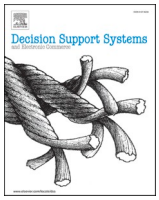
Decision Support Systems 196 (2025) 114493

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Handling imperfection: A taxonomy for machine learning on data with data quality defects

Michael Hagn, Bernd Heinrich *, Thomas Krapf, Alexander Schiller

Department of Information Systems, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

ARTICLE INFO

Keywords:
Taxonomy
Machine learning
Data quality
Data uncertainty

ABSTRACT

In recent years, machine learning (ML) has become ubiquitous in sectors including transportation, security, health, and finance to analyze large amounts of data and support decision-making. However, real-world datasets used in ML often exhibit various data quality (DQ) defects that can significantly impair the performance and validity of ML models and thus also the decisions derived from them. Therefore, a plethora of methods across various research strands have been proposed to address DQ defects and mitigate their negative impact on ML-based data analysis and decision support. This has resulted in a fragmented research landscape, where comparisons and classifications of methods dealing with ML on data with DQ defects are very challenging for both researchers and practitioners. Thus, based on a structured design process, we develop and present a taxonomy for this research field. The taxonomy serves as a systematic framework to classify and organize existing research and methods according to relevant dimensions and facilitates future work in this area. Its reliability, understandability, completeness, and usefulness are supported by an evaluation with external researchers and practitioners. Finally, we identify current trends and research gaps and derive challenges and directions for future research.

1. Introduction

In recent years, machine learning (ML) has been adopted in various sectors such as transportation, security, health, and finance to analyze large amounts of data and support decision-making. As an important example, ML is used to assist physicians in critical tasks such as the diagnostic classification of medical images for cancer screenings [1]. In finance, ML facilitates tasks such as algorithmic trading, credit scoring, and fraud detection [2]. In addition, transportation companies such as Uber use ML algorithms to allocate rides, set fares, and recruit drivers [3].

However, the performance and validity of ML models heavily depend on the quality of the data that is used for both training and use [4,5]. In many studies and applications, data quality (DQ) is assumed to be near-perfect; however, real-world datasets often exhibit various DQ defects [6] which can significantly affect the performance and validity of ML models, and thus the decisions derived from them when applied in practice [7]. Several incidents have been traced back to this issue in recent years, some with severe consequences such as the collision between a self-driving car and a truck in Florida in 2017. As a result of inaccurate data labeling during its training process, the ML model failed to recognize the truck as an obstacle, resulting in a fatal accident [8].

DQ defects can arise from various causes: For example, measurement errors lead to noisy or inaccurate data [9], defect sensors [10] or incomplete survey responses [11] can cause missing data, previously accurate data values can become outdated over time [12,13], or privacy concerns may result in distributed, incomplete data [14]. Furthermore, the same data may be affected by multiple causes simultaneously. Importantly, such data is not only defect in the sense that, e.g., certain data values are inaccurate or missing, but it is also uncertain since the true underlying values are unknown and can, e.g., be modeled using probability distributions. As a result of such uncertain data values or instances, the predictions of ML models trained on or applied to these data instances with DQ defects are also subject to predictive uncertainty. Predictive uncertainty describes the uncertainty of the model prediction itself. For example, when a classification model is used to classify an uncertain data instance, the resulting class prediction is not deterministic, but, e.g., a probability distribution over the set of possible classes [15]. Therefore, when ML models are trained on or applied to defect data, the impact of the defect data must be considered when supporting decisions based on their uncertain predictions.

Such issues have been recognized in literature and various methods

* Corresponding author.
E-mail addresses: michael.hagn@ur.de (M. Hagn), bernd.heinrich@ur.de (B. Heinrich), thomas.krapf@ur.de (T. Krapf), alexander.schiller@ur.de (A. Schiller).

https://doi.org/10.1016/j.dss.2025.114493
Received 7 November 2024; Received in revised form 11 June 2025; Accepted 11 June 2025
Available online 16 June 2025

have been proposed to account for DQ defects and mitigate their negative impact on ML-based data analysis and decision support. The range of possible strategies for such methods is very broad as, e.g., DQ defects could already be targeted before the ML model is trained or applied (e.g., by an imputation method in the case of missing data values), or the training procedure of an ML model could be adjusted to handle a specific DQ defect (e.g., by data label correction during model training). Due to this abundance of proposed methods across various research strands, a fragmented research landscape has emerged in which comparisons and classifications of the methods tackling ML on data with DQ defects are very challenging. Furthermore, this is not only the case from a methodological point of view; a valid assessment and comparison of the performance of the different proposed methods is also hardly possible, since the methods are often evaluated on (very) customized datasets (due to a lack of benchmark datasets with labeled DQ defects) [16]. A further complicating factor is the terminology and concepts used across papers, which—although describing the same or similar ideas—vary depending on the research strand. These challenges underscore the need for a framework to classify and organize existing research and methods according to relevant dimensions describing the field of ML on data with DQ defects.

Related studies [17,18] have already reviewed specific subsets of DQ defect types and associated DQ handling methods (see Section 2.2 on Related Work) and thus address selected parts and aspects of DQ and its role in ML applications. However, to the best of our knowledge, there is no comprehensive taxonomy comprising relevant dimensions of ML on data with DQ defects. Therefore, we aim to close this gap by addressing the research question: *How can existing works dealing with ML on data with DQ defects be systematized and classified with respect to relevant dimensions of a taxonomy?*

We present a comprehensive taxonomy for the research field of ML on data with DQ defects. Our aim is not to provide an exhaustive, fully comprehensive overview or survey on all existing literature in the field or to detail specific methods, but rather to provide a structured framework for this important area of research. Our main contributions can be summarized as follows:

- Following an established taxonomy design methodology, we develop a comprehensive taxonomy for the field of ML on data with DQ defects, enabling the classification of DQ defects and strategies for addressing these DQ defects in ML-based data analysis with respect to relevant dimensions.
- Based on this taxonomy and the literature we analyzed during the taxonomy building process, we identify current trends and research gaps in this ever-evolving research field and deduce challenges and directions for future research.

The remainder of the paper is structured as follows: Section 2 provides background on DQ and its manifold influence on ML before reviewing related studies on the role of DQ defects in ML in detail. Section 3 outlines the general taxonomy design process proposed by Kundisch et al. [19] which we follow to develop our taxonomy. In Section 4, the final taxonomy for the field of ML on data with DQ defects is presented in detail. In Section 5, we evaluate our taxonomy and present the results. We conclude with a discussion of our contributions to research and practice, and potential starting points for future research.

2. Background and related work

2.1. Background

DQ has been the subject of extensive research, yielding important theoretical and practical insights. Various works have established theoretical foundations for DQ and DQ dimensions [20,21]. Building on these works, general frameworks and methodologies for DQ have been introduced [22,23]. It has been established that DQ can be characterized and analyzed with respect to multiple DQ dimensions [21] such as completeness, accuracy, consistency and currency. Based on this, research has proposed methods to identify and measure DQ defects of various DQ dimensions with the aim of managing DQ and supporting decision-making, e.g., by designing metrics for the consistency [24] and currency [12,13] of data in order to identify data of poor quality before their further use. Going beyond both the identification of DQ defects and the measurement of DQ, multiple works have introduced data cleaning methods and frameworks to address and resolve DQ defects [25]. Others have examined the causes of DQ defects in the data collection process and have proposed ideas for curating DQ by improving this process, e.g., by adding incentives for higher DQ [26,27]. In addition to DQ defects in data values, research has also addressed DQ defects in data labels by providing methods for handling inaccurate or missing labels, for example by a confident learning methodology [28].

Literature has also discussed the variety of ways in which DQ influences ML models trained on or applied to data with DQ defects. It has been established that DQ defects usually negatively affect model performance and validity [4,5] and induce predictive uncertainty [15]. In contrast, addressing DQ defects may have negative effects as well, e.g., if the imputation of missing values introduces bias into the data which then affects ML model training or use. In addition, model fairness, i.e., the absence of bias regarding, e.g., discrimination or social inequalities [29] can also be influenced by DQ [30]. Indeed, research has shown that, depending on how DQ defects are addressed, improving DQ can actually be detrimental to fairness. For example, Guha et al. [31] demonstrate that automated data cleaning techniques may negatively impact fairness in ML models and are even more likely to worsen fairness than to improve it. Another aspect related to DQ is the explainability of ML models, with DQ defects potentially influencing not only the ML model itself, but also the validity of explanations generated for it [32]. Moreover, DQ also is relevant for privacy-enhancing techniques in ML such as federated learning, where data with poor DQ needs to be identified to improve performance [33]. Given these various ways in which DQ influences ML models, we focus on the training and use of ML models on data with DQ defects in our work. In this setting, we analyzed research works proposing methods which aim to address, and thus resolve or mitigate problems caused by DQ defects in ML-based data analysis in an application context such as finance or health (see Section 3.2 for further details on our focus). In particular, we do not consider stages before data preprocessing (e.g., the data collection) or stages beyond the output of the ML model (e.g., the interpretation and adoption of the model predictions by humans). We also do not consider cases in which new DQ defects may influence ML models after the training process of the model is completed, e.g., when previously used data is reclaimed due to privacy concerns, resulting in the need for machine unlearning [14].

2.2. Related work

In this section, we present related studies aiming to organize the field of ML on data with DQ defects. Importantly, we do not focus on works that just provide an overview of DQ without ML context.

A first category of works discusses DQ and its role in ML-based data analysis at a more general level as one of several important challenges in the field of ML. L'Heureux et al. [34] outline ML challenges arising from big data properties such as the volume, velocity, variety, and veracity of big data, which inherently include a number of DQ defects resulting from these properties. Paleyes et al. [35] examine DQ issues appearing during the process of data collection and data (pre)processing for ML-based analysis. Other works [36,37] provide an overview of ML methods for learning from class-imbalanced data, which is partly the result of DQ defects such as inaccurate or missing data. None of these works, however, aim to provide a comprehensive overview of ML on data with DQ defects. Moreover, they do not focus on addressing these

DQ defects in ML-based data analysis in the application context such as health, especially in the training, use, and output of the ML model.

A second category of works specifically focuses on overviews of DQ defects and their influence on ML training and use. Gong et al. [4] discuss DQ and its impact on ML models and give an overview of DQ dimensions and DQ metrics. Mohammed et al. [7] empirically examine how a variety of DQ defects affect ML model performance. Priestley et al. [38] also analyze DQ in different stages of an ML-based data analysis process and list relevant DQ issues in each stage of the process. Gudivada et al. [39] discuss a selection of DQ defects such as missing, heterogeneous, and duplicate data in the context of ML and present a selection of available DQ management tools. Similarly, Li et al. [18] review common DQ issues before focusing on the potential of Active Learning-based approaches for handling DQ defects. However, all of these works only discuss specific aspects of DQ defects and their handling in ML applications. They in particular ignore the possibility of addressing DQ defects by modifications of the ML model itself. Furthermore, they do not aim to organize and systematize the relevant dimensions of the research field of ML on data with DQ defects.

A third category of works includes studies that incorporate the possibility of addressing DQ defects by modifying the ML model. Das et al. [17] present an overview of various distribution-based and feature-based data irregularities and discuss selected approaches to improve classifier robustness against such irregularities. However, they limit their scope to four types of data irregularities, excluding DQ defects such as inaccurate or outdated data. Gawlikowski et al. [40] review methods to handle DQ defects by incorporating DQ-related (data) uncertainty into different types of neural network (NN) models. Similarly, Wang et al. [41] discuss techniques for dealing with various forms of uncertainty, including data uncertainty caused by DQ defects in graph NNs. Yet, none of them examine DQ defects in detail or cover ML models beyond NNs and graph NNs, respectively. Therefore, these works also do not provide a comprehensive systematization of the field of ML on data with DQ defects.

In summary, while some studies and overviews address selected aspects of DQ defects in ML-based data analysis, a thorough systemization of this research field is still missing. In particular, there is no comprehensive taxonomy of ML on data with DQ defects that allows researchers and practitioners to organize and systematize approaches and methods in the field. As a result, previous discussions of future research agendas are also limited to the selected DQ aspects covered in existing works. We aim to close these gaps by addressing our research question.

3. Methodology

3.1. Research method

We now describe the extended taxonomy design process (ETDP) by Kundisch et al. [19] which we follow for the development of our taxonomy. A taxonomy is a classification system that helps with the conceptualization of objects (i.e., in our case, papers) [19]. It comprises a number of dimensions and a set of characteristics for each dimension. It is also possible for a dimension to contain multiple subdimensions, each with a set of characteristics, and the dimension can then be seen as a group of subdimensions. When an object is classified according to the taxonomy, one or multiple characteristics are selected for each dimension or subdimension respectively. The ETDP is based on the well-established taxonomy design process by Nickerson et al. [42] that is widely used in research fields such as computer science, information systems, and finance. Its iterative nature enables taxonomy designers to update, extend and refine taxonomies without the need to repeat the entire taxonomy design process from scratch. We also follow the 26 taxonomy design recommendations of Kundisch et al. [19] (in addition to the ETDP) with operational support and examples of good practice. The 18 steps of the ETDP for designing and evaluating the taxonomy are summarized in Table 1.

Kundisch et al. [19] recommend that at least one empirical-to-conceptual step and at least one conceptual-to-empirical step should be conducted. In the following Section 3.2, we describe the first 14 steps of the ETDP for our taxonomy for the field of ML on data with DQ defects. In Section 4, the final taxonomy is presented and communicated (Step 18). Finally, we discuss the conceptualization and execution of the taxonomy evaluation (Steps 15–17) in Section 5.

3.2. Taxonomy design approach

Following the ETDP as described in the previous section, we began by defining the observed phenomenon (Step 1). We focused on the training and use of ML models on data with DQ defects. In this setting, we analyzed research works proposing methods which aim to address, and thus resolve or mitigate problems caused by DQ defects in ML-based data analysis in an application context such as finance or health. To define and clarify which works are relevant to us, we considered a simplified process of ML-based data analysis that comprises three general stages and is shown in Fig. 1.

In Stage 1, the data is assessed and, in most cases, preprocessed and prepared. The DQ defects may already be identified and potentially addressed in this stage, e.g., by imputing missing data values. In Stage 2, the preprocessed data from Stage 1 is analyzed by an ML model in the context of an application, e.g., a diagnostic classification of healthcare data or a decision to grant or refuse a credit based on finance data. Typically, an ML model performs a regression, classification, clustering, or similar task in this second stage. In Stage 3, the ML model and its outputs (predictions) are analyzed and potential information about performance, validity and predictive uncertainty can be examined, e.g., by applying confidence measurements. Papers are relevant to our work if and only if they include an ML-based data analysis (Stage 2) in an application context as described above, i.e., if they comprise at least two consecutive stages of this three-stage pipeline. In particular, we exclude cases in which a method (e.g., an ML method) is only used to correct or address DQ defects (e.g., by ML-based imputation of missing values) without a further analysis of the preprocessed data in an application context. Furthermore, we consider the already existing raw data that contains DQ defects, but do not explicitly focus on the real-world causes of these DQ defects prior to Stage 1 (e.g., malfunctioning sensors during data collection), as this constitutes a broad own field of research. Note that some existing methods for addressing DQ defects can also change the type of DQ defects in the data. For example, imputing missing values can turn a problem of missing data into a problem of inaccurate data as the imputed values are potentially incorrect. In such cases, we only analyze the DQ defects initially addressed by the method and consider any further changes to the data and its DQ defects to be an inherent part of the method.

In the next steps (2–4) of the ETDP, we defined the target user groups and intended purpose of our taxonomy and formulated its meta characteristic. The target groups of the taxonomy are researchers and practitioners working with ML on data with DQ defects. The purpose of the taxonomy is to provide a comprehensive structure for classifying and systematizing various methods on ML on data with DQ defects in both theoretical contexts and practical applications. We further aim to use our taxonomy to identify different research strands, foci, and research gaps in this area. We define the meta-characteristic of the taxonomy to be *dimensions of ML on data with DQ defects*. Step 5 includes determining the ending criteria for the iterative taxonomy building process and specifying the evaluation goals. For the ending criteria, we chose the criteria presented in [19,42]. They postulate that no new dimensions must be added or removed during the last iteration. After these conditions were met, we conducted a rigorous evaluation with external researchers and practitioners to assess whether the taxonomy is reliable,

Table 1
Extended Taxonomy Design Process (ETDP).

Step	Description	
1	The observed phenomenon is defined.	
2	The target user group(s) are specified.	
3	The intended purposes of the taxonomy are specified.	
4	The meta-characteristic of the taxonomy is formulated. It determines the perspective on the observed phenomenon (delimiting non-considered phenomena). All dimensions and characteristics of the taxonomy must relate to the meta-characteristic.	
5	Ending conditions for the (iterative) taxonomy building process and evaluation goals are defined.	
	While ending conditions are not satisfied	
6	The approach for the development step of the taxonomy is chosen	
	Empirical-to-conceptual	Conceptual-to-empirical
7	Real-life objects (e.g., papers addressing methods for ML on data with DQ defects) are identified.	Taxonomy dimensions and characteristics are conceptualized by the authors.
8	Taxonomy dimensions and characteristics are recognized based on the identified real-life objects.	Real-life objects are examined regarding the conceptualized dimensions and characteristics.
9	Taxonomy dimensions and characteristics are grouped by the authors.	[This step is skipped for conceptual-to-empirical iterations.]
10	The state of the taxonomy is created (in the first iteration) or revised (in later iterations) by applying taxonomy operations such as adding, updating, or deleting (sub)dimensions and characteristics.	
11-14	It is tested whether the taxonomy satisfies the defined ending conditions	
	Yes	No
	The process continues with Step 15.	A new iteration starting at Step 6 is performed.
15	The taxonomy evaluation is conceptualized and operationalized.	
16	The taxonomy evaluation is executed.	
17	Based on the evaluation, it is assessed how well the taxonomy fulfills its intended goals for the identified target user group(s) (cf. Step 2).	
18	The taxonomy building process and the final taxonomy are reported in an appropriate manner.	

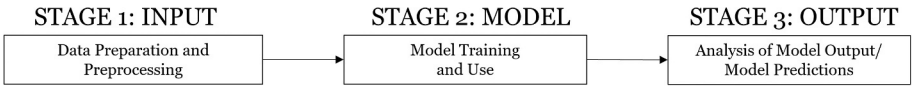


Fig. 1. Three-stage pipeline of ML-based data analysis.

understandable, complete, and useful (cf. Section 5 and Supplement 2[1]). For this evaluation, we randomly selected 16 of 150 identified relevant papers as 'unseen' test papers and used the remaining 134 papers for the literature analysis in the taxonomy building process.

For our taxonomy, we conducted four iterations (Steps 6–10) as shown in Fig. 2. More precisely, we alternately traversed two conceptual-to-empirical and two empirical-to-conceptual iterations, incorporating expertise from the authors and external researchers, as well as additional insights obtained from the identified works on ML on data with DQ defects. In Iteration 4, no further changes were proposed. Therefore, the ending conditions defined in Step 5 were met and the iterative taxonomy building process was completed. Each iteration and the development of the taxonomy throughout the iterations are described in detail in Supplement 1. The subsequent evaluation showed that the evaluation goals were fulfilled, and no further adjustments of the taxonomy dimensions and characteristics were necessary.

4. Taxonomy for ML on data with DQ defects

In the following, we present the dimensions and characteristics of our proposed taxonomy (cf. Fig. 3). In general, the characteristics of each dimension and subdimension are non-exclusive, as papers may address different data with different DQ defects as well as different tasks or models. The only exceptions are characteristics of the form *no use* that are obviously exclusive to the other characteristics of that dimension.

The **first dimension** is the **DQ-related Problem Setting** which explicates how the paper describes the DQ problem it addresses. This dimension encompasses three subdimensions, the first of which is **DQ Dimension**. Its characteristics are the established data value-based DQ dimensions *accuracy*, *currency*, *completeness*, *consistency*, a characteristic *other* (dimension) containing rarely used DQ dimensions (e.g., uniqueness) as well as *not specified*. The last characteristic is only selected if the DQ dimension cannot be extracted or inferred from the paper, i.e., if the paper simply describes the data as defect or deficient (as opposed to using words such as "missing", indicating the characteristic *completeness*).

The second subdimension is the **DQ-related Task**, i.e., the objective with respect to handling DQ defects. This subdimension has eight characteristics. The first characteristic is *outlier correction*, i.e., the detection, removal and/or correction of outliers (outlier instances) in the data. The second characteristic is *noise correction*, i.e., the detection, removal and/or correction of inaccurate data instances, features or feature values. In literature, such inaccurate instances are often referred to as noisy instances, where for example both the feature values and the labels of the instances may be corrupted. The third characteristic is *inconsistency correction,* i.e., the detection, correction and/or removal of inconsistencies, e.g., by violations of consistency rules in the data. Such violations of consistency may comprise, e.g., domain value violations where a feature value is outside the feasible domain of the feature, violations of rules describing relations between different instances or violations of syntactic or semantic rules in the data. The fourth characteristic is *imputation,* i.e., the imputation or removal of missing data which includes not only, e.g., the completion of missing data values but also missing data handling such as the listwise deletion of incomplete instances or the deletion of incomplete features from the dataset. While the first four characteristics are associated to DQ-related tasks typically performed during the input stage, the fifth characteristic *defect data in training/use* refers to the handling of defect data during training or use of an ML model, i.e., by modifying an existing or designing a new

[1] The Supplement is available at https://zenodo.org/records/15631630.

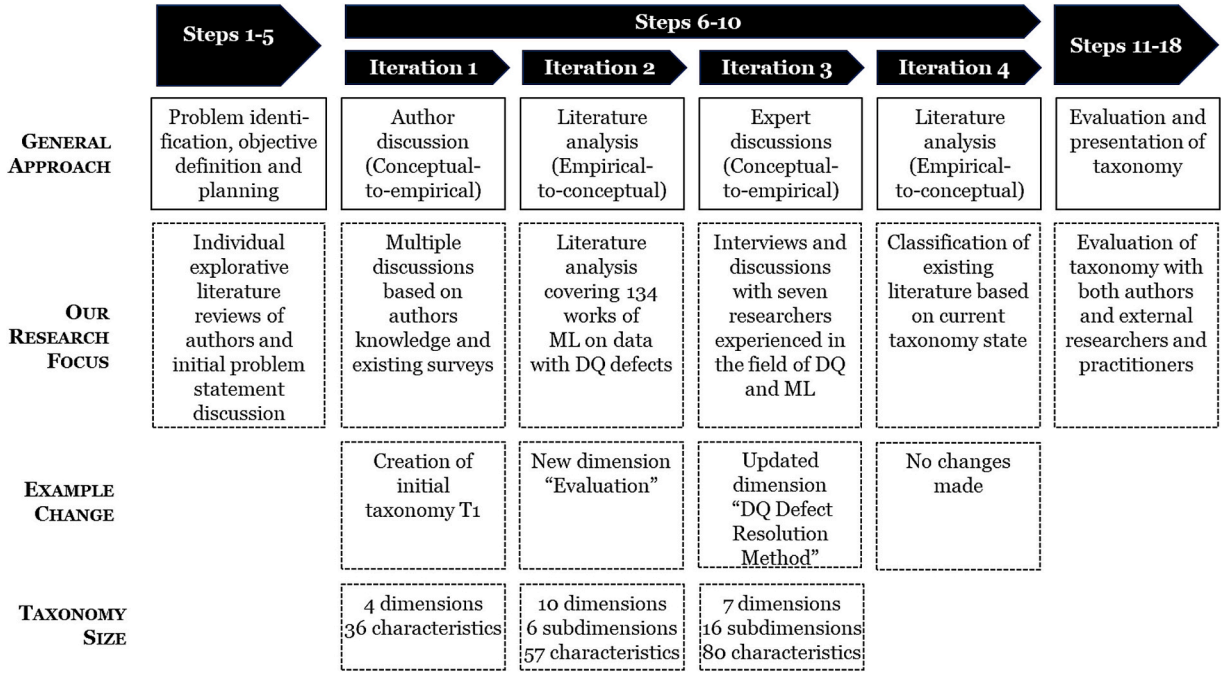


Fig. 2. Steps of the taxonomy building process.

ML model, DQ defects are directly handled without addressing them during the input stage of the pipeline. Examples for this characteristic include modifications of ML models to handle missing data or methods to make the training and use of ML models more robust against noisy data. The sixth characteristic is *data augmentation*, i.e., the creation of new data instances which are assumed to be free of DQ defects. Note that this differs from imputation as data augmentation typically creates new instances whereas imputation aims to fill in missing values in existing data instances [43]. The seventh characteristic is *duplicate detection,* i.e., the detection and removal of data duplicates. The final characteristic *other* DQ-related task comprises rare DQ-related tasks that do not fit into the other characteristics, such as knowledge transfer between domains or automated error detection.

The third subdimension is **Uncertainty Modeling/Quantification**, that is, if and how the paper formalizes and models the uncertainty resulting from DQ defects. More precisely, since the true data usually is not known (otherwise the DQ defects could be directly resolved), data affected by DQ defects is often referred to as uncertain data. By uncertainty modeling, we refer to quantitative, formal descriptions of data uncertainty based on a theoretical foundation. The characteristics of this dimension are, first, *probabilistic,* i.e., the modeling of uncertainty using probability theory, second, *belief-based,* i.e., the modeling of uncertainty using belief theory, third, *fuzzy theory,* i.e., the modeling of uncertainty using fuzzy (set) theory, and fourth, *other* theories/methods of uncertainty modeling. Moreover, if no such treatment of uncertainty is employed, the characteristic *no uncertainty modeling* is chosen. Since we require theoretical foundations for uncertainty modeling, the use of measures which may be interpreted as quantifications of uncertainty but lack theoretical foundations – such as noise scores or Softmax scores which are normalized on the interval [0,1] but must not be interpreted as probabilities [44] – is not considered as uncertainty modeling and thus belongs to this last characteristic.

The **second dimension** is the **Pipeline Position** in which the DQ defects are addressed. The characteristics of this dimension comprise the three pipeline stages of ML-based data analysis *input, model,* and/or *output* (cf. Fig. 1). Note that while we require the paper to consider at least two consecutive stages to be relevant for our taxonomy, addressing the DQ defects does not necessarily take place in all of the stages considered. For example, a paper may include the first two steps (i.e., input and model), but only address DQ in the *input* stage, e.g., by imputing missing values without further consideration of the DQ defects in the subsequent use and evaluation of the model.

The **third dimension** focuses on the data and encompasses multiple subdimensions describing the **Properties of the Data** and the DQ defects considered in the paper. The first subdimension is the **Structure** of the data, i.e., whether the data are *structured, semi-structured,* and/or *unstructured.* The second subdimension is the **Bias Type** of the data. In literature, the bias type or missingness mechanism of the data is a common notion when discussing the DQ dimension of completeness [45,46] where data can be missing not at random, missing at random or missing completely at random (cf. Supplement 1). However, we argue that these different bias types may also be defined and analyzed for other DQ dimensions such as accuracy (e.g., systematic versus random measurement errors). Therefore, we propose the notions of data being *defect completely at random (DCAR)*, *defect at random (DAR)*, or *defect not at random (DNAR).* In addition, the bias type may also be *not specified* in case it cannot be extracted or inferred from the paper. Another subdimension is the **Availability** of the data, i.e., whether the data/dataset used in the paper is *publicly available* or *not publicly available.* The third characteristic for this subdimension is that the availability of the data is *not specified.* A fourth subdimension is the **Extent** of uncertainty/DQ defects in the data, that is, whether the DQ defects are contained in *training, validation*, and/or *test/use data.* A fifth subdimension is **Object** of uncertainty/DQ defects, that is, whether the DQ defects appear in *feature values, labels, instances*, and/or *other* data objects such as graphs, networks, or unstructured data (e.g., textual data). The characteristic *feature value* is selected if any feature values in the data are defect. The characteristic *instance* is selected in addition to *feature value* if and only if all feature values of instances are affected by DQ defects. Examples of such a case are papers where entire instances are randomly deleted—so that the values of all features are missing—to test the robustness of the ML model to this defect. Thus, *instance* can never be an exclusive characteristic of this subdimension but poses a useful characteristic by indicating whether the entire instance is affected instead of just some of its feature values. Note that the characteristics of all subdimensions of this dimension are generally non-exclusive, as any paper may not only

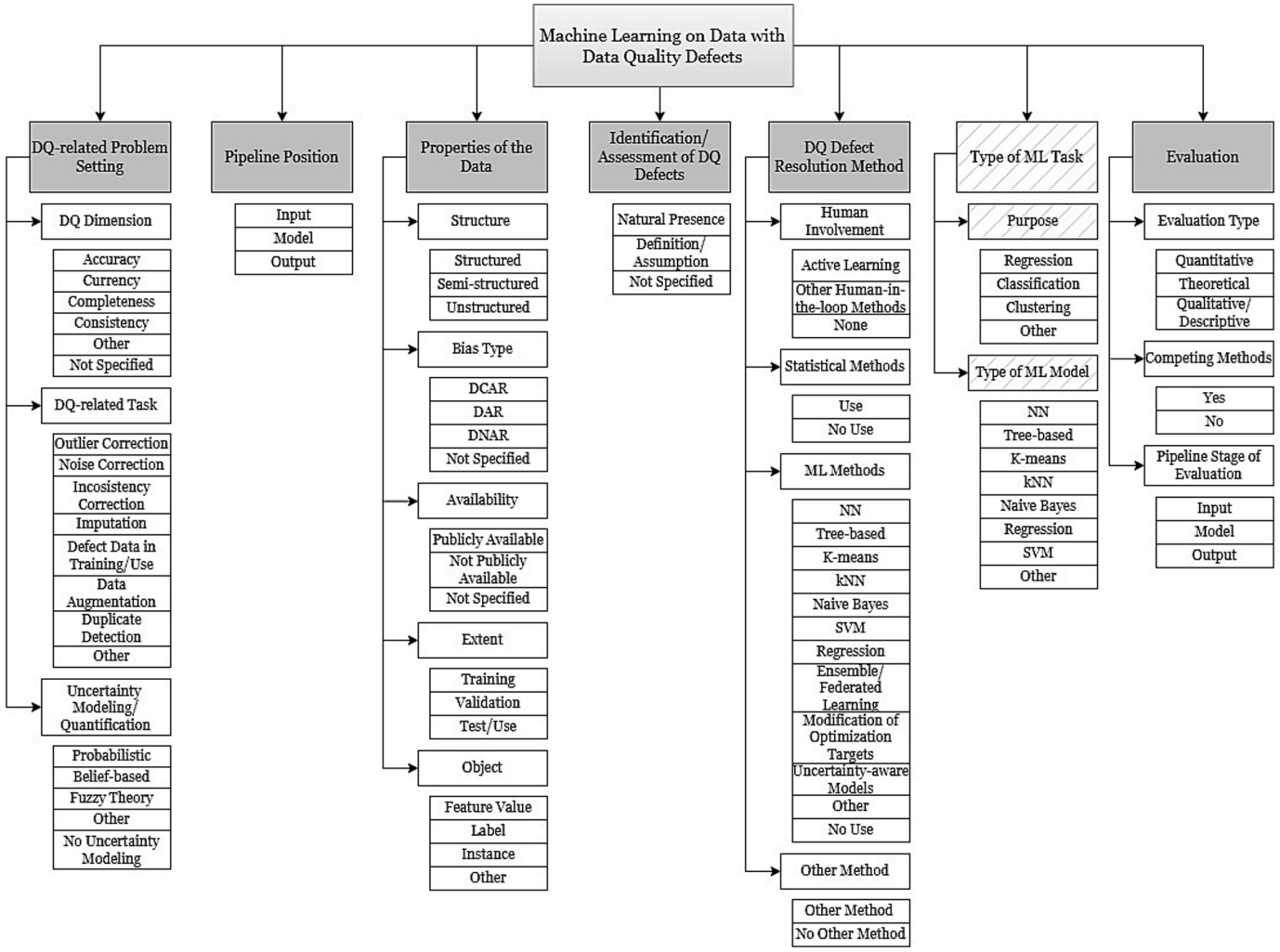


Fig. 3. Taxonomy of ML on data with DQ defects.

address data with multiple defects, but also different types of data (e.g., structured and unstructured data).

The **fourth dimension** is **Identification/Assessment of DQ Defects**. This dimension has two characteristics based on why the data is known to be defect. On the one hand, DQ defects may be *naturally present* in the data. In this case, information about the presence of DQ defects must be obtained or made available externally. On the other hand, DQ defects can be obtained by *definition or assumption*, e.g., by assuming measurement data as uncertain/noisy (by definition), or by manually deleting feature values or instances to create incomplete data. The third characteristic is that the identification of DQ defects is *not specified* in the paper. Note that while the identification and assessment of DQ defects usually takes place in the input stage of the data analysis pipeline, DQ defects may also be identified in later stages, i.e., during the model use itself or even in the model output.

The **fifth dimension** is the **DQ Defect Resolution Method**, i.e., the method used in the paper to address or resolve the DQ defects. This dimension has a total of four subdimensions. The first subdimension is **Human Involvement** to address DQ defects. Its characteristics are *active learning methods*, *other human-in-the-loop methods,* and *none*. The second subdimension is the use of **Statistical Methods** to address DQ Defects. Its characteristics are *use* of statistical methods and *no use* of statistical methods. The third subdimension is the use of **ML Methods** to address DQ Defects. The characteristics of this subdimension are various well-known standard ML models that can be used to address DQ defects as well as characteristics describing more specialized models for addressing DQ defects. The characteristics describing the use of standard ML models are use of *neural networks (NNs)*, *tree-based models* such as decision trees (DTs) or random forests (RFs), *support vector machines (SVMs)*, *Naïve Bayes (NB)*, *k-means*, *k-nearest neighbors (kNN)*, and *regression* models. Another characteristic is the use of *ensemble and/or federated learning* models, e.g., training multiple models on different imputed datasets and using ensemble learning to minimize the influence of inaccurately imputed data. The use of *uncertainty-aware models* is a characteristic which not only includes the use of existing, uncertainty-aware model architectures such as Bayesian NNs (BNNs) whose weights are probability distributions to account for uncertainty in the weights trained on potentially defect data, but also methods which modify or extend deterministic ML models to account for uncertainty. An example of the latter would be modifications of ML models to take probability distributions as input data and propagate the uncertainty information given by the distribution to the output. A further characteristic is the *modification of optimization targets* to address DQ defects. For ML models, different types of optimization targets exist and can be modified in various ways to account for DQ defects. Examples of such modifications are NNs with modified loss functions to account for DQ defects in the training data (e.g., by adding regularization) or DTs with modified entropy criteria in the tree building process. The final characteristics of this subdimension are *other* use of ML methods (e.g., Expectation Maximization or genetic programming) and *no use* of ML methods to address DQ defects. The fourth subdimension **Other Method** signals whether another method of DQ defect resolution, e.g., a fuzzy

approach not using ML or statistical methods, is employed.

The **sixth dimension** is the **Type of ML Task** for the data analysis in the application context (and not for the handling of DQ defects as considered by the previous dimension). This dimension has two subdimensions, the first of which focuses on the main **Purpose** of the ML model and encompasses the characteristics *classification, clustering, regression*, or *other* type of task (e.g., segmentation). The second subdimension is the more detailed **Type of ML Model** used for the task, that is, whether an *NN* (or one of its variants, such as a CNN), a *tree-based* model (e.g., DT or RF), *SVM*, *k-means, kNN, NB,* a *regression model* (e.g., linear or logistic regression), or *other* ML model is employed. To obtain a reasonable number of characteristics for this subdimension, we propose to distinguish only the types of ML models listed above. Unlike other dimensions, this dimension strictly relates to the ML model but not to the DQ defects (yet it is very informative for users of the taxonomy). We therefore marked it in hatched style in Fig. 3.

The **seventh** and final **dimension** is the **Evaluation** of the method that is conducted in the paper. This dimension comprises three subdimensions. The first subdimension is **Evaluation Type**, that is, whether the evaluation of the method is *quantitative* (i.e., by providing measures such as accuracy, precision, F1-measure, mean average error etc.), *theoretical* (i.e., by providing theoretical proofs and/or a deeper theoretical analysis for the method) and/or *qualitative/descriptive* (i.e., the evaluation discusses a wider range of purposes provided by the new method or its ability to handle special types of input data or DQ defects). The second subdimension is whether the evaluation contains a comparison to **Competing Methods** from existing literature or not. The third subdimension is the **Pipeline Stage of Evaluation**, i.e., the stage (s) of the ML-based data analysis pipeline that the evaluation focuses on. The characteristics of this subdimension are the evaluation of the method in the *input* stage, the *model* stage, and/or the *output* stage. An example for an evaluation at the input stage is a performance assessment of the method addressing DQ defects only in the preprocessing, e.g., by measuring imputation errors, regardless of the resulting performance of the ML model. An example for an evaluation at the model stage would be the assessment of the performance of the ML model in terms of, e.g., accuracy, without versus with addressing the DQ defects. An example of an evaluation at the output stage would be when additional confidence levels for all model outputs are provided and their validity is analyzed.

5. Evaluation

Following the ETDP, we evaluate our taxonomy in terms of reliability, understandability, completeness, and usefulness. More precisely, since the taxonomy is designed for researchers and practitioners working with ML on data with DQ defects, it is a key requirement that the taxonomy is reliable and understandable to both target groups. Reliability constitutes the proportion of joint judgment in which there is an agreement [47]. Understandability describes whether the taxonomy can be understood and employed by its target groups and can be measured by task performance [48]. Furthermore, the taxonomy should be complete und useful [42,49]. To assess these four criteria for our taxonomy, we performed a two-part evaluation with external experts, which is described in the following. A detailed evaluation protocol is provided in Supplement 2.

5.1. Evaluation methodology

For the evaluation, we used 16 papers which had been randomly selected from the 150 papers identified in Iteration 2 of the taxonomy building process (cf. Section 3). We were able to recruit at panel of eight external experts (called external raters), each with expertise in the fields of DQ and ML and between 3 and 13 years of experience in both academic research and industry. Each external rater received six randomly assigned papers and classified them according to the taxonomy. Thus, in total, each of the 16 papers was classified by three external raters (for details cf. Supplement 2). We first evaluated the reliability and understandability of the taxonomy based on these results. Subsequently, all eight external raters were asked to fill out a questionnaire on the completeness and usefulness of the taxonomy.

To evaluate the reliability of the taxonomy, we measured the agreements between the external raters and a gold standard (in line with [50]) for the classification of the 16 papers for each dimension and over all dimensions on average. To create the gold standard, each test paper was also classified by the four authors of the taxonomy (internal raters). The authors then discussed their classifications to reach a consensus on the classification of all 16 papers (cf. Supplement 2). The agreements between each internal rater and the gold standard were measured as well. In addition, we measured the agreements of the groups of researchers and practitioners respectively (both as parts of external raters). This allowed us to assess whether the description of the taxonomy together with its dimensions and characteristics is understandable for experts with different backgrounds, as substantial agreement is only possible if the taxonomy is well understood.

A primary concern for any evaluation is that it should be designed in a such a way that high-quality answers can be expected. Well-established research on the appropriate length and complexity of evaluations suggests that survey quality suffers when respondents have difficulty completing complex tasks [51]. Similarly, in evaluations that require high effort, loss of interest among participants may lead to ill-considered and unreliable answers or high non-response to questions, potentially invalidating the entire results of the study [52]. Such concerns also apply for our evaluation as each external rater would have to read, understand, and classify six full research papers describing complex ML methods and applications and, in most cases, containing a number of mathematical formulas. Thus, we concluded that the volume and complexity of the task should be reduced to allow for complete and reliable answers by each rater, which was supported by discussions between the authors and other practitioners prior to the evaluation. Therefore, the raters did not receive the full papers, but automatically generated summaries with a length of no more than two pages per paper. These summaries could not be created by the authors, as the authors' knowledge about the taxonomy could certainly bias the paper summaries in a way that could lead the external raters to classify the papers in the same way as the authors. To avoid this, we used GPT-4o to create a summary of each paper based on a pre-defined prompt which contains a brief description of the taxonomy's objectives, but not its explicit dimensions and characteristics. For a *post-hoc* evaluation of the quality of these paper summaries, we conducted a taxonomy classification using four randomly selected test papers. This classification was independently performed by the authors once based on the full papers and once based on the respective summaries generated by GPT-4o. A comparison of both taxonomy classifications for each paper revealed a near-perfect agreement (cf. Supplement 2.9 for details). Furthermore, these summaries serve as a consistent and comprehensible basis for all internal and external raters, regardless of time constraints or individual knowledge. Thus, both internal and external raters only read the summaries, but not the full papers for their classifications. In particular, the gold standard is also based on the summaries to ensure the comparability between the gold standard and the test persons' classifications (cf. Supplement 2).

To measure and analyze the agreement between groups of raters, we used the concept of inter-rater reliabilities. This approach has proven to be appropriate for taxonomy evaluations as it measures the consensus between multiple classification results of the same papers [49,53]. There are several metrics for measuring inter-rater reliability. In our evaluation, we use two well-established metrics, the "joint probability of agreement" and a Kappa statistic [54,55], namely the Prevalence and Bias adjusted Kappa (PABAK) [55]. Both metrics are discussed in detail in Supplement 2. Completeness and usefulness were evaluated by the average questionnaire results on a 5-point Likert scale.

5.2. Results

We first discuss the results of the evaluation of reliability and understandability based on the classifications by the test persons. To measure inter-rater reliability among the (internal) raters, we calculated inter-rater reliability between rater and gold standard for each internal rater and aggregated the results across the four raters. Across all characteristics, we received an average joint probability of agreement of 0.909 and an average PABAK of 0.840, showing almost perfect agreement between (internal) raters according to established guidelines [56]. Similarly, we received a joint probability of agreement between external raters and the gold standard of 0.821 and a PABAK value of 0.691, indicating substantial agreement [56]. Moreover, we also compared the average results between the group of four researchers and the group of four practitioners among the eight external raters to assess whether the taxonomy could be equally understood and applied by both user groups. The group of researchers achieved an average PABAK of 0.708 with respect to the gold standard whereas the group of practitioners achieved 0.674, showing comparable performance of both groups (cf. Table 2).

The results show that the criterion of reliability is supported and that the descriptions of the taxonomy are understandable not only to the authors, but also for external persons not involved in the taxonomy building process. Naturally, the authors achieved higher agreement with the gold standard (which emerged as the result of the authors' classifications and discussion) in every dimension. For a more detailed analysis of the results, we now discuss the results for each (sub)dimension separately. For this analysis, we refer to the joint probability of agreement between raters and gold standard as these values are very intuitive and easy to interpret. The results for each (sub)dimension are shown in Table 3.

Agreement was high for all subdimensions of the dimension *DQ-related Problem Setting*, especially for the subdimension *DQ Dimension* with a strong agreement of 0.920 from the external raters and near-perfect agreement of 0.995 from the authors. For the dimension *Pipeline Position*, the agreement between external raters and the gold standard was relatively low at 0.736, which was due to some differences in the assessment of whether DQ defects were addressed in the output stage. More precisely, only mentioning the addressed DQ defects as the cause for improved model performances was sufficient for some raters to tag the output stage, while for others it was not. These differences did not exist among the authors, which explains their higher agreement. The subdimensions of the dimension *Properties of the Data* generally showed strong agreement across raters, particularly the first subdimension *Structure* and the last subdimension *Object*. For the subdimension *Bias Type*, the agreement was also generally high. Disagreements occurred in a few cases where the bias type of the used dataset was only implicit or the paper discussed all three possible bias types in general, although the data in the evaluation did not exhibit all three types. For the third subdimension *Availability* the average agreement was 0.847, with disagreements arising mostly because some raters knew whether specific datasets are publicly available and classified the paper accordingly, while others classified the availability as *not specified* because they had no further knowledge about the availability of the mentioned datasets. For *Extent*, the lower agreement of 0.667 was due to the fact that the presence of DQ defects in the test data was only implicit in some papers. In the gold standard, the characteristic *Test/Use* was selected in all of these cases, whereas some external raters only selected it when defects of the test data were explicitly mentioned. A dimension with lower agreement was *Identification/Assessment of DQ Defects*, with results of 0.653 among external raters and 0.724 among authors. In many papers, this information was only implicit or mentioned very briefly which may have led to oversights and subsequent classifications as *not specified*. Furthermore, in some cases it was not entirely clear to the raters whether the DQ defects described in a paper were existing defects identified by measurements, or whether the defects were simply assumed. The subdimensions of *DQ Defect Resolution Method* showed varying levels of agreement. The agreement on the subdimension *Human Involvement* was very high while the subdimension *Statistical Methods* showed a lower value of 0.708 as methods such as linear regression or k-means were considered statistical by some raters and ML methods by others. The agreement on *ML Methods* was generally high, especially considering that this subdimension encompassed several non-exclusive characteristics, with many proposed methods belonging to multiple characteristics. Disagreements occurred in cases where some raters found the distinction between ML methods used for DQ defect improvement and ML methods used for the ML-based data analysis in the application context to be not entirely clear. For *Other Method*, the external raters achieved a lower agreement of 0.604 compared to the authors' agreement of 0.836, because individual external raters sometimes classified uncommon approaches (e.g., graph learning techniques) they were unfamiliar with as *Other Method* whereas other raters with more knowledge of these particular approaches recognized them as statistical or ML methods. Both subdimensions of the dimension *Type of ML Task* showed high agreement between the external raters. Agreement on *Evaluation Type* and *Competing Methods* was also high, with almost all raters agreeing that the majority of papers featured a quantitative evaluation. The lowest agreement between external raters and the gold standard was recorded in the final subdimension *Pipeline Stage of Evaluation*. In the taxonomy, an evaluation of the ML model performance in Stage 2 was defined as an evaluation in the model stage. An evaluation in the output stage was defined as an evaluation that explicitly used DQ or uncertainty/confidence measures for the model output, or an evaluation of an output-based method of addressing the DQ defects. However, this definition partly differs from the also intuitive understanding that any evaluation of model performance is an evaluation in the output stage. Therefore, four external raters classified the evaluation of ML model performance as an output stage evaluation, contrary to the definition used by the authors, leading to a lower agreement. This difference of

Table 2
Aggregated inter-rater reliability between gold standard and raters.

	Internal raters	External raters		
	Authors	All external raters	Practitioners	Researchers
Joint probability of agreement	0.909	0.821	0.816	0.825
PABAK	0.840	0.691	0.674	0.708

Table 3
Detailed inter-rater reliability between raters and gold standard.

Dimension	Average joint probability of agreement with external raters	Average joint probability of agreement with authors
DQ Dimension	0.920	0.995
DQ-related Task	0.880	0.938
Uncertainty Modeling/Quantification	0.867	0.938
Pipeline Position	0.736	0.906
Structure	0.938	0.948
Bias Type	0.854	0.934
Availability	0.847	0.896
Extent	0.667	0.828
Object	0.880	0.910
Identification/Assessment of DQ Defects	0.653	0.724
Human Involvement	0.931	0.969
Statistical Methods	0.708	0.773
ML Methods	0.861	0.908
Other Method	0.604	0.836
Purpose	0.901	0.969
Type of ML Model	0.927	0.953
Evaluation Type	0.965	0.969
Competing Methods	0.917	0.969
Pipeline Stage of Evaluation	0.541	0.901

Table 4
Average questionnaire results on likert scale (lower is better).

Question	All test persons	Researchers	Practitioners
Q1: Does the taxonomy include **all relevant dimensions and characteristics** to classify the six papers and is the taxonomy in your opinion complete in general?	2.17	2.00	2.25
Q2: Are the **dimensions and characteristics detailed enough** to allow differentiation between similar objects in the six papers and in general?	1.67	1.50	1.75
Q3: Is the **taxonomy useful** for the structuring and classification of the ML methods in the six papers and for ML methods on data with DQ defects in general?	1.17	1.50	1.00
Q4: Are **all dimensions and characteristics necessary** to classify the six papers and to classify ML methods on data with DQ defects in general?	1.50	1.00	1.75

interpretation did not occur among the authors, resulting in a much higher agreement of 0.901 with the gold standard. In summary, the reliability and understandability of the taxonomy was supported for both internal and external raters across all dimensions. Lower agreement between raters on certain dimensions could be attributed to the raters' differing background knowledge about particular aspects and methods discussed in some of the more specialized papers.

We now discuss the results of the questionnaire-based evaluation of the taxonomy regarding completeness und usefulness. Each external rater received a questionnaire with four questions, where answers on a 5-point Likert scale (1 = strongly agree, 5 = strongly disagree) were possible (cf. Supplement 2). In addition, the external raters were asked to make suggestions for taxonomy updates, especially if they did not agree that the taxonomy fulfilled one of the criteria. The results are shown in Table 4.

Almost all of the external raters agreed that the taxonomy is complete (cf. Questions 1 and 2), with the exception of one of the practitioners. However, this practitioner suggested that the dimensions *Reason for DQ Defect* and *Goal of Real-World ML Application* were missing from the taxonomy. The first suggested dimension is out of scope by the definition of our taxonomy (cf. Section 2.1 and Section 3) as it relates to the data collection process before the input stage. The latter dimension is also out of scope, as it relates to the application of the results beyond the output stage (cf. Section 2.1 and Section 3). Importantly, all external raters consistently agreed that the taxonomy is useful (cf. Question 3). The external raters also agreed that all dimensions and characteristics were necessary (cf. Question 4), with one practitioner being neutral on this question, reasoning that he was only able to assess this for the six reviewed papers but not in general. Since the practitioner did not suggest removing any dimensions or characteristics and agreed that other papers may indeed discuss and require all dimensions, we decided not to update the taxonomy based on this response. Finally, we note that both researchers and practitioners agree that the taxonomy satisfies the evaluated criteria of completeness and usefulness.

6. Discussion

In this section, we discuss the results and contributions of our work and provide guidance for future research in the field of ML on data with DQ defects. We proceed as follows: We first discuss our evaluation results. Then, we outline the contributions of our taxonomy to both research and practice by discussing potential applications of the taxonomy in both areas. We also reflect how future research in the field of ML on data with DQ defects could potentially influence our taxonomy. Finally, we identify research gaps and potential for future research based on our taxonomy and its building process.

The evaluation results support the reliability, understandability, completeness, and usefulness of our taxonomy. Both researchers and practitioners were able to apply the taxonomy to various papers and achieve substantial inter-rater agreement. While most taxonomy dimensions showed very strong agreement between raters, a few characteristics of certain dimensions were more difficult to distinguish, especially for external raters, resulting in slightly lower inter-rater agreement for these dimensions. The dimension *Identification/Assessment of DQ Defects* proved to be particularly challenging, as raters were asked to decide whether DQ defects in a paper were artificially introduced or assumed, or whether defects were naturally present. Often, the identification of such (natural) DQ defects poses a challenge in itself, for which DQ metrics [57] (initially mentioned in the description of *Natural Presence*) potentially offer a solution. However, using DQ metrics to indicate potential DQ defects is often part of the method addressing the defects itself. Thus, mentioning DQ metrics in this description could lead raters to incorrectly classify a paper without natural presence of DQ defects into this characteristic simply because DQ metrics are used for addressing the defects. Therefore, we updated the taxonomy by slightly revising the initial description of this characteristic for our final description of the taxonomy in this work. While the subdimensions *Other Method* and *Pipeline Stage of Evaluation* also proved to be challenging for some external raters, we found that the lower agreement among external raters was due to different knowledge about the methods used in the papers and the subdimensions are thus generally comprehensible. All other dimensions and characteristics, especially the central dimensions *DQ-related Problem Setting, DQ Defect Resolution Method* (particularly *ML Methods*), and *Type of ML Task* satisfied the criterion of very high reliability. Overall, the taxonomy has shown to be reliable and understandable to users from both research and practice. In addition, both researchers and practitioners agreed that the taxonomy is complete and useful. We therefore propose to not further adapt its dimensions, characteristics, and their descriptions.

Our taxonomy provides a clear and intuitive framework for ML on data with DQ defects, which is a highly relevant area in both research and practice. Researchers developing new methods in this field can use our taxonomy to position their work within existing literature and can exactly define the research gaps they aim to address. Moreover, the dimensions and characteristics of our taxonomy allow researchers to identify additional relevant information that should be provided in their work, for example, clear and comprehensive descriptions of the analyzed data and its DQ defects. Similarly, the dimensions of our taxonomy support researchers in describing the DQ-related problem that they aim to address with their method (e.g., the extent of DQ defects from training to test data which must be considered in the design of any method), thereby providing guidance to improve their work. Moreover, possible adaptations of methods to related problems can be identified in this manner, thus broadening their contribution to research. For instance, a method for addressing DQ defects by modifying the optimization criterion in NNs may also be suited for other types of ML models in which similar optimization criteria are used. For practitioners, the taxonomy allows them to structure, search, and filter existing methods dealing with ML on data with DQ defects according to their individual needs and preferences. In particular, an overview of available methods for specific use cases can be provided and the most suitable option can be selected, considering the exact DQ defect(s), the data properties, and requirements for the ML-based analysis. The taxonomy dimensions can also help practitioners to identify all relevant dimensions that need to be considered when working with ML on data with DQ defects.

In the future, research in the field of ML on data with DQ defects may result in new ML models and new methods for addressing DQ defects. Our taxonomy will remain implicitly valid without adjustments as any

new models and methods can be classified into the characteristics "other". However, it may be more informative for users of the taxonomy to add new characteristics for models and techniques becoming more relevant in the future, thus making them explicit. Such updates will be possible iteratively without the need for major changes or a new taxonomy building process. Therefore, we argue that our taxonomy can not only support researchers and practitioners in their current work but can also be iteratively adapted, ensuring continued support as the research field advances.

Furthermore, while we do not aim to present an exhaustive, fully comprehensive overview or survey that describes all existing literature in the field of ML on data with DQ defects in detail, the taxonomy and the literature that was researched and classified during the taxonomy building process (cf. Supplement 1) has shed light on current research foci and research gaps. In the following, we will discuss six key findings which we identified as underrepresented and offer potential for future research directions.

First, the taxonomy building process demonstrated that many papers in the field of ML on data with DQ defects lack a conceptional analysis of the DQ defects and provide little detail on them. The DQ defects are often discussed very briefly, and relevant dimensions such as the *Extent* and the *Bias Type* are not mentioned at all and ignored, despite being essential for the functionality of a method. Of the 150 analyzed papers (cf. Iteration 2 and Evaluation), only 36 explicitly include the bias type of the analyzed data. Furthermore, the identification of DQ defects is often not discussed in detail, even though it is highly relevant to practical applicability whether the defects occur naturally or whether they are artificially introduced. Of the 150 analyzed papers, 48 do not specify how the DQ defects were identified. For datasets with natural presence of DQ defects, it is also relevant how they are identified and what criteria (e.g., the value of a DQ metric exceeding a certain threshold) are used to decide whether a DQ defect is present. When DQ defects are introduced artificially, the specific mechanism is critical as it can introduce different biases and patterns, thus making some methods unsuitable. For example, if data uncertainty is introduced by imputation of randomly deleted values [58], biases of the imputation method (e.g., using predictions of an ML model for imputation that is biased towards the majority class in its training data) can change the distribution of feature values (compared to the original, deleted values), thereby adding significant bias to the data. The taxonomy building process also showed that the DQ defects and the resulting data uncertainty are rarely measured or quantified for a more detailed analysis, e.g., by explicitly modeling uncertain values probabilistically and analyzing the properties of the probability distribution.

Second, DQ defects and uncertainty in the training or test data of an ML model may lead to significant uncertainty in its predictions. Not only initial data uncertainty, but also the resulting predictive uncertainty is often not analyzed in detail. Quantifying and analyzing predictive uncertainty are crucial for the responsible use of ML models and valid risk assessment in ML-assisted decision-making. Otherwise, a user may become overconfident in an ML model's decisions and blindly trust them, ignoring any potential risks which could be accurately assessed by quantifying predictive uncertainty. However, most papers studied during the taxonomy building process address DQ defects either at the input or the model stage, but rarely in the output stage where predictive uncertainty is typically analyzed. Of the 150 analyzed papers, only five address DQ defects and the resulting uncertainty in the output stage. Only some papers on BNNs do discuss predictive uncertainty, but they usually focus on model uncertainty—i.e., the uncertainty of the trained model weights—as its primary cause, while mostly ignoring data uncertainty (with the exception of a few selected papers, e.g., [59]).

The taxonomy building process also revealed that research in ML on data with DQ defects often focuses on certain types of data and DQ defects, while others are underrepresented. For DQ defects, the majority of papers addresses the DQ dimensions completeness and accuracy, whereas other important dimensions such as consistency and currency are rarely discussed, even though their relevance is highlighted in a few studies [38]. Addressing DQ defects from these dimensions in the training and use of ML methods should thus be studied carefully in future research. It may also be promising to analyze if methods developed for incomplete or inaccurate data can be adapted for data suffering from defects related to other DQ dimensions. Regarding the type of the analyzed data, the most commonly addressed type is structured data (121 of the 150 papers). With respect to unstructured data, image data is focused on most often, whereas other types are rarely studied in the context of ML analysis with DQ defects. However, it can be expected that the widespread use and rapid progress of large language models will lead to a significantly greater focus on investigating DQ defects in textual data in the future.

Another finding is that existing research focuses on specific types of methods to address DQ defects. While addressing DQ defects using statistical methods or ML methods is very common, methods based on human involvement such as active learning are discussed in only seven of the 150 identified papers. Active learning methods have been recognized in literature as useful approaches to improve DQ by including human decisions on strategically selected data instances (to avoid exhaustive workloads for human users), thereby improving model performance compared to fully automated models [38]. Another type of method which is discussed in 16 of the 150 papers is the modification of ML models for data with explicitly modeled data uncertainty, e.g., by probability distributions. While most research focuses on addressing DQ defects or data uncertainty at the input stage, it has also proved promising to adapt ML models to handle probability distributions representing uncertain data instances (e.g., [15]). Such methods require changes to the ML model design but allow explicit quantification of the predictive uncertainty in the model output by probability distributions. Therefore, they enable mathematically sound confidence estimations, making them a promising focus of future research.

The taxonomy building process also revealed several issues regarding the evaluation of the proposed methods. Firstly, evaluations often do not address the (sub)dimension *Extent* of the data with DQ defects and its potential impact on the evaluation. In particular, when methods are trained on data with DQ defects, we observed many papers in which it is not discussed explicitly whether the test data are actually affected by the same or similar DQ defects. In some cases, test data is even assumed to be free of DQ defects even though the training data suffer from them. The existence of test data without DQ defects in such cases is either an unrealistic setting or the result of non-random test data selection (e.g., only data without missing values is selected as test data), which often introduces biases into the test data. Therefore, the extent of the uncertain data and its potential impact on the evaluation should be determined and analyzed. Secondly, evaluations are often restricted to an evaluation of the ML model performance. However, if DQ defects are addressed in the input stage, the respective results should also be directly evaluated before training and using the ML model, as the performance of the DQ defect improvement method is only evaluated implicitly by measuring the ML model performance. While 98 of the 150 identified papers address DQ defects in the input stage, only 14 of these 98 papers also perform an evaluation in the input stage. For such an evaluation in the input stage, suitable metrics for measuring whether the DQ defects were successfully addressed are required. For some types of DQ defects such metrics already exist, for example, imputation performance on artificially introduced missing data can be measured by comparing the imputed values with the known original values. However, depending on the aspects of the DQ defects being addressed in the input stage, suitable metrics still need to be developed by future research.

Finally, to perform thorough, representative, and reproducible evaluations of proposed methods for ML on data with DQ defects, it is essential to have benchmark datasets containing specific and known DQ defects. For example, such datasets could contain erroneous feature values together with the information of which feature values are

erroneous, or they could represent known data uncertainty by probability distributions. With such benchmark datasets available, authors do not have to resort to assuming or artificially introducing DQ defects into datasets without known defects. Furthermore, benchmark datasets allow different researchers to evaluate their proposed methods on a common basis, thus enabling standardized evaluations and improving comparability between the performances of different methods. In addition, the existence of benchmark datasets could also encourage the study of underrepresented data types or DQ defect types. As a positive example for the task of duplicate detection, Naumann [60] created such benchmark datasets that include ground truth information about the actual duplicates in the data, providing a valuable contribution to duplicate detection research. However, for most types of data and DQ defects, such benchmark datasets including information about DQ defects/data uncertainty are hardly available. Although platforms such as the UCI ML Repository[2] or Kaggle[3] do provide benchmark datasets with DQ defects (e.g., missing or inaccurate values), those datasets neither include a ground truth for DQ defects nor quantify resulting data uncertainty, e.g., through probability distributions for inaccurate values. This leads researchers to use their own methods to impute missing values or to generate probability distributions representing data uncertainty rather than relying on standardized benchmarks. Creating and providing benchmark datasets for ML on data with DQ defects should therefore be a primary goal for future research in this field.

7. Conclusion

DQ defects pose a critical challenge that must be addressed in ML-based data analysis as they can severely affect the validity and performance of ML models and the (human) decisions based on them, as recent examples highlight (cf. Introduction). To structure and systematize this field for both researchers and practitioners, we created a taxonomy for the field of ML on data with DQ defects covering relevant dimensions of both DQ defects and methods that aim to address these defects in both training and use of ML models. Based on this taxonomy, we identified existing gaps and challenges in this research field. Despite the fact that we developed and evaluated the taxonomy according to the ETDP [19], the resulting taxonomy is not without limitations. First of all, not all characteristics of each dimension can be arbitrarily combined with all characteristics of the other dimensions, i.e., the dimensions are not perfectly orthogonal. For example, the use of uncertainty-aware ML models such as BNNs to address DQ defects already implies that probabilistic uncertainty modeling was employed. Although this represents a typical limitation of taxonomies, we have taken careful measures to ensure that most characteristics across different dimensions are independent and can be combined in papers. A further limitation to acknowledge is that we only considered the process of ML-based data analysis from the input data to the model output stage, and hence deliberately excluded the human user who receives the model output and derives decisions based on it. Similarly, we do not address details of the data collection process that may initially cause the DQ defects in the input data. Examinations and systematizations of these stages may be subjects of future research.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (grant number 494840328).

CRediT authorship contribution statement

Michael Hagn: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Bernd Heinrich:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Thomas Krapf:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Alexander Schiller:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2025.114493.

Data availability

The data used for the research is available in the Supplementary Material.

References

[1] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis,, Lancet Digit. Health 4 (2022) e507–e519.
[2] M. El Hajj, J. Hammoud, Unveiling the influence of artificial intelligence and machine learning on financial markets: a comprehensive analysis of AI applications in trading, risk management, and financial operations, J. Risk Financ. Manag. 16 (2023) 434.
[3] M. Möhlmann, L. Zalmanson, O. Henfridsson, R.W. Gregory, Algorithmic management of work on online labor platforms: when matching meets control, MISQ 45 (2021) 1999–2022.
[4] Y. Gong, G. Liu, Y. Xue, R. Li, L. Meng, A survey on dataset quality in machine learning, Inf. Softw. Technol. 162 (2023) 107268.
[5] N. Polyzotis, S. Roy, S.E. Whang, M. Zinkevich, Data lifecycle challenges in production machine learning, SIGMOD Rec. 47 (2018) 17–28.
[6] P. Krasikov, C. Legner, Introducing a data perspective to sustainability: how companies develop data sourcing practices for sustainability initiatives, Commun. Assoc. Inf. Syst. 53 (2023) 162–188.
[7] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch, The effects of data quality on machine learning performance on tabular data, Inf. Syst. 102549 (2025).
[8] Sama, Top 3 Examples of How Poor Data Quality Impacts ML. https://www.sama.com/blog/top-3-examples-of-how-poor-data-quality-impacts-ml, 2024 (accessed 10-30-2024).
[9] P.D. Wentzell, S. Hou, Exploratory data analysis with noisy measurements, J. Chemom. 26 (2012) 264–281.
[10] Y. Li, T. Bao, H. Chen, K. Zhang, X. Shu, Z. Chen, Y. Hu, A large-scale sensor missing data imputation framework for dams using deep learning and transfer learning strategy, Meas. 178 (2021) 109377.
[11] J. Peng, J. Hahn, K.-W. Huang, Handling missing values in information systems research: a review of methods and assumptions, Inf. Syst. Res. 34 (2023) 5–26.
[12] B. Heinrich, M. Klier, Assessing data currency — a probabilistic approach, J. Inf. Sci. 37 (2011) 86–100.
[13] Y. Zak, A. Even, Development and evaluation of a continuous-time Markov chain model for detecting and handling data currency declines, Decis. Support Syst. 103 (2017) 82–93.
[14] H. Zhang, T. Nakamura, T. Isohara, K. Sakurai, A review on machine unlearning, SN Comput. Sci. 4 (2023).
[15] T. Krapf, M. Hagn, P. Miethaner, A. Schiller, L. Luttner, B. Heinrich, Piecewise linear transformation – Propagating aleatoric uncertainty in neural networks, in: Proc. of the AAAI Conf. on Artif. Intell 38, 2024, pp. 20456–20464.
[16] M. Pattanodom, N. Iam-On, T. Boongoen, Clustering data with the presence of missing values by ensemble approach, in: Second Asian Conf. on Defence Technol, 2016, pp. 151–156.
[17] S. Das, S. Datta, B.B. Chaudhuri, Handling data irregularities in classification: foundations, trends, and future challenges, Pattern Recogn. 81 (2018) 674–693.
[18] N. Li, Y. Qi, C. Li, Z. Zhao, Active learning for data quality control: a survey, J. Data Inf. Qual. 16 (2024) 1–45.
[19] D. Kundisch, J. Muntermann, A.M. Oberländer, D. Rau, M. Röglinger, T. Schoormann, D. Szopinski, An update for taxonomy designers, Bus. Inf. Syst. Eng. 64 (2022) 421–439.

[2] https://archive.ics.uci.edu/datasets
[3] https://www.kaggle.com/datasets

[20] R. Price, G. Shanks, A semiotic information quality framework: development and comparative analysis, J. Inf. Technol. 20 (2005) 88–102.
[21] Y. Wand, R.Y. Wang, Anchoring data quality dimensions in ontological foundations, Commun. ACM 39 (1996) 86–95.
[22] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. 41 (2009) 1–52.
[23] C. Cichy, S. Rass, An overview of data quality frameworks, IEEE Access 7 (2019) 24634–24648.
[24] B. Heinrich, M. Klier, A. Schiller, G. Wagner, Assessing data quality – a probability-based metric for semantic consistency, Decis. Support Syst. 110 (2018) 95–106.
[25] S. Krishnan, J. Wang, E. Wu, M.J. Franklin, K. Goldberg, ActiveClean, Proc. VLDB Endow. 9 (2016) 948–959.
[26] Y.M. Herrera, D. Kapur, Improving data quality: actors, incentives, and capabilities, Polit. Anal. 15 (2007) 365–386.
[27] D. Peng, F. Wu, G. Chen, Data quality guided incentive mechanism design for crowdsensing, IEEE Trans. Mobile Comput. 17 (2018) 307–319.
[28] C.G. Northcutt, L. Jiang, I.L. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2021) 1373–1411.
[29] I. Valentim, N. Lourenco, N. Antunes, The impact of data preparation on the fairness of software systems, in: IEEE 30th Int. Symp. on Softw. Reliab. Eng, 2019, pp. 391–401.
[30] A. Barry, L. Han, G. Demartini, On the impact of data quality on image classification fairness, in: Proc. of the 46th Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Retr, 2023, pp. 2225–2229.
[31] S. Guha, F.A. Khan, J. Stoyanovich, S. Schelter, Automated data cleaning can hurt fairness in machine learning-based decision making, IEEE Trans. Knowl. Data Eng. 36 (2024) 7368–7379.
[32] B. Heinrich, T. Krapf, P. Miethaner, EXPLORE: A novel method for local explanations, in: Proc. of the Int. Conf. on Inf. Syst. (ICIS), 2024.
[33] Z. Zhang, G. Chen, Y. Xu, L. Huang, C. Zhang, S. Xiao, FedDQA: a novel regularization-based deep learning method for data quality assessment in federated learning, Decis. Support Syst. 180 (2024) 114183.
[34] A. L'Heureux, K. Grolinger, H.F. Elyamany, M.A.M. Capretz, Machine learning with big data: challenges and approaches, IEEE Access 5 (2017) 7776–7797.
[35] A. Paleyes, R.-G. Urma, N.D. Lawrence, Challenges in deploying machine learning: a survey of case studies, ACM Comput. Surv. 55 (2023) 1–29.
[36] G. Rekha, A.K. Tyagi, V.K. Reddy, A wide scale classification of class imbalance problem and its solutions: a systematic literature review, J. Comput. Sci. 15 (2019) 886–929.
[37] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.
[38] M. Priestley, F. O'donnell, E. Simperl, A survey of data quality requirements that matter in ML development pipelines, J. Data Inf. Qual. 15 (2023) 1–39.
[39] V.N. Gudivada, A. Apon, J. Ding, Data quality considerations for big data and machine learning: going beyond data cleaning and transformations, Int. J. Adv. Softw. (2017) 1–20.
[40] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X.X. Zhu, A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (2023) 1513–1589.
[41] F. Wang, Y. Liu, K. Liu, Y. Wang, S. Medya, P.S. Yu, Uncertainty in Graph Neural Networks: A Survey, Preprint arXiv:2403.07185, 2024.
[42] R.C. Nickerson, U. Varshney, J. Muntermann, A method for taxonomy development and its application in information systems, Eur. J. Inf. Syst. 22 (2013) 336–359.
[43] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019).
[44] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: Int. Conf. on Learning Represent, 2017.
[45] F.Z. Poleto, J.M. Singer, C.D. Paulino, Missing data mechanisms and their implications on the analysis of categorical data, Stat. Comput. 21 (2011) 31–43.
[46] M.C. Tremblay, K. Dutta, D. Vandermeer, Using data mining techniques to discover bias patterns in missing data, J. Data Inf. Qual. 2 (2010) 1–19.
[47] A.Y. Nahm, S.S. Rao, L.E. Solis-Galvan, T.S. Ragu-Nathan, The Q-sort method: assessing reliability and construct validity of questionnaire items at a pre-testing stage, J. Mod. Appl. Stat. Methods 1 (2002) 114–125.
[48] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: a comparison of rule-based and example-based explanations, Artif. Intell. 291 (2021) 103404.
[49] D. Szopinski, T. Schoormann, D. Kundisch, Because your taxonomy is worth it: towards a framework for taxonomy evaluation, in: Proc. of the 27th Eur. Conf. on Inf. Syst. (ECIS), 2019.
[50] H. Gimpel, D. Rau, M. Röglinger, Understanding FinTech start-ups – a taxonomy of consumer-oriented service offerings, Electron. Mark. 28 (2018) 245–264.
[51] K. Brosnan, B. Grün, S. Dolnicar, Cognitive load reduction strategies in questionnaire design, Int. J. Mark. Res. 63 (2021) 125–133.
[52] H. Sharma, How short or long should be a questionnaire for any research? Researchers dilemma in deciding the appropriate questionnaire length, Saudi J. Anaesth. 16 (2022) 65–68.
[53] K.L. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters, Advanced Analytics, LLC, 2014.
[54] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[55] T. Byrt, J. Bishop, J.B. Carlin, Bias, prevalence and kappa, J. Clin. Epidemiol. 46 (1993) 423–429.
[56] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biom. 33 (1977) 159–174.
[57] B. Heinrich, D. Hristova, M. Klier, A. Schiller, M. Szubartowicz, Requirements for data quality metrics, J. Data Inf. Qual. 9 (2017) 1–32.
[58] C. Cappiello, F. Cerutti, C. Sancricca, R. Zanelli, About the effects of data imputation techniques on ML uncertainty, in: Joint Workshops at 49th Int. Conf. on Very Large Data, 2025.
[59] W.A. Wright, Bayesian approach to neural-network modeling with input uncertainty, IEEE Trans. Neural Netw. 10 (1999) 1261–1270.
[60] U. Draisbach, F. Naumann, DuDe: the duplicate detection toolkit, in: Proc. Int. Workshop on Qual. in Databases, 2010.

Michael Hagn is a Research Assistant at the Department of Information Systems at the University of Regensburg, Germany. His research focuses on Data Quality, Uncertainty in Machine Learning, and Explainable Artificial Intelligence. He has published his work in conferences such as the AAAI Conference on Artificial Intelligence.

Bernd Heinrich is Chair Professor of Information Systems at the Faculty of Computer Science and Data Science of the University of Regensburg, Germany. His research interests include Data Quality, Explainable Artificial Intelligence, Machine Learning, Process Planning, and Context-Aware Services with a methodological emphasis on mathematical and analytical modeling. Bernd Heinrich has published his research in various journals, including MIS Quarterly, Decision Support Systems, Business & Information Systems Engineering, Electronic Markets, Information Systems Frontiers, INFORMS Service Science and Journal of Decision Systems. He has also presented his work at leading international conferences such as the International Conference on Information Systems and the European Conference on Information Systems.

Thomas Krapf is a Research Assistant at the Department of Information Systems at the University of Regensburg, Germany. His research interests include the topics Data Quality, Data Uncertainty in Machine Learning, and Explainable Artificial Intelligence. He has published his works in conferences such as the AAAI Conference on Artificial Intelligence and the International Conference on Information Systems.

Alexander Schiller is a Postdoctoral Researcher and Research Group Leader in Information Systems at the Faculty of Computer Science and Data Science at the University of Regensburg, Germany. His research interests include Data and Business Analytics, Data Quality, Explainable Artificial Intelligence, Machine Learning as well as Business Process Management. His research has been published in journals such as MIS Quarterly, Decision Support Systems, Electronic Markets, Information Systems Frontiers, and Journal of Decision Systems. Moreover, he has presented the results of his work at international conferences such as the International Conference on Information Systems, the European Conference of Information Systems and the AAAI Conference on Artificial Intelligence.