Data Management Framework for Risk Estimate of Electronic Boards in Drilling and Measurement Tools ⋆

Jinlong Kang [*,**] **Zeina Al Masry** [*] **Christophe Varnier** [*]
Ahmed Mosallam [**] **Noureddine Zerhouni** [*]

[*] *SUPMICROTECH, CNRS, institut FEMTO-ST, Besançon 25000 France (e-mail: jinlong.kang, zeina.almasry, christophe.varnier, noureddine.zerhouni@femto-st.fr).*
[**] *Schlumberger Riboud Product Center, SLB, Clamart 92140 France (e-mail: jkang5, amosallam@slb.com)*

Abstract: With computer science and technology development in today's world, many traditional industries, such as the oil and gas industry, are beginning to transform to digitalization. In this transformation process, many data-driven models are often necessary; e.g., a data-driven model, based on existing data, is used to estimate the risk associated with drilling tools. Before building this model, the preliminary work needs to assess how much data are available at this stage, what is the quality of the data, whether the existing data are suitable for building the model, and if not, what measures can be taken to improve the data quality. To answer these questions, this paper presents a data management framework that includes data preparation, data quality assessment, and data-based knowledge acquisition. An actual case study result demonstrates that the framework can answer these questions.

Keywords: data management, data quality, risk estimate, electronic board, environmental exposure

1. INTRODUCTION

Drilling and measurement (D&M) tools used for oil well drilling are complex systems that contain multiple electronic boards. These boards play a vital role in the tool's functions, such as signal acquisition, processing, and operation control. However, the challenging downhole operating conditions, i.e., high temperature, vibration, and shocks, cause the electronic boards to fail in complex ways, resulting in drilling job failures and significant economic losses. One way to enhance drilling job success and efficiency is to plan proactive tool maintenance activities. Furthermore, estimating the risk of electronic board failures is essential to improve tool maintenance decision-making. In the context of maintenance, the risk is defined as the probability of failure multiplied by the consequence of the failure. Assuming the consequence of the failure is constant for the same type of electronic board; then the risk is equivalent to the probability of failure.

Due to the specificity of D&M tools, there are only a few research works about the risk estimate of electronic boards in D&M tools. For example, Kale et al. (2014) used a stress function with two reliability functions to estimate the risk of electronic boards. The stress function is built based on exposure time at different environmental levels. Kang et al. (2022) proposed a risk level estimate method based on the hidden Markov model. These methods are data-driven and assume that the data are available and of high quality. However, the assumed high-quality data are usually unavailable for the D&M tool electronic boards for many reasons, such as tool lost in the hole because of drilling accidents, computerized maintenance management system transformation, communication errors during drilling operations, and field engineers forgetting to dump the data. Ignoring the data quality issues, e.g., missing values and outliers, might result in misinformed analyses. Therefore, data quality assessment is of great significance before constructing data-driven models.

Given the data quality challenges for prognostics and health management applications, researchers have put forth various data quality management methodologies. For instance, Omri et al. (2021) proposed a data quality requirement model for fault diagnosis based on empirical classification results of public classification datasets. Jia et al. (2018) presented a data suitability assessment method for machine prognosis using maximum mean discrepancy. Chen et al. (2013) introduced a new method to evaluate and improve data quality for system health diagnosis modeling. However, to our knowledge, there are no studies on data quality management for risk estimate, especially for electronic boards or systems. To fill this research gap, this paper proposes a data management framework to deal with data quality issues in risk estimation for electronic boards in D&M tools.

The remainder of the paper is organized as follows. In Section 2, the data management framework is presented as well as the description of data quality dimensions. A case

⋆ This work is supported by the EIPHI graduate school (contract "ANR-17-EURE-0002")

study is then presented in Section 3. A discussion and a conclusion are presented in Sections 4 and 5, respectively.

2. DATA MANAGEMENT FRAMEWORK

Figure 1 shows the proposed data management framework for estimating the risk of electronic boards. The framework consists of three main segments, including data preparation, data quality, and data-based knowledge. These segments are described in detail in the following sections.

2.1 Data Preparation

The data preparation consists of three steps, i.e., data acquisition and storage, channel selection, and feature extraction.

Data Acquisition and Data Storage In drilling operations, each D&M tool measures many analog signals through built-in sensors. The analog signals are then converted to digital signals via signal acquisition circuits or analog-to-digital converters; however, limited data are transmitted to the surface due to bandwidth limitations. Most of the data are processed in real time by a field programmable gate array or a digital signal processor inside of the tool. The processed information is then input into a central processing unit (CPU) for automatic control of the tool or stored on a memory board for offline data analysis. After the drilling job is completed, the tool is pulled out of the well. The data stored on the memory board will then be copied on the hard disk and subsequently uploaded to a data cloud when the Internet is available. The raw data from the cloud can be retrieved and used for building the risk estimate model of the electronic boards in the D&M tool.

Channel Selection As previously mentioned, the D&M tool acquires numerous information channels containing data during a drilling job; however, only a few channels are used to build the risk estimate model. In general, channels containing environmental data, such as temperature and vibration are selected because environmental exposure significantly impacts the electronic board condition (Michael G. Pecht, Myeongsu Kang, 2018). It is important to note that humidity is not considered in modeling the risk of the electronic boards. Because the boards are mounted inside the tool, and nitrogen is filled into the tool before each drilling job, it is unlikely that the boards will be exposed to moisture.

Feature Extraction Histogram features are commonly used for the risk estimate of electronic boards (Kang et al., 2022; Kumar et al., 2012). In estimating the risk of the electronic boards in D&M tools, the histogram features are the board environmental exposure time under different environmental levels. In this case, the environmental levels are equivalent to the histogram bins, and the exposure time is the product of the corresponding histogram frequencies and the data recording rate.

2.2 Data Quality

Assessing the data quality is an essential procedure before creating the risk estimate model. Through data quality assessment and data-based knowledge of data quality vs. model performance, which will be described Section 2.3, it is possible to assume the data quality requirement for building a risk estimate model with desired performance. If the data do not satisfy the condition, then data quality improvement techniques should be made. In this section, we will initially introduce and define data quality followed by describing the metrics used to indicate data quality and different ways to improve the data quality.

Definition There is no definitive definition of data quality. Information quality describes the extent to which information is fit for purpose (Lee et al., 2002; Fadahunsi et al., 2021). Data quality refers to how well data meet the requirements of data consumers (Karkouch et al., 2016). Data are usually regarded as being high quality if it is suitable for its intended use in business, decision-making, and planning. In this paper, the data quality for the risk estimate of electronic boards in D&M tools is defined as "*Data quality refers to the fitness of use for risk estimate*".

Metrics The data quality metric is also termed by various researchers as data quality dimension (Wang and Strong, 1996), data quality indicator (Wang et al., 2019), or data quality characteristics (Gualo et al., 2021). Similar to the data quality definition, many types of data quality metrics have been proposed and presented in the literature. For example, in International Standard ISO/IEC 25012 (ISO/IEC 25012:2008), five data quality metrics, namely, accuracy, completeness, consistency, credibility, and currentness. Klein and Lehner (2009) used five metrics to represent the data quality in sensor data streaming environments, including accuracy, confidence, completeness, data volume, and timeliness. Rekatsinas et al. (2015). appraised the quality of data sources using coverage, accuracy, timeliness, and position bias. Wang and Strong (1996) proposed four categories of data quality metrics; i.e., intrinsic data quality, contextual data quality, representational data quality, and accessibility data quality, in addition to several subcriteria for each category. Although many data quality metrics are introduced, it is not recommended that all of them be used because based on the definition of data quality, data quality highly depends on the intended application scenario. In this paper, three metrics (data volume, accuracy, and completeness) are used to quantify the quality of the environmental exposure data used for building the risk estimate model. These three metrics are also widely used in prognostic and health management applications (Omri et al., 2021).

Data volume is often regarded as the most important data quality metric. For the risk estimate of electronic boards in D&M tools, the data volume relates to the two following aspects:

- n: the number of failed electronic boards, $n > 0$
- N_i: the number of observations of failed board i, $N_i > 0$

Each observation of the data represents the histogram features extracted using the environmental exposure data of the selected channel from a drilling job.

Accuracy is the degree to which the data values reflect the actual event state in a specific context of use (ISO/IEC

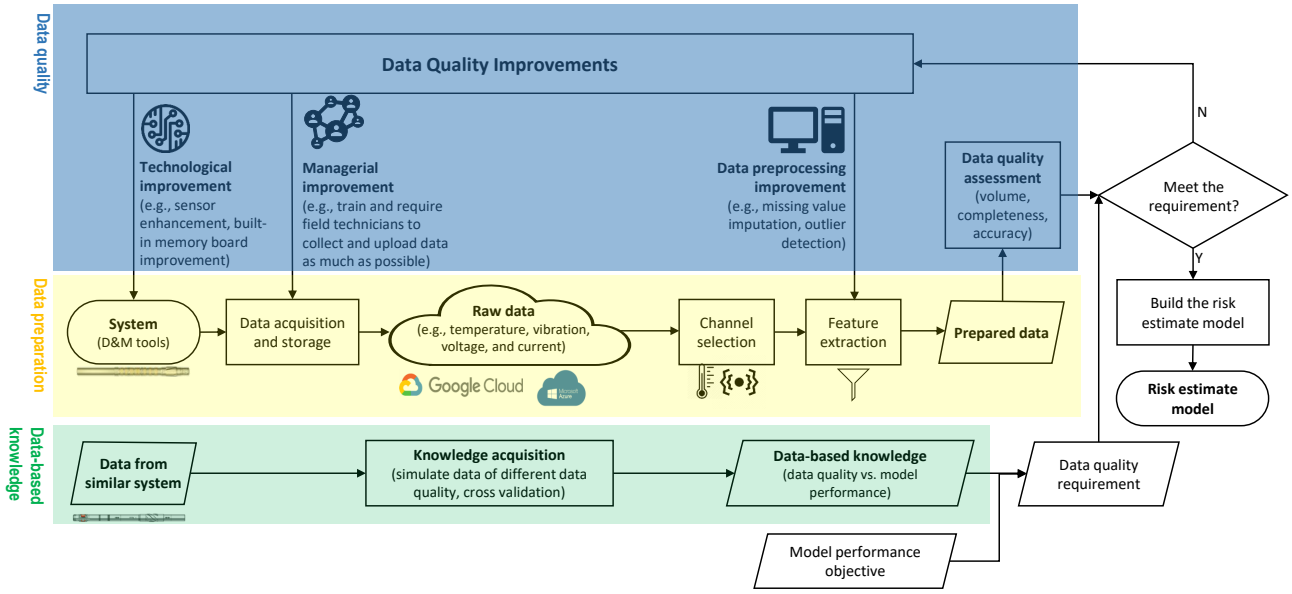

Fig. 1. The data management framework.

25012:2008). In our context, the event is a drilling job, the state is the operational condition, and the data values are the extracted histogram features; i.e., the observations. Assume that an outlier is an observation that does not accurately represent the corresponding real operating environment. As a result, the data accuracy of the electronic board i is defined as follows:

$$Accuracy_i = 1 - \frac{|O_i|}{N_i} \tag{1}$$

where O_i denotes the ensemble of outliers of the electronic board i, $|\bullet|$ refers the cardinality of the data space, and $0 \leq |O_i| \leq N_i$.

Completeness is the data quality metric that handles the problem of a missing value (Omri et al., 2021). There is only one type of missing data in our situation. The environmental exposure data of some drilling jobs are completely missing because the field engineers might not have uploaded the data to the cloud, or the memory board was damaged or lost in hole due to drilling accidents. Similar to data accuracy, the completeness of the electronic board i is expressed as follows:

$$Completeness_i = 1 - \frac{M_i}{N_i} \tag{2}$$

where M_i denotes the number of drilling jobs without collecting environmental exposure data of the electronic board i and $0 \leq M_i \leq N_i$.

Improvements Using the previously mentioned metrics makes it possible to assess the data quality of the available data. Furthermore, it is also conceivable to obtain knowledge about the relationship between data quality and risk assessment model performance during simulations using the data collected from similar systems. A detailed description of acquiring the knowledge will be presented later in Section 2.3. Given the knowledge and model performance goals, one can conclude the data meets the data quality required to build a risk estimate model with expected performance. Data quality improvement activities should be conducted if the data does not satisfy the requirement. The data quality can be enhanced from three aspects; i.e., technological improvement, managerial improvement, and data preprocessing improvement. Technological improvement is to improve the data quality from a technology perspective; e.g., enhance sensors to increase signal precision, which may well improve the accuracy of the data quality. Managerial improvement is to enhance the quality of data from a management perspective. As mentioned in the previous section, missing data are likely caused by field engineers because they failed to move the data from the memory board to the hard disk and upload it to the data cloud. Thus, training of field engineers should be enhanced and a key performance indicator of data collection ratio be set for the engineers to increase their awareness of data collection, which will improve the data quality of completeness. Both technological and managerial improvement measures are costly because they demand additional investment of time and budget. Furthermore, technological and managerial improvements require long-term continuous input, and it is not easy to realize a significant improvement in data quality in the short term. Data preprocessing improvement, however, would achieve quality improvement immediately. In this report, the authors will focus on data preprocessing improvement.

Data preprocessing improvement uses data preprocessing methods such as outlier detection and missing value imputation to improve data quality.

Outlier detection also known as anomaly detection or novelty detection, focuses on discovering those observations that are significantly different from most of the data (Zimek and Schubert, 2017). The outlier detection methods can be grouped into four categories, including statistical-based methods, distance-based methods, density-based methods, and clustering-based methods (Smiti, 2020). For additional details on outlier detection methods, refer to two review articles by Smiti (2020) and Chandola et al. (2009). It should be noted that the outlier detection method used depends heavily on

the context of use. In our context, outliers are attributed to measurement or recording errors. More specifically, in some drilling jobs, the total environmental exposure time measured by the tool differs from the drilling time due to recording errors. In reality, these two parameters should be equal because if the tool is drilling, the environmental data are being measured. There are two ways to handle outliers; one is to remove them directly, which is the easiest but reduces data completeness. The second way is to substitute the outliers with other values (e.g., mean, median).

Missing value imputation seeks to replace missing data with estimated values. The missing value imputation methods can be generally classified into two categories; i.e., statistical and machine learning-based methods (Lin and Tsai, 2020) (Hasan et al., 2021). The commonly used statistical-based methods for missing value imputation are expectation-maximum, linear regression, least squares, and mean or mode. Machine learning techniques, such as regression tree, random forest, support vector regression, and k-nearest neighbor are used widely in missing value handling methods. Additional details about missing value imputation are contained in review articles by Lin and Tsai (2020) and Hasan et al. (2021).

2.3 Data-based Knownledge

Typically, if the system under study is a recent implementation or the system is not commonly used, it is not easy to accumulate sufficient high-quality data to gain knowledge about the relationship between data quality and model performance. In this case, an option is to use a similar system containing more data than the system under study. These data can be used to accumulate the knowledge. In this paper, this knowledge is referred to as data-based knowledge and denoted by $\mathcal{K}(\boldsymbol{Q}, \ell)$, $\boldsymbol{Q}$ is a vector of data quality metrics, and ℓ is the loss to be defined later.

Loss Function It is difficult or impossible to know the actual risk level of an electronic board after each drilling job is completed. Therefore, traditional model evaluation metrics, such as classification error and mean squared error are not suitable for evaluating electronic board risk estimate model performance. However, for boards that have failed, their actual lifetime is known. Based on this information, the loss function (3) indicates the risk estimate model performance.

$$\ell(\boldsymbol{t}, \boldsymbol{T}) = \frac{\sum_{i=1}^{n} c_1 \mathbb{1}(t_i \geq T_i) + c_2 (T_i - t_i) \mathbb{1}(t_i < T_i)}{n} \quad (3)$$

where

- n: the number of failed electronic boards.
- t_i: the time when the electronic board i is replaced, assuming all electronic boards' lives begin at time 0. Specifically, t_i equals the time when the electronic board is estimated as being at the highest risk level by the risk estimate model.
- T_i: actual life of the electronic board i; i.e., the time when the electronic board actually failed.
- c_1: unit failure cost.
- c_2: premature replacement cost per unit of time.
- $\mathbb{1}$: an indicator function. In other words, $t_i \geq T_i$ means the electronic board is replaced too late, which involves failure cost. On the other hand, $t_i < T_i$ means the electronic board is replaced too early, which includes premature replacement cost.

Knowledge Acquisition Removing and modifying portion of the data in a similar system can simulate data with different data quality. It is possible to train risk estimate models based on these simulated data. Then based on these models, the same test data are used to estimate the lifetime of test boards and calculate the loss function. In addition, to better characterize the model losses, cross-validation techniques can be adopted. As a result, the knowledge $\mathcal{K}(\boldsymbol{Q}, \ell)$ can be obtained. A case study presented in Section 3 will demonstrate how to acquire the knowledge in more detail.

In

3. CASE STUDY

This section will use historical tool measurement data collected in the field from actual D&M tools to demonstrate how the data management framework functions. The system being studied is the CPU board of a specific logging-while-drilling tool. A similar system is the CPU board of a particular rotary steerable system tool. The risk estimate model used in this paper is based on the hidden Markov model (HMM). More details are contained in work by Kang et al. (2022).

3.1 Data-based Knowledge Acquisition

Due to space limitations of this paper and the difficulty of simulating data with different accuracy, this paper will only present how to acquire the knowledge of data volume and completeness vs. model performance in this case study. In addition, the paper presents only one aspect of data volume; i.e., the number of failed electronic boards n. Because electronic boards have different numbers of observations, randomly removing observations from the overall sample might result in uneven completeness among the electronic boards. The procedures for acquiring the knowledge are shown in Algorithm 1, where $\boldsymbol{Q} = [n, Completeness]$. In this case study, $m = 20$, $\boldsymbol{L}_1 = [5, 10, \ldots, 50]$, $\boldsymbol{L}_2 = [0.40, 0.45, \ldots, 1.00]$, $c_1 = 10000, c_2 = 3$ and $n_{test} = 50$.

3.2 Data Quality Assessment

The data from 14 electronic boards in the system under study were collected using the data preparation procedures. To better demonstrate the data quality effect on the risk estimate model, seven boards with low data completeness were selected as training data, and their data quality metrics are shown in Table 1. The other seven boards with data completeness close to 1 were used as test data. Based on the data quality metrics of the training data and the knowledge presented in Fig. 2, we can roughly foresee the loss would be more than 5500 if the risk estimate model is trained using this training data. If the model performance goal is that the loss should be less than 5500, then the data quality does not meet the requirement. In turn, data quality improvement measures need to be implemented.

Algorithm 1 Knowledge $\mathcal{K}(\boldsymbol{Q}, \ell)$ acquisition

Input: entire dataset, simulation times m, sequence $\boldsymbol{L}_1$ contains the numbers of training electronic boards, sequence $\boldsymbol{L}_2$ contains the completeness of each training electronic board, c_1, c_2, the number of test electronic boards n_{test}
Output: $\mathcal{K}(\boldsymbol{Q}, \ell)$

for $i \in \{1, 2, 3, \ldots, m\}$ **do**
  sampling observations of n_{test} boards from the entire dataset without replacement as the test data
  for n in $\boldsymbol{L}_1$ **do**
    sampling observations of n boards from the remaining data without replacement as temporary dataset $Temp$
    for $Completeness$ in $\boldsymbol{L}_2$ **do**
      remove some observations from $Temp$ to make the completeness of each board equal to $Completenss$, use the data after removal operation as the training data
      train the risk estimate model $\boldsymbol{\Omega}$ using the training data
      predict the lifetime of test boards using the model $\boldsymbol{\Omega}$
      calculate the loss function and store the result in $\mathcal{K}(\boldsymbol{Q}, \ell)$
    end for
  end for
end for

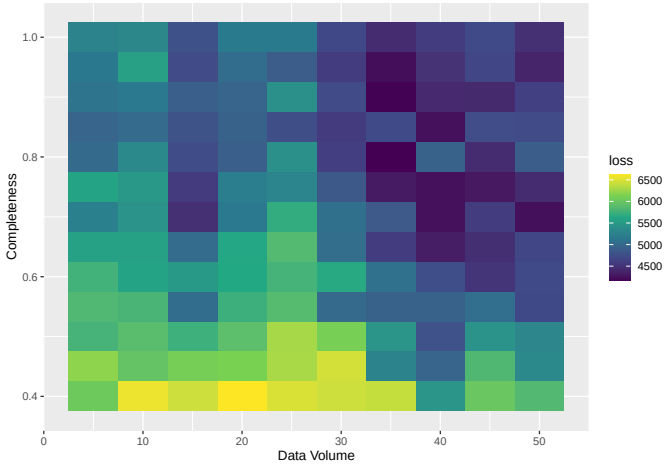

Fig. 2. Data-based knowledge $\mathcal{K}(\boldsymbol{Q}, \ell)$ visualization

Table 1. Data quality metrics of the training boards

Board i	1	2	3	4	5	6	7
N_i	69	3	8	27	11	21	14
$Completeness_i$	0.52	0.67	0.38	0.67	0.64	0.67	0.64

3.3 Risk Estimate Model Performance After Improving Data Quality

Mean imputation is adopted to fill in the missing values of the training data. The model performances before and after missing value handling are compared based on the loss function on the test data. Specifically, the loss before missing value handling is 8994.2, whereas, after missing value handling, it is 6098.6. The loss is reduced by about 3000, which proves that even simple data quality improvement techniques, such as mean imputation, can improve the model performance if the data quality of the training data is low.

4. DISCUSSION

This paper presents a data management framework to determine whether the collected data meet the data quality requirements to build a risk estimate model with the expected performance. The goal of this framework is to assist in making improved maintenance decisions, which explains why the failure cost and early replacement cost is applied in defining the loss function. Thus, this framework can also be extended to predictive model selection based on data quality. Specifically, data-based knowledge can be obtained about different predictive models through a simulation study. Then, based on this knowledge and the data quality assessment of the available data, it is possible to inform the decision maker the current best predictive model is. For example, in Fig. 3, the difference in loss between HMM and mean time to failure (MTTF) is shown. Notice that the MTTF model predicts the lifetime as the MTTF of training data. Because the MTTF model is uncomplicated, it has a larger bias but less variance than HMM. As a result, one should choose the MTTF model if the difference between the two models is insignificant and the decision maker is relatively conservative. Moreover, it is possible to label which model is the best for different data quality coordinates based on the performance of each model. Then, based on these labels, a classification model can be trained for more accurate model selection. In the case study in Section 3, the loss function for 10 data volumes and 13 completeness was presented. This knowledge accuracy can be enhanced through additional simulations. There are two main methods to perform this. First, increase the number of simulations. Second, increase the size and density of the grid for the data quality dimension.

5. CONCLUSION

This paper has presented a framework for data management that addresses data quality requirements for electronic board risk estimate models. The model builder can effectively decide whether the data meet the quality requirements by applying this framework. If the data do not meet the criteria, one can use the three methods mentioned (i.e., technological improvement, managerial improvement, and data preprocessing improvement) to improve the data quality. This paper successfully applied the framework to a real-world case in the oil and gas industry. The case study results show that the framework can effectively guide the construction of improved electronic board risk estimate models from a data quality perspective. Furthermore, extended uses of the framework (e.g., model selection) can

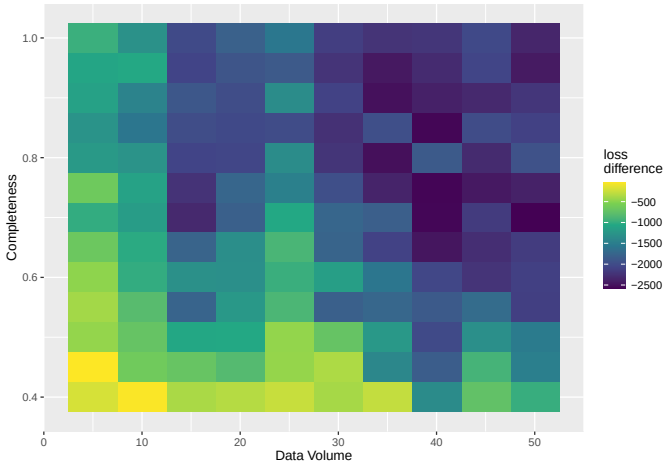

Fig. 3. Loss difference between HMM-based model and MTTF-based model

be used to acquire more accurate data-based knowledge through additional simulations.

ACKNOWLEDGEMENTS

This work is supported by the EIPHI graduate school (contract "ANR-17-EURE-0002").

REFERENCES

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3). doi:10.1145/1541880.1541882.

Chen, Y., Zhu, F., and Lee, J. (2013). Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method. *Computers in Industry*, 64(3), 214–225. doi: 10.1016/j.compind.2012.10.005.

Fadahunsi, K.P., O'Connor, S., Akinlua, J.T., Wark, P.A., Gallagher, J., Carroll, C., Car, J., Majeed, A., and O'Donoghue, J. (2021). Information quality frameworks for digital health technologies: Systematic review. *Journal of Medical Internet Research*, 23(5), e23479. doi: 10.2196/23479.

Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., and Piattini, M. (2021). Data quality certification using iso/iec 25012: Industrial experiences. *Journal of Systems and Software*, 176, 110938. doi: 10.1016/j.jss.2021.110938.

Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T., and Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, 27, 100799. doi: https://doi.org/10.1016/j.imu.2021.100799.

ISO/IEC 25012:2008 (2008). Software engineering — software product quality requirements and evaluation (square) — data quality model. Standard, International Organization for Standardization, Geneva, CH.

Jia, X., Zhao, M., Di, Y., Yang, Q., and Lee, J. (2018). Assessment of data suitability for machine prognosis using maximum mean discrepancy. *IEEE Transactions on Industrial Electronics*, 65(7), 5872–5881. doi: 10.1109/TIE.2017.2777383.

Kale, A.A., Carter-Journet, K., Falgout, T.A., Heuermann-Kuehn, L., and Zurcher, D. (2014). A probabilistic approach for reliability and life prediction of electronics in drilling and evaluation tools. *Annual Conference of the PHM Society*, 6(1). doi:10.36001/phmconf.2014.v6i1.2492.

Kang, J., Varnier, C., Mosallam, A., Zerhouni, N., Youssef, F.B., and Shen, N. (2022). Risk level estimation for electronics boards in drilling and measurement tools based on the hidden markov model. In *2022 Prognostics and Health Management Conference (PHM-2022 London)*, 495–500. doi:10.1109/PHM2022-London52454.2022.00093.

Karkouch, A., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73, 57–81. doi: 10.1016/j.jnca.2016.08.002.

Klein, A. and Lehner, W. (2009). Representing data quality in sensor data streaming environments. *Journal of Data and Information Quality*, 1(2). doi: 10.1145/1577840.1577845.

Kumar, S., Vichare, N.M., Dolev, E., and Pecht, M. (2012). A health indicator method for degradation detection of electronic products. *Microelectronics Reliability*, 52(2), 439–445. doi: https://doi.org/10.1016/j.microrel.2011.09.030.

Lee, Y.W., Strong, D.M., Kahn, B.K., and Wang, R.Y. (2002). Aimq: a methodology for information quality assessment. *Information & Management*, 40(2), 133–146. doi:10.1016/S0378-7206(02)00043-5.

Lin, W.C. and Tsai, C.F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2), 1487–1509. doi: 10.1007/s10462-019-09709-4.

Michael G. Pecht, Myeongsu Kang (2018). *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things.* John Wiley and Sons Ltd.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., and Zerhouni, N. (2021). Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in Industry*, 127, 103414. doi:10.1016/j.compind.2021.103414.

Rekatsinas, T., Dong, X.L., Getoor, L., and Srivastava, D. (2015). Finding quality in quantity: The challenge of discovering valuable sources for integration. In *7th Biennial Conference on Innovative Data Systems Research (CIDR '15)*. Citeseer.

Smiti, A. (2020). A critical overview of outlier detection methods. *Computer Science Review*, 38, 100306. doi: https://doi.org/10.1016/j.cosrev.2020.100306.

Wang, R.Y. and Strong, D.M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. doi: 10.1080/07421222.1996.11518099.

Wang, Z., Fu, Y., Song, C., Ge, W., Qiao, L., and Zhang, H. (2019). A data quality improvement method based on the greedy algorithm. In X.B. Zhai, B. Chen, and K. Zhu (eds.), *Machine Learning and Intelligent Communications*, 256–266. Springer International Publishing, Cham.

Zimek, A. and Schubert, E. (2017). Outlier detection. In L. Liu and M.T. Özsu (eds.), *Encyclopedia of Database Systems*, 1–5. Springer New York, New York. doi: 10.1007/978-1-4899-7993-3_80719-1.