Highlights

A data quality management framework for equipment failure risk estimation: application to the oil and gas industry

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

- Describe and illustrate the metrics used to quantify the data quality for industrial equipment failure risk estimation.
- Introduce a novel indicator to measure the effect of data quality on equipment failure risk estimation model performance.
- Propose a comprehensive and practical data quality management framework for equipment failure risk estimation. The framework includes a decision-making model to determine if the data quality meets the requirements and the best risk estimation model.
- The proposed framework is validated on electronic boards in drilling tools using actual data collected during drilling operations in multiple fields worldwide.

A data quality management framework for equipment failure risk estimation: application to the oil and gas industry

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

[a] *SUPMICROTECH, CNRS, institut FEMTO-ST, 25 Rue Alain Savary, Besançon, 25000, France*
[b] *SLB, 1 Rue Henri Becquerel, Clamart, 92140, France*

Abstract

In engineering asset management, accurate failure risk estimation is essential for averting equipment breakdowns and optimizing risk-based maintenance strategies. Data quality and model fitting are two critical sources of prediction uncertainty in risk estimation. While much attention has been devoted to model fitting for risk estimation, the critical role of data quality has often been overshadowed. To bridge this research gap, this paper presents a novel data quality management framework tailored for industrial equipment failure risk estimation. The framework covers the steps from data to model. It consists of the following main phases: data development, data quality assessment, data quality requirement decision-making, data quality improvement, and risk estimation model development. The framework provides detailed guidelines that can facilitate data practitioners to build individualized data quality requirement decision-making models for failure risk estimation of their equipment. The decision-making model can measure the adequacy of existing data to build a risk estimation model that meets the specified requirements and further determines the best risk estimation model given the available data. A case study using actual data collected during oil well drilling operations from multiple oil fields demonstrates the practicality and effectiveness of the framework. In this case study, four risk estimation models are compared, including two baseline models (mean time to failure and median time to failure) and two machine-learning models (quantile regression and hidden Markov model). In addition, a decision tree-based decision model is developed to determine whether the data quality meets the requirements and the best risk estimation model in case the data quality meets the requirements.

Keywords: data quality, risk estimate, field data, oil and gas, downhole tools

1. Introduction

Risk is commonly characterized as the possibility or likelihood of a potential event occurring(Society for Risk Analysis, 2018). In the specific context of equipment failure risk estimation within the realm of engineering asset management, the risk of failure can be defined as the likelihood of a failure in an industrial system that can lead to costly consequences such as downtime, maintenance costs, and even safety hazards. By quantifying the risk of failure associated with industrial equipment, organizations can prioritize maintenance tasks, reduce unplanned downtime, and extend the life of critical assets(Martínez-Galán Fernández et al., 2022). Therefore, accurate equipment failure risk estimation is critical to making informed asset management decisions.

With the rapid advancement of Internet of Things (IoT) technology, computer science, and artificial intelligence, failure risk estimation for industrial equipment has increasingly relied on data-driven models powered by machine learning algorithms. For instance, (Mazumder et al., 2021) conducted a comparative study of eight machine-learning algorithms to estimate the failure risk in steel oil and gas pipelines. (Betz et al., 2023) introduced an innovative risk estimation model for building equipment, leveraging condition inspection data and a neural network algorithm. (Wang et al., 2023) utilized historical failure mode and effect analysis data to predict component or product failure risk using various machine learning classifiers. In machine learning applications, three primary sources of prediction uncertainty emerge scope compliance, data quality, and model fitting (Kläs and Vollmer, 2018).

Scope compliance-related uncertainty arises because of disparities between the context in which the model is developed and the real-world application context. Data quality-related uncertainty stems from limitations in data quality when applying the model, encompassing issues like missing values and noisy data. Lastly, model fitting-related uncertainty is a consequence of the inherent limitations of the learned model.

In the context of equipment risk estimation, scope non-compliance is often deemed impossible as the scope is clearly defined – it is to estimate the risk of equipment failures. Thus, two predominant factors significantly influence failure risk estimation accuracy: data quality and model fitting. However, much of the research effort in academia has been focused on solving problems related to model fit, and little research has been done on data quality (Omri et al., 2021)(Gitzel et al., 2015)(Jia et al., 2022). In fact, in practical machine learning applications, a significant amount of time and resources invested are devoted to data collection, cleaning, and preparation (Press, 2023)(Gupta et al., 2021). Based on these considerations, this paper proposes an extended data quality management framework for equipment risk estimation based on a prior work presented at the IFAC World Congress 2023 (Kang et al., 2023).

The framework aims to address key issues related to data quality in industrial equipment risk estimation:

1. How can data quality from industrial equipment be assessed? This question explores indicators for assessing the quality of data collected from industrial equipment, such as data volume, completeness, accuracy, and consistency.
2. How to measure the effect of data quality on the performance of equipment risk estimation models given actual risk are partially known? Understanding the relationship between data quality and model performance is critical. This question explores techniques to quantify the effect of data quality on the accuracy of risk estimates.
3. How can it be determined whether data are sufficient for modeling risk estimates? This question addresses the process of determining whether the data collected meets the data quality required for modeling.
4. If the data are sufficient, which model is the best for modeling risk estimates? This question addresses the decision-making process of model selection.
5. If the data are substandard and insufficient for modeling risk estimates, what measures are in place to improve the quality of the data? This question explores the data quality improvement techniques including data preprocessing (e.g., missing value imputation, and outlier detection), and other data management practices.

The several key enhancements compared to previous work are summarized as follows:

- **Scope expansion**: The framework is broadened to encompass equipment risk estimation more broadly, moving beyond its previous confinement to electronic systems within downhole tools.

- **Decision-making model**: A novel decision-making model is introduced within the framework. This model serves the crucial function of assessing whether the available data quality meets specific requirements and, equally important, determines the best risk estimation model if the data meets the requirement.

- **Enhanced loss function**: The loss function characterizing risk model performance is improved. It now incorporates a new parameter termed the ”cost ratio”, which captures the influence of the maintenance cost difference of different equipment on the loss function.

- **Practical guidance**: This paper offers in-depth details and comprehensive explanations, providing valuable guidelines for data practitioners to build data quality management for risk estimates of their equipment.

These extensions enhance the applicability and effectiveness of the data quality management framework, making it a valuable resource for data professionals engaged in risk estimation for a wide range of industrial equipment.

The rest of the paper is organized as follows: Section 2 reviews the existing data quality management related works. Section 3 presents a comprehensive framework for managing data quality in equipment risk assessment, including data quality assessment, data quality requirement decision-making, data quality improvement, and risk estimation model development. Section 4 presents a case study to validate the effectiveness of the proposed framework, utilizing field data collected from real-world drilling operations conducted globally. Moreover, the validation is restricted to the electronic boards in drilling tools used for oil well construction. Section 5 summarizes the key takeaways, highlights the importance of data quality in industrial equipment risk assessment, and suggests future research directions in this critical area.

2. Related works

2.1. Data quality definitions

Data quality has attracted significant attention and research in various domains, including, but not limited to, information management, IoT, digital manufacturing, healthcare, and prognostics and health management (PHM). Many definitions of data quality have emerged in these domains. Table 1 briefly summarizes these different definitions.

Table 1: Data quality definitions

Domain	**Reference**	**Definition**
Information management	(Wang and Strong, 1996)	Data that are fit for use by data consumers.
IoT	(Karkouch et al., 2016)	How suitable the gathered data (from the smart things) are for providing ubiquitous services for IoT users.
	(Martín et al., 2023)	Data quality defines the degree of compliance with the requirements enforced by data consumers
Digital manufacturing	(Wang et al., 2008)	A measure of the agreement between the data views presented by an information system and that same data in the real world.
Healthcare	(Ehsani-Moghaddam et al., 2021)	Data quality is the degree to which a given dataset meets a user's requirements.
PHM	(Chen et al., 2013)	The data quality should reflect the suitability of data to satisfy the modeling for purposes of failure detection, diagnosis and prediction.

In addition to the data quality definitions found in research articles, the international standard ISO 8000 series defines data quality as degree to which a set of inherent data characteristics fulfills requirements (ISO 8000-2:2022). Another international standard (ISO/IEC 25012:2008) defines data quality as degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions. Moreover, the company IBM defines that data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness, and fitness for purpose (IBM, 2024).

Although there are various definitions of data quality in different domains, a common consensus is that data is considered high quality when it is well suited for the intended purposes in the application context.

2.2. Data quality metrics

In the research literature, the term "data quality metrics" is synonymous with a variety of terms, including "data quality dimensions" (Wang and Strong, 1996), "data quality indicators" (Wang et al., 2019), and "data quality characteristics" (Gualo et al., 2021). The data quality metrics are a set of data quality attributes that represent a single aspect or construct of data quality (Wang and Strong, 1996). There are two widely recognized international standards related to data quality. The first one is (ISO/IEC 25012:2008), in this international standard, the data quality metrics are classified into two categories: inherent and system-dependent.

- Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions.

- System dependent data quality refers to the degree to which data quality is reached and preserved within a computer system when data is used under specified conditions.

For inherent data quality, five data quality metrics are defined: accuracy, completeness, consistency, credibility, and currentness. On the other hand, system-dependent data quality is assessed using three metrics: availability, portability, and recoverability. Additionally, there are seven metrics that jointly consider inherent and system-dependent data quality.

The other international standard is (ISO 8000-8:2015). This standard identifies three categories to measure data quality, which are

- syntactic quality: degree to which data conforms to its specified syntax;

- Semantic quality: degree to which data corresponds to what it represents;

- pragmatic quality: degree to which data is found suitable and worthwhile for a particular purpose.

In addition to the two international standards, researchers have identified various data quality dimensions from different perspectives or application contexts. For instance, In the sensor data streaming environment, (Klein and Lehner, 2009) adopted a set of five metrics to represent data quality. These metrics include accuracy, confidence, completeness, data volume, and timeliness. (Rekatsinas et al., 2015) used metrics such as coverage, accuracy, timeliness, and position bias to assess the quality of data sources, which are well suited for applications involving data integration and fusion. (Purnomoadi et al., 2023) applied six dimensions of data quality for asset health indexing. The six dimensions are accuracy, completeness, validity, consistency, uniqueness, and timeliness. (Díaz Iturry et al., 2021) found that in health records, most data quality problems are related with completeness, followed by consistency, correctness and accuracy. (Xu et al., 2022) defined five data quality metrics for imbalanced data in multiple products manufacturing process: free of error, appropriate amount of data, ease of manipulation, relevance, and imbalance level. (Merino et al., 2016) proposed the three data quality characteristics for assessing the levels of data quality-in-use in big data projects: contextual adequacy, operational adequacy and temporal adequacy. From a product perspective, (Wang, 1998) proposed a comprehensive framework consisting of four categories of data quality metrics: intrinsic data quality, contextual data quality, representational data quality, and accessibility data quality, each of which includes several subcriteria. The framework provides a detailed taxonomy for assessing data quality from different perspectives.

Given the wide variety of data quality indicators, careful consideration must be given to selecting these indicators, because not all have uniform applicability or relevance in different contexts. Similar to the contextual relevance of data quality, the selection of data quality metrics should be based on the precise requirements and objectives inherent in the intended application (Buelvas et al., 2023).

Table 2 summarizes definitions of a few widely used data quality metrics. It is noteworthy to acknowledge that the enumeration of metrics presented herein is not comprehensive; rather, it serves to provide a representative overview of the research endeavors undertaken in this domain.

2.3. Data quality related works in engineering asset management

While the field of data quality management has seen significant developments, it is worth noting that the majority of research efforts have been directed towards data quality measurements and monitoring within databases or information management contexts (Batini et al., 2009)(Ehrlinger and Wöß, 2022)(Cichy and Rass, 2019). In contrast, the domain of engineering asset management, which deals with critical industrial systems, has received relatively less attention in terms of data quality management research.

Table 2: Definitions of widely used data quality metrics

Metric	**Definition**
Accuracy	The degree to which the data values reflect the actual event state in a specific context of use (ISO/IEC 25012:2008).
Completeness	The ratio of complete elements. An element can refer to any data unit, e.g., an attribute, a record, or a table (Batini et al., 2009).
Data volume	The number of raw data items (values) available for use to compute a result data item (Karkouch et al., 2016).
Consistency	The degree to which the data's format and value conform to the predefined schema (Behkamal et al., 2014).
Timeliness	The probability that an attribute value is still up-to-date (Kaiser et al., 2007).
Currency	Time difference between when data are stored in the system and when data are updated in the real world (Batini et al., 2009).
Relevancy	Extent to which information is applicable and helpful for the task at hand (Wang and Strong, 1996).
Interpretability	To extend to which data is appropriate languages, symbols, and units and the definition are the clear (Pipino et al., 2002).
Accessibility	Extent to which information is available, or easily and quickly retrievable (Wang and Strong, 1996).

Existing research work in this area mainly focuses on different data quality issues in various engineering asset management tasks, including failure detection, fault diagnosis, degradation assessment, maintenance decision-making, and structural health monitoring. A comprehensive search of engineering-related literature published after 2010 was conducted on Scopus using keywords related to data quality management (e.g., data quality, data requirements, data management, data suitability) as well as keywords related to engineering asset management tasks (e.g., failure detection, failure diagnosis, degradation, prognosis, maintenance, risk estimation, RUL prediction). A total of 243 relevant articles were identified and their abstracts were thoroughly researched to further confirm their relevance to data quality management for engineering asset management tasks. Ultimately, 15 articles were selected as a representative sample from a larger pool of articles. It is worth noting that this selection was not exhaustive, but rather a screening of relevant literature. Table 3 provides a concise overview of the explored data quality issues of these data quality-related works.

These works have predominantly focused on specific aspects of data quality management, with limited discussions regarding comprehensive frameworks that span the entire data lifecycle from data generation to data-driven model development. Moreover, there has been relatively less emphasis on data quality in the context of equipment failure risk estimation.

3. Proposed data quality management framework

The proposed data quality management framework for equipment failure risk estimation is illustrated in Figure 1. This comprehensive framework is composed of five phases: data development, data quality assessment, data quality requirement, data quality improvement, and risk estimation model development. Each of these phases is explained in detail in the following subsections.

3.1. Assumptions

There are several key assumptions that should be taken into account before implementing the framework. These assumptions are detailed below:

- Only the start and end times of life are known for failed equipment; actual risks of failure over time or health states over time are unknown.
- The studied equipment is assumed to be newly implemented or infrequently used, resulting in limited availability of high-quality data.
- The data quality of the studied equipment is assumed not to exceed the data quality of the similar equipment.
- The data quality of the similar equipment is assumed high.
- Unit failure cost and premature replacement cost per unit time are known and considered constants for the equipment.

Table 3: Summary of data quality related works in engineering asset management

Reference	**Task**	**Explored data quality issue**
(Jia et al., 2018);(Zhou et al., 2021);(Ji et al., 2022);(Yao et al., 2023)	Fault diagnosis	Data quality assessment
(Omri et al., 2021)	Fault diagnosis	Data quality requirement
(Xie et al., 2023)	Machinery health monitoring	Data quality improvement
(Chen et al., 2013)	Fault diagnosis	Data quality assessment and improvement
(Jia et al., 2018);(Jia et al., 2022)	Fault detection, fault diagnosis, and degradation assessment	Data quality assessment
(Lukens et al., 2022)	Degradation assessment	Data quality assessment and improvement
(Lukens et al., 2019)	Maintenance decision-making	Data quality improvement
(Madhikermi et al., 2016)	Maintenance decision-making	Data quality assessment
(Koziel et al., 2021)	Maintenance decision-making	Investment decision-making for data quality improvement
(Makhoul, 2022)	Structural health monitoring	Data quality metric selection
(Deng et al., 2023)	Structural health monitoring	Data quality assessment

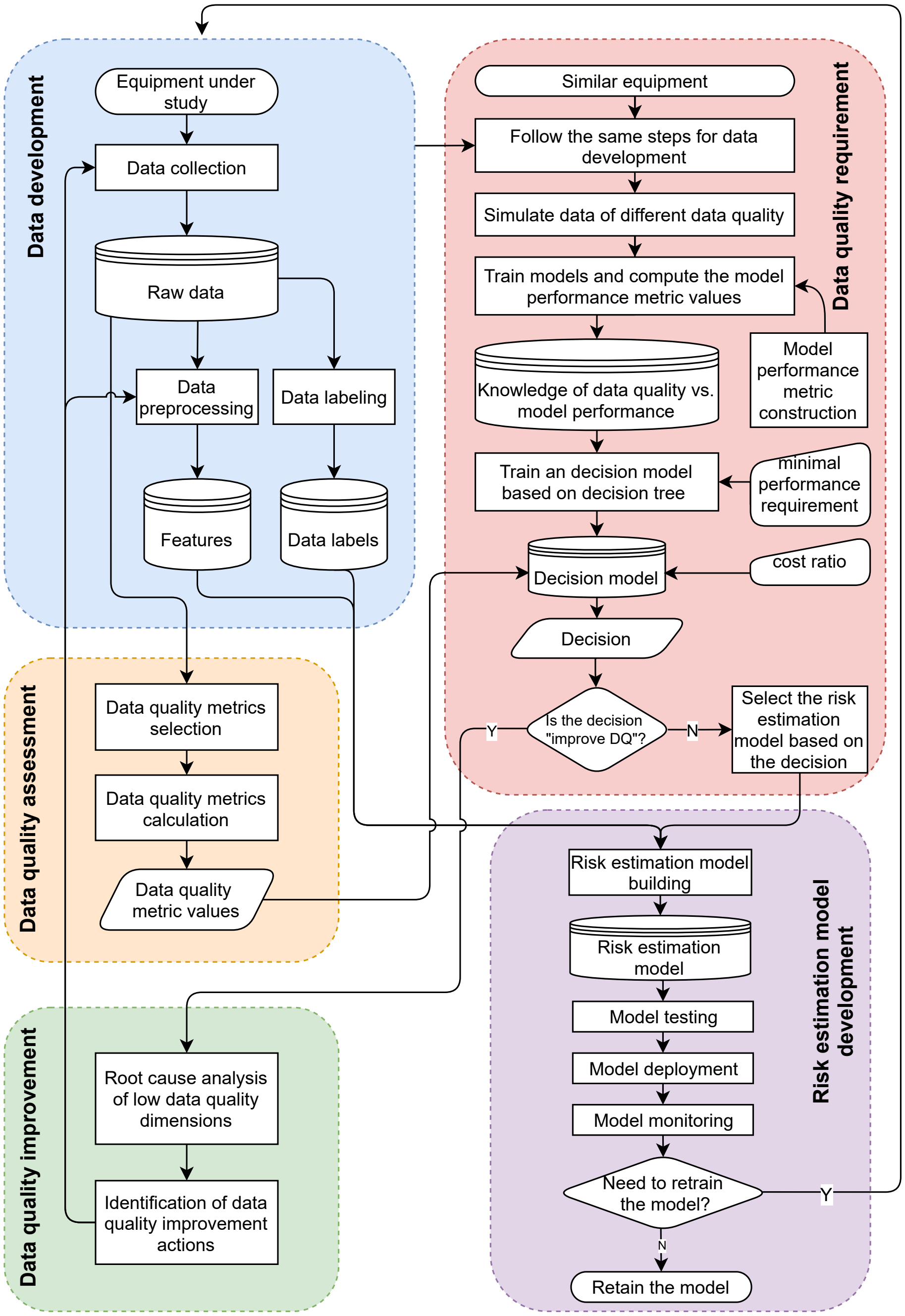
Data development
Equipment under study
Data collection
Raw data
Data preprocessing
Data labeling
Features
Data labels
Data quality requirement
Similar equipment
Follow the same steps for data development
Simulate data of different data quality
Train models and compute the model performance metric values
Knowledge of data quality vs. model performance
Model performance metric construction
Train an decision model based on decision tree
minimal performance requirement
Decision model
cost ratio
Decision
Is the decision "improve DQ"?
Y
N
Select the risk estimation model based on the decision
Data quality assessment
Data quality metrics selection
Data quality metrics calculation
Data quality metric values
Risk estimation model development
Risk estimation model building
Risk estimation model
Model testing
Model deployment
Model monitoring
Need to retrain the model?
Y
N
Retain the model
Data quality improvement
Root cause analysis of low data quality dimensions
Identification of data quality improvement actions

Figure 1: The data quality management framework

3.2. Data development

As mentioned in the Introduction, the cornerstone of building data-driven models for estimating equipment failure risk lies in the data, because the efficacy and usefulness of these models depend heavily on the data quality. This section on data development outlines the three key steps required to formulate and process the data, that is, data collection, data preprocessing, and data labeling.

3.2.1. Data collection

Data-driven equipment failure risk estimation begins with the collection of data, typically sourced from the computerized maintenance management system (CMMS) linked to the equipment. A CMMS is a specialized software that centralizes maintenance information, optimizing the use and availability of physical equipment (IBM, 2023). The CMMS database holds a wide range of information, including equipment details, operational data, work orders, and materials inventory. The operational data encompasses readings from sensors on the equipment, such as temperature sensors and accelerometers, which measure temperature and vibration. Work orders cover maintenance, equipment orders, shipments, and related tasks.

While the CMMS database is rich in data, it is neither practical nor necessary to use all of it for equipment failure risk estimation. In practice, specific subsets of data are selectively extracted, focusing on equipment identity, run history, operational environment data (e.g., temperature, vibration), and relevant maintenance work orders.

3.2.2. Data preprocessing

Data preprocessing is a process of refining and reshaping raw data into a format suitable for subsequent risk estimation model training. Traditionally, this process requires significant engineering effort and is characterized by iterative improvement through rigorous trial and error handling (Zha et al., 2023). Fig. 2 illustrates the four essential steps in data preprocessing: data cleaning, feature extraction, feature transformation, and feature reduction.

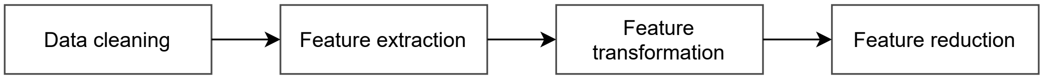

Figure 2: Data preprocessing steps

Data cleaning addresses errors like missing values, outliers, and duplicates. Section 3.5 provides detailed insights into missing-value handling, outlier detection, and data deduplication. Industrial equipment operational data cleansing often requires additional steps guided by Subject Matter Experts (SMEs) in fields like reliability, electrical, and chemical engineering. SMEs contribute specialized knowledge to tasks such as determining stable operation phases. Unsupervised methods, like those explored by (Mosallam et al., 2011), can complement SME expertise.

Feature extraction focuses on obtaining discriminative features from the raw data for failure risk estimation models. Though widely used statistical features extracted from time-domain, frequency-domain, and time-frequency-domain can be applied to general-purpose equipment such as gearboxes and motors (Atamuradov et al., 2017). Specialized equipment, like drilling tools in the oil and gas industry, requires SME guidance. Despite deep learning's feature-learning capabilities, its limited interpretability and computational complexity in real industrial applications make conventional feature extraction techniques preferable for interpretability and information sensitivity reduction (Fink et al., 2020)(Zha et al., 2023).

Feature transformation enhances model performance by converting original features. Common methods include normalization (scaling to [0,1] or [-1,1]) and standardization (zero mean and unit variance). These techniques ensure uniform data magnitude, equal feature contribution, and prevent dominance by larger-value features. Additional methods, such as Box-Cox transformation for skewed data, and feature multiplication, are employed for improved representation.

Feature reduction includes two approaches: feature selection and dimensionality reduction. Feature selection is a process that entails the choice of a subset of input features from a dataset. Feature selection methods can be classified into two categories: unsupervised and supervised (Cai et al., 2018). Supervised feature selection methods can be further categorized into wrapper, filtering, and embedded methods (Meisenbacher et al., 2022). Dimensionality reduction aims to transform high-dimensional features into a low-dimensional space while retaining salient information. One of the most widely used methods of dimensionality reduction is principal component analysis.

3.2.3. Data labeling

Data labeling is an essential process in data development. It involves attaching one or more meaningful and informative labels to raw data, usually time series sensor data, in the context of industrial equipment failure risk estimation. These labels typically convey information regarding the equipment's health status (or fault mode) and the underlying failure mechanisms, enabling data practitioners (e.g., data scientists) to select correct data for model training.

Maintenance work orders can also be initiated in response to suspected equipment failures; notably, maintenance technicians rather than maintenance experts often record the failure description and shop analysis provided in maintenance work order data. As a result, there can be uncertainty regarding the accuracy of the failure reports in the maintenance work order and the identification of the failure root cause.

Labeling industrial equipment sensor data is a complex and costly task, distinct from more common annotation tasks like image or text annotation. It requires a profound understanding of equipment operations, maintenance protocols, and failure mechanisms. SMEs, with their specialized knowledge, are typically entrusted with this responsibility. They review sensor data, failure descriptions, and shop analyses in maintenance work order data to validate failure occurrences and identify root causes. In some cases, failed equipment may undergo a more extensive investigation at the technology center for a detailed analysis of the failure and its contributing factors.

3.3. Data quality assessment

Data quality assessment is one of the five key phases in data quality management. It aims to evaluate the suitability of a dataset for its intended purpose. Section 2.2 describes widely used quantitative data quality metrics. These metrics enable data practitioners to calculate values that offer insights into the data's fitness for use. The process of data quality assessment consists of the following two key steps:

Data quality metric selection: This initial step involves identifying and selecting pertinent data quality metrics. The choice of metrics should align with the specific application context and the data characteristics. (Heinrich et al., 2018) proposed five requirements for data quality metrics, namely, the existence of minimum and maximum metric values (R1), the interval scaling of the metric values (R2), the quality of the configuration parameters, and the determination of the metric values (R3), the sound aggregation of the metric values (R4), and the economic efficiency of the metric (R5). These criteria support both decision-making under uncertainty and economically oriented data quality management.

Data quality metric calculation: Once the relevant data quality metrics are chosen, the next step is to apply these metrics to the dataset. This step involves calculating the metric values based on the dataset's characteristics and the definitions of the selected metrics. For example, completeness metrics involve calculating the percentage of missing values, while data volume metrics assess the number of samples.

3.4. Data quality requirement

Once the data quality assessment has been completed, it is necessary to determine whether the data quality meets the data quality requirements. A straightforward approach is establishing thresholds for data quality metrics, but determining these thresholds remains challenging.

This paper proposes a new method for determining if the data quality meet specific requirements. The method is based on the relationship between model performance and data quality, and a decision tree model. In addition, model performance is evaluated based on the average maintenance cost. Thus, this method for determining data quality requirements takes costs into account. This section will first present an indicator for assessing the performance of the equipment risk estimation model and then describe how to acquire knowledge of the relationship between model performance and data quality.

3.4.1. Risk estimation model performance metric

When assessing the risk of equipment failure in an industrial environment, it is often challenging to determine the actual risk of failure over time, which is especially true for complex equipment with particularly complex failure mechanisms. In such cases, available data usually only provide information on when equipment failures occurred.

On the other hand, cost factors play a pivotal role in maintenance decisions guided by risk estimation. Consider the scenario where a critical component is the primary driver of equipment failures, and this component cannot be repaired but only replaced. In this case, the risk-based maintenance decision-making process involves three principal costs: component replacement cost, cost associated with undetected failures, and cost associated with premature component replacement. The first of these costs is often deterministic, while the latter two depend on the accuracy of the risk assessment model.

Specifically, suppose the risk estimation model can predict the component failures accurately. In that case, the components can be replaced at the optimal time, thus avoiding any losses due to failures or premature replacement. However, if the model predicts the failure too late, it can incur costs associated with equipment failures. Conversely, if the model predicts failures too early, it may lead to unnecessary component replacement costs.

Given the above analysis, the authors proposed a loss function for assessing the performance of a risk estimation model in their prior work (Kang et al., 2023):

$$\ell = \frac{\sum_{i=1}^{N} \left[c_1 \mathbb{I}(\hat{T}_i \geq T_i) + c_2 (T_i - \hat{T}_i) \mathbb{I}(\hat{T}_i < T_i) \right]}{N} \tag{1}$$

where

- N: the number of components.
- $\hat{T}_i$: the time when the component i is replaced based on failure risk estimation, assuming all components' lives start at time 0; specifically, the component i is replaced when the failure risk estimate reaches a certain level.
- T_i: actual life of component i, i.e., the time when the component actually failed.
- c_1: unit failure cost, i.e., the cost caused by one undetected failure.
- c_2: premature replacement cost per unit of time.
- $\mathbb{I}$ is an indicator function. In other words, $\hat{T}_i \geq T_i$ means the component i is replaced too late, which incurs failure cost, while $\hat{T}_i < T_i$ means the component i is replaced too early, which incurs premature replacement cost.

Different components could have different unit failure costs and premature replacement costs per unit of time. To capture the effect of these cost differences on the loss function and reduce the number of cost parameters, this paper enhances the loss function by introducing a new parameter called cost ratio (denoted as $r = c_1/c_2$). Firstly, Eq. (1) can be written as Eq. (2).

$$\ell = c_2 \left[\frac{c_1}{c_2} \times \frac{\sum_{i=1}^{N} \mathbb{I}(\hat{T}_i \geq T_i)}{N} + \frac{\sum_{i=1}^{N} (T_i - \hat{T}_i)\mathbb{I}(\hat{T}_i < T_i)}{N} \right]. \tag{2}$$

Then, since the cost parameters c_1 and c_2 are constants for the component, Eq. (2) can be further reformulated as Eq. (3) by substituting r for c_1/c_2.

$$\ell \propto r \times \underbrace{\frac{\sum_{i=1}^{N} \mathbb{I}(\hat{T}_i \geq T_i)}{N}}_{\text{term 1}} + \underbrace{\frac{\sum_{i=1}^{N} (T_i - \hat{T}_i)\mathbb{I}(\hat{T}_i < T_i)}{N}}_{\text{term 2}}. \tag{3}$$

This new expression consisting of two different terms as shown in Eq. (3). The first term can be interpreted as the average undetected failures, while the second term can be interpreted as the average premature replacement time. Both these terms are influenced by the data quality, while the parameter r is inherent to the component itself. The inclusion of r is essential as different components exhibit distinct values for ratio between unit failure cost and premature replacement cost per unit of time.

3.4.2. Knowledge of data quality vs. model performance

Failure risk estimation of industrial equipment is inherently contextual. Different types of equipment have different characteristics and monitoring parameters and, therefore, cannot be cross applied. For example, it would be unwise to attempt to use a model trained on gearbox data for electronic boards, because the fundamental nature of these equipment types and their associated data varies widely. Consequently, the knowledge gained through simulation studies on publicly available datasets, which are often not correlated with the equipment under study, cannot reflect the true relationship between the quality of the data from the equipment under study and the performance of the model.

Adding to this challenge is that obtaining large amounts of high-quality data for newly implemented or infrequently used equipment can be daunting. This paper proposes a new approach for indirectly acquiring knowledge about the relationship between data quality and model performance, as illustrated in Fig. 1. That is, using data from similar equipment, which have more data than the equipment under study. Based on the data from similar equipment, knowledge about the relationship between data quality and model performance can be obtained through simulation studies. In this paper, this knowledge is succinctly represented as $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$, where $\mathbf{Q}$ denotes a vector containing data quality metrics, $\mathbf{\Omega}$ represents risk estimation models, r is the cost ratio, and ℓ corresponds to the previously defined loss function in Eq. (3).

The simulation studies are based on carefully processing data from similar equipment. By selectively removing or modifying data segments, data with different levels of data quality can be modeled. These synthetic datasets can train risk assessment models, thus effectively exploring the relationship between data quality and model performance. Subsequently, using the same test dataset, these trained models are used to estimate the lifetime of the test boards, which helps to compare the loss function values. For more robust assessments of model performances, it is recommended that cross-validation techniques are used.

3.4.3. Decision model

Once this knowledge is obtained, along with the minimum performance requirement, it becomes feasible to develop a decision model using the decision tree algorithm. The choice of the decision tree algorithm in this paper is motivated by its simplicity, ease of understanding, and the capacity to visualize the decision model. The decision model can be mathematically expressed as in Eq. (4). The developed decision model can then determine whether the data quality of the equipment under study should be improved and, if not, which risk estimation model is the best.

$$D = g(C, \mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)), \tag{4}$$

where C is the minimum performance requirement. It can be determined based on the average cost requirement thanks to the definition of the loss function. D is the decision predicted by the decision model.

A detailed case study will be presented in Section 4 to illustrate the practical application of the method to acquire the knowledge $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$ and make decision based on the decision model.

3.5. Data quality improvement

Data quality improvement involves two key steps: analyzing the root causes of low data quality dimensions and identifying data quality improvement actions.

Root cause analysis of low data quality dimensions: The first step in the data quality improvement process is to examine the dimensions of low data quality to discover the root causes of low data quality. This step requires a comprehensive understanding of the data generation mechanisms and collection process. The root causes can be grouped into three categories: hardware-related issues, software-related issues, and human factors. Hardware problems may include insufficient sensor accuracy, capacity limitations and physical damage of memory storage boards, and communication errors during data transfer from lower to upper computer systems. These problems are rooted in the design and infrastructure of the hardware system. Software issues include data loss or inconsistency due to CMMS system migration or limitations. In addition, data loss issues may occur during the data collection process; which often stems from human error; e.g., inadvertent data overwriting and field engineers neglecting to upload data to the server.

Identification of data quality improvement actions: After thoroughly analyzing the root causes of data quality deficiencies, the next step is to develop corresponding data quality improvement measures. Improvement measures may include a variety of strategies and initiatives, each aligned with the specific causes and dimensions of data quality that need to be improved. Based on the above analysis of the root causes of low data quality, the data quality of industrial equipment can be improved in the following three area.

- System upgrades: This area focuses on improving data quality from a hardware and software technology perspective. For example, deploying enhanced sensors with higher accuracy and upgrading communication systems can reduce measurement errors and inaccuracies, directly affecting data quality. Choosing a powerful, stable, and mature CMMS can also help manage equipment data, avoiding data loss or inconsistency caused by frequent CMMS migrations.

- Management improvement: Management improvement centers on optimizing data quality from a management and human factors perspective. Many data quality issues often stem from human error or negligence. As analyzed above, data loss can be due to field technicians forgetting to transfer data from the memory board to the hard drive and upload it to the data cloud. To minimize such issues, increased training and introduction of Key Performance Indicators (KPIs), such as data collection ratios, for field engineers can increase their awareness of the importance and responsibility of data collection.

Both technical and management improvements require a significant investment of time and resources. In addition, they require sustained commitment and ongoing efforts to significantly improve data quality. They are essential for continuous data quality improvement but may not produce immediate results and/or business value.

- Data preprocessing improvements: In contrast, data preprocessing improvements offer an immediate and practical approach to improving data quality. One can quickly resolve some data quality issues by employing data preprocessing techniques such as deduplication, outlier detection, and missing value imputation.

  - Data deduplication aims to compress data through removing duplicated data items and replacing them with a pointer to the unique remaining copy (Karkouch et al., 2016). Intrinsically, it reduces the amount of data and affects the data quality of data volume.
  - Outlier detection focuses on finding observations significantly different from most data (Zimek and Schubert, 2017). Outlier detection methods can be categorized into four groups: statistical-based, distance-based, density-based, and clustering-based methods (Smiti, 2020). Statistical-based methods rely on statistical techniques to identify outliers. Distance-based methods assess the dissimilarity or distance between data points to determine outliers. Density-based methods focus on the density distribution of data points. They identify outliers as data points existing in regions of low data density. Clustering-based methods seek to partition data into clusters, with outliers being data points that do not conform to any cluster or belong to small, isolated clusters. For more details on outlier detection methods, see two review articles (Smiti, 2020) and (Chandola et al., 2009).
  - Missing value imputation attempts to replace missing data with

estimated values. Missing value estimation methods can be categorized into two types: statistical-based and machine learning-based methods (Lin and Tsai, 2020) (Hasan et al., 2021). Statistical-based methods rely on statistical measures and patterns within the data to impute missing values. Widely used statistical-based techniques include expectation maximization, linear regression, and imputation using the mean or mode of the available data. Machine learning leverages algorithms and models to predict and impute missing values. Typical machine learning-based techniques for missing value imputation include regression trees, random forests, support vector regression, and k-nearest neighbor. For more details on missing value estimation, see the review articles Lin and Tsai (2020) and Hasan et al. (2021).

3.6. Risk estimation model development

Once it has been determined that the data quality meets the minimum requirements, developing a risk assessment model can begin. This stage involves well-defined steps, including model building, testing, deployment, and monitoring. These critical steps in the model development process are described in detail below.

Model building: The foundation of risk assessment lies in building the model. In this initial phase, data scientists use the features and data labels from the data development phase to create a robust, accurate model that effectively captures the relationship between input data and risk estimates. The risk estimation model can be formulated as in Eq. (5).

$$Risk = \mathbf{\Omega}(\mathbf{X}, \mathbf{y}), \tag{5}$$

where **X** are the extracted features from the data development phase, and **y** are the data labels. It is important to note that the data labels here are not the equipment's failure risk, because the actual risk is often difficult to access, as described earlier. The data label here is more of a failure mode or failure mechanism analysis for each device, which helps data scientists select the right data.

Model testing: Rigorous testing is critical to assess the performance and reliability of the model. In this phase, the model built above is delivered for field testing to a small group of users who apply the model to new unseen data. Testing helps to identify any issues, such as over- or under-fitting, and ensures that the model generalizes well to the new data.

Model deployment: Once the model has performed satisfactorily during testing, the machine learning engineer or software engineer can deploy it into a production environment, including integrating the model into a system or application for real-time or batch processing of risk estimates. Factors to consider when deploying include scalability, reliability, and ensuring the model is synchronized with the latest data. It is critical to monitor the performance of the model in the production environment to identify and address performance drift over time.

Model monitoring: Model monitoring is a continuous process to ensure that models continue to perform accurately and reliably in the production environment, including tracking KPIs, detecting deviations from expected behavior, and initiating corrective action where necessary. Monitoring may also include periodically retraining the model with new data to adapt to changing patterns and maintain its prediction accuracy. To monitor the performance of the risk estimation model, KPIs such as mean time between repair, average maintenance cost, and number of service quality events can be tracked on a monthly or quarterly basis. In the event of significant changes in the KPIs, it is necessary to investigate the underlying factors and potentially retrain the model.

In summary, developing a risk estimation model is a comprehensive process involving multiple players, starting with sufficiently good-quality data and then building, testing, deploying, and monitoring the model. Each of these steps is critical to ensure that the model performs well in the initial stages and maintains its validity and fairness in real-world applications.

4. Case study

The drilling system shown in Fig. 3 is used for oil well construction and consists of a drilling rig, a drillpipe, and a bottomhole assembly. The bottomhole assembly is a crucial part of the drilling system. Oftentimes the assembly includes a drill bit, a rotary steering system tool, a measurement-while-drilling tool, a logging-while-drilling tool, and other mechanical components such as drill collars, and stabilizers. (SLB, 2023).

The drilling tools in the bottomhole assembly contain many electronic boards to enable them to achieve the required functions such as data acquisition, signal processing, operation control, and data storage. The equipment under study is the central processing unit (CPU) board of a specific logging-while-drilling tool as shown in Fig. 4. The similar equipment chosen to

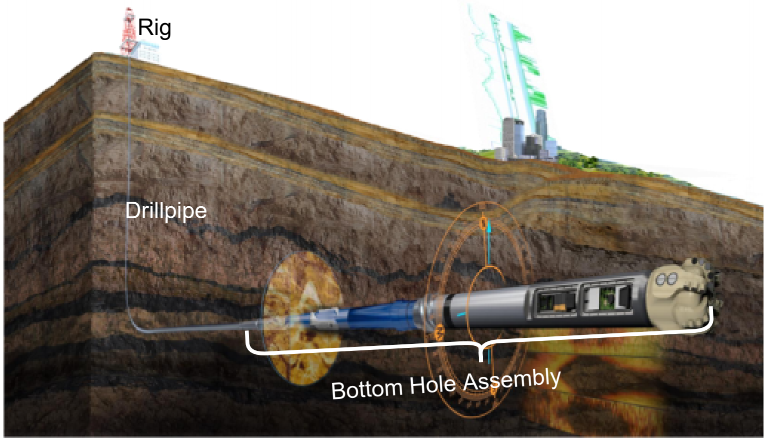

Figure 3: Drilling system schematic

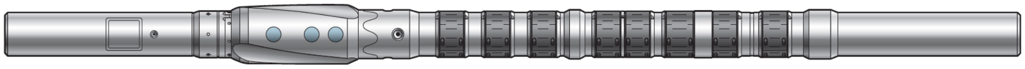

Figure 4: Logging-while-drilling tool

build the knowledge $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$ is the CPU board of a particular rotary steerable system tool as shown in Fig. 5. Both CPUs achieve similar functions and have the same measured parameters to characterize the operational environment: temperature, shock, and vibration.

4.1. Data development

The raw data used in this case study were collected during drilling operations in multiple fields worldwide. These drilling operations varied in terms of duration and operating environment. Specifically, operating environment data of 554 failed CPU boards of the rotary steerable system (similar equipment) were collected. The lifetimes of these boards span a range from 500

Figure 5: Rotary steerable system tool

to 3000 hours. For the CPU board of the logging-while-drilling tool (equipment under study), data were gathered from 18 failed boards. These boards had lifetimes ranging from 700 to 3700 hours. Among these, 12 boards were utilized as training data, while the remaining six boards, which had nearly complete data, were designated for use as test data. All the boards mentioned have been confirmed as failed by SMEs, and their raw data are stored in the CMMS.

The data preprocessing steps are adapted from the authors' previous work in (Kang et al., 2022). However, due to space limitations and data preprocessing is not the core of this paper, the authors will not delve into the details of data preprocessing here. Interested readers can refer to (Kang et al., 2022) for a comprehensive understanding of how features are extracted from the raw time series data of temperature, shock, and vibration for the CPU board.

4.2. Data quality assessment

As mentioned earlier, the selection of data quality metrics should align with the specific application context and the characteristics of the data. In this case study, the equipment failure risk estimation is conducted offline, meaning it occurs after the tool has been pulled up from the oil well and is not during drilling operations. Therefore, metrics related to timeliness and currency are not crucial. Additionally, the sensors in the tool are assumed to be robust, and the readings are considered correct. As a result, there is no need to specifically evaluate the accuracy of the raw data. Instead, the operational environment data of the drilling tool may have missing values. Moreover, data volume is typically considered important for data-driven models. Hence, in this case study, the selected data quality metrics are completeness (denoted as $Comp$) and data volume (denoted as n). Data volume is the number of failed CPU boards, while completeness is the life cycle data coverage (computed as operation time with data collected/total operation time) of the board.

Based on these two metrics, the data quality of the training data for the equipment under study is calculated to be $\mathbf{Q} = [12, 0.76]$. Similarly, the data quality of the test data for the equipment under study is calculated to be $\mathbf{Q} = [6, 1]$, where $\mathbf{Q} = [n, Comp]$.

4.3. Data quality requirement

4.3.1. Acquire the knowledge

The procedures for acquiring the knowledge $\mathcal{K}(\mathbf{Q}, \boldsymbol{\Omega}, r, \ell)$ are shown in Algorithm 1. In this case study,

- the number of simulations $m = 60$,
- the data volume sequence $\mathbf{L}_1 = [2, 3, 4, \ldots, 30]$,
- the completeness sequence $\mathbf{L}_2 = [0.50, 0.55, 0.60, \ldots, 1.00]$,
- the cost ratio sequence $\mathbf{L}_3 = [50, 100, 200, 400, 500, 1000, 1500, 2000, 3000, 4000, 5000, 6000]$,
- the number of test boards $n_{test} = 50$.

Four models are compared in this case study, namely, mean time to failure (MTTF), median time to failure (MeTTF), quantile regression (QR), and hidden Markov model (HMM). When calculate the loss function values for MTTF, the test CPU board is replaced when it reaches the mean time to failure of training boards/data; for MeTTF, the test CPU board is replaced when it reaches the median time to failure of training boards; for QR, the test CPU board is replaced when the predicted remaining lifetime is less than 200 hours; for HMM, the test CPU board is replaced when the estimated risk level reaches the highest risk level.

Table 4 gives an excerpt of the acquired knowledge $\mathcal{K}(\mathbf{Q}, \boldsymbol{\Omega}, r, \ell)$. One can also store the average undetected failure, i.e., first term in Eq. 3, and premature replacement time, i.e., second term in Eq. 3 to preserve the knowledge and then calculate the loss using Eq. 3. This way, one can save storage space.

4.3.2. Build the decision model

Based on the acquired knowledge, one can decide which model is the best by comparing the loss function values if the data volume, completeness, and cost ratio are known. For example, based on the acquired knowledge shown in the first row of Table 4, it can be inferred that the best model is MTTF when the data volume, the completeness, and the cost ratio are 5, 0.5, and 400, respectively, because MTTF obtains the minimal loss function value among the four models. Fig. 6 shows the best model under different data volumes, completeness, and cost ratios. The figure indicates that the QR

Algorithm 1 Knowledge $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$ acquisition

```
Input: entire dataset of the CPU in the rotary steerable system tool, the number of simulations $m$, sequence $\mathbf{L}_1$ containing the numbers of training CPU boards, sequence $\mathbf{L}_2$ containing the completeness of each training CPU board, sequence $\mathbf{L}_3$ containing the cost ratios, the number of test boards $n_{test}$
Output: $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$
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 $\mathbf{L}_1$ do
        sampling observations of $n$ boards from the remaining data without replacement as temporary dataset $Temp$
        for $Comp$ in $\mathbf{L}_2$ do
            remove some observations from $Temp$ to make the completeness of each board equal to $Comp$, use the data after removal operation as the training data
            train the four candidate risk estimation models $\mathbf{\Omega}$ using the training data
            predict the replacement time of test boards using the models $\mathbf{\Omega}$
            for $r$ in $\mathbf{L}_3$ do
                calculate the loss function and store the result in $\mathcal{K}_i(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$
            end for
        end for
    end for
end for
calculate the average loss over $m$ simulations, i.e., $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{K}_i(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$
```

Table 4: Excerpt of the acquired knowledge $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$

n	$Comp$	r	ℓ_{HMM}	ℓ_{MTTF}	ℓ_{MeTTF}	ℓ_{QR}
5.00	0.50	400.00	618.04	479.89	507.74	783.38
5.00	0.50	500.00	652.42	538.09	559.49	798.96
5.00	0.50	1000.00	824.29	829.14	818.24	876.88
5.00	0.50	1500.00	996.17	1120.18	1076.99	954.79
5.00	0.50	2000.00	1168.04	1411.22	1335.74	1032.71
5.00	0.50	3000.00	1511.79	1993.30	1853.24	1188.54
5.00	0.50	4000.00	1855.54	2575.39	2370.74	1344.38
5.00	0.50	5000.00	2199.29	3157.47	2888.24	1500.21
5.00	0.50	6000.00	2543.04	3739.55	3405.74	1656.04
5.00	0.55	50.00	455.13	276.16	326.61	656.34
5.00	0.55	100.00	472.76	305.26	352.49	667.06

model tends to perform best when the data volume or the cost ratio is high, the HMM model excels when the completeness is high, and MTTF is favored when the completeness and the cost ratio are low. MeTTF emerges as the top choice in very few scenarios. To enhance precision in model selection, a decision tree is constructed on top of the information presented in Fig. 6. The decision model for determining the best model without minimal performance requirement is shown in Fig. 7. In this figure, the term "unused" in the legend denotes that the class exists in the training data for training the decision tree model. However, its amount is so minimal that it exerts negligible influence on estimating decision tree model parameters. The node on the split indicates that the corresponding class is the dominant one at that specific split point in the decision tree. In other words, it signifies that the majority of data points at that split belong to that particular class. The decision process follows the left branch if the condition on the node is true, and the right branch otherwise. The node on the leaf (i.e., the most bottom) represents the final decision in the decision tree model. These meanings also apply to the subsequent visualizations of decision trees.

Furthermore, if the minimum model performance requirement is also given, e.g., a requirement that the loss must be not greater than C, then the data quality must be improved if the $\ell^* > C$ (ℓ^* is the loss function value of the best model). Otherwise, the best model can be determined. For instance, if the minimum model performance requirement C is set at 500, then referring to the information in the initial row of Table 4, it can be de-

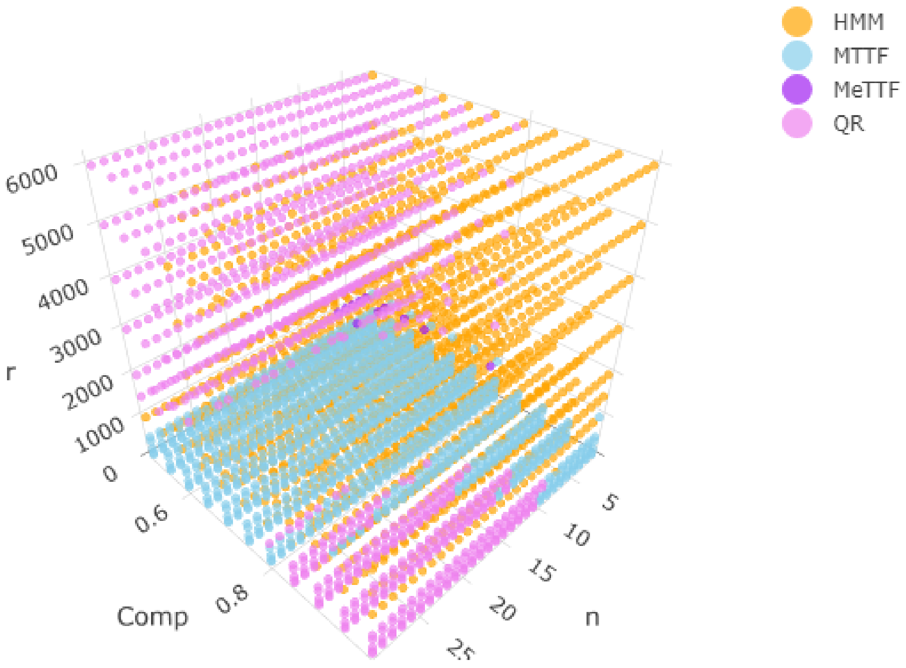

Figure 6: Best model under different data volume, completeness, and cost ratio

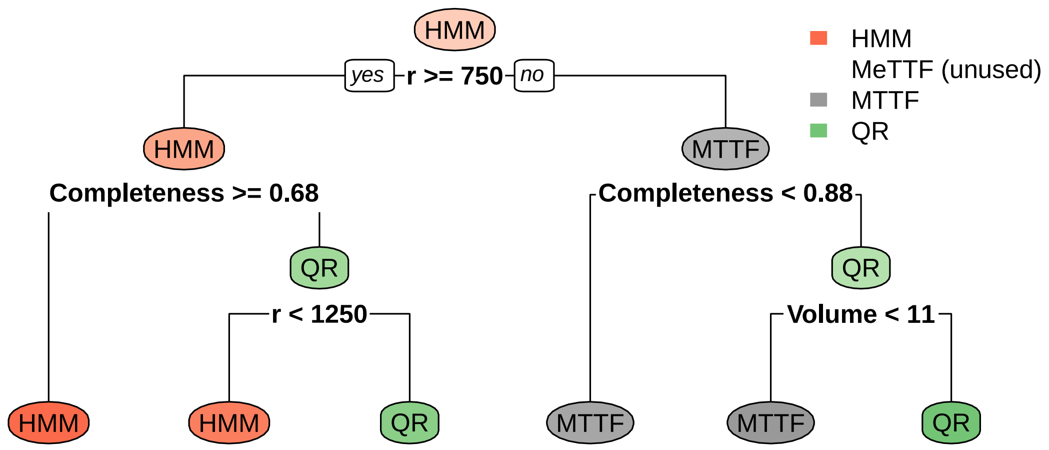

Figure 7: The decision model for determining the best model without minimal performance requirement

duced that MTTF is the best model when the data volume, the completeness, and the cost ratio are 5, 0.5, and 400, respectively. This is because MTTF achieves the minimum loss function value among the four models, and its loss is below 500. However, in cases where the data volume, the completeness, and the cost ratio are 5, 0.5, and 500, respectively, as indicated in the second row of the table, then the decision is to improve the data quality. This is because MTTF obtains the minimum loss function value among the four models, but its loss exceeds the specified threshold C of 500.

The decision models for several values of C are illustrated in Fig. 8 through 12. In these figures, "Improve DQ" means "Need to improve data quality". Additionally, it is observed that for small values of C, the prevailing decision tends to be "Improve DQ." Conversely, when C is significantly large, the decision model tends to align with the decision model that does not impose a minimum performance requirement. This observation is logical because, with small C values, all four models typically fall short of meeting the minimum performance standard. In contrast, with large C values, at least one of the four models satisfies the minimum performance requirement.

4.4. Make the decision and act

To facilitate a comparison of model performance before and after data quality improvement, 10 boards are utilized as training data before data quality enhancement. The average completeness score for these 10 boards is

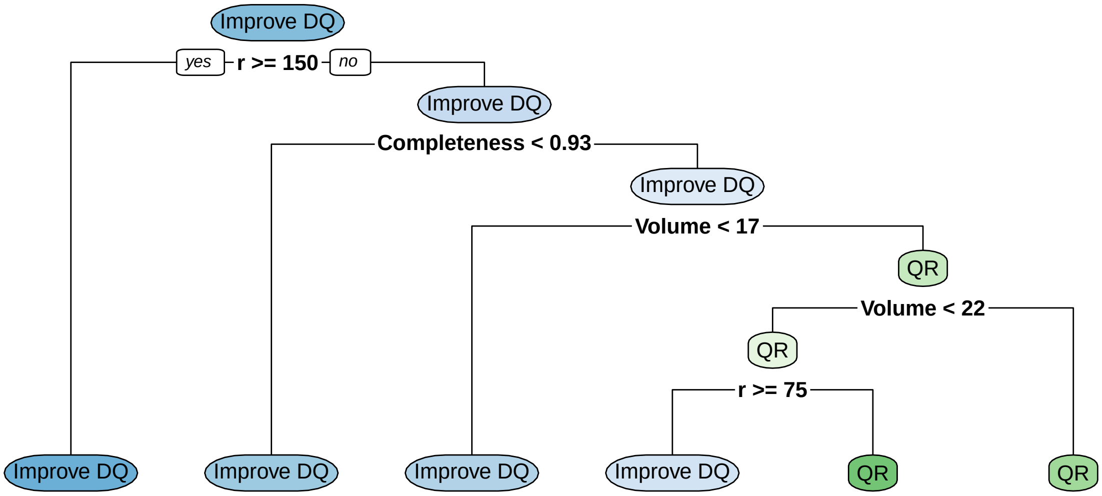

Figure 8: The decision model under $C = 250$

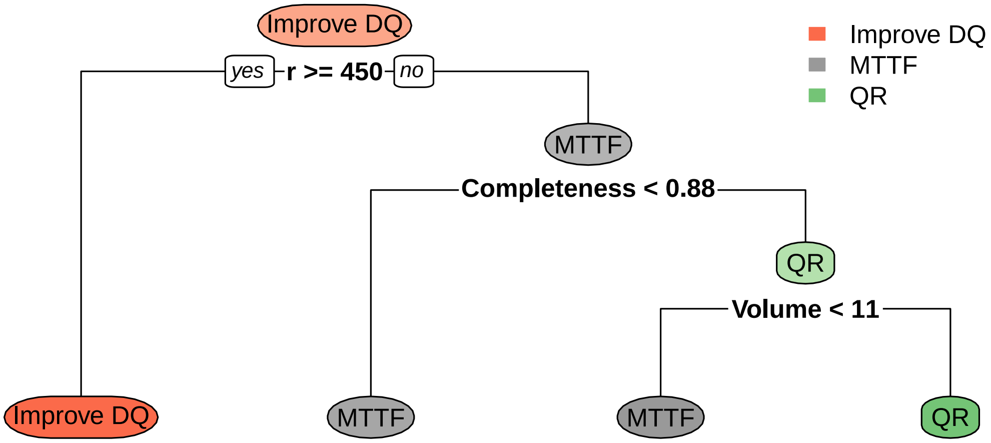

Figure 9: The decision model under $C = 500$

0.76. Then, all 12 boards are used as training data after data quality improvement by filling in the missing data using mean imputation. Consequently, the data quality metrics before data quality improvement can be represented as $\mathbf{Q} = [10, 0.76]$. In contrast, the six test boards remain unchanged, ensuring the models are tested on the same dataset.

As previously mentioned, the decisions made by the decision tree-based model depend on the knowledge $\mathcal{K}(\mathbf{Q}, \mathbf{\Omega}, r, \ell)$ and the minimum performance

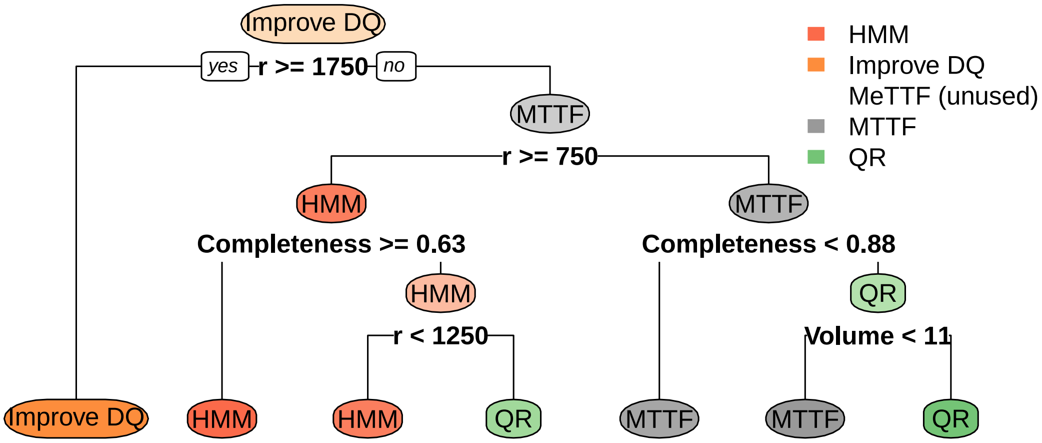

Figure 10: The decision model under $C = 1000$

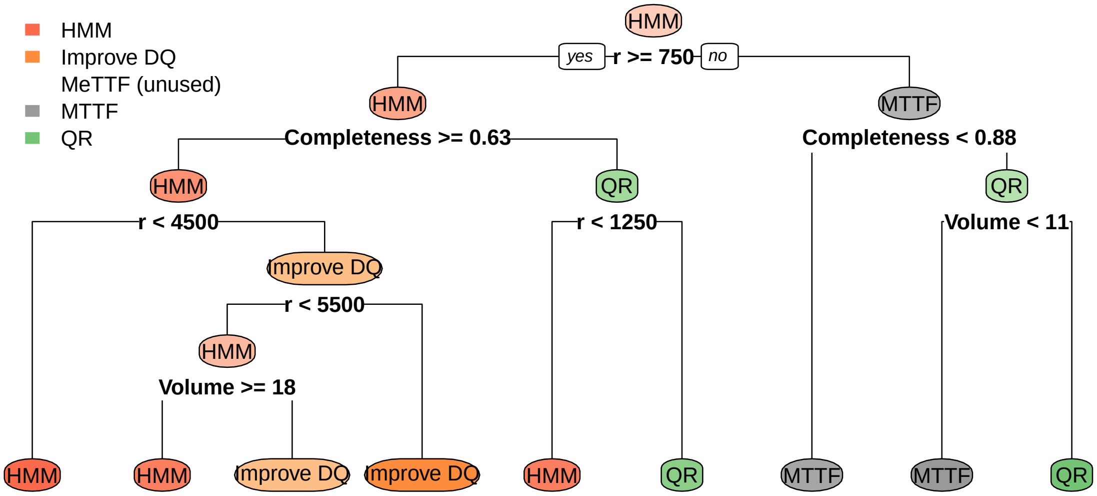

Figure 11: The decision model under $C = 2000$

requirement C. To rigorously validate the proposed framework, confusion matrices are generated to compare the predicted decisions and actual decisions. The predicted decisions are derived from the decision models based on decision tree models. On the other hand, the actual decisions are inferred from the test results on the test data. Specifically, the four trained models are applied to the test data, and the model losses are calculated and

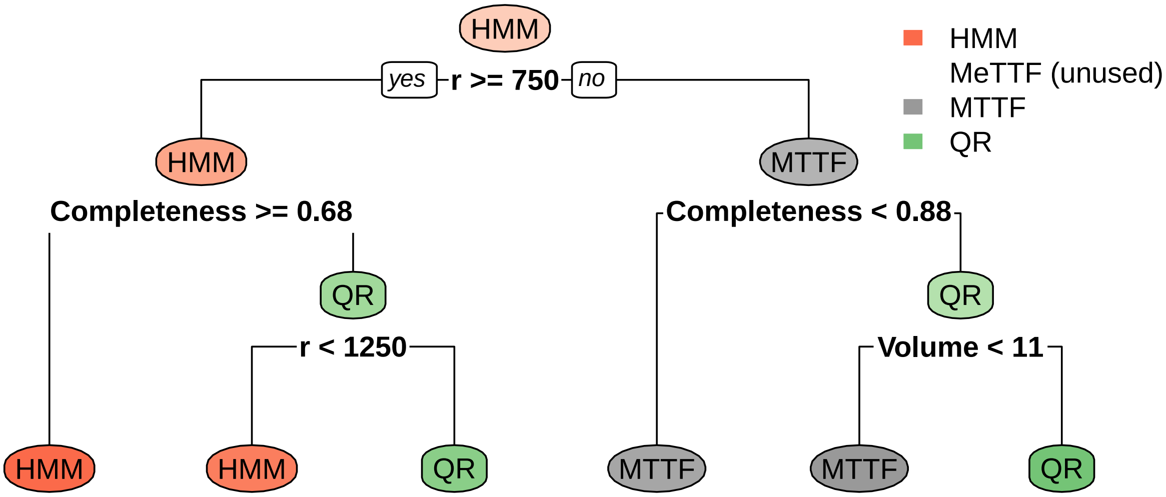

Figure 12: The decision model under $C = 3000$

compared to the minimum model performance requirement C. If the loss of all four models is greater than C, then actual decision is “Improve DQ”, otherwise, the corresponding model with the smallest loss is the best model. These comparisons are made under varying cost ratios r and C. The cost ratios used in this subsection are the same as defined above in the sequence $\mathbf{L}_3$, while the values of C are consistent with the decision models shown in Fig. 8 through 12

The result of the confusion metrics under different C is shown in Table 5. From the table, the average decision accuracy can be computed, that is,

$$\frac{10+9+10+9+9+10}{12\times 6}\times 100\% = 79.17\%, \tag{6}$$

which proves the effectiveness of the framework.

Given the cost ratio and minimum model performance requirement, the need to improve data quality can be determined with the help of the corresponding decision model. As an example, suppose the minimum model performance requirement C is fixed at 1000. The authors examine two scenarios involving the cost ratio r of the equipment under study.

In the first scenario, when the cost ratio r is set at 1000, the decision model depicted in Fig. 10 suggests that the data quality (i.e., $\mathbf{Q} = [10, 0.76]$) does not require improvement, and as a result, the HMM is chosen to construct

Table 5: Confusion metrics under no minimal performance requirement and different C

No minimum performance requirement					
		Actual decision			
		Improve DQ	HMM	MeTTF	MTTF
Predicted decision	HMM		7		
	MeTTF				
	MTTF			2	3
C=250					
Predicted decision	Improve DQ	9			3
	HMM				
	MeTTF				
	MTTF				
C=500					
Predicted decision	Improve DQ	7		1	
	HMM				
	MeTTF				
	MTTF			1	3
C=1000					
Predicted decision	Improve DQ	4	1		
	HMM		2		
	MeTTF				
	MTTF			2	3
C=2000					
Predicted decision	Improve DQ	1	1		
	HMM		5		
	MeTTF				
	MTTF			2	3
C=3000					
Predicted decision	Improve DQ				
	HMM		7		
	MeTTF				
	MTTF			2	3

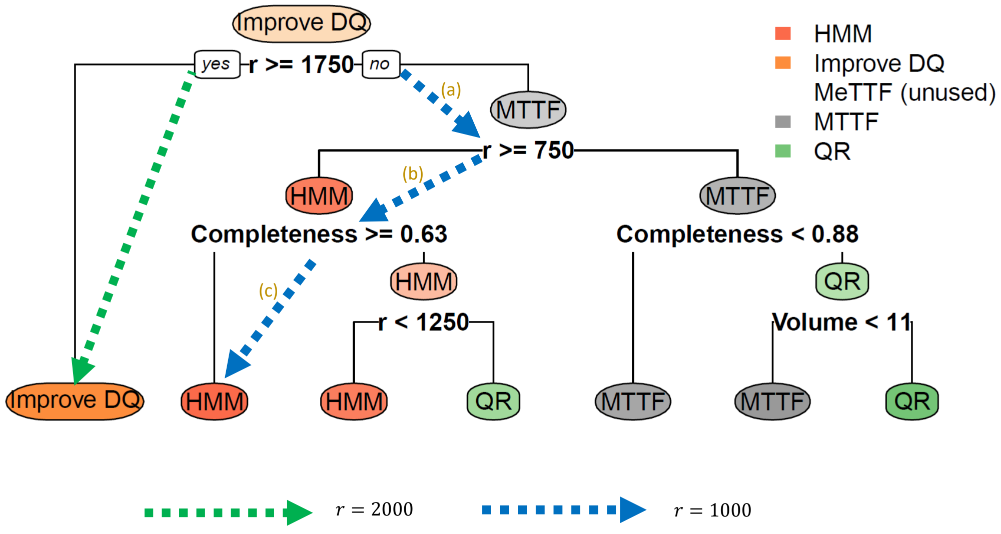

Figure 13: Decision derivation process for two scenarios ($r = 2000$ and $r = 1000$) with C fixed at 1000 and [Volume, Completeness] $= [10, 0.76]$

the risk estimation model.

In the second scenario, with the cost ratio is r set at 2000, the decision model indicates that the data quality needs enhancement. The decision derivation process for both scenarios is illustrated in Fig. 13.

Additionally, Fig. 14 showcases the model's performance before and after the data quality improvement, achieved by including two additional failed boards and addressing missing values.The data quality after data quality improvement is $\mathbf{Q}' = [12, 1]$. From the figure, it can be seen that the data quality improvement leads to a reduction in losses, especially when the cost ratios are large.

5. Conclusions and future works

This paper introduces an innovative framework for managing data quality, which is specifically tailored for estimating the risk of industrial equipment failure. This comprehensive framework encompasses data development, data quality assessment, decision-making for data quality requirements, data quality enhancement, and model development. It furnishes valuable guidance for

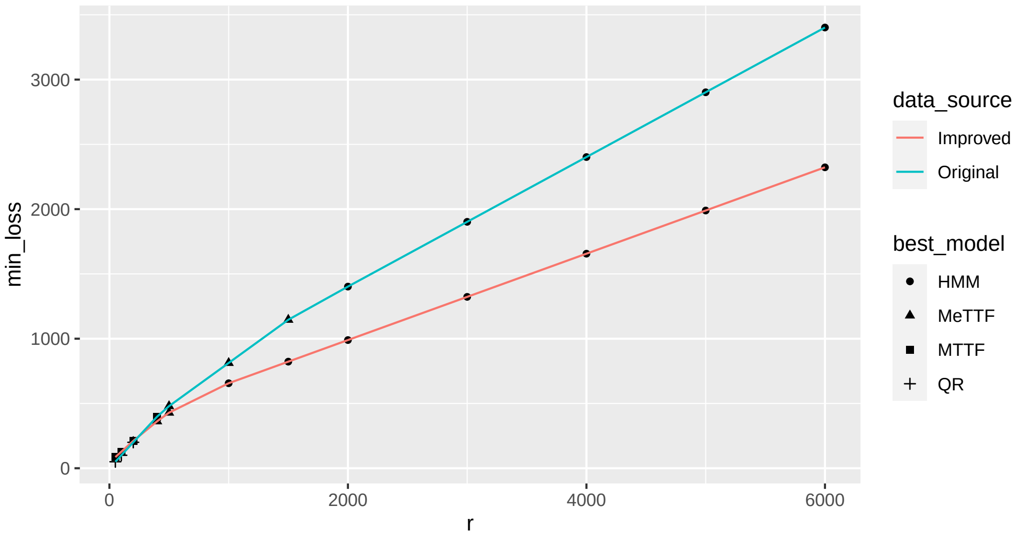

Figure 14: Comparisons of the best model and loss before and after data quality improvement under different cost ratios

data practitioners seeking to manage the data quality for risk estimation. Noteworthy advancements in this framework include incorporating a decision tree-based model for evaluating data quality compliance and selecting the best risk estimation model. Additionally, it introduces an improved loss function featuring a “cost ratio” parameter, enabling the model to accommodate equipment with varying failure costs versus early replacement costs.

The efficacy of this framework is exemplified through a case study utilizing actual data from oil well drilling operations. The proposed framework’s practical utility is showcased by comparing four risk estimation models, including baseline and machine learning. The framework’s validation employs a confusion matrix across different cost ratios and minimum performance requirements, revealing an average decision accuracy of 79.17%, confirming the effectiveness of the decision-making approach in this framework. In real-world situations, the decision-making model aids data practitioners in making informed decisions, enabling them to determine if data quality meets specific criteria and if not, guiding improvements. The presented case study results underscore the tangible advantages of enhancing data quality, especially when the cost ratio is substantial.

In conclusion, this paper delivers a robust data quality management framework that includes decision-making, empowering data practitioners in engineering asset management to make well-informed decisions regarding data quality and risk estimation models.

While the proposed framework has provided valuable insights into risk estimation, there are still several promising avenues for future research that warrant exploration.

- **Extend to repairable system**: The proposed loss function for evaluating the performance of a risk estimation model is presently constrained to non-repairable systems, such as electronic systems, with validation limited to electronic boards of drilling tools. To expand the scope of the data quality management framework and its application to a wider range of systems, future work could involve extending the loss function to repairable systems as well.

- **Incorporating economic considerations**: One key direction for improvement is the integration of cost and gain considerations related to data quality enhancement within the decision-making model. Understanding the economic implications of data quality improvements can lead to the development of more cost-effective data quality management strategies. This approach can help organizations make informed decisions about allocating resources for data improvement.

- **Assessing data label quality**: The quality of data labels, especially those related to failure modes and causes, significantly affects the accuracy of risk estimation. Investigating the influence of label quality on model performance is essential. In practical engineering asset management scenarios, obtaining accurate labels can be a complex and resource-intensive process, especially when dealing with complex equipment. Addressing this challenge is crucial for more realistic data quality management.

- **Diverse data formats**: Real-world data quality management often involves dealing with diverse data formats including time series, tables, text, images, and more. Expanding the current understanding of methodologies and frameworks for assessing data quality across these heterogeneous formats is vital. Each format poses unique challenges, and future research should aim to develop versatile quality assessment techniques to handle this diversity effectively.

- **Feature extraction improvement**: Feature extraction remains a critical aspect of the risk estimation model, profoundly affecting model performance. When a risk estimation model performs poorly, prioritizing the refinement of feature extraction methods should be considered. Improving the model by extracting more informative features can lead to more meaningful insights. Consequently, focusing on feature extraction enhancement may take precedence over data quality management efforts in some cases.

- **Deep learning-based data augmentation:** In addition to the three data preprocessing approaches mentioned—outlier detection, missing value handling, and data deduplication—deep learning methods, such as generative adversarial network (Li et al., 2022), can be employed to generate synthetic data and in turn improve the data quality. These methods can further enhance data quality and expand the repertoire of data quality improvement strategies.

Acknowledgment

This work has been supported by the EIPHI graduate school (contract "ANR-17-EURE-0002") and the CIFRE fellowship (n°2020/0127). The authors thank Rod Hotz from Schlumberger and the two anonymous reviewers from the conference whose comments helped improve and clarify this manuscript.

References

Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., Zerhouni, N., 2017. Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. International Journal of Prognostics and Health Management 8. doi:10.36001/ijphm.2017.v8i3.2667.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41. doi:10.1145/1541880.1541883.

Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z., 2014. A metrics-driven approach for quality assessment of linked open data. Journal of Theoretical and Applied Electronic Commerce Research 9, 64–79. doi:10.4067/S0718-18762014000200006.

Betz, T.S., El-Rayes, K., Grussing, M.N., Landers, K.E., Bartels, L.B., 2023. Parametric estimation of equipment failure risk with machine learning and constrained optimization. Journal of Performance of Constructed Facilities 37. doi:10.1061/jpcfev.cfeng-4284.

Buelvas, J.H., Múnera, D., Gaviria, N., 2023. Dq-man: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems. Internet of Things 22, 100769.

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79. doi:10.1016/j.neucom.2017.11.077.

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

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

Cichy, C., Rass, S., 2019. An overview of data quality frameworks. IEEE Access 7, 24634–24648. doi:10.1109/ACCESS.2019.2899751.

Deng, Y., Ju, H., Zhong, G., Li, A., Ding, Y., 2023. A general data quality evaluation framework for dynamic response monitoring of long-span bridges. Mechanical Systems and Signal Processing 200, 110514. doi:10.1016/j.ymssp.2023.110514.

Díaz Iturry, M., Alves-Souza, S.N., Ito, M., da Silva, S.A., 2021. Data quality in health records: A literature review, in: 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. doi:10.23919/CISTI52073.2021.9476536.

Ehrlinger, L., Wöß, W., 2022. A Survey of Data Quality Measurement and Monitoring Tools. Frontiers in Big Data 5. doi:10.3389/fdata.2022.850611.

Ehsani-Moghaddam, B., Martin, K., Queenan, J.A., 2021. Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. Health Information Management Journal 50, 88–92. doi:10.1177/1833358319887743.

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.J., Ducoffe, M., 2020. Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications of Artificial Intelligence 92, 103678. doi:10.1016/j.engappai.2020.103678.

Gitzel, R., Turring, S., Maczey, S., 2015. A data quality dashboard for reliability data, in: 2015 IEEE 17th Conference on Business Informatics, pp. 90–97. doi:10.1109/CBI.2015.24.

Gualo, F., Rodriguez, M., Verdugo, J., Caballero, I., 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.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R., Munigala, V., 2021. Data quality for machine learning tasks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. p. 4040–4041. doi:10.1145/3447548.3470817.

Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T., 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:10.1016/j.imu.2021.100799.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., Szubartowicz, M., 2018. Requirements for data quality metrics. J. Data and Information Quality 9. doi:10.1145/3148238.

IBM, 2023. What is a CMMS? Definition, how it works and benefits | IBM. URL: https://www.ibm.com/topics/what-is-a-cmms.

IBM, 2024. What is data quality? URL: https://www.ibm.com/topics/data-quality.

ISO 8000-2:2022, 2022. Data quality — Part 2: Vocabulary. Standard. International Organization for Standardization. Geneva, CH.

ISO 8000-8:2015, 2015. Data quality - Part 8: Information and data quality: Concepts and measuring. Standard. International Organization for Standardization. Geneva, CH.

ISO/IEC 25012:2008, 2008. Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Standard. International Organization for Standardization. Geneva, CH.

Ji, R., Hou, H., Sheng, G., Jiang, X., 2022. Data quality assessment for electrical equipment condition monitoring, in: 2022 9th International Conference on Condition Monitoring and Diagnosis (CMD), pp. 1–4. doi:10.23919/CMD54214.2022.9991385.

Jia, X., Ji, D.Y., Minami, T., Lee, J., 2022. Data quality and usability assessment methodology for prognostics and health management: A systematic framework. IFAC-PapersOnLine 55, 55–60. doi:10.1016/j.ifacol.2022.09.183. 5th IFAC Workshop on Advanced Maintenance Engineering, Services and Technologies AMEST 2022.

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

Kaiser, M., Klier, M., Heinrich, B., 2007. How to measure data quality? - a metric-based approach, in: Proceedings of the 28th International Conferenceon Information Systems (ICIS).

Kang, J., Masry, Z.A., Varnier, C., Mosallam, A., Zerhouni, N., 2023. Data Management Framework for Risk Estimate of Electronic Boards in Drilling and Measurement Tools, in: IFAC World Congress 2023, IFAC-PapersOnLine, Yokohama, Japan.

Kang, J., Varnier, C., Mosallam, A., Zerhouni, N., Youssef, F.B., 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), pp. 495–500. doi:10.1109/PHM2022-London52454.2022.00093.

Karkouch, A., Mousannif, H., Al Moatassime, H., 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.

Kläs, M., Vollmer, A.M., 2018. Uncertainty in machine learning applications: A practice-driven classification of uncertainty, in: Gallina, B., Skavhaug, A., Schoitsch, E., Bitsch, F. (Eds.), Computer Safety, Reliability, and Security, Springer International Publishing, Cham. pp. 431–438.

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

Koziel, S., Hilber, P., Westerlund, P., Shayesteh, E., 2021. Investments in data quality: Evaluating impacts of faulty data on asset management in power systems. Applied Energy 281, 116057. doi:10.1016/j.apenergy.2020.116057.

Li, X., Metsis, V., Wang, H., Ngu, A.H.H., 2022. TTS-GAN: A transformer-based time-series generative adversarial network, in: Michalowski, M., Abidi, S.S.R., Abidi, S. (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham. pp. 133–143.

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

Lukens, S., Naik, M., Saetia, K., Hu, X., 2019. Best Practices Framework for Improving Maintenance Data Quality to Enable Asset Performance Analytics. Annual Conference of the PHM Society 11. doi:10.36001/phmconf.2019.v11i1.836. number: 1.

Lukens, S., Rousis, D., Baer, T., Lujan, M., Smith, M., 2022. A Data Quality Scorecard for Assessing the Suitability of Asset Condition Data for Prognostics Modeling. Annual Conference of the PHM Society 14. doi:10.36001/phmconf.2022.v14i1.3188. number: 1.

Madhikermi, M., Kubler, S., Robert, J., Buda, A., Främling, K., 2016. Data quality assessment of maintenance reporting procedures. Expert Systems with Applications 63, 145–164. doi:10.1016/j.eswa.2016.06.043.

Makhoul, N., 2022. Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring. Advances in Bridge Engineering 3, 17. doi:10.1186/s43251-022-00068-9.

Martín, L., Sánchez, L., Lanza, J., Sotres, P., 2023. Development and evaluation of artificial intelligence techniques for iot data quality assessment and curation. Internet of Things 22, 100779. doi:10.1016/j.iot.2023.100779.

Martínez-Galán Fernández, P., Guillén López, A.J., Márquez, A.C., Gomez Fernández, J.F., Marcos, J.A., 2022. Dynamic risk assessment for cbm-based adaptation of maintenance planning. Reliability Engineering & System Safety 223, 108359. doi:10.1016/j.ress.2022.108359.

Mazumder, R.K., Salman, A.M., Li, Y., 2021. Failure risk analysis of pipelines using data-driven machine learning algorithms. Structural Safety 89, 102047. doi:10.1016/J.STRUSAFE.2020.102047.

Meisenbacher, S., Turowski, M., Phipps, K., Rätz, M., Müller, D., Hagenmeyer, V., Mikut, R., 2022. Review of automated time series forecasting pipelines. WIREs Data Mining and Knowledge Discovery 12, e1475. doi:10.1002/widm.1475.

Merino, J., Caballero, I., Rivas, B., Serrano, M., Piattini, M., 2016. A data quality in use model for big data. Future Generation Computer Systems 63, 123–130. URL: https://www.sciencedirect.com/science/article/pii/S0167739X15003817, doi:https://doi.org/10.1016/j.future.2015.11.024. modeling and Management for Big Data Analytics and Visualization.

Mosallam, A., Byttner, S., Svensson, M., Rögnvaldsson, T., 2011. Nonlinear relation mining for maintenance prediction, in: 2011 Aerospace Conference, IEEE. pp. 1–9.

Omri, N., Al Masry, Z., Mairot, N., Giampiccolo, S., 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.

Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data quality assessment. Commun. ACM 45, 211–218. URL: https://doi.org/10.1145/505248.506010, doi:10.1145/505248.506010.

Press, G., 2023. Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-s
?sh=687198b46f63.

Purnomoadi, A., Sari, I.M., Anna Maria, J., Fiddiansyah, D.B., Saputro, N.E., Sofan Hadi, M., 2023. A method to quantify data quality in asset health indices model, in: 2023 4th International Conference on High Voltage Engineering and Power Systems (ICHVEPS), pp. 16–20. doi:10.1109/ICHVEPS58902.2023.10257386.

Rekatsinas, T., Dong, X.L., Getoor, L., 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.

SLB, 2023. BHA. URL: https://glossary.oilfield.slb.com/en/terms/b/bha.

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

Society for Risk Analysis, 2018. Society for Risk Analysis Glossary. URL: https://www.sra.org/wp-content/uploads/2020/04/SRA-Glossary-FINAL.pdf.

Wang, K.Q., Tong, S.R., Roucoules, L., Eynard, B., 2008. Analysis of data quality and information quality problems in digital manufacturing, in: 2008 4th IEEE International Conference on Management of Innovation and Technology, pp. 439–443. doi:10.1109/ICMIT.2008.4654405.

Wang, R.Y., 1998. A product perspective on total data quality management. Communications of the ACM 41, 58–65.

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

Wang, Z., Du, H., Tao, L., Javed, S.A., 2023. Risk assessment in machine learning enhanced failure mode and effects analysis. Data Technologies and Applications ahead-of-print. doi:10.1108/DTA-06-2022-0232.

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

Xie, Q., Tao, G., Xie, C., Wen, Z., 2023. Abnormal data detection based on adaptive sliding window and weighted multiscale local outlier factor for machinery health monitoring. IEEE Transactions on Industrial Electronics 70, 11725–11734. doi:10.1109/TIE.2022.3231279.

Xu, D., Zhang, Z., Shi, J., 2022. A data quality assessment and control method in multiple products manufacturing process, in: 2022 5th International Conference on Data Science and Information Technology (DSIT), pp. 1–5. doi:10.1109/DSIT55514.2022.9943883.

Yao, Y., Wu, L., Xie, B., Lei, L., Wang, Z., Li, Y., 2023. A two-stage data quality improvement strategy for deep neural networks in fault severity estimation. Mechanical Systems and Signal Processing 200, 110588. doi:10.1016/j.ymssp.2023.110588.

Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., Hu, X., 2023. Data-centric Artificial Intelligence: A Survey. doi:10.48550/arXiv.2303.10158.

Zhou, Z., Wegner, L.D., Sparling, B.F., 2021. Data quality indicators for vibration-based damage detection and localization. Engineering Structures 230, 111703. doi:10.1016/j.engstruct.2020.111703.

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