Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality

Gilberto Recupito
Sesa Lab - University of Salerno
Salerno, Italy
grecupito@unisa.it

Raimondo Rapacciuolo
University of Salerno
Salerno, Italy
r.rapacciuolo1@studenti.unisa.it

Dario Di Nucci
Sesa Lab - University of Salerno
Salerno, Italy
ddinucci@unisa.it

Fabio Palomba
Sesa Lab - University of Salerno
Salerno, Italy
fpalomba@unisa.it

ABSTRACT

Artificial Intelligence (AI) is rapidly advancing with a data-centered approach suitable for various domains. Nevertheless, AI faces significant challenges, particularly in data quality. Data collection from diverse sources can introduce quality issues that may threaten the development of AI-enabled systems. A growing concern in this context is the emergence of *data smells* – issues specific to the data used in building AI models, which can have long-term consequences. In this paper, we aim at enlarging the current body of knowledge on data smells, by proposing a two-step investigation into the matter. First, we updated an existing literature review in an effort of cataloguing the currently existing data smells and the tools to detect them. Afterward, we assess the prevalence of data smells and their correlation with data quality metrics. We identify a novel set composed of 12 data smells distributed across three additional categories. Secondly, we observe that the correlation between data smells and data quality is notably impactful, exhibiting a pronounced and substantial effect, especially in highly diffused data smell instances. This research sheds light on the complex relationship between data smells and data quality, providing valuable insights into the challenges of maintaining AI-enabled systems.

CCS CONCEPTS

• **Software and its engineering → Software maintenance tools**.

KEYWORDS

AI Technical Debt; Data Smells; Data Quality; Software Engineering for Artificial Intelligence, Empirical Software Engineering.

ACM Reference Format:
Gilberto Recupito, Raimondo Rapacciuolo, Dario Di Nucci, and Fabio Palomba. 2024. Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality. In *Conference on AI Engineering Software Engineering for AI (CAIN 2024), April 14–15, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3644815.3644960

CAIN 2024, April 14–15, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0591-5/24/04.
https://doi.org/10.1145/3644815.3644960

1 INTRODUCTION

Artificial Intelligence (AI) is more and more diffused nowadays, being used by individuals and companies to make informed decisions [54] and automate tasks that would typically done by humans [37]. Indeed, AI-intensive systems, i.e., systems that embed artificial intelligence models and algorithms, have been recently deployed in multiple domains, with some recent applications showing highly efficient and accurate performance [30, 33].

However, the development of artificial intelligence-enabled systems differs from other types of software because the program and its effectiveness in solving a specific task heavily rely on the data and observations used to train models [2].

More specifically, AI-enabled systems are defined as *a system consisting of various software components, out of which at least one is an AI-specific component* [29]. In this context, data represents the primary source of producing business-oriented AI-enabled systems. Performing data analysis and validation is a crucial initial step for designers of machine learning components. Failure to properly analyze the training data may lead to model degradation [26]. Therefore, it is crucial to prioritize data quality to build a reliable and effective AI-enabled system. Data quality issues can arise for various reasons, *e.g.*, data entry errors, inadequate data cleaning, or bias in the data. Addressing these issues requires a well-defined data quality management strategy that includes profiling, cleansing, and data enrichment [32]. While different tools and practices are available to support feature engineering and data transformation for managing AI pipelines [38], the need to improve the practices related to quality assurance is continuously increasing [9].

Data quality degradation could also lead to technical debt for the whole system [12]. Data debt, primarily when introduced by data quality issues or data anomalies, can strongly impact AI-enabled systems, degrading model performance and causing problems to all the subsequent phases involved in the pipeline [1].

In analyzing technical debt specific to data, *data smells* are represented using the analogy of code smells. As code smells are defined as *symptoms of poor design and implementation choices* [14], data smells are *data value-based indications of latent data quality issues caused by poor practices that may lead to problems in the future* [13]. While other types of data quality issues are investigated in research to be addressed [16, 18, 27, 49], the knowledge about the effect and the presence of data smells is still limited to the catalog provided by Foidl et al. [13].

This limitation confines the possibility of spending research efforts to increase the data quality management process of the system. Since the actual state of the definition of data smells is preliminary, it has been difficult to conduct studies to find management strategies for this type of data quality issue. Increasing the knowledge base of data smells, this study aims to provide researchers with a better understanding of how data smells can be identified and managed. Therefore, this study first explores the novel literature to update and extend the catalog of data smells. Subsequently, through empirical methods, analyzed the prevalence and impact of data smells in 19 real-world datasets on data quality aspects. In particular, we updated the literature review and introduced three new types of data smells (*i.e.*, *Redundant Value Smells*, *Distribution Smells*, and *Miscellaneous Smells*) and 12 new data smells to the existing catalog. We identify data smells in the most frequently used dataset and examine their relationship with data quality.

To sum up, our work provides the following main contributions:

- An updated catalog of data smells, which advances the state of the art by providing the definition of novel data smells, along with the information on how to detect them;
- Results of an empirical investigation into the prevalence and impact of data smells on data quality metrics, which can be used by researchers and practitioners to understand how data smells may impact the training of AI solutions;
- A publicly available replication package [39], which includes the whole set of data and scripts used in the study and that can be used by researchers to build on top of our findings.

Structure of the paper. The article continues with the following sections: Section 2 provides information about the state of data smells and discusses related work. Section 3 explains the main goal and the overall method to address our research questions. Section 4 details the method and the results obtained to produce the new catalog of data smells. Section 5 details the method and the results of the empirical analysis for the correlation of data smells and data quality. Section 6 discusses the threats faced and mitigated in this study. Section 7 outlines the implications and discusses the study results. Section 8 conclude the article with the key findings.

2 BACKGROUND AND RELATED WORK

In this section, we first provide a background on AI technical debt, discussing the main research advances in the area. Secondly, we elaborate on the related works, explaining how our study compares and advances them.

2.1 Technical Debt

The term "technical debt" was coined by Cunningham [7] as a metaphor for describing the adoption of suboptimal solutions to achieve short-term benefits, which are expected to be repaid with greater costs in the medium or long term. Similar to financial debt, technical debt can pose serious challenges to the maintainability of a system, leading to increased maintenance costs and reduced quality [23, 42]. The concept of technical debt has been explored through various studies. Tom et al. [47] discussed the different types of granularity that highlight the level of the effect of each technical debt, including code debt, architectural debt, environmental debt, knowledge distribution debt, and testing debt. Li et al. [24] extended the classification to include requirements and infrastructure debt. Specifically for code debt, code smells are symptoms of poor design and implementation choices that can significantly impact the maintainability of a software system [14]. Structural metrics and historical metrics were explored, resulting in the proposal of various tools and methodologies [28, 31, 34, 35, 48]. Therefore, the research effort invested in addressing technical debt increases the overall quality of software systems and allow practitioners to automatically detect code smells to guarantee high system's quality.

In traditional software systems the effort to investigate technical debt has been enough to align several benefits that practitioners could use to increase the system's overall quality. With the rise of AI-enabled systems, exploring technical debt opens new challenges. Sculley et al. [41] brings the definition of technical debt inside AI-enabled systems, highlighting all the potential issues that can arise when using machine learning models. For instance, changes to the data distribution over time can also lead to the unpredictable impact to the whole system, defined as *CACE principle* (Changing Anything Changes Everything). The significant contribution that Sculley et al. [41] give to define technical debt in AI-enabled systems, opens the road to conduct several studies. Tang et al. [45] carried out an empirical study that analyzed 26 machine learning (ML) projects. They identified seven new ML-specific technical debt types, focusing on aspects related to the model code. Their findings shed light on unique challenges and debt types specific to ML projects, helping researchers and practitioners understand the nature of technical debt in ML projects. In a recent study, Bogner et al. [1] analyzed technical debt in AI-enabled systems using a systematic mapping study. The analysis identified 72 antipatterns, most related to data and model debt. The study's findings offer valuable insights into the specific areas requiring attention to manage technical debt in AI-enabled systems effectively.

Several studies explored the definition of data debt in AI-enabled systems. Sculley et al. [41] defined the concept of technical debt in the context of the data used for building AI models. In detail, they explored the dependencies of the data, warning for the dependencies which changes could provoke unpredictable consequences to the whole system (*Unstable data dependencies*) and the underutilized data dependencies in an AI pipeline. Bogner et al. [1] subsequently conducted a systematic mapping study to explore the types of data smells. Data debt is the most recurrent type of AI-specific debt of all the types of technical debt explored for AI-enabled systems. Munappy et al. [32] employed a case study to explore data management issues in Deep-Learning systems. They discovered as main issues related to the data structure the critical challenges related to the deduplication of the data and management of heterogeneous data in terms of encoding and format. Bosu and MacDonell [3] conducted a systematic literature review of data quality research in empirical software engineering. They reported that only a few studies (23) considered the three essential activities related to data quality management (data collection reporting, data pre-processing, and data quality issues). Yoon and Doo [52] evaluated six different techniques to face outliers anomalies in the context of software project data, discovering that data cleaning techniques on artificial data sets are a considerable solution for this type of data quality issue. Liebchen and Shepperd [25] updated the conducted literature

review to discover new challenges related to data quality management in software engineering, highlighting an increasing interest of the practitioners in exploring techniques that automatically detect data quality issues. These studies put the basis by leveraging the need to explore further issues related to data debt.

2.2 Related Work

The current state of research on data smells is in its preliminary stages, with limited exploration and documentation in existing literature. Foidl et al. [13] conducted a literature review to present the definition of data smells and define a catalog with 36 data smells in three main categories: *Believability Smells, Understandability Smells*, and *Consistency Smells*. Moreover, they proposed two tools to detect part of the smell defined. Similarly, Shome et al. [43] explored and analyzed commonly used datasets in Kaggle[1] to identify and define data smells. They defined four categories of specific data smells: *Redundant value smells, Categorical value smells, Missing value smells*, and *String value smells*. While these contributions provide valuable insights, the current research reveals a scattered comprehension of the ramifications of data smells. The fragmented nature of this knowledge implies a need for further integration and synthesis to extract a unified catalog, useful for practitioners and researchers to have a complete overview of all the data smells defined in the literature. Subsequently, even if the literature provides a clear definition that eases the understanding of the implication of data smells, there is still a lack of knowledge about how these smells affect the overall data quality. This study aims to fill this gap regarding data smells and their impact on data quality. To achieve this, we created an updated and unified catalog of data smells, consolidating categorizations for a more cohesive understanding. Then, we conducted empirical analysis to uncover the intricate relationship between different data smells and various aspects of data quality. Our research aims to substantially contribute to advancing our comprehension of data smells and their implications in the broader context of data quality assurance.

3 RESEARCH QUESTIONS AND METHODS

The *goal* of the study is to address the existing gaps and limitations in understanding data smells and their impact on data quality. Specifically, it aims to contribute to the field by undertaking two primary objectives. On the one hand, it seeks to create an updated and unified catalog of data smells, building upon the previous literature review conducted by Foidl et al. [13]. On the other hand, the study intends to understand the relationship between data smells and data quality. To formalize and address the main goal of our study, we applied the Goal-Question-Metric approach proposed by Caldiera et al. [6]. In detail, we defined the goal of our study:

Our Goal.
Purpose: Understand
Issue: the characteristics of
Object: data smells, their prevalence, and their relationship with data quality aspects
Viewpoint: from the points of view of researchers and data engineers.

[1]Kaggle: https://www.kaggle.com/

From the goal, we defined three main research questions. First, since there is a lack of a unified definition of the data smells defined in the literature, we wanted to elicit a complete and unified catalog of the main data smells defined in the literature with a name and a description of their features. In particular, we asked:

RQ$_1$. *What specific data smells are documented in the existing literature and what are the tools to detect data smells?*

Additionally, we needed a comprehensive analysis of data smell prevalence as a foundational step in developing effective data quality management strategies. Identifying common data smells and their frequency allows for prioritizing efforts in data cleaning, validation, and improvement. Therefore, we identified the following research question:

RQ$_2$. *What is the prevalence of data smells in real-world datasets?*

Finally, practitioners and researchers need to understand the relationship between data smells and data quality metrics to understand their severity and develop targeted strategies for data quality enhancement. By exploring significant relationships, we pinpointed the specific data quality dimensions most affected by certain data smells. With this aim, we formulated the following research question:

RQ$_3$. *How do these data smells contribute to the degradation of data quality?*

We leveraged two methods to address the main goal of the study and the related research questions. First, we followed the method by Wohlin [51] to update systematic literature reviews to explore data smells. This method provides a structured framework to collect, evaluate, and synthesize existing literature on data smells. Following the systematic literature review process, we aimed to identify and consolidate comprehensive information on documented data smells, their categorizations, and tools proposed for their detection, starting from the basis of the knowledge defined by Foidl et al. [13].

The tools delineated in the review process served as the foundation for the analytical phase to address the second and third research questions. Identifying data smells, we conducted a prevalence analysis within widely utilized datasets. In the final stage, we employed the metrics established by Elouataoui et al. [10] to undertake a correlation analysis to delve into the intricate relationships between the identified data smells and various data quality metrics.

4 ON THE EXPLORATION OF DATA SMELLS

To collect a comprehensive overview of the data smells defined by researchers so far, we extended the taxonomy provided by Foidl et al. [13]. In particular, we followed the guidelines by Kitchenham et al. [21] and Wohlin et al. [51], conducting four main steps, namely (i) database search, (ii) snowballing, (iii) application of exclusion/inclusion criteria, and (iv) quality assessment.

Database Search. As a first step, we identified the key terms to use within the queries, extracting them from the research questions. Then, we found alternative spellings and synonyms for these terms and applied boolean operators as conjunctions, in particular, the

"OR" operator for the union of alternative spellings and synonyms and the *"AND"* operator for the concatenation of the key terms.

These steps led to defining the following search query:

("data smell*" *OR* "data defect*" *OR* "data debt*") *AND* (("tool" *OR* "technique*" *OR* "strateg*" *OR* "identification" *OR* "refactoring") *OR* ("dataset*") *OR* ("definition*" *OR* "catalog*") *OR* (("machine learning" *OR* "artificial intelligence" *OR* "ai" *OR* "deep learning" *OR* "dl")))

The search query was executed against three databases, namely IEEEExplore, ACM Digital Library, and Scopus.

Inclusion & Exclusion Criteria. The study selection criteria are intended to identify the primary studies that may address the research questions. Inclusion and exclusion criteria were based on the research questions and piloted to ensure they can be reliably interpreted to classify primary studies [21]. In our case, we filtered out resources based on the following exclusion criteria:

EC1. Articles not written in English.
EC2. Restricted license papers, so papers whose full-text read was not available for free.
EC3. Articles that were not peer-reviewed.

As for inclusion criteria, we defined the following:

IC1. Articles defining data smells.
IC2. Articles assessing tools about data quality and smells.

Snowballing. According to the guidelines formulated by Wohlin [50], we applied the exclusion and inclusion criteria defined before and after each iteration of backward and forward snowballing. The snowballing process has been iterated until a state of saturation (*i.e.*, the snowballing process continued until the last iteration does not allow the inclusion of new articles).

Quality Assessment. Before data extraction, we defined a list of questions to help assess the quality of the selected paper with the final goal of having high-quality resources and discarding the papers that did not provide enough details. Following, we report the set questions:

Q1. Does the paper define one or more data smells?
Q2. Does the paper provide instances of the impact of the data smell on a machine learning system?
Q3. Does the paper clearly define how the method is conducted to assess the presence of data smells?
Q4. Does the paper define one or more tools or techniques to identify or refactor the presence of data smells in a dataset?
Q5. Are the main aspects of the paper clearly defined?

The first four questions can be considered mutually exclusive. All the questions can be answered with "Yes" (Score = 1), "Partially" (Score = 0.5), or "No" (Score = 0); the final quality score for each paper was computed by summing up the score of the answers to the two questions, to be accepted the article should have at least a score of 1.5.

4.1 Research Method Execution

We began our exploration by comprehensively searching databases, which yielded an initial collection of 197 articles. We applied the inclusion and exclusion criteria to refine this pool and were left with a core set of just four papers. Following the methodology recommended by Wohlin et al. [51], we initiated a process of forward snowballing from the seminal work of Foidl et al. [13], as well as backward and forward snowballing for newly identified articles during the literature review. This method allowed for adding 14 other articles during the first iteration and eight more during the second iteration, bringing our total to 26 papers. The third iteration was significant in helping us determine the saturation state of our search. By applying quality assessment criteria, we identified a final set of 12 papers that were eligible for inclusion. These selected contributions augment and refine the taxonomy, thus enhancing our understanding of the domain. This process allowed us to effectively narrow down a large set of articles to a smaller, more relevant set.

4.2 Analysis of the Results

4.2.1 Data Smells Catalog. Starting from the initial catalog defined by Foidl et al. [13], we collected data on smell definitions and types from 12 articles. To answer the first research question, we reported all the main data smells defined in the literature in a unified catalog divided by classes of smells containing a name and description.

In detail, starting from the original catalog defined, we found three new categories of data smell and 12 new data smells.

Figure 1 outlines various types of data quality issues known as "smells" within different categories such as *Believability*, *Understandability*, *Consistency*, *Redundancy*, *Distribution*, and *Miscellaneous*. These smells are early indicators of potential problems in datasets, which can affect data analysis and machine learning models. In the following, we describe each category in more detail:

Believability smells encompass issues related to the credibility and trustworthiness of data. New smells identified in this category are *Multiple Value Smell* [44] that describe multiple data in a unique value (*e.g.*, the age and the gender of a specific person are represented as a unified value). In contrast, *Splitted Value Smell* represents a single data split in more columns [44].

Understandability Smells deals with issues affecting data comprehensibility, and it involves two subcategories. *Encoding Smells* highlight problems like representing integers as strings or vice versa, which can lead to confusion during data processing [13, 15]. *Syntactic Smells* cover issues such as special characters, spacing inconsistencies, and data values that are too long to understand. New smells retrieved and grouped in this category are *String in human-friendly format* [43] and *Missing value Smell* [4, 8, 11, 17, 36, 40, 43, 44].

Consistency smells relates to the uniformity and consistency of data elements. Problems like *Syntax Inconsistency* and *Special Character Inconsistency* show discrepancies in how data values are used. Inconsistencies in spacing, casing, and unit measurements can also be problematic. Identifying and rectifying these inconsistencies is crucial for accurate data analysis and reporting. A new smell identified in this category is *Value Length inconsistency* [20], related to some of the values of a set that are represented with a significant difference in terms of length. *Syntactic smells* represent the inappropriate use of values that increase the difficulty of interpreting and using the data [13]. A new smell identified under this category is *String in human-friendly format*, representing the use of human-friendly values. This increases the complexity of data analysis, as

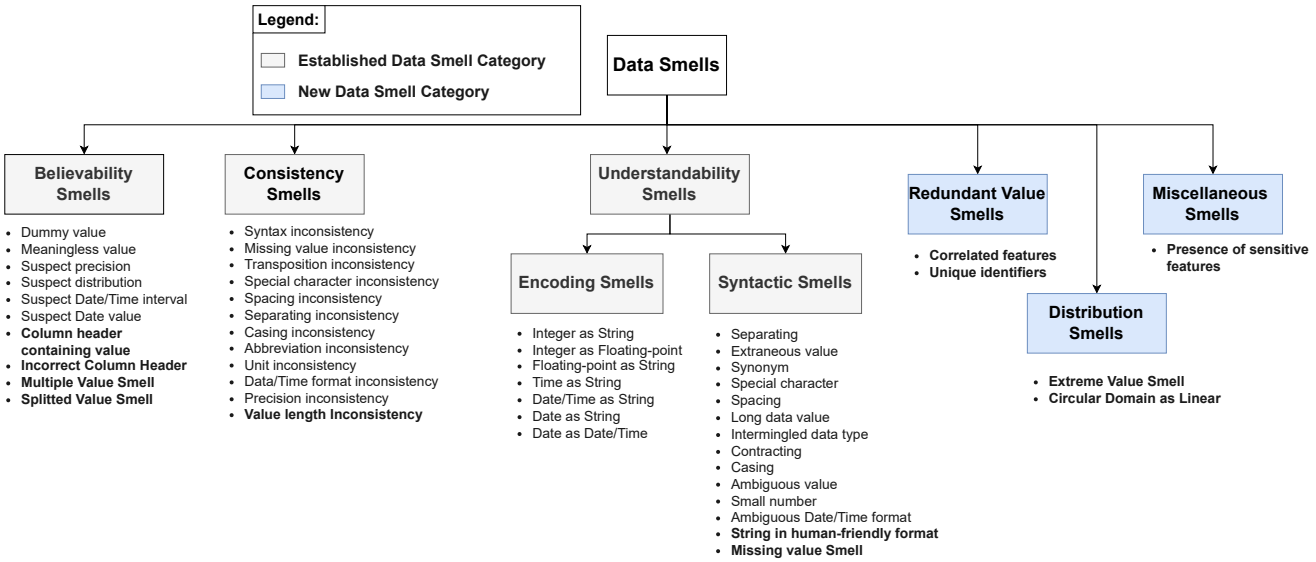

Figure 1: An extended version of the catalog of data smells defined by Foidl et al. [13]. The new data smells are highlighted in bold.

such values may pose challenges in automated processing. *Redundant value smells* point to data elements or features that provide little to no additional information for the training of AI models. *Correlated Features* [43] indicate redundant data when two features have a linear relationship, potentially introducing noise in models. *Unique Identifiers* [43] signifies redundant data that can be removed to improve data quality and model performance.

Distribution smells are related to the values in terms of the whole range of values represented. New smells under this category are *Extreme Value Smell* [4, 20], in which some of the values strongly differ from the distribution, and *Circular Domain as Linear* [20], in which a feature that has a range of values limited in a set of values (*i.e.*, days of the week) are represented as linear. A possible solution to refactor this data smell is data binning, grouping values in intervals to select a single value representing the whole range [20].

Finally, *Miscellaneous Smells* covers various data quality issues that do not fit into the previous categories. The presence of *Sensitive Features* warns against including high-impact features that may introduce bias and unfairness in predictions [43].

Addressing these data smells is essential for ensuring the quality and reliability of data, which is fundamental for making informed decisions and building accurate machine learning models. Data cleaning and preprocessing techniques can help mitigate these issues, making the data more suitable for analysis and model training.

4.2.2 Data Smells Identification Tools and Strategies. The results highlight three main tools helpful in identifying data smells. To get further information about the characteristics of these tools, we analyzed the documentation to understand their characteristics and which data smells can be detected using them.

The first tool is *Rule-Based Data Smell Detection* [13]. This tool uses an open-source data validation approach focusing on rule-based data smell detection. This tool is designed to identify smells such as Long Data Value, Casing, and several Encoding Smells. It includes a user-friendly graphical interface for uploading CSV files, enabling users to adjust the suspicion level based on predefined settings. Moreover, users can define parameters for each detection method individually. Subsequently, a machine learning-based version of data smell detection tools is implemented and proposed by Foidl et al. [13]. This version uses machine learning algorithms to detect data smells. In detail, different models are implemented and used to detect several *Inconsistency Smells* and *Encoding Smells*. Furthermore, to detect smells that rely on the semantics of the value, the tool uses an NLP approach (*i.e.*, Word2Vec) to detect a *Believability Smell* (*i.e.*, Synonyms). Finally, *Data Validator* is a component inside the framework TensorFlow Extended (TFX) that automatically collects information about the data schemas used for training machine learning models and reports quality issues to the user. From the set of quality issues able to define, Data Validator reports the presence of *Encoding Smells* (*i.e.*, *Non-boolean value for boolean feature type* and *Extreme Value Smell*).

> 👉 **Answer to RQ$_1$.** Data smells can be classified into eight categories, namely, *Believability Smells*, *Encoding Smells*, *Syntactic Smells*, *Consistency Smells*, *Redundant Value Smells*, *Distribution Smells*, and *Miscellaneous Smells*, with a final catalog of 50 data smells. From the original study, 12 more data smells are added to the catalog, and three new data smell categories are defined. Finally, the results show three data smell detection tools to identify a significant part of the data smells.

5 ON THE PREVALENCE AND IMPACT OF DATA SMELLS ON DATA QUALITY

We conducted several steps to answer **RQ$_2$** and **RQ$_3$**. First, we collected datasets that allowed us to address our objectives. Subsequently, we selected one of the tools retrieved by the systematic

literature review to identify data smells and assess their prevalence. Afterwards, we selected a set of data quality metrics measurable to understand the quality properties of the data and perform correlation analysis between the presence of the data smells and the data quality metrics retrieved.

5.1 Data Collection

The first step of this phase was data collection to gather a set of datasets studied in the literature in the form and structure of tabular data to be analyzable by the detection tools. We based our research on the data studied by Le Quy et al. [22] that overviews several real-world, tabular datasets used for fairness-aware machine learning and analyzes correlations between the different attributes using a Bayesian network. We added the datasets reported by Hirzel et al. [19] to enhance this initial set. The datasets are divided by their application domain into different categories, namely: financial datasets, criminological datasets, healthcare and social datasets, educational datasets, and miscellaneous datasets; from the datasets reported in the papers, we extracted those related to classification tasks, resulting in a total of 19 datasets.

For each dataset, we provided a description and a set of metadata, including name, path, protected attribute, privileged classes, and favorable labels, to use in the data analysis phase.

5.2 Data Smell and Data Quality Metrics Collection

The second step of this phase consisted of analyzing the collected datasets. To carry out this phase, we built a tool wrapping the data quality tool DSD[2] and implemented a module to compute metrics about data quality.

5.2.1 Data Quality Metrics. This module implements the classes to compute the quality metrics of our datasets. We decided to rely on the metrics described by Elouataoui et al. [10]; since they introduced a set of metrics to assess the quality in the context of big data processes, we decided to consider only a subset of metrics, excluding all the time-related metrics and the process-related metrics.

The final set includes the following metrics:

- **Completeness (Com)**: In big data environments, the collected raw data are usually incomplete and lack contextual information. Thus, data completeness is a crucial criterion when assessing data quality [10]. It can be defined as:

$$\frac{Number\_of\_non\_empty\_values}{Total\_values} \times 100 \tag{1}$$

- **Uniqueness (Uni)**: Large-scale datasets are usually redundant since the data are gathered from multiple sources; therefore, the same information can be recorded more than once in a different format [10]. It can be defined as:

$$\frac{Number\_of\_unique\_rows}{Total\_rows} \times 100 \tag{2}$$

- **Consistency (Con)**: Consistent data should be defined as data presented in the same structure and types and coherent with data schemas and standards [10]. It can be defined as:

$$\frac{Number\_of\_values\_with\_consistent\_types}{Total\_values} \times 100 \tag{3}$$

- **Readability (Read)**: Data validity is not limited to data format but also refers to data semantics. Indeed, raw data may contain misspelled words or even nonsense words, especially when the database is overwhelmed by human data entries [10]. It can be defined as:

$$\frac{Number\_of\_non\_mispelled\_values}{Total\_values} \times 100 \tag{4}$$

[2]https://github.com/mkerschbaumer/rb-data-smell-detection

5.3 Data Analysis to Address RQ$_2$

To address RQ$_2$, we run our wrapper of the DSD tool against the collected datasets. More specifically, the tool reports the presence of each of the considered data smells on each dataset, hence allowing us to assess the prevalence of data smells.

5.4 Data Analysis to Address RQ$_3$

To address RQ_3, we performed a statistical analysis based on hypothesis testing to test the effect of data smells on the data quality metrics. We defined two sets relative to the data smells analyzed (DS) and the data quality metrics retrieved (DQ).

$$DS = \{EV, MV, CA, SS, FS\} \tag{5}$$

$$DQ = \{Com, Uni, Con, Read\} \tag{6}$$

Then, we formulated the null hypotheses as follows, considering each combination of data smells and data quality metrics:

> $H_0(ds, dq)$: There is no statistically significant relationship between the data smell $ds \in DS$ and the data quality metric $dq \in DQ$.

When a null hypothesis can be rejected with high confidence, the alternative hypothesis could be acceptable, admitting the negative effect that the data smells have on the data quality metrics:

> $H_a(ds, dq)$: There is a statistically significant relationship between the data smell $ds \in DS$ and the data quality metric $dq \in DQ$.

After selecting the tool to retrieve data smells and define the relative null hypotheses, we designed the statistical analysis to investigate the correlation between data smells and data quality. In detail, we defined x_{ds} as the *independent variable* representative of the number of occurrences of the data smell $ds \in DS$ and y_{dq} as the relative *dependent variable* representing the value of the data quality metric $dq \in DQ$. We analyzed the association between each independent variable and each single dependent variable to check for potential significant correlations. In particular, we applied Spearman's correlation rank coefficient [5]. The decision to use Spearman's correlation was based on our observation of the non-normal distribution of each data smell.

Guided by the observed correlations, we decided to deepen our understanding of the underlying relationships by employing a generalized linear model (GLM) [46]. Acknowledging that correlations do not imply causation, a GLM allows us to explore the dependencies between variables more nuanced, accommodating various distributional assumptions and potential nonlinearities. Before

Table 1: Distribution of Data Smells

Data Smell	Total Count
Extreme Value Smell (EV)	137
Missing Value Smell (MV)	110
Casing Smell (CA)	24
Suspect Sign Smell (SS)	20
Floating Point Number As String Smell (FS)	2

building the model, it is necessary to understand if multicollinearity could be present among all the variables [53]. In our study, we predefined a significance level (α) at 0.05 for hypothesis testing.

5.5 Analysis of the Results

5.5.1 Prevalence of Data Smells. Table 1 summarizes the results we achieved. Considering the smells we analyzed, the most common is the *Extreme Value Smell* with a total count of 143 smelly instances, the second one is the *Missing Value Smell* with a total number of 119 instances, and the third is *Casing Smell* with 24 instances.

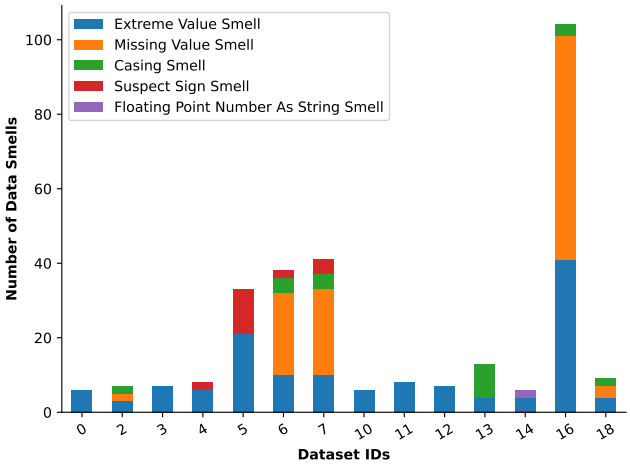

Figure 2: Prevalence of Data Smells in the analyzed datasets

Figure 2 illustrates the number of features affected by smells for each dataset. The datasets that do not have instances of data smell are not included in the plot to increase interpretability. Therefore, 14 out of 19 datasets present at least an attribute affected by data smells. The most diffused data smell in the several datasets is *Extreme Value Smell*, present in all the datasets affected by at least one data smell. While the prevalence of data smell is high, the distribution of the instances is not equally distributed across all the datasets. One of the datasets analyzed ("speed dating", with Dataset ID *16*) presents 104 instances of data smells, composed of 60 instances of *Missing Values*, 41 instances of *Extreme Value Smell*, and three instances of *Casing Smells*. *Floating Point Number As String Smell* is present only in one of the datasets analyzed (*heart disease*, with DatasetID 14).

Table 2: Spearman Test Results

Variable	Independent Variable	Spearman Statistic	P-value
Uniqueness	EV	0.2038	***
	MV	0.2672	***
	CA	0.1667	***
	SS	0.1150	**
	FS	0.0578	
Consistency	EV	0.1533	***
	MV	-0.1833	***
	CA	0.0576	
	SS	0.0525	
	FS	0.2078	***
Readability	EV	0.3978	***
	MV	0.0213	
	CA	-0.2038	***
	SS	0.1376	**
	FS	0.1051	*
Completeness	EV	-0.1357	**
	MV	-0.9524	***
	CA	-0.0746	
	SS	-0.0002	
	FS	0.0311	

Significance levels: $p < 0.001$ (***),$p < 0.01$ (**),$p < 0.05$ (*)

Answer to RQ$_2$. To summarize the results, the most common data smell detected is the Extreme Value Smell with a total number of 137 smelly instances, and the second one is the Missing Value Smell with 110 instances. Most datasets have at least one attribute affected by data smells, resulting in 6 data smells found.

5.5.2 Correlation between Data Smells and Data Quality. We first analyzed the correlation between each data smell and quality metric to answer the last research question.

Table 2 reports the results of the Spearman Correlation Rank test. Asterisks denote the significance levels, and no asterisks denote a p-value higher than 0.05, meaning low statistical significance.

The strength and significance of these correlations vary across measures. *Extreme Value Smells*, *Missing Value Smells*, *Casing Smells*, and *Suspect Sign Smells* exhibit statistically significant correlations with *Uniqueness*, suggesting a robust relationship. Regarding *Consistency*, *Extreme Value Smell*, and *Floating Point Number as String Smell* demonstrate a significant positive correlation, while *Missing Value Smell* displays a significant negative correlation. *Readability* shows strong associations with almost all the proposed smells, confirming a significant correlation with *Extreme Value Smells*. Finally, *Completeness* is highly correlated to the *Missing Value Smells*, highlighting a strong negative correlation, followed by a statistically significant but slight correlation with the *Extreme Value Smell*.

The observed variations in correlation strength and statistical significance emphasize the nuanced relationships between these smells and data quality aspects, contributing valuable insights to our understanding of the factors influencing text characteristics.

Table 3: GLM Coefficient Statistics

Model	Variable	Coefficient (significance)
Uniqueness	EV	9.17×10^{-6}
	MV	2.65×10^{-7}
	SS	6.23×10^{-7}
	CA	8.67×10^{-5} (***)
Completeness	EV	4.03×10^{-6}
	MV	-0.0002 (***)
Consistency	EV	4.90×10^{-6}
	MV	-1.02×10^{-5} (***)
	FS	-6.29×10^{-6}
Readability	EV	0.0001 (**)
	SS	3.62×10^{-5}
	CA	-0.0001 (***)
	FS	-0.0025 (**)

Significance levels: p < 0.001 (***),p < 0.01 (**),p < 0.05 (*)

From the outcome of the Spearman correlation ranks, we built the GLM model, selecting the independent variables that have a statistically significant association with the relative dependent variable. The results obtained from the applied Generalized Linear Model (GLM), as presented in Table 3, offer valuable insights into the significant relationships between data smells and data quality, revealing several correlations.

For each data smell, the reported relative coefficient represents the impact on the data quality metric, accompanied by the corresponding significance level. These correlation coefficients are calculated based on different proportions of the dataset, indicating how a unit increase in the specified data smell influences the associated quality attribute. It is important to note that even if the model indicates a low impact for the presence of a single instance of the data smell, the cumulative effect of multiple instances can be substantial. Thus, understanding the potential effects of data smells becomes crucial, especially in high quantities. The presence of *Missing Value Smells* exhibits a negative correlation, emphasizing its slight but statistically significant effect on *Completeness*.

The model related to the *Consistency* exhibits a highly significant negative correlation with *Missing Value Smells* but presents a very low coefficient, indicating that even if the variable could have a statistically significant effect, the effect of the data smells could be critical for this data quality aspect if the number of occurrences of *Missing Value Smells* is high. Differently, the effect of this data smell on *Completeness* is higher with a very low p-value, leading to a strong effect. Regarding *Uniqueness* of the data, *Casing Smells* have a slight but statistically significant effect. More interestingly, the presence of *Extreme Value Smells* is positively correlated to *Readability*, leading to adding values that differ significantly from the distribution increases the amount of information the data gives. *Floating Point Number as String Smells* have a highly significant effect on *Readability* of the data. *Casing Smells* also slightly affects *Readability*, in which the relationship is extremely significant.

In summary, while the effect of a single instance of a data smell may appear negligible to data quality, the noteworthy significance levels associated with each correlation underscore the gravity of having data smells in high quantities. The cumulative impact of multiple instances becomes increasingly pronounced, emphasizing the potential detrimental effects on data quality.

Answer to RQ$_3$. The analysis reveals nuanced relationships between data smells and data quality metrics. While individual data smells may exhibit a seemingly low impact, the significance levels associated with each correlation signal the severity of having these smells in high quantities.

6 THREATS TO VALIDITY

Threats to the External Validity. This study's primary concern is the potential bias introduced by selecting database sources. To mitigate this threat, we adhered to the guided practices outlined by Wohlin et al. [50] by including main database sources. Further, we conducted two iterative rounds of snowballing to ensure a comprehensive exploration of articles addressing data quality issues.

While the study aimed to explore deeply data quality issues, specifically focusing on identifying data smells, there was a crucial decision to balance the scope. A more extensive exploration could have introduced various other types of data issues unrelated to the concept of smells. Although our broad investigation led to the identification of new data smells and even novel categories of such smells, it is acknowledged that certain data quality issues associated with the concept of smells may remain undefined.

As for selecting the 19 datasets for studying the correlation between data smells and data quality aspects, careful consideration was given to address potential biases. The choice of these datasets was guided by the need to use representative datasets commonly used in AI applications and research. It is important to note that while the chosen datasets provide a robust basis for examining correlations, the variability among datasets may influence the generalizability of specific findings.

Threats to the Conclusion Validity. The choice of instrumentation is critical to conclusion validity, particularly in detecting data smells and measuring data quality aspects. Firstly, utilizing the tool published by Foidl et al. [13] provides a foundation for identifying and assessing data smells. By leveraging an established tool, the study benefits from a standardized and structured approach to data smell detection. Moreover, using metrics defined and validated by Elouataoui et al. [10] adds another layer of rigor to the measurement process. From the set of metrics defined, we deliberately chose to focus exclusively on metrics tied to the structural elements of the data. With this decision, we intentionally omitted metrics associated with external factors, such as how the data are applied in the context of an application. The study gains access to well-established measurement tools by aligning with these metrics. While this study

does not explore other aspects in the context of the data quality, it enhances the replicability of the research as a starting point.

Threats to the Internal Validity. While our study explores the correlation between data smells and data quality metrics, we recognize the threat related to internal validity—the potential influence of confounding variables. Although we observe correlations between data smells and metrics, it is essential to acknowledge the existence of other factors that may be tightly correlated and impact the metrics used in our study. These confounding variables could introduce complexities, as correlations found may be influenced by factors related to the data-gathering process or the presence of different types of data quality issues beyond the scope of this investigation.

7 DISCUSSION AND IMPLICATIONS

Lack of Data Quality Assurance Instruments. In the context of monitoring data quality issues in AI-enabled systems, Breck et al. [4] shows the absence of a standardized tool for this purpose. Recognizing this gap, Foidl et al. [13] started addressing the problem by introducing innovative tools designed to identify data smells. This work represents a pioneering effort in the field, leading to the opportunity to explore the effect of data smells on data quality, as this study aimed. However, the tool proposed is still limited, being able to detect only part of the smells that are defined. This limitation implies still an open challenge for the research, improving the actual state of automatic data smell detection and data quality monitoring.

Data Smells: Do They Really Smell Bad? According to our findings, we may conclude that only some data smells influence data quality aspects related to the data structure. However, it would still be possible that the effects of data smells might be observed in the long run, namely when considering evolutionary aspects of data like data change-proneness, data engineering, and the data governance processes. More interesting, the outcome of the generalized linear model for the relationship between data smells and readability highlights a positive correlation between the presence of *Extreme Value Smell* and *Readability*. In this case, extreme values inside a distribution seem to correlate to the amount of information a single attribute could give. This finding also underscores the importance of considering data smells in the broader context of data analysis, as they may not only indicate potential issues but also reveal interesting patterns and relationships. As this positive correlation aligns with the notion that extreme values might enhance the amount of information conveyed by a single attribute, it invites deeper investigation into the mechanisms driving this relationship. Future research could delve into the practical implications of this correlation, providing insights into how data practitioners and analysts can leverage or manage extreme values to improve the interpretability and utility of their datasets.

Structural Data Quality Metrics: Are They Enough? Regarding the data quality metrics set, we selected the structural data quality metrics defined by Elouataui et al. [10]. This set of metrics allows the understanding of the data quality based on the structural characteristics related to the readability of the distributions, completeness, consistency, and uniqueness. While these metrics are well-defined and easy to use to evaluate the quality of a dataset, there could be the need to extend the definition of actual data quality metrics. The relationship and the analysis of data smells could help to this goal. Considering the new smells identified, it could be possible to define new metrics related to *Multiple Value Smell*, understanding the number of values a single attribute's value contains. Similarly, defining *Column Header Containing Value* could lead to defining new metrics related to the explainability of the value. To sum up, while the structural metrics selected are robust and user-friendly, considering the analysis of data smells introduces the prospect of refining and expanding the metrics landscape. By incorporating insights from data smells, we can develop new metrics that delve deeper into the subtleties of data quality, enhancing our ability to evaluate and ensure the integrity of datasets comprehensively.

8 CONCLUSION

This study outlines the state of the definition of data smells in AI-enabled systems. First, we defined a new taxonomy with 12 new data smells and three new categories to extend the catalog of data smells. Then, we analyzed the presence of data smells to understand their relationship with data quality. The outcomes highlight the emerging severity of these types of data quality issues, that while the effect could be irrelevant in small amounts, the impact of such issues can be significant when introduced in high quantity. Therefore, this research enriches data smells and emphasizes their tangible implications on data quality. The high severity recognized serves as a call to increase the awareness of researchers and practitioners and leverage the need to institute robust strategies and best practices in data governance and quality assurance. As AI plays a pivotal role in various domains, understanding and addressing data smells becomes crucial for ensuring AI-enabled systems' reliability, trustworthiness, and effectiveness. As implications of these results, it is necessary to define new detection and refactoring strategies to support the practitioners in the data quality management process. On the one hand, such detection strategies could empower practitioners to identify and isolate specific instances of data smells, providing a more granular understanding of their presence and facilitating targeted interventions. On the other hand, defining new refactoring strategies to fix data smells effectively could support practitioners in guaranteeing high data quality and allow for investigating the effect on the overall AI-enabled systems.

DATA AVAILABILITY

The data collected as part of the systematic literature review, the scripts used to analyze and generate data, charts, and plots discussed when addressing our research goals, are publicly available at [39].

ACKNOWLEDGMENT

This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects PRIN 2022 "QualAI: Continuous Quality Improvement of AI-based Systems", grant n. 2022B3BP5S , CUP: H53D23003510006.

REFERENCES

[1] Justus Bogner, Roberto Verdecchia, and Ilias Gerostathopoulos. 2021. Characterizing technical debt and antipatterns in AI-based systems: A systematic mapping study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*. IEEE, 64–73.

[2] Jan Bosch, Helena Holmström Olsson, Björn Brinne, and Ivica Crnkovic. 2022. AI Engineering: Realizing the Potential of AI. *IEEE Software* 39, 6 (2022), 23–27.

[3] Michael Franklin Bosu and Stephen G MacDonell. 2013. Data quality in empirical software engineering: a targeted review. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*. 171–176.

[4] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning.. In *MLSys*.

[5] Lionel Briand, Khaled El Emam, and Sandro Morasca. 1996. On the application of measurement theory in software engineering. *Empirical Software Engineering* 1 (1996), 61–88.

[6] Victor R Basili1 Gianluigi Caldiera and H Dieter Rombach. 1994. The goal question metric approach. *Encyclopedia of software engineering* (1994), 528–532.

[7] Ward Cunningham. 1992. The WyCash portfolio management system. *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications, OOPSLA* Part F129621 (1992), 29 – 30. https://doi.org/10.1145/157709.157715

[8] Lisa Ehrlinger, Thomas Grubinger, Bence Varga, Mario Pichler, Thomas Natschläger, and Jürgen Zeindl. 2018. Treating missing data in industrial data analytics. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. IEEE, 148–155.

[9] Lisa Ehrlinger and Wolfram Wöß. 2022. A survey of data quality measurement and monitoring tools. *Frontiers in big data* 5 (2022), 850611.

[10] Widad Elouataoui, Imane El Alaoui, Saida El Mendili, and Youssef Gahi. 2022. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data and Cognitive Computing* 6, 4 (2022), 153.

[11] Harald Foidl and Michael Felderer. 2019. Risk-based data validation in machine learning-based software systems. In *proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation*. 13–18.

[12] Harald Foidl, Michael Felderer, and Stefan Biffl. 2019. Technical Debt in Data-Intensive Software Systems. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 338–341. https://doi.org/10.1109/SEAA.2019.00058

[13] Harald Foidl, Michael Felderer, and Rudolf Ramler. 2022. Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*. 229–239.

[14] Martin Fowler and Kent Beck. 1997. Refactoring: Improving the design of existing code. In *11th European Conference. Jyväskylä, Finland.*

[15] V. Golendukhina, H. Foidl, M. Felderer, and R. Ramler. 2022. Preliminary findings on the occurrence and causes of data smells in a real-world business travel data processing pipeline. *SEA4DQ 2022 - Proceedings of the 2nd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things, co-located with ESEC/FSE 2022* (2022), 18–21. https://doi.org/10.1145/3549037.3561275 cited By 0.

[16] Lina Gong, Shujuan Jiang, and Li Jiang. 2019. Tackling Class Imbalance Problem in Software Defect Prediction Through Cluster-Based Over-Sampling With Filtering. *IEEE Access* 7 (2019), 145725–145737. https://doi.org/10.1109/ACCESS.2019.2945858

[17] Ulrike M Graetsch, Hourieh Khalajzadeh, Mojtaba Shahin, Rashina Hoda, and John Grundy. 2023. Dealing with data challenges when delivering data-intensive software solutions. *IEEE Transactions on Software Engineering* (2023).

[18] David Gray, David Bowes, Neil Davey, Yi Sun, and Bruce Christianson. 2011. The misuse of the NASA metrics data program data sets for automated software defect prediction. In *15th Annual Conference on Evaluation and Assessment in Software Engineering (EASE 2011)*. 96–103. https://doi.org/10.1049/ic.2011.0012

[19] Martin Hirzel and Michael Feffer. 2023. A Suite of Fairness Datasets for Tabular Classification. *arXiv preprint arXiv:2308.00133* (2023).

[20] Nick Hynes, D Sculley, and Michael Terry. 2017. The data linter: Lightweight, automated sanity checking for ml data sets. In *NIPS MLSys Workshop*, Vol. 1.

[21] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26.

[22] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.

[23] Valentina Lenarduzzi, Terese Besker, Davide Taibi, Antonio Martini, and Francesca Arcelli Fontana. 2021. A systematic literature review on Technical Debt prioritization: Strategies, processes, factors, and tools. *Journal of Systems and Software* 171 (2021), 110827.

[24] Zengyang Li, Paris Avgeriou, and Peng Liang. 2015. A systematic mapping study on technical debt and its management. *Journal of Systems and Software* 101 (2015), 193–220. https://doi.org/10.1016/j.jss.2014.12.027

[25] Gernot Liebchen and Martin Shepperd. 2016. Data Sets and Data Quality in Software Engineering: Eight Years On. In *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering* (Ciudad Real, Spain) *(PROMISE 2016)*. Association for Computing Machinery, New York, NY, USA, Article 7, 4 pages. https://doi.org/10.1145/2972958.2972967

[26] Lucy Ellen Lwakatare, Aiswarya Raj, Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. 2019. A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. In *Agile Processes in Software Engineering and Extreme Programming: 20th International Conference, XP 2019, Montréal, QC, Canada, May 21–25, 2019, Proceedings 20*. Springer International Publishing, 227–243.

[27] Taha Mansouri, Mohammad Reza Sadeghi Moghadam, Fatemeh Monshizadeh, and Ahad Zareravasan. 2023. IoT data quality issues and potential solutions: a literature review. *Comput. J.* 66, 3 (2023), 615–625.

[28] R. Marinescu. 2004. Detection strategies: metrics-based rules for detecting design flaws. In *20th IEEE International Conference on Software Maintenance, 2004. Proceedings.* 350–359. https://doi.org/10.1109/ICSM.2004.1357820

[29] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, and Stefan Wagner. 2022. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–59.

[30] Claire Cain Miller. 2015. Can an algorithm hire better than a human. *The New York Times* 25 (2015).

[31] Naouel Moha, Yann-Gael Gueheneuc, Laurence Duchien, and Anne-Francoise Le Meur. 2010. DECOR: A Method for the Specification and Detection of Code and Design Smells. *IEEE Transactions on Software Engineering* 36, 1 (2010), 20–36. https://doi.org/10.1109/TSE.2009.50

[32] Aiswarya Munappy, Jan Bosch, Helena Holmström Olsson, Anders Arpteg, and Björn Brinne. 2019. Data Management Challenges for Deep Learning. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 140–147. https://doi.org/10.1109/SEAA.2019.00030

[33] Parmy Olson. 2011. The algorithm that beats your bank manager. *CNN Money March* 15 (2011).

[34] Fabio Palomba, Gabriele Bavota, Massimiliano Di Penta, Rocco Oliveto, Denys Poshyvanyk, and Andrea De Lucia. 2015. Mining Version Histories for Detecting Code Smells. *IEEE Transactions on Software Engineering* 41, 5 (2015), 462–489. https://doi.org/10.1109/TSE.2014.2372760

[35] Fabio Palomba, Annibale Panichella, Andrea De Lucia, Rocco Oliveto, and Andy Zaidman. 2016. A textual-based technique for Smell Detection. In *2016 IEEE 24th International Conference on Program Comprehension (ICPC)*. 1–10. https://doi.org/10.1109/ICPC.2016.7503704

[36] Xavier Pleimling, Vedant Shah, and Ismini Lourentzou. 2022. [Data] Quality Lies In The Eyes Of The Beholder. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. 118–124.

[37] Jörg Rech and Klaus-Dieter Althoff. 2004. Artificial intelligence and software engineering: Status and future trends. *KI* 18, 3 (2004), 5–11.

[38] Gilberto Recupito, Fabiano Pecorelli, Gemma Catolino, Sergio Moreschini, Dario Di Nucci, Fabio Palomba, and Damian A. Tamburri. 2022. A Multivocal Literature Review of MLOps Tools and Features. In *2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 84–91. https://doi.org/10.1109/SEAA56994.2022.00021

[39] Gilberto Recupito, Raimondo Rapacciuolo, Dario Di Nucci, and Fabio Palomba. 2023. *Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality. Online Appendix*. https://figshare.com/s/31a3742977548f96f506

[40] Joseph L Schafer and Maren K Olsen. 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research* 33, 4 (1998), 545–571.

[41] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).

[42] Carolyn Seaman and Yuepu Guo. 2011. Measuring and monitoring technical debt. In *Advances in Computers*. Vol. 82. Elsevier, 25–46.

[43] Arumoy Shome, Luis Cruz, and Arie Van Deursen. 2022. Data smells in public datasets. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*. 205–216.

[44] Dina Sukhobok, Nikolay Nikolov, and Dumitru Roman. 2017. Tabular data anomaly patterns. In *2017 International Conference on Big Data Innovations and Applications (Innovate-Data)*. IEEE, 25–34.

[45] Yiming Tang, Raffi Khatchadourian, Mehdi Bagherzadeh, Rhia Singh, Ajani Stewart, and Anita Raja. 2021. An empirical study of refactorings and technical debt in machine learning systems. In *2021 IEEE/ACM 43rd international conference on software engineering (ICSE)*. IEEE, 238–250.

[46] Henri Theil. 1969. A multinomial extension of the linear logit model. *International economic review* 10, 3 (1969), 251–259.

[47] Edith Tom, Aybüke Aurum, and Richard Vidgen. 2013. An exploration of technical debt. *Journal of Systems and Software* 86, 6 (2013), 1498–1516. https://doi.org/10.1016/j.jss.2012.12.052

[48] Michele Tufano, Fabio Palomba, Gabriele Bavota, Rocco Oliveto, Massimiliano Di Penta, Andrea De Lucia, and Denys Poshyvanyk. 2015. When and Why Your Code Starts to Smell Bad. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. 403–414. https://doi.org/10.1109/ICSE.2015.59

[49] Heng Wang and Zubin Abraham. 2015. Concept drift detection for streaming data. In *2015 International Joint Conference on Neural Networks (IJCNN)*. 1–9. https://doi.org/10.1109/IJCNN.2015.7280398
[50] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 1–10.
[51] Claes Wohlin, Emilia Mendes, Katia Romero Felizardo, and Marcos Kalinowski. 2020. Guidelines for the search strategy to update systematic literature reviews in software engineering. *Information and Software Technology* 127 (2020), 106366. https://doi.org/10.1016/j.infsof.2020.106366
[52] Kyung-A Yoon and Doo-Hwan Bae. 2010. A pattern-based outlier detection method identifying abnormal attributes in software project data. *Information and Software Technology* 52, 2 (2010), 137–151. https://doi.org/10.1016/j.infsof.2009.08.005
[53] Han Yu, Shanhe Jiang, and Kenneth C. Land. 2015. Multicollinearity in hierarchical linear models. *Social Science Research* 53 (2015), 118–136. https://doi.org/10.1016/j.ssresearch.2015.04.008
[54] Jianlong Zhou and Fang Chen. 2018. *Human and Machine Learning*. Springer.