masry, N. Mairot, S. Giampiccolo, N. Zerhouni, Industrial data management strategy towards an sme-oriented phm, Journal of Manufacturing Systems 56 (2020) 23–36.

[6] I. Trabelsi, M. Zolghadri, B. Zeddini, M. Barkallah, M. Haddar, Fmeca-based risk assessment approach for proactive obsolescence management, in: IFIP International Conference on Product Lifecycle Management, Springer, 2020, pp. 215–226.

[7] Z. Al Masry, N. Omri, C. Varnier, B. Morello, N. Zerhouni, Operating approach for fleet of systems subjected to predictive maintenance, in: Euro-Mediterranean Conference on Mathematical Reliability, 2019.

[8] E. L. S. Teixeira, B. Tjahjono, S. C. A. Alfaro, A novel framework to link prognostics and health management and product–service systems using online simulation, Computers in Industry 63 (2012) 669–679.

[9] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, D. Siegel, Prognostics and health management design for rotary machinery systems—reviews, methodology and applications, Mechanical systems and signal processing 42 (2014) 314–334.

[10] S. Datta, S. Sarkar, A review on different pipeline fault detection methods, Journal of Loss Prevention in the Process Industries 41 (2016) 97–106.
[11] X. Jia, M. Zhao, Y. Di, Q. Yang, J. Lee, Assessment of data suitability for machine prognosis using maximum mean discrepancy, IEEE transactions on industrial electronics 65 (2017) 5872–5881.
[12] Y. Chen, F. Zhu, J. Lee, Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method, Computers in industry 64 (2013) 214–225.
[13] ISO/IEC, Iso 80008:2015 data quality part 8: Information and data quality: Concepts and measuring, in: ISO/IEC, Tech. Rep. ISO/IEC 8000, 2015, 2015.
[14] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93.
[15] ISO/IEC, Software engineering software product quality requirements and evaluation (square) data quality model, in: ISO/IEC, Tech. Rep. ISO/IEC 25012, 2008, 2008.
[16] L. Sebastian-Coleman, Measuring data quality for ongoing improvement: a data quality assessment framework, Newnes, 2012.
[17] D. McGilvray, Executing data quality projects: Ten steps to quality data and trusted information (TM), Elsevier, 2008.
[18] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha, Data quality: A survey of data quality dimensions, in: 2012 International Conference on Information Retrieval & Knowledge Management, IEEE, 2012, pp. 300–304.
[19] T. C. Redman, Data Quality for the Information Age, 1st ed., Artech House, Inc., Norwood, MA, USA, 1997.
[20] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., 1988.
[21] C. Batini, M. Scannapieco, Data and information quality: Concepts, methodologies and techniques, 2016.
[22] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artificial intelligence review 22 (2004) 85–126.
[23] T. Roy, S. Dey, Fault detectability conditions for linear deterministic heat equations, IEEE control systems letters 3 (2018) 204–209.
[24] S. X. Ding, Model-based fault diagnosis techniques: design schemes, algorithms, and tools, Springer Science & Business Media, 2008.
[25] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[26] S. Nembrini, I. R. König, M. N. Wright, The revival of the gini importance?, Bioinformatics 34 (2018) 3711–3718.
[27] I.-C. Yeh, C.-h. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications 36 (2009) 2473–2480.
[28] M. Filannino, Dbworld e-mail classification using a very small corpus, The University of Manchester (2011).
[29] W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology., Proceedings of the national academy of sciences 87 (1990) 9193–9196.
[30] M. Bohanec, V. Rajkovic, Knowledge acquisition and explanation for multi-attribute decision making, in: 8th Intl Workshop on Expert Systems and their Applications, 1988, pp. 59–78.
[31] B. Ross, T. Shultz, G. Silverstein, E. Wisniewski, et al., The influence of prior knowledge on concept acquisition: Experimental and computational results (1990).
[32] N. Hooda, S. Bawa, P. S. Rana, Fraudulent firm classification: a case study of an external audit, Applied Artificial Intelligence 32 (2018) 48–64.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] J. Czerniak, H. Zarzycki, Application of rough sets in the presumptive diagnosis of urinary system diseases, in: Artificial intelligence and security in computing systems, Springer, 2003, pp. 41–51.
[38] Y. Wang, I. H. Witten, Modeling for optimal probability prediction (2002).
[39] I.-C. Yeh, K.-J. Yang, T.-M. Ting, Knowledge discovery on rfm model using bernoulli sequence, Expert Systems with Applications 36 (2009) 5866–5871.
[40] M. Amin, A. Ali, Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions, Wavy AI Research Foundation: Lahore, Pakistan (2018).
[41] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, S. Nahavandi, An expert system for selecting wart treatment method, Computers in biology and medicine 81 (2017) 167–175.
[42] A. Fischer, A special newton-type optimization method, Optimization 24 (1992) 269–284.
[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[44] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, K. G. Moons, A gentle introduction to imputation of missing values, Journal of clinical epidemiology 59 (2006) 1087–1091.
[45] N. Jaques, S. Taylor, A. Sano, R. Picard, Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 202–208.
[46] Z. Zheng, Y. Cai, Y. Li, Oversampling method for imbalanced classification, Computing and Informatics 34 (2016) 1017–1037.
[47] H. Mallinson, A. Gammerman, Imputation using support vector machines, Department of Computer Science. Royal Holloway, University of London. Egham, UK (2003).
[48] S. Jin, R. Kikuuwe, M. Yamamoto, Real-time quadratic sliding mode filter for removing noise, Advanced Robotics 26 (2012) 877–896.
[49] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, C. Malossi, Bagan: Data augmentation with balancing gan, arXiv preprint arXiv:1803.09655

(2018).
[50] G. King, J. Honaker, A. Joseph, K. Scheve, List-wise deletion is evil: what to do about missing data in political science, in: Annual Meeting of the American Political Science Association, Boston, 1998.
[51] X. Zhu, X. Wu, Class noise vs. attribute noise: A quantitative study, Artificial intelligence review 22 (2004) 177–210.