POLITECNICO DI TORINO

School of Management and Production Engineering

Master's degree in engineering and management

Master's Degree's Thesis

Artificial Intelligence for Manufacturing Data Quality: A Systematic Review of Trends, Techniques and Challenges

Supervisor:

Prof. Domenico A. MAISANO

Co-Supervisor:

Ing. Lucrezia FERRARA

Candidate:

Sara DEANDREIS

October 2025

Abstract

The integration of artificial intelligence (AI), particularly machine and deep learning, in manufacturing has accelerated in recent years, making data quality a critical prerequisite for reliable and trustworthy AI systems. This thesis systematically investigates how AI is applied to assess and enhance data reliability in manufacturing, aiming to consolidate existing knowledge, identify key gaps, and provide guidance for robust AI approaches.

A systematic literature review is conducted according to the PRISMA 2020 guidelines, chosen for its methodological rigor, transparency, and reproducibility. The review is grounded in a comprehensive retrieval of documents carried out using the Scopus and Web of Science databases. Additionally, an innovative Python-based semantic filtering step is applied to screen documents according to conceptual similarity with predefined keywords, enabling the review of over 22,000 records and resulting in a selected corpus of 164 studies. The filtering achieves an accuracy of approximately 86%, ensuring a robust assessment.

Full-text analysis of the final corpus shows an evolution in data quality conceptualization, shifting from intrinsic attributes such as accuracy and completeness to more advanced dimensions, including dimension such as fairness and cross-domain generalizability. In parallel, AI methods evolve from rule-based methods to deep learning and hybrid architectures. Nevertheless, major challenges persist, particularly the absence of standardized benchmarks, class imbalance and high labelling cost.

Overall, the research highlights AI as indispensable for data quality management in manufacturing, while also acknowledging structural and methodological limitations. Although there is consensus on the centrality of data quality, considerable variability persists across industrial sectors in the definition and implementation of quality metrics. The recent ISO/IEC 5259 standard represents a promising step toward harmonization and provides a foundation for future unified frameworks for trustworthy, data-centric AI in complex industrial contexts.

List of Figures

List of Tables

Table of Contents

Chapter 1 - Introduction: trustworthy AI and data quality

The integration of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), into a broad range of domains, including manufacturing, agriculture, healthcare, and transportation, has accelerated significantly over the past decade (Vaswani et al. 2023; Deng 2018; Silver, David 2017; He et al. 2016; Redmon et al. 2016; McKinsey 2023). This rapid development has been driven by progress in computational capabilities and neural network architecture. As AI systems increasingly influence critical societal processes, public dialogue has shifted toward concerns surrounding their transparency, fairness, and reliability (Esteva et al. 2017; Jumper et al. 2021; Teoh e Kidd 2017; UK Governement 2023). In this context, the presence of a trustworthy AI becomes essential. This term encompasses several qualities such as security, robustness, fairness, interpretability and accountability (Adadi e Berrada 2018; Liu et al. 2022; Li et al. 2023; Kale et al. 2023; Alzubaidi et al. 2023; Moody's 2024; European Commission, High-Level Expert Group on AI 2020; Moody's 2024).

One of the determining factors in the reliability of AI systems is the quality of the data on which they are trained. The expression *garbage in, garbage out* (Geiger et al. 2020) illustrates the principle that faulty and biased training data inevitably lead to faulty AI behavior. Indeed, biases embedded in training data can be amplified during inference, resulting in discriminatory or otherwise unfair results (Suresh e Guttag 2021; Mehrabi et al. 2022). Consequently, data quality is of paramount importance in ensuring ethical, reliable and trustworthy AI systems (Zhao et al. 2017; Whittlestone J. et al. 2019).

The concept of data quality has been investigated for several decades. A foundational framework was proposed by Wang and Strong (Richard Y. Wang e Diane M. Strong 1996), who defined data quality in terms of distinct dimensions, such as accuracy, completeness, consistency, and accessibility. Since then, these dimensions have been adapted and refined across various domains. However, no single standard and definition has yet emerged, especially in complex and evolving fields like AI (Andrew Black e Peter van Nederpelt 2020).

While data integrity refers to maintaining the accuracy and consistency of data over its lifecycle, data quality is concerned with the data's suitability for its intended use. From the perspective of AI development, this means evaluating whether a dataset is comprehensive, accurate, timely, relevant, and representative. These aspects, often called data quality dimensions, are critical not only for the performance of AI models, but also for their interpretability and societal impact.

Data quality assessment procedures will likely become an integral part of the AI certification process, particularly in sectors like healthcare and industrial automation. For this reason, methodologies that assess and guarantee the reliability of training data are increasingly necessary, and it is on these that standards bodies and developers now focus their attention.

In addition, ethical and social considerations further complicate the evaluation of data. Data collection and processing are not neutral activities, as they inherently reflect human assumptions regarding what is valuable or relevant. This implies that no data set is completely objective or complete (Jess Whittlestone e Stephen Cave 2019). Moreover, the digitization of data facilitates its replication, sharing and transformation on unprecedented scales, raising concerns about privacy, consent and accountability. Ethical frameworks often emphasize principles such as beneficence, non-maleficence, justice, autonomy and explicability, but these are not always easy to apply in practice due to the inherent trade-offs among them.

As more decision-making is delegated to AI systems, the societal impact of these technologies continues to grow (Richard Y. Wang e Diane M. Strong 1996). According to Wang and Strong (Wang e Strong 1996), *fitness for purpose* should be the guiding concept for assessing whether data meet user expectations and requirements.

In summary, trustworthy AI can not be achieved without a rigorous focus on data quality. For AI applications in manufacturing and other high-risk environments, this means establishing robust frameworks to define, measure, and improve data quality. These frameworks must consider both technical and ethical dimensions to ensure that AI systems are not only high performing, but also well-reasoned and accountable.

1.1- Problem Identification and Research Objective

As previously mentioned, AI systems reliability is deeply related to the quality of the data on which they are based. Although the concept of data quality has long been studied in all sectors, its evaluation and correlation in the context of AI-driven applications, particularly in the manufacturing sector, remains an underdeveloped and fragmented area.

Despite the increasing use of machine learning and deep learning techniques in industrial environments, a standardized approach to assessing data reliability for applications such as process monitoring and predictive maintenance is still lacking. Manufacturing processes are inherently complex and dynamic, characterized by high volumes of heterogeneous data generated by sensors, machines and human input. As manufacturing environments become increasingly data-driven, the ability to ensure the trustworthiness of collected data has become essential, not only for the technical performance of AI models but also for ensuring safety, transparency, and compliance with emerging regulations. For this reason, it is necessary to ensure data consistency, accuracy, completeness and representativeness. Furthermore, industrial artificial intelligence systems are often deployed in critical environments, where errors can lead to significant security risks, operational inefficiencies or financial losses. Therefore, the issue of data reliability is not only a technical concern but also a matter of trust, accountability, and regulatory compliance.

Although several frameworks for data quality assessment have been developed in fields such as healthcare and medicine, the consolidation of knowledge specific to the manufacturing domain remains limited. In the medical field, for instance, recent research has proposed frameworks, such as the METRIC framework, to assess data quality in ways that align with the broader goals of trustworthy AI. These frameworks recognize that high-quality data is a prerequisite for AI systems that are safe, fair, and transparent. A systematic review conducted under PRISMA guidelines reveals that, while there is agreement on the importance of data quality, there is a variability in how different sectors define and implement data quality metrics (Schwabe et al. 2024).

Furthermore, the role of AI in evaluating or improving data reliability itself is a relatively novel concept. This raises the need to systematically investigate how AI techniques can be used not only to process manufacturing data, but also to critically assess its trustworthiness before being used in decision-making pipelines.

The objective of this thesis is to provide a comprehensive and structured literature review of the current contributions and research efforts in the analysis and assessment of data reliability within the manufacturing sector.

From this objective, the research question (*RQ*) is formulated as follows:

(RQ) How is artificial intelligence currently being applied to assess and enhance the reliability of data in manufacturing processes, and what are the key challenges and gaps in the existing literature?

This central question is supported by the following sub-questions:

- *(RQa) Which dimensions of data quality are most frequently addressed in AI applications for manufacturing? Along which characteristics should data quality be evaluated when employing a dataset for trustworthy AI in manufacturing?*
- *(RQb) What are the main techniques used to evaluate or improve data reliability in this context?*

By answering these questions, synthesizing current knowledge and identifying existing gaps, the thesis aims to provide a comprehensive overview of the current state of the art, identify opportunities for future research, and contribute to the development of more robust and trustworthy AI systems in the manufacturing domain.

The remainder of the thesis is structured as follows. Chapter 2 introduces the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, which provides a systematic approach to conducting transparent and reproducible literature reviews. In Chapter 3 the methodological pipeline developed in accordance with the PRISMA principles, detailing each phase of the data collection, processing, and filtering strategy implemented to address the research questions is presented. Chapter 4 discusses the results obtained through the review process, highlighting key trends,

findings, and gaps in the literature on AI and data quality in manufacturing. In conclusion, Chapter 5 summarizes the main insights and outlines potential directions for future research and methodological refinements.

Chapter 2 - PRISMA

While trustworthiness in AI concerns various aspects, including ethical considerations, transparency, and safety requirements, this study focuses on the critical role of data quality in ML and DL. Since data quality significantly influences the behaviour of ML models, assessing data quality becomes a pivotal component. To address the research questions, a systematic review was conducted following the PRISMA methodology, which provides guidelines for performing quantitative analyses of documents.

The objective of the review is to systematically collect, condense, and expand the existing body of knowledge in the selected research area, thereby advancing the understanding of data quality in ML applications. Specifically, the research question aims at combining insights from general data quality frameworks with the impact of data quality on ML applications within production processes.

2.1- PRISMA guidelines

The PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) were developed to enhance the transparency and completeness of reporting in systematic reviews and meta-analyses. The guidelines assist authors in clearly presenting the rationale for their review, the methods applied, and the results obtained.

First published in 2009, the PRISMA statement aimed to standardize the reporting process in a way that promotes reproducibility and clarity. However, with the evolution of systematic review methodologies and terminology over the past decade, an updated version became necessary: PRISMA 2020. The revised PRISMA statement replaces the original 2009 version, introducing updated guidance that reflects advancements in the identification, selection, appraisal, and synthesis of studies, thus ensuring greater methodological rigor and clarity in systematic reviews.

2.2 - PRISMA: History and development

Systematic reviews and meta-analyses were originally adopted in healthcare and medical field as a starting point for developing clinical practice guidelines. Physicians use them to keep up to date (Oxman et al. 1994; Swingler et al. 2003) and even funding agencies may require a systematic review to ensure that further research is justified (Moher et al. 2009). In recent year, editors in health journals have been moving in this direction too (Young e Horton 2005).

Systematic reviews play many critical roles. Firstly, they can provide summaries of the state of knowledge in a field from which future research priorities can be identified. They can address questions that otherwise could not be answered by individual studies, and they can identify problems in primary research that should be corrected in future analysis. Additionally, they can generate or evaluate theories about how or why phenomena occur. To ensure that a systematic review is valuable, authors must provide a transparent, complete, and accurate explanation of the purpose, methods, and findings of the study.

For instance, several studies have assessed the quality of review reports. In 1987, Mulrow examined 50 review articles published in four major medical journals in 1985 and 1986 and found that none fulfilled the explicit reporting criteria, such as assessing the quality of the included studies (Mulrow 1987). In 1987, the adequacy of reporting of 83 meta-analyses on 23 characteristics across six domains was assessed (Sacks et al. 1987). Reporting was generally poor; between one and fourteen characteristics were adequately reported. A 1996 update of this study found little improvement (Sacks et al. 1996).

In 1999, to address the problem of sub-optimal reporting of meta-analyses, an international group developed a guide called QUOROM Statement (Quality Of Reporting Of Meta-analyses), focusing on reporting of meta-analyses of randomized controlled trials (Moher et al. 1999).

In 2009, the guideline was updated to consider various conceptual and practical advances in the science of systematic reviews and was renamed PRISMA. The original PRISMA statement was published in several journals (Moher et al. 2009) and accompanied by an explanation and elaboration document (Liberati et al. 2009).

The significant advancements in systematic review methodology and terminology over the past decade prompted an international group to update the original PRISMA statement in 2017. The PRISMA 2020 statement was initially published as a preprint on MetaArXiv in September 2020 and subsequently released in March 2021 (Matthew J. Page et al. 2021). Since the release of the PRISMA 2009 statement, the systematic review process has been transformed by technological advancements such as natural language processing and machine learning, which have enhanced the identification of relevant studies. Additionally, new methods for synthesizing findings in the absence of feasible meta-analyses have been developed (Matthew J. Page et al. 2021; Campbell et al. 2020), alongside updated tools for assessing risk of bias in included studies (Sterne et al. 2019; 2016). The shift from evaluating *quality* to assessing *certainty* in evidence further reflects the evolution in terminology (Hultcrantz et al. 2017). The publishing landscape has also expanded, offering more avenues for registering protocols, disseminating review findings, and ensuring data accessibility (Hutton et al. 2016). These cumulative developments underscored the need for a comprehensive update to the original PRISMA guidelines, ensuring their continued relevance and applicability in contemporary research contexts.

2.3 - Scope of the guidelines

PRISMA was initially developed with the objective of enhancing the transparency and completeness of reporting in systematic reviews and meta-analyses, primarily within the context of health and medical research (Hutton et al. 2016). However, over time, its scope has significantly broadened, extending to various other disciplinary fields, including social sciences, education, engineering, and technology.

The PRISMA 2020 statement was specifically updated to reflect methodological and terminological advances in systematic reviews over the past decade. Although it maintains a primary focus on reviews assessing the effects of health interventions, its guidelines have been structured to be applicable to systematic reviews of other types of interventions as well as to reviews with broader objectives. Moreover, PRISMA 2020 is relevant not only for systematic reviews that include a synthesis component (e.g., meta-analysis) but also for those that do not (e.g., when only one eligible study is identified).

The checklist also applies to mixed-method systematic reviews that incorporate both quantitative and qualitative evidence, although in these cases, additional guidelines on qualitative data synthesis should be consulted. Furthermore, PRISMA 2020 can be applied to original systematic reviews, updated reviews, or living (continuously updated) reviews. Nevertheless, it is not intended to guide the actual conduct of systematic reviews, for which comprehensive methodological resources are recommended. Importantly, PRISMA 2020 is not designed to assess the conduct or methodological quality of systematic reviews but rather to ensure a transparent and comprehensive report of the methods and findings (Matthew J. Page et al. 2021).

The expansion of PRISMA to areas beyond health research highlights its flexibility and utility in promoting clear and rigorous scientific communication across various fields.

The overall process described above is visually summarized in the PRISMA 2020 flow diagram (Figure 1), which outlines the study identification, screening, and inclusion phases, and allows for transparent documentation of the selection process in systematic reviews.

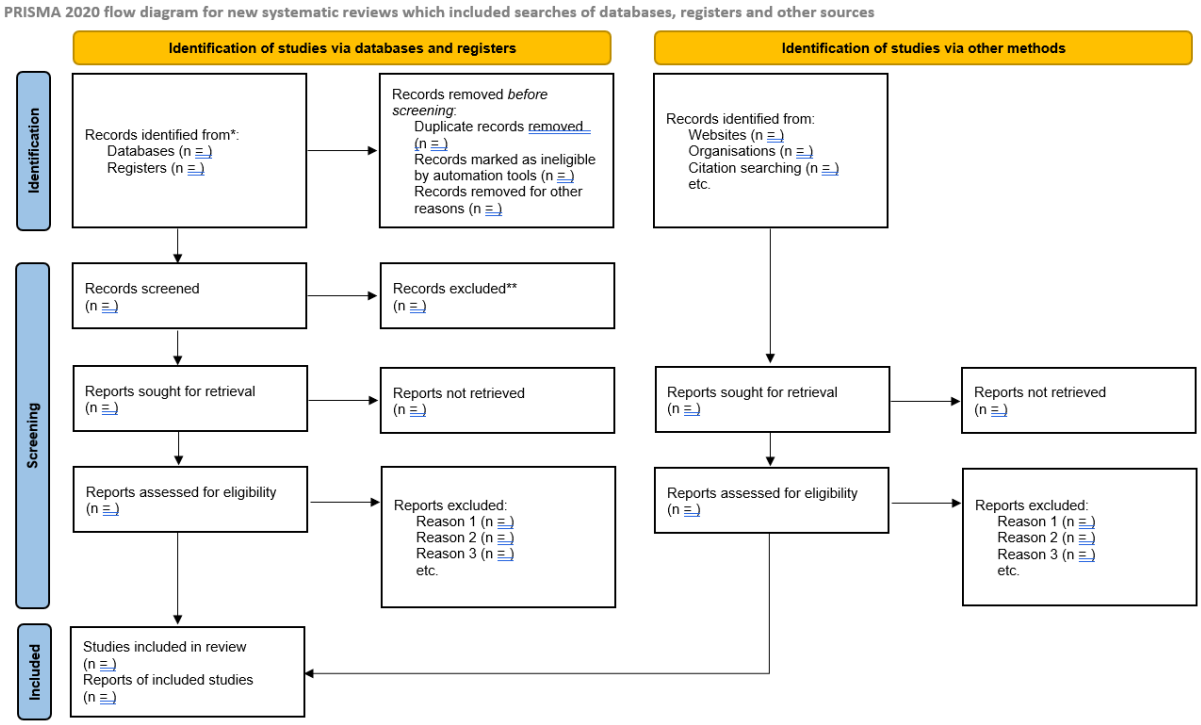

Figure 1 : PRISMA guidelines flowchart (Matthew J Page et al. 2021)

Chapter 3 – Methodology

A semi-automatic, unregistered systematic review was carried out in accordance with PRISMA guidelines to identify data quality criteria relevant to trustworthy AI applications in manufacturing. The methodological approach of this study was designed to ensure a comprehensive and systematic review of existing literature relevant to the research topic.

3.1 – Systematic Search

The data collection process was implemented using two major bibliometric databases: Scopus and Web of Science (WoS). The rationale for selecting Scopus and WoS over other databases such as PubMed, ACM Digital Library, and IEEE Xplore, lies in their broader and multidisciplinary coverage. While databases like PubMed are highly specialized in biomedical and life sciences research and ACM Digital Library and IEEE Xplore are predominantly focused on computer science and engineering, Scopus and WoS provide a more comprehensive spectrum of disciplines, including engineering, social sciences, management studies, and environmental sciences. This broader scope aligns more closely with the interdisciplinary nature of the current research. Additionally, both Scopus and WoS offer advanced search functionalities, citation tracking, and robust filtering options that facilitate more comprehensive and systematic literature reviews.

The following search string in pseudo-code was executed on Web of Science and Scopus. The search query is structured to identify academic literature that addresses data quality concerns within the context of artificial intelligence (AI), machine learning (ML), and deep learning. The query is divided into two main logical segments connected by the OR operator.

```
1. (("data quality" OR "data-quality"
2.
3. OR "data qualities" OR "quality of data"
4.
5. OR "quality of the data" OR "qualities of data"
6.
7. OR "qualities of the data" OR "quality of training
8.
9. data"
10.
11.       OR "quality of the training data" OR "quality of ML
12.
13.       data"
14.
15.       OR "data bias" OR "data biases"
16.
17.       OR "bias in the data" OR "biases in the data"
18.
19.       OR "data problem" OR "data problems"
20.
21.       OR "problem in the data" OR "problem with the data"
22.
23.       OR "problems with the data" OR "data error"
24.
25.       OR "data errors" OR "error in the data"
26.
27.       )
28.
29.       AND
30.
31.       ("dimension" OR "dimensions"
32.
33.       OR "AI" OR "artificial intelligence"
34.
35.       OR "ML" OR "machine learning"
```

```
36.
37.      OR "deep learning"
38.
39.      OR "neural network" OR "neural networks"))
40.
41.      OR
42.
43.      ("data quality framework" OR "data quality frame
44.
45.      works"
46.
47.      OR "framework of data quality" OR "framework for data
48.
49.      quality")
```

The first segment targets various expressions and synonyms associated with data quality issues. It includes a comprehensive set of terms such as *data quality, data biases*, *data problems* and *data errors*, ensuring broad coverage of potential data quality concerns. Additionally, it incorporates specific references to the quality of training data, highlighting its critical role in the development of AI/ML models.

The second segment contextualizes these data quality concerns by specifying relevant technological frameworks. Keywords such as *AI*, *ML*, *deep learning*, and *neural networks* are included to narrow the search to literature that discusses data quality issues in the context of these specific computational fields.

Moreover, a third segment, separated by the OR operator, is dedicated to capturing references to frameworks for assessing data quality. This includes expressions like *data quality framework* and *framework for data quality*, ensuring that systematic approaches to data quality are also considered in the search results.

The logical structure of the query is designed to ensure comprehensive coverage by combining data quality issues with technological contexts while also capturing systematic frameworks for assessing data quality.

For the data retrieval from Scopus, performed in March 2025, an advanced search query was utilized to refine the dataset according to specific keywords and research criteria. The search process yielded over 88,000 results. To further analyze and categorize these documents, the *Analyze results* function was employed, focusing specifically on the subject area of Engineering, which accounted for 13.3% (Figure 2) of the total documents, corresponding to about 20,000 entries.

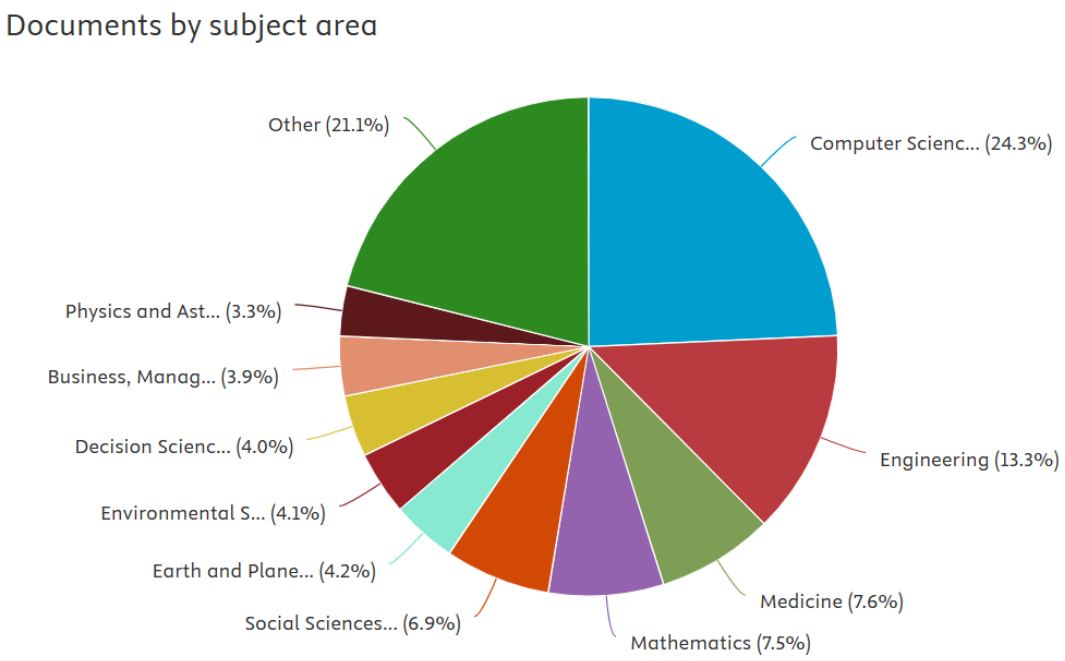

Figure 2: Documents distribution by subject area (Scopus 2025)

Analysing the search results, it is possible to observe the distribution of publications of interest over time, with a notable peak in the year 2024 (Figure 3).

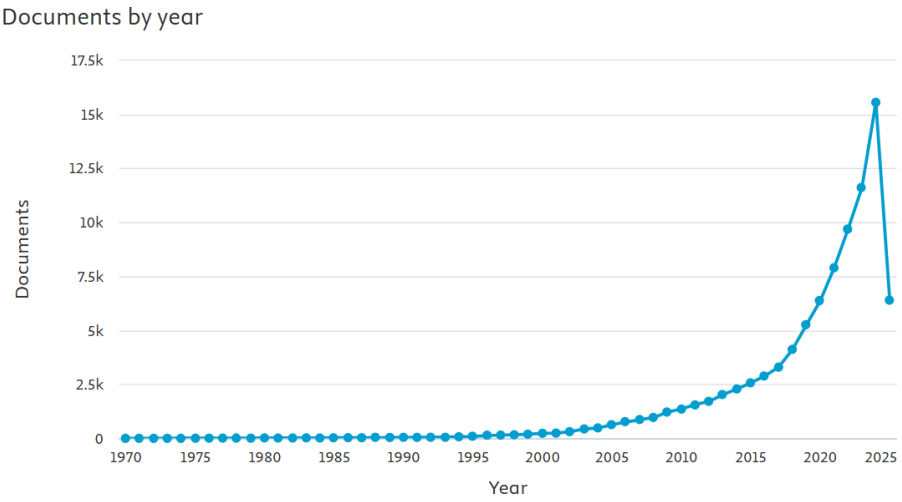

Figure 3: Annual publications distributions (Scopus 2025)

For the purposes of this study, metadata from the first 20,000 documents were extracted from Scopus and exported in CSV format. The exported metadata included the following fields: *Title*, *Year*, *DOI*, *Link*, *Abstract*, *Keywords*, *Publisher*, *Language of Original Document*, *and Document Type*, which were then organized as columns in the Excel tables. The CSV file was then formatted to remove delimiters and organize the data into structured columns for subsequent analysis. However, a significant issue emerged during this process: due to the CSV format, the presence of commas within certain metadata fields, such as the title or abstract, led to misalignment across columns. For instance, if a comma appeared in the title, it could shift all subsequent data fields, causing the *Year* value to appear under *DOI*, the *DOI* under *Link*, and so on. This issue was solved by implementing the following Python script to correctly parse the metadata and realign them into the intended column structure.

```
1. import pandas as pd
2.
3. def correggi_refusi(input_file, output_file):
4.     # Carica il file Excel
5.     df = pd.read_excel(input_file, engine="openpyxl")
6.
7.     # Controlla ogni riga per verificare se la colonna B contiene testo
```

```
    for index, row in df.iterrows():
        if pd.notna(row["Year"]) and not str(row["Year"]).isdigit():
                # Unisci il titolo (colonna A e B)
                df.at[index, "Title"] = str(row["Title"]) + " " + str(row["Year"])
                df.at[index, "Year"] = None  # Rimuove il testo errato dalla colonna B

                # Sposta gli altri dati se necessario
                if pd.notna(row["DOI"]) and str(row["DOI"]).startswith(("19", "20")):
                    df.at[index, "Year"] = row["DOI"]
                    df.at[index, "DOI"] = None

                if pd.notna(row["Link"]) and str(row["Link"]).startswith("10."):
                    df.at[index, "DOI"] = row["Link"]
                    df.at[index, "Link"] = None

                if pd.notna(row["Abstract"]) and str(row["Abstract"]).startswith("https://www"):
                    df.at[index, "Link"] = row["Abstract"]
                    df.at[index, "Abstract"] = None

                # Scala tutte le colonne successive
                df.at[index, "Abstract"] = row["Author Keywords"]
                df.at[index, "Author Keywords"] = row["Publisher"]
                df.at[index, "Publisher"] = row["Language of Original Document"]
                df.at[index, "Language of Original Document"] = row["Document Type"]
                df.at[index, "Document Type"] = row["Source"]
                df.at[index, "Source"] = None

            # Controlla se anche la colonna C contiene parte del titolo
            if pd.notna(row["DOI"]) and not str(row["DOI"]).startswith(("19", "20")) and not str(row["DOI"]).startswith("10."):
                df.at[index, "Title"] = str(df.at[index, "Title"]) + " " + str(row["DOI"])
                df.at[index, "DOI"] = None
```

```
                if pd.notna(row["Link"]) and str(row["Link"]).startswith(("19", "20")):
                    df.at[index, "Year"] = row["Link"]
                    df.at[index, "Link"] = None

                if pd.notna(row["Abstract"]) and str(row["Abstract"]).startswith("10."):
                    df.at[index, "DOI"] = row["Abstract"]
                    df.at[index, "Abstract"] = None

                if pd.notna(row["Author Keywords"]) and str(row["Author Keywords"]).startswith("https://www"):
                    df.at[index, "Link"] = row["Author Keywords"]
                    df.at[index, "Author Keywords"] = None

                # Scala tutte le colonne successive
                df.at[index, "Abstract"] = row["Author Keywords"]
                df.at[index, "Author Keywords"] = row["Publisher"]
                df.at[index, "Publisher"] = row["Language of Original Document"]
                df.at[index, "Language of Original Document"] = row["Document Type"]
                df.at[index, "Document Type"] = row["Source"]
                df.at[index, "Source"] = None

        # Salva il file corretto
        df.to_excel(output_file, index=False, engine="openpyxl")
        print(f"Correzione completata! File salvato come: {output_file}")

    # Esegui la funzione
    if __name__ == "__main__":
        correggi_refusi("input.xlsx", "output_corretto.xlsx")
```

Below is a detailed breakdown of the script's logic and implementation steps, illustrating how the data were programmatically corrected and reorganized.

1. Library Import and Function Definition:

The script begins by importing the pandas library, essential for data manipulation and analysis in Python. The function *correggi_refusi()* takes two arguments: *input_file*, the path to the Excel file to be corrected, and *output_file*, the path where the corrected file will be saved.

2. Loading the Excel File:

The function uses the *read_excel()* method to load the data into a DataFrame (*df*). The *engine="openpyxl"* parameter is specified to ensure compatibility with .xlsx files.

3. Iterating Through Rows:

The function iterates over each row of the DataFrame using *iterrows()*. This method allows the function to access both the index and the data in each row, enabling the manipulation of specific cell values.

4. Checking and correcting the *Year* Column:
    - The function checks whether the *Year* column contains a non-numeric entry. If so, it is assumed that this entry is part of the *Title* column.
    - The *Title* is then updated by concatenating the current *Title* value with the erroneous *Year* value, effectively combining both into a single text entry.
    - The *Year* column is then cleared by setting it to None.
5. Reassigning Data to Appropriate Columns:
    - The function checks subsequent columns to verify if any data has been incorrectly placed in the *DOI, Link,* or *Abstract* columns.
    - If the *DOI* column contains a date-like entry (e.g., starting with *19* or *20*), it is moved to the *Year* column.
    - If the *Link* column contains a DOI-like entry (e.g., starting with *10*), it is moved to the *DOI* column.
    - If the *Abstract* column contains a URL (e.g., starting with *https://*), it is reassigned to the *Link* column.

6. Shifting Columns:

After reassigning the *Year, DOI,* and *Link* columns, the function proceeds to shift data in the remaining columns to maintain logical consistency.

- The *Abstract* column is updated with the content from *Author Keywords*.
- *Author Keywords* is updated with the content from *Publisher*.
- *Publisher* is updated with the content from *Language of Original Document*.
- *Language of Original Document* is updated with the content from *Document Type*.
- *Document Type* is updated with the content from *Source*.
- The *Source* column is then cleared by setting it to *None*.

7. Handling Additional Misalignments in the *DOI* Column:

If the *DOI* column contains data that is not a valid DOI or date, it is treated as part of the *Title*. The function concatenates this text to the existing *Title* content and clears the *DOI* column.

8. Saving the Corrected File:

After processing all rows and adjusting the data as necessary, the corrected *DataFrame* is saved as a new Excel file using the *to_excel()* method. The file is saved without the index column, and the engine *openpyxl* is specified to ensure compatibility. A confirmation message is then printed to indicate the successful completion of the operation.

9. Function Execution:

At the end of the script, the function is executed within the *if __name__ == "__main__"* block. This ensures that the function only runs when the script is executed directly, not when it is imported as a module.

In the case of WoS, a standard search query was employed, resulting in a total of 10,155 records. The same fields were exported as in Scopus. Unlike Scopus, WoS provides direct export functionality in Excel format, facilitating the subsequent data processing and organization for analysis.

3.2 - Data cleaning and exclusion/inclusion criteria

Following the data export and initial formatting, the CSV files from Scopus and WoS were converted into Excel tables, allowing for more efficient data cleaning and filtering processes. The same procedure was applied to both datasets.

Starting with the Scopus dataset (i.e., 20,000 initial documents), the following filtering steps were conducted:

1. DOI filtering

During this phase, we made a crucial assumption: if an article lacked a DOI[1] and a link (URL DOI) or either of these fields, it was removed. Without a DOI code or its link, in fact, it is not possible to access the publication and analyse the document. Consequently, the Scopus dataset was reduced from 20,000 to 18,599 documents.

2. Language filtering

The *Language of Original Document* column was used to retain only those articles whose language was indicated as English or whose language field was empty. This step further refined the Scopus dataset to 18,583 documents.

3. Document Type filtering

The final filtering step targeted the 'Document Type' column. Records categorized as *article* or *review* or those with empty fields were retained, reducing the dataset to 17,413 documents. This selection was carried out to include only primary research articles and literature reviews, as these represent the most substantial and peer-reviewed sources of scientific evidence for a literature review. Other document types (e.g., *editorials, letters, conference abstracts* and *technical notes*) were excluded because they typically lack rigorous peer review and are usually non-open access (i.e., readable for free, without a particular additional subscription for the user).

[1] DOI, an acronym for Digital Object Identifier, allows digital objects to be uniquely identified and reliably accessed https://www.doi.org/index.html

A similar filtering procedure was applied to the Web of Science dataset, initially consisting of 10,155 documents. Following the removal of entries lacking a DOI or link (i.e., 1,219 documents), the dataset was reduced to 8,936 records. The subsequent language filtering step, focusing exclusively on English or empty language fields, further refined the dataset to 8,850 documents. Finally, by retaining only articles, reviews, and empty document type fields, the dataset was narrowed down to 6,427 documents, aligning the selection criteria with those applied to the Scopus dataset.

3.3 - Merging

After the filtering processes, the two datasets from Scopus and WoS were merged into a single table. Thus, a merger operation based on the DOI field was performed to identify and remove duplicates.

The combined dataset (i.e., from Scopus and WoS records), obtained after the filtering process, consisted of 23,840 documents. Records with equal DOIs were considered redundant and eliminated. In total, 1,775 duplicates were detected and deleted, resulting in a final dataset of 22,065 articles that served as the basis for subsequent screening and analysis. The flow diagram below (Figure 4) provides a visual summary of the data processing steps, from initial datasets to the final merged one.

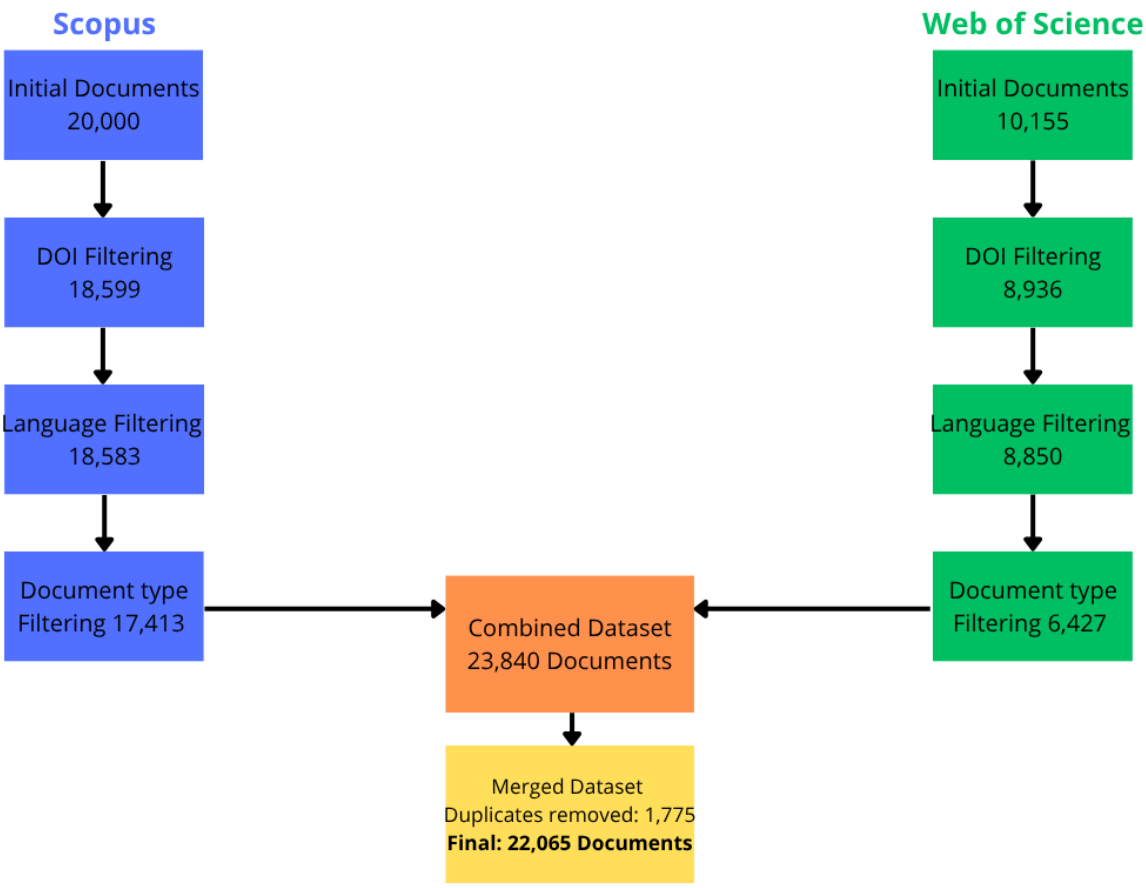

Figure 4: Document filtering and merging workflow (own elaboration)

3.4 - Screening

3.4.1 - Semantic Filtering Using AI-Based Text Analysis

At this stage, we obtained a consolidated Excel table containing 22,065 rows, each representing a unique document. The objective was to further filter these entries, retaining only those articles that exhibit semantic relevance to specific keywords associated with the research focus. Traditional keyword-based filtering would have been insufficient due to variations in terminology and phrasing. Thus, we employed a more advanced, AI-driven approach to semantic filtering using Python, leveraging libraries such as pandas and sentence-transformers.

As mentioned, the primary objective was to identify and retain only those documents that are semantically aligned with the predefined set of keywords (e.g., *machine learning*, *artificial intelligence, etc.*). However, a simple direct keyword search could potentially exclude relevant documents that discuss these topics using different terminology or contextual framing. To address this, we adopted a semantic filtering approach.

Rather than relying on explicit keyword matches, vector embeddings were utilized to capture the contextual meaning of text. This process is analogous to having a virtual assistant (hereafter referred to as the *agent*) who reads each abstract and assesses its conceptual relevance to the specified keywords.

The agent reads each abstract and automatically converts the text into a high-dimensional vector representation, capturing the semantic meaning rather than just the words themselves. Similarly, the keywords are also embedded as vectors, representing their conceptual meanings in the same vector space (for further theoretical background on embedding, see Section 3.4.3.1). The model then compares the embedding of each abstract with the embeddings of the keywords to assess semantic similarity.

The implementation was conducted using BERT, Bidirectional Encoder Representations from Transformers, (a brief theoretical overview of BERT is provided in Section 3.4.3.2), a transformer-based language model renowned for its ability to grasp contextual nuances in text. The specific automatic implementation performed by BERT model involved the following steps:

1. Data Preprocessing: Each abstract was extracted from the Excel table and converted into a text string for processing.

2. Embedding Generation: Using the BERT model, each abstract was converted into a 768-dimensional vector, representing its semantic content.

3. Similarity Calculation: The cosine similarity between each abstract's embedding and the keyword embeddings was calculated. This metric quantifies the degree of semantic alignment, with values ranging from -1 (completely dissimilar) to 1 (identical meaning).

4. Thresholding: A similarity threshold was defined to determine whether a document should be retained or not. If the similarity score exceeded the threshold, the document was considered semantically relevant and retained in the dataset, otherwise, it was excluded.

BERT model was selected due to its ability to capture bidirectional contextual meaning, as opposed to traditional unidirectional models. This capability enables it to effectively discern nuanced semantic connections between text segments, thereby improving the accuracy of the filtering process.

Bidirectional contextual meaning refers to the model's ability to consider both the preceding and following words in a sentence when interpreting the meaning of a specific word or phrase. Unlike traditional models that process text in a single direction (left-to-right or right-to-left), BERT simultaneously analyzes the context to the left and right of a target word, allowing it to understand nuanced meanings more accurately. This bidirectional approach enables BERT to better capture the full context and semantic relationships between words, resulting in more accurate language understanding and improved performance in tasks like text classification, question answering, and information retrieval.

3.4.2 - Implementation and Explanation of the Python Code

```
#1
!pip install torch sentence-transformers
!pip install tf-keras
#2
import pandas as pd
import torch
import os
import matplotlib.pyplot as plt
from sentence_transformers import SentenceTransformer, util

    # 3.1 Carica il file Excel
    df = pd.read_excel("WIP_merged.xlsx", sheet_name="merged")

```

```
    # 3.2 Crea una colonna "full_context" combinando titolo, abstract e keyword
    df['full_context'] = (
        df['Title'].fillna('') + '. ' +
        df['Abstract'].fillna('') + '. ' +
        df['Author Keywords'].fillna('')
    )

    # 4. Parole chiave da confrontare semanticamente
    parole_chiave = [
        "Data Quality", "data-quality", "Data qualities", "quality of data", "Quality of the data",
        "qualities of data", "qualities of the data", "Quality of the training data", "Quality of ML data",
        "Data bias", "data biases", "Bias in the data", "biases in data", "Data problem", "data problems",
        "problem in data", "problem with the data", "data error", "data errors", "error in data",
        "error in the data", "Dimension", "dimensions", "AI", "artificial intelligence", "ML",
        "machine learning", "Deep learning", "Neural network", "neural networks", "Data quality framework",
        "data quality frame works", "framework of data quality", "framework for data quality",
        "data reliability", "data integrity", "data consistency", "data accuracy", "data completeness",
        "trustworthy data", "quality of input data", "label quality", "skewed data", "sampling bias",
        "machine intelligence", "data-driven models", "deep neural networks", "automated learning",
        "ML for process optimization", "deep learning for visual inspection", "AI in supply chain",
        "AI for defect detection", "latency in decision making", "lack of explainability"
    ]

    # 5. Crea il modello per embedding semantico
    model = SentenceTransformer('all-MiniLM-L6-v2')

    # 6. Calcola gli embedding
```

```
40.     embedding_keywords = model.encode(parole_chiave, convert_to_tensor=True)
41.     embedding_context = model.encode(df['full_context'].tolist(),
    convert_to_tensor=True)
42.
43.     # 7. Calcola la similarità tra ogni documento e le parole chiave
44.     similarity_matrix = util.cos_sim(embedding_context, embedding_keywords)
45.
46.     # 8. Prendi la similarità massima per ogni documento
47.     max_similarities, _ = torch.max(similarity_matrix, dim=1)
48.     df['similarità'] = max_similarities.cpu().numpy()
49.
50.     # 9. Filtra i documenti con similarità ≥ soglia
51.     soglia = 0.45
52.     df_filtrato = df[df['similarità'] >= soglia]
53.     df_esclusi = df[df['similarità'] < soglia]
54.
55.     # 10. Statistiche
56.     print(f"Documenti totali: {len(df)}")
57.     print(f"Documenti rilevanti (similarità ≥ {soglia}): {len(df_filtrato)}")
58.     print(f"Documenti esclusi: {len(df_esclusi)}")
59.
60.     # 11. Visualizza la distribuzione delle similarità
61.     plt.figure(figsize=(8, 4))
62.     plt.hist(df['similarità'], bins=50, color='skyblue', edgecolor='black')
63.     plt.axvline(soglia, color='red', linestyle='--', label=f"Soglia = {soglia}")
64.     plt.title("Distribuzione delle similarità semantiche")
65.     plt.xlabel("Similarità")
66.     plt.ylabel("Numero di documenti")
67.     plt.legend()
68.     plt.tight_layout()
69.     plt.show()
70.
71.     # 12. Salva i risultati
72.     df_filtrato.to_excel("articoli_filtrati.xlsx", index=False)
73.     df_esclusi.to_excel("articoli_esclusi.xlsx", index=False)
74.
75.     # 13. Mostra il percorso dei file
```

```
76.     print("File salvati in:")
77.     print(os.path.abspath("articoli_filtrati.xlsx"))
78.     print(os.path.abspath("articoli_esclusi.xlsx"))
```

The following section outlines the structure and functioning of the implemented Python script, providing a step-by-step explanation of its main components and operations.

1. Library Installation

PyTorch is a comprehensive deep learning library developed by Facebook's Research Lab, widely utilized for tensor computations and model training due to its robust and flexible framework for building and deploying neural networks. The sentence-transformers library is an extension of *PyTorch*, specifically designed to generate semantic embeddings of sentences using pre-trained models, enabling efficient similarity calculations and clustering of textual data. Additionally, the script includes the installation of tf-keras, a high-level neural networks API that provides a simplified interface for constructing and training neural networks, although it is not actively employed in the current implementation.

2. Import Libraries and Read Data

The script leverages several key libraries to facilitate data processing and analysis. *Pandas* is employed for data manipulation and analysis, enabling the reading, transformation, and organization of data within structured DataFrames. *PyTorch* is utilized for tensor operations and model handling, providing a powerful framework for deep learning tasks and seamless integration with other machine learning libraries. The *os* module manages file paths and directory operations, ensuring efficient file handling and data storage throughout the script. For data visualization, *matplotlib.pyplot* is used to generate graphical representations of similarity scores, aiding in the interpretation of semantic analysis results. Lastly, the *sentence-transformers* library is employed to generate sentence embeddings and perform semantic similarity calculations, leveraging pre-trained models to effectively measure contextual alignment between textual data and predefined keywords.

3. Load Excel File and Create Full Context Column

The script begins by reading data from the *merged* sheet within the Excel file called *WIP_merged.xlsx*. To facilitate semantic analysis, a new column (*full_context*) is created by concatenating the content of three specific columns: *Title*, *Abstract*, and *Author Keywords*. This approach consolidates all relevant textual information into a single column, providing a comprehensive representation of each document's context. To prevent potential errors during the concatenation process, the. fillna*('')* method is applied to each column, replacing any missing values with empty strings. This ensures that the concatenation operation proceeds smoothly without generating null-related errors.

4. Define Keywords for Semantic Analysis

A comprehensive list of keywords designed to capture a broad range of expressions and terminologies associated with data quality, artificial intelligence (AI), and machine learning (ML). This list includes specific terms, synonyms, and variations to ensure comprehensive semantic coverage, enabling the identification of relevant content even when different phrasing or terminology is used across documents.

The keywords are carefully selected to cover essential aspects of data quality, such as *data accuracy*, *data consistency*, *data completeness*, *data integrity*, and *data reliability*. Additionally, it includes terms related to common data quality issues, such as *data bias, data errors***,** and *data problems*, ensuring that various types of data-related concerns are adequately represented.

In the context of AI and ML, the list incorporates phrases related to data usage in model training, such as *Quality of training data*, *Quality of ML data*, and *Label quality*. Furthermore, broader AI/ML concepts such as *deep learning*, *neural networks*, *machine intelligence*, and *automated learning* are included to capture the intersection of data quality within advanced computational frameworks.

The list also extends to more specific applications of AI and ML in industrial contexts, such as *ML for process optimization*, *AI for defect detection*, and *deep learning for visual*

inspection, reflecting scenarios where data quality issues can significantly impact model performance and decision-making.

By encompassing both general and context-specific terms, the keyword list ensures a comprehensive semantic analysis that not only identifies explicit references to data quality but also captures implicit mentions related to AI and ML applications. These keywords will be used to compute semantic similarity scores against the document content, allowing the script to effectively assess the relevance of each document based on its contextual alignment with the defined concepts.

5. Initialize Sentence Embedding Model

MiniLM is a transformer-based model that leverages the architecture of BERT but in a more compact and computationally efficient format. Unlike the full BERT model, which consists of hundreds of millions of parameters, MiniLM is designed to achieve similar semantic understanding with significantly fewer parameters, making it faster and less resource intensive. The *all-MiniLM-L6-v2* model is a compact, pre-trained model optimized for semantic similarity tasks. It generates dense vector embeddings for textual content, allowing for cosine similarity calculations between sentences. This makes it particularly suitable for large-scale semantic analysis, as it maintains robust contextual representation capabilities while minimizing computational overhead, aligning well with the objectives of the implemented filtering process.

6.Compute Embeddings for Keywords and Document Context

The *model.encode()* function plays a crucial role in the semantic analysis process by converting each textual input into a high-dimensional vector representation, known as an *embedding*. This transformation enables the script to capture semantic meaning in a numerical format, facilitating the calculation of similarity scores between texts.

In the context of this script, the *model.encode()* function is applied to two distinct datasets. First, it processes the list of keywords, transforming each keyword or phrase in *parole_chiave* into a tensor—a structured array of numerical values representing the semantic meaning of each keyword. This step effectively converts linguistic content into a mathematical form that can be systematically compared with other text embeddings.

Similarly, the *full_context* column, which consolidates the *Title*, *Abstract*, and *Author Keywords* for each document, is also processed using the same encoding function. Each document is converted into a corresponding tensor, creating a vector representation that encapsulates the semantic context of the entire text.

By embedding both the keywords and document content in the same high-dimensional space, the script enables direct comparison of their semantic proximity, allowing for the calculation of cosine similarity scores between them. This alignment of text and keywords in a shared vector space is fundamental to the subsequent similarity analysis.

7. Calculate Semantic Similarity

util.cos_sim() is a function provided by the sentence-transformers library. It computes the cosine similarity between two sets of embeddings, such as document embeddings and keyword embeddings. This results in a similarity matrix, where each row represents a document and each column represents a keyword. The values in the matrix range from -1 to 1, where 1 indicates maximum similarity and -1 indicates maximum dissimilarity.

8. Extract Maximum Similarity Score for Each Document

After calculating the similarity matrix using *util.cos_sim()*, the script proceeds to identify the highest similarity score for each document. This is achieved using the *torch.max()* function, which plays a crucial role in extracting the most relevant similarity score across all keyword embeddings.

The function *torch.max()* is specifically applied along the keyword dimension (columns) of the similarity matrix, as indicated by the parameter *dim=1*. This parameter instructs the function to locate the maximum value along each row, effectively identifying the highest similarity score for each document across all keywords.

The function returns a tuple containing two elements. The first element contains the maximum similarity scores for each document, representing the highest degree of semantic similarity with any of the predefined keywords. The second element contains the index of the corresponding keyword that generated the highest similarity score, though this index is not utilized in this particular implementation.

The maximum similarity scores are then extracted and stored in a new column named *similarità* in the DataFrame. The *.cpu()* method is used to convert the tensor to a NumPy array, enabling seamless integration with the DataFrame structure. The resulting similarity scores provide a quantitative measure of how closely each document aligns with the defined set of keywords, forming the basis for further filtering and analysis.

9. Filter Documents Based on Similarity Threshold

A similarity threshold of 0.45 is established to effectively filter relevant documents based on their semantic similarity to the predefined set of keywords. This threshold was determined through iterative testing and evaluation, during which different threshold values were assessed to identify the optimal balance between precision and recall.

A threshold of 0.45 was chosen as it provided a sufficient level of semantic alignment, allowing for the inclusion of documents that were contextually relevant to the targeted keywords without being overly restrictive. Lower thresholds, such as 0.3 or 0.4, tended to include too many unrelated documents, reducing the overall quality of the filtered dataset. Conversely, higher thresholds, such as 0.5 or 0.6, were too stringent and excluded potentially relevant documents that shared moderate but significant contextual similarities with the keywords.

Therefore, setting the threshold at 0.45 ensures the inclusion of a comprehensive yet manageable corpus of documents, optimizing the trade-off between capturing relevant content and minimizing noise. Documents with similarity scores equal to or above 0.45 are considered relevant and are saved in the *DataFrame df_filtrato*. Those with scores below the threshold are deemed less relevant and are saved in *df_esclusi*.

10. Display Statistical Information

The script provides statistical feedback regarding the filtering process. Specifically, a total of 2,763 documents were identified as relevant, having similarity scores equal to or above the threshold, and were saved in the DataFrame *df_filtrato*. Conversely, 19,302 documents with similarity scores below the threshold were classified as less relevant and saved in *df_esclusi*. This output highlights the effectiveness of the chosen threshold

in refining the dataset to focus on contextually aligned documents while excluding those less pertinent to the targeted keywords.

11. Visualize Similarity Distribution

A histogram is generated to visualize the distribution of similarity scores across all documents. A vertical red dashed line is drawn to indicate the similarity threshold (Figure 5).

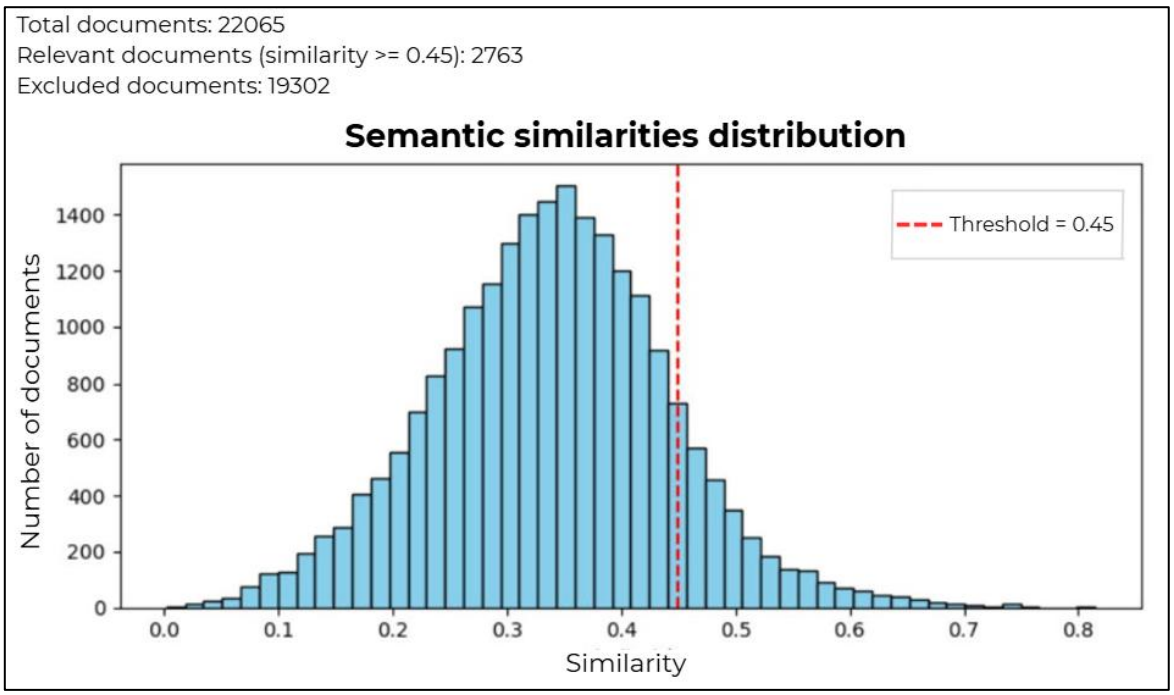

Figure 5: Semantic similarity distribution with threshold = 0.45 (Python code)

12. Save the Filtered Data

The filtered datasets (*df_filtrato* and *df_esclusi*) are saved as separate Excel files named *articoli_filtrati.xlsx* and *articoli_esclusi.xlsx* directly from the Python code.

13. Display Output File Paths

The absolute file paths for the saved files are printed to confirm the location of the output files.

3.4.3 - Theory recall: Embedding and BERT

To ensure a comprehensive understanding of the semantic filtering process adopted in this study, the following section outlines the theoretical foundations of the embedding technique and the BERT model, both of which constitute the key elements of the text analysis methodology applied.

3.4.3.1- Embedding

Unlike humans, who are capable of reasoning through abstract concepts and complex semantic structures, machines can only operate on numerical data represented in binary form. Therefore, enabling computers to process textual or otherwise non-numerical inputs requires an effective transformation of such data into numerical representations. This transformation must be performed at a level of granularity that allows the capture of semantic relationships embedded in natural language. Consequently, one of the fundamental challenges in natural language processing lies in developing representations of word meaning that are both computationally tractable and semantically informative.

Word embeddings emerged as a transformative solution. The technique embeds words in a space with a lot of dimensions, where each single word is encoded as a vector (a list of numbers). The key aspect of this technique is placing words with similar meanings in proximity to each other in this multi-dimensional space. The two words having similar vectors will likely be semantically close together. Interestingly, the direction from sets of close words (e.g., *king* to *queen*) can embed underlying relationships (e.g., *regality*).

An embedding is nothing more than a vector (N-dimensions) that tries to capture the meaning of a word or sentence, placing it in a vector space, also called semantic space (Almeida e Xexéo 2023). As vectors, they obey the inherent mathematics: they have a length, norm, and direction, and can be compared using measurement methods.

Moreover, embedding models can operate at various levels of granularity:

- Word embeddings (e.g., Word2Vec) focus on individual words or tokens.

- Sentence or document embeddings (e.g., BERT) aim to capture the full semantic content of longer inputs.

A clearer understanding of the functioning and relevance of word embeddings can be gained by examining how they are typically visualized and interpreted. In the absence of embedding techniques, a rudimentary approach to numerical representation would involve assigning a unique integer index to each word within a vocabulary of, for example, 10,000 terms. Based on this mapping, each word could be expressed as an *n*-dimensional vector, where *n* corresponds to the vocabulary size. In such a representation, known as *one-hot encoding* or 1-*of-N encoding* (Naseem et al. 2020), each word vector consists entirely of zeros except for a single element set to one, located at the index position assigned to that word. Figure 6 exemplifies how this might work.

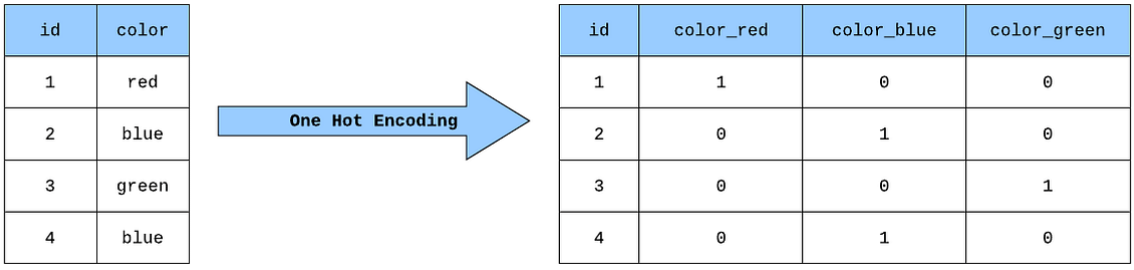

Figure 6: One-hot encoding example (Novack 2020)

However, this type of text representation, being purely symbolic rather than an embedding, suffers from major limitations in generalization and contextual relevance. In such a framework, a machine can only recognize whether a word exists in the vocabulary, without understanding its meaning or its relationships to other words. As a result, the representation is largely inadequate for real-world applications.

The primary goal of embeddings (as vector representations) and embedding models (which map text inputs to such vectors) is to overcome the limitations of sparse and context-free methods like one-hot encoding. These models aim to capture semantic content, contextual dependencies, and inter-word relationships by leveraging patterns

learned during training. As a result, machines are not only able to recognize a word's presence but also infer its meaning relative to other words in the language. By projecting words into a continuous, high-dimensional semantic space, embedding models enable generalization across similar linguistic units, supporting more sophisticated and context-aware processing of language (Naseem et al. 2020). For example, *apple* and *pear* (types of fruits) or *hammer* and *wrench* (types of tools) will be grouped together in that space. This capacity to abstract and reason about relationships is what has made embeddings so crucial to real-world NLP tasks. Figure 7plots a simplified representation of a vector space.

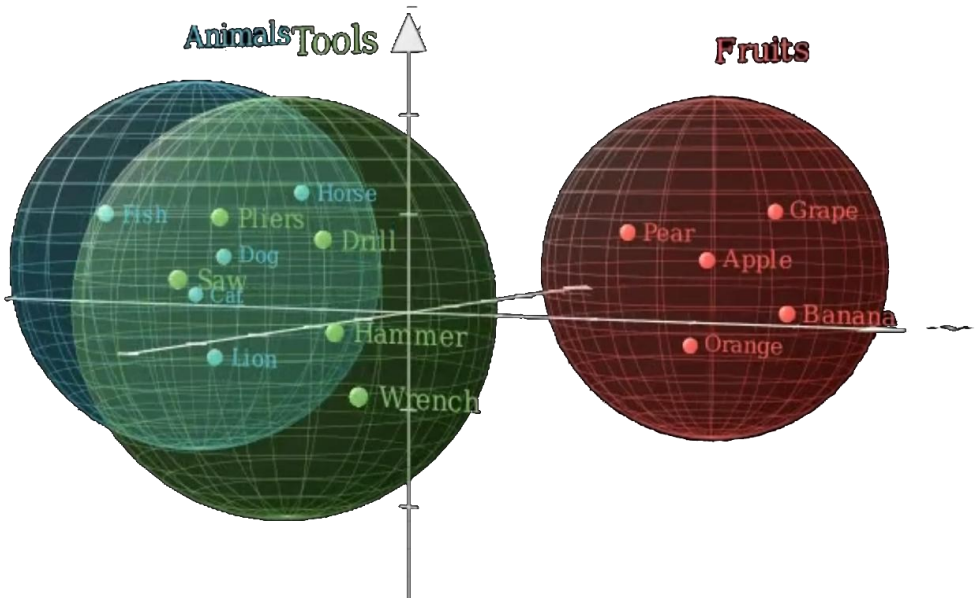

Figure 7: 3D semantic space with clustered categories (own elaboration)

A complete set of word embeddings exhibits several useful and non-trivial properties, enabling not only the recognition of semantically similar words but also the capture of complex linguistic relationships. One of the most notable features of a trained embedding space is its ability to group similar words in close proximity within an N-dimensional vector space. For example, terms such as *car*, *vehicle*, and *van* tend to cluster together, while remaining distant from unrelated terms like *moon*, *tree*, or *space*.

This spatial similarity can be quantified using metrics such as *Euclidean distance*, which measures the straight-line distance between two vectors, or *cosine similarity*, which evaluates the angle between vectors.

In addition to capturing *word similarity*, word embeddings can also model more abstract *linguistic relationships* through vector arithmetic. A classic example involves gender-based analogies: the vector difference between *man* and *woman* is similar to that between *king* and *queen*, or *uncle* and *aunt*. Such transformations illustrate how word embeddings encode not just semantic proximity, but also structured, interpretable relationships between concepts.

Word embeddings are produced by models, statistical or neural networks based, learning to represent words as vectors based on the patterns that occur in large collections of text data (Naseem et al., 2020). Such models are typically trained using unsupervised or self-supervised methods, i.e., they do not require labelled data. Instead, they leverage the distributional assumption that words with similar contexts would also have similar meanings. The datasets typically contain books, websites, human conversations, etc.

By discovering how to express every word as a point in higher-dimensional space, in which semantic relationships are represented in the distances and directions between vectors, these models learn to predict word co-occurrence in text.

3.4.3.2 - BERT: complex embedding model

At the end of 2018, a group of scientists from the Google AI Language laboratory presented a new linguistic model called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019). BERT is a pre-trained language model on a large corpus that uses the masked language modelling and next-sentence prediction objectives. The BERT family models are developed upon the Transformer encoder-decoder architecture (Vaswani et al. 2023). An encoder reads and understands input text by converting it into a numerical representation that captures its meaning. A decoder takes this representation and generates new text based on it, such as a summary or a

translation. Thus, while encoder-only models generate word embeddings, the decoder-only models can generate text (Figure 8).

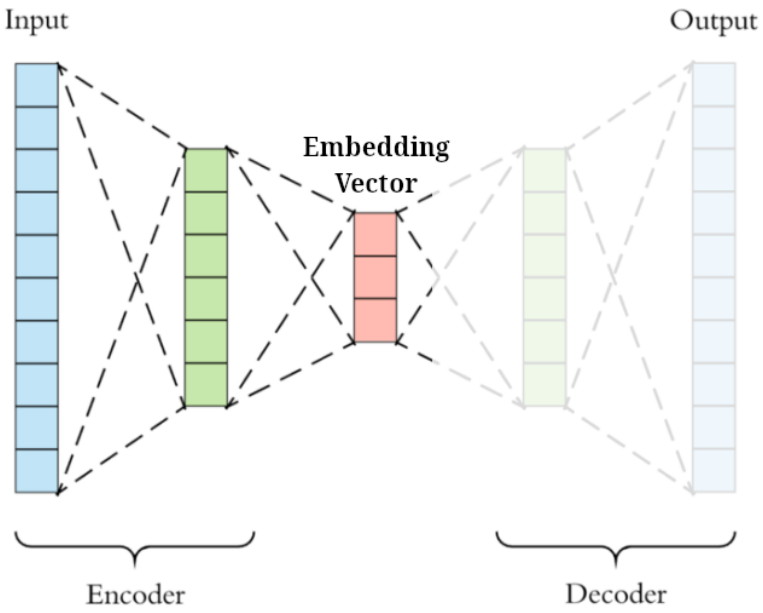

Figure 8: Encoder- Decoder architecture example (Yaron 2019)

However, BERT models have an encoder only architecture. In fact, they can serve various low-level NLP tasks: semantic search, clustering, sentence similarity, classification. While Word2Vec assigns each word a fixed vector based on its general usage across a corpus - meaning it doesn't change depending on context - BERT dynamically generates embeddings for words based on the entire sentence they appear in. This allows BERT to capture in-context nuances and understand the semantic relationships between words more accurately. Here is an example:

Imagine the word *bank* in two different sentences:

1. *"He sat on the river bank and watched the water flow."*
2. *"She deposited money into the* bank *yesterday."*

A traditional model might treat the word *bank* the same way in both sentences; however, BERT understands context bidirectionally — it examines the entire sentence (left and right of the word) to determine meaning.

So, BERT will interpret:

- In sentence 1, *bank* = riverbank.
- In sentence 2, *bank* = financial institution.

As previously mentioned, one of the major applications of word embeddings are *semantic searches*. Imagine you have a collection of articles about various topics, and you want to find articles that are semantically related to a search query. For example, consider the search query:

Tips for teaching my dog commands.

The traditional web search engine process, based on keywords matching, may fail to retrieve relevant information, leading to unrelated or poorly related articles based on your query.

Semantic search using models like BERT, however, understands the meaning of the query. It does not just search for the word *commands* but understands that your query is asking for training tips related to dogs. This will result in a more precise answer to the user.

To demonstrate the practical application of semantic search using transformer-based language models, the following Python implementation leverages BERT to encode both user queries and textual data into dense vector representations. These embeddings are then used to compute semantic similarity, enabling the retrieval of contextually relevant results beyond simple keyword matching.

3.5 - Performance Metrics of Semantic Filtering Process

To evaluate the effectiveness of the semantic filtering process applied to the document corpus, a two-step validation approach was adopted, combining algorithmic selection with human judgment. In particular, the aim of *Step 1* is to validate the algorithm in

selecting the relevant papers, while the objective of *Step 2* is to check manually all the relevant selected papers from the model. The semantic filter was based on a truncated Gaussian distribution applied to the similarity scores produced by a BERT-based model, which excluded documents in the left portion of the distribution and retained only those exceeding a predefined threshold on the right tail (Figure 5).

The initial dataset comprised 22,065 documents, resulting from the integration and deduplication of bibliographic records extracted from Scopus and Web of Science. The semantic filtering algorithm excluded 19,302 documents, while 2,763 documents were retained as potentially relevant based on their similarity to a predefined set of keywords related to data quality, artificial intelligence, and machine learning.

In order to validate the algorithm (*Step 1*), a manual screening was conducted on approximately 10% of the total corpus, amounting to 2,370 documents randomly sampled across both included and excluded sets (i.e., $\frac{2{,}370}{22{,}065} \approx 10\%$). Each document was assessed by examining its Title, Abstract, and Author Keywords.

Among manually examined documents, 38 were identified as relevant, while the remaining 2,332 were classified as irrelevant (i.e., $337 + 1{,}995 = 2{,}332$ irrelevant documents).

This first phase provided an initial estimation of the alignment between the semantic filtering algorithm and human evaluation, serving as a key reference point for the calculation of classification metrics.

The second validation phase (*Step 2*) consisted of a complete manual review of the 2,763 documents selected by the Python semantic filtering algorithm. Among these, 375 documents had already been reviewed during the first manual evaluation phase (*Step 1*). In particular, 337 had been classified as irrelevant and 38 had been confirmed as relevant to the purpose of the search, as in Figure 9. To avoid redundancies, these previously analyzed documents were excluded from the second screening phase.

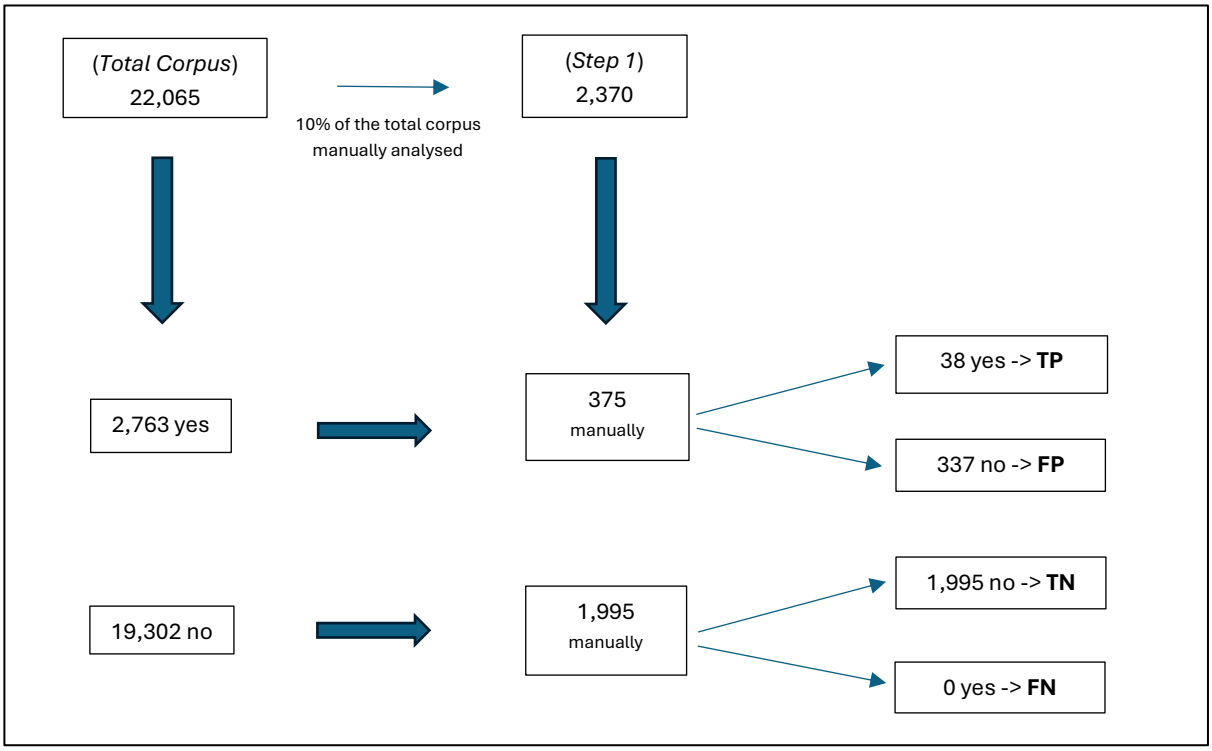

Figure 9: Model output with manual validation on 10% of the total corpus (own elaboration)

The remaining 2,388 documents (among those included by the Python algorithm), which had never been evaluated before, were subjected to a manual evaluation, applying the same assessment criteria as in *Step 1*. The manual analysis resulted in confirmation that 126 documents were correctly included in the analysis, while the remaining 2,262 were considered irrelevant to the defined inclusion criteria.

Combining the results from both steps, the total number of documents correctly identified as relevant is:

$$38\ (Step\ 1) + 126\ (Step\ 2) = 164\ documents$$

Accordingly, the remaining 2,599 documents included by the filter were manually assessed as not relevant.

To evaluate the performance of the semantic filter using standard classification metrics, the following definitions were adopted:

- True Positives (TP): Documents included by the semantic filter and confirmed as relevant during manual review → 38
- False Positives (FP): Documents included by the filter but manually rejected as irrelevant → 337
- True Negatives (TN): Documents excluded by the filter and manually confirmed as irrelevant → 1,995
- False Negatives (FN): Documents excluded by the filter but manually judged relevant → 0

This classification enables a detailed quantitative assessment of the filter's behaviour.

Table 1: Confusion matrix components from manual analysis

Category	Count [documents]
True Positives (TP)	38
False Positives (FP)	337
True Negatives (TN)	1,995
False Negatives (FN)	0

Based on the classification above, the following metrics are derived:

- Accuracy $= \frac{TP + TN}{\mathrm{TP + FP + FN + TN}} = \frac{38 + 1{,}995}{2{,}370} \approx 85.78\ \%$
- Precision $= \frac{TP}{TP + FP} = \frac{38}{38 + 337} \approx 10.13\ \%$
- Recall $= \frac{TP}{TP + FN} = \frac{38}{38 + 0} \approx 100\ \%$

The performance metrics obtained reveal a high accuracy (≈ 85.78%) and a perfect recall (100%), indicating that the semantic filter was highly effective in identifying nearly all documents considered relevant by human evaluation. Accuracy, which can be defined as the proportion of correct classifications over the total number of instances in the

dataset, provides a general indication of overall model performance. However, it does not distinguish between different types of mistakes.

Conversely, recall, defined as the ratio of true positives to the sum of true positives and false negatives, captures the model's ability to retrieve all relevant items. The perfect recall observed here suggests that the filter successfully retained every document judged relevant by human screening.

On the other hand, precision, calculated as the proportion of true positives out of all positive predictions, is notably low (≈ 10.13%). This indicates that, although the filter retrieved all relevant documents, it also included a high number of irrelevant ones, which were subsequently rejected during manual validation. This outcome is consistent with the choice made during the implementation of the semantic filtering algorithm, where a relatively broad inclusion threshold (set at 0.45 on the similarity score distribution) was deliberately adopted. While this threshold focuses on the right tail of the truncated Gaussian distribution (Figure 5), it still retains a relatively large portion of documents in order to maximize recall. This strategic choice was guided by the objective of ensuring that potentially relevant documents would not be prematurely excluded, thus favouring a more inclusive filtering phase that could subsequently be refined through manual screening. If a more restrictive threshold had been applied, the algorithm would probably have achieved greater precision by reducing the number of irrelevant documents incorrectly included. However, this would have occurred at the cost of a lower *accuracy* and potentially a loss of relevant documents, an outcome perceived less desirable given the exploratory nature of the literature review and the need to ensure comprehensive coverage of the topic.

To further assess the reliability of the semantic filtering process, a statistical interpretation of the algorithm's performance was carried out by estimating Type I and Type II error rates. In this context, the null hypothesis (H_0) is: *A document included by the semantic filter is actually relevant to the analysis*.

Rejecting this hypothesis when it is true constitutes a Type I error (α), or a false positive; in this case, the algorithm incorrectly classifies a document as relevant even though it is

not actually pertinent to the analysis. Conversely, a Type II error (β) occurs when the null hypothesis is false, but it is not rejected: this results in the algorithm failing to identify and includes a document that is indeed pertinent to the research objectives.

Based on the manually validated dataset, the estimated Type I error (α) reflects the proportion of documents incorrectly classified as relevant among those included by the filter (the false positives). Out of a total of 375 documents included by the semantic filter and manually reviewed, 337 were identified as false positives. Therefore, the Type I error rate is calculated as:

$$\alpha = \frac{FP}{TP + FP} = \frac{337}{38 + 337} \approx 89.87\ \%.$$

This high value indicates that a large proportion of the documents selected by the algorithm were not considered relevant upon manual inspection. However, this outcome is consistent with the design strategy adopted in the filtering phase, which intentionally prioritized inclusivity to reduce the likelihood of missing relevant content.

The Type II error (β), on the other hand, corresponds to the proportion of relevant documents that were excluded by the filter (the false negatives). In this case, no relevant documents are missed as all documents identified as relevant through manual validation have already been included by the semantic filter. As a result, the number of false negatives (FN) is zero, and the Type II error rate is calculated as:

$$\beta = \frac{FN}{FN + TP} = \frac{0}{0 + 38} \approx 0\ \%.$$

This outcome confirms that the semantic filter successfully captured all relevant documents within the corpus. The *recall* of the system is therefore maximized, which was precisely the intended effect of setting a relatively inclusive similarity threshold (i.e., 0.45).

Although this strategy resulted in a substantial number of false positives, reflected in a low precision, it effectively guaranteed that no relevant literature was inadvertently excluded. Such an outcome represents an intended and acceptable trade-off considering the review's exploratory nature and its aim to achieve comprehensive coverage of the topic.

3.6 - Full Text Analysis

In accordance with PRISMA methodology, the next phase of the document screening process corresponds to the *Eligibility* step, which involves a full-text analysis of the documents previously identified as relevant (i.e., 164 documents).

Following the screening process, a total of 164 documents are retained based on the combined outcomes of the manual validation procedures: 126 documents are identified during the second-level screening (*Step 2*) and 38 are confirmed during the initial sample assessment (*Step 1*).

Each of these documents is subjected to a full-text eligibility check to verify their actual suitability for inclusion in the final dataset. This step is aimed at ensuring that methodological content, thematic alignment, and level of detail provided by the studies were consistent with the research objectives. During this process, it was found that 28 documents were not accessible in full text due to access restrictions. The unavailability was primarily imputable to technical issues related to DOI resolution and the presence of paywalls requiring additional subscriptions. As a result, the number of documents eligible for in-depth analysis was reduced to 136.

These 136 documents represent the final set of sources on which the qualitative and content-based analyses are conducted in the subsequent stages of the research.

The purpose of this step is not to examine the methodological design or research strategies employed by the authors, but rather to extract meaningful insights concerning data-related issues within the scope of artificial intelligence and machine learning in manufacturing settings.

More specifically, the analysis is aimed at identifying two distinct but often overlapping dimensions within each paper. The first concerns problems and challenges reported in management, quality, or structure of data: these are issues which tend to persist over time and technological evolutions (*RQa*). The second dimension relates to the solutions, techniques, and frameworks proposed or adopted by the authors to address those issues (*RQb*). This separation is critical, as it enables a clearer understanding of which

obstacles are structurally rooted, and which are being actively mitigated through evolving technological solutions.

This dual-level reading of each article was guided by a structured analysis framework, implemented in the form of a spreadsheet, where each column corresponded to a specific analytical variable (Figure 10).

The columns were designed to capture the following dimensions:

- *Relevance*: A binary classification indicating whether the paper was considered relevant (*yes*) or not (*no*) for answering the research questions.
- *Problems/Challenges*: This column identifies the main issues or limitations addressed in the paper, which may concern various aspects such as data, methodology, implementation, or application context.
- *Techniques/Solutions/Tools*: Techniques, methods, or tools proposed or discussed by the authors to address the identified challenges.
- *Research Questions (RQ)/Aim of the paper*: This field summarizes the research objectives or explicit questions stated by the authors, offering alignment with our *RQa* and *RQb*.
- *Results*: A brief synthesis of the main findings or conclusions reached by the study, focusing on data-related aspects.

Is the paper relevant for the topic under investigation?	Problems/Challenges	Techniques/Solutions/Tools	RQ(s)/Aim of the paper	Results
no			"what data is dirty?," "why is it dirty?," and "how does a particular piece of data contribute to the overall dirtiness?	
si	• Types of data • Data heterogeneity • Data imbalance • Data scarsity • Labeling data (to train ML models)	• Data augmentation: artificially expanding the data volume and possibly enhancing the data diversity for better model generalization • Active learning • Adaptive sampling	RQ: *How can data quality be evaluated and improved in ML-based design and manufacturing?* *RQ1 What definitions, concepts, frameworks, and techniques are frequently utilized in this domain?* *RQ2 What are the dominant data challenges in this domain?* *RQ3 What are the concepts developed to investigate and evaluate data quality regarding the identified data challenges?* *RQ4 What are the techniques developed to resolve the identified data challenges?* *RQ5 How do the techniques compare and how applicable are they in this domain?* *RQ6 What are the status, trends, and future directions for the surveyed data quality improvement techniques?*	(1) The data handling techniques, terminologies, and challenges in ML-based design and manufacturing are investigated. (2) The data quality concepts, such as data quality, data readiness, and information quality (InfoQ), as well as their applications in this domain, are reviewed. (3) The data imbalance and biases in design and manufacturing data sets are analyzed. (4) We present and discuss data quality improvement and bias mitigation techniques, focusing on data augmentation and active learning.

Figure 10: Example of Excel spreadsheet for Full text analysis (own elaboration)

This allowed for a systematic and replicable review of each document, facilitating the identification of recurring themes, emerging trends, and gaps in current practices. The collected evidence serves as the basis for mapping common data-related problems, contextualizing them across different manufacturing environments, and observing how the field is conceptually and practically responding to such challenges.

Among the 136 documents selected for in-depth analysis, several metadata fields extracted from bibliometric databases were found to be incomplete or missing, such as *document type*. Although the initial inclusion and exclusion criteria were clearly defined, some documents, such as short conference papers, were nevertheless retained in the dataset due to their thematic relevance. However, these documents often lacked sufficient in-depth analysis and did not provide substantial analytical value.

Of the 136 documents, 25 were analysed entirely manually, while the remaining 111 were first examined with the support of ChatGPT, followed by manual validation. This methodological choice aimed to establish a sufficiently robust baseline of manually analysed documents to serve as a benchmark for subsequent comparison. The underlying objective was to verify the degree of consistency between human interpretation and the outputs generated by ChatGPT. To this end, approximately one fifth of the overall corpus (25 out of 136 documents) was initially examined exclusively through manual review and subsequently processed using the same prompts in ChatGPT. Once a satisfactory correspondence between the manual and AI-assisted analyses was observed, the remaining documents were analysed directly with the support of ChatGPT and subjected to subsequent validation.

ChatGPT was employed exclusively as a support tool, with each document being individually processed and the tool was used to extract structured information corresponding to predefined Excel columns: *Application Area*, *Problems/Challenges*, *Techniques/Solutions/ Tools*, *RQ(s)/Aims of the paper* and *Results* (as per columns title in Figure 10).

The reason behind the choice of using ChatGPT as a support tool is that large language models have recently demonstrated strong effectiveness in tasks comparable to literature analysis, such as abstract screening in systematic reviews. In this context,

ChatGPT v4.0 achieved excellent performance, with overall accuracy above 90% and balanced levels of sensitivity and specificity, while drastically reducing the time and cost of manual evaluation. According to Li (Michael Li et al. 2024), these results highlight the potential of ChatGPT as a reliable assistant in supporting, rather than replacing, human evaluation. This process helped guide manual reading and improve review efficiency without compromising critical evaluation.

In accordance with the PRISMA guidelines, the full-text analysis was conducted by applying a set of predefined eligibility criteria to assess the relevance of each document. The inclusion criteria required that the study:

- Addresses generalized and transferable concepts of data quality, particularly in relation to its impact on AI systems in manufacturing contexts.
- Explicitly evaluates or discusses how data quality influences the trustworthiness or performance of AI-driven applications.

The exclusion criteria, consistent with those previously applied during earlier screening phases (as detailed in previous sections), included:

- Studies whose primary focus was not data quality.
- Contributions lacking sufficient methodological or conceptual depth to support the research objectives (e.g. short or unstructured abstracts, editorial notes, or promotional content).

Based on the application of these criteria, the full-text analysis of the 136 selected documents led to the following categorization:

62 documents were considered irrelevant or only marginally related to the research questions and were therefore excluded from further analysis.

33 documents were classified as case studies or highly specialized researches, typically focused on narrowly defined applications or sector-specific implementations. Although thematically related to the broader topic of AI and data quality, these studies exhibited a high degree of verticality, namely, they addressed highly specific use cases, technologies, or industrial contexts that limited their generalizability. For example,

several short conference papers included in this category explored niche applications with limited methodological transferability or theoretical depth.

41 documents were deemed fully relevant and aligned with the research objectives. These studies addressed both *RQa* and *RQb* in a clear and substantive way and provided meaningful insights into data-related challenges as well as the corresponding AI-based solutions. In addition to thematic alignment, these contributions were characterized by a sufficient level of generality and abstraction, which made their findings applicable across a range of manufacturing contexts. Hence, they represent the empirical foundation of the analysis presented in Chapter 4.

Chapter 4 - Results

The 41 papers identified as relevant for this analysis span a period from 1995 to 2025, offering a comprehensive overview of how definitions, technologies, and artificial intelligence approaches to data quality in manufacturing have evolved over the last three decades. The chronological and thematic examination of these works allows for a nuanced understanding of both conceptual developments and practical implementations, as well as the persistent challenges that continue to shape the field.

This analysis is guided by the following research question:

How is artificial intelligence currently being applied to assess and enhance the reliability of data in manufacturing processes, and what are the key challenges and gaps in the existing literature?

In addressing this overarching question, two sub-questions are considered:

- *(RQa) Which dimensions of data quality are most frequently addressed in AI applications for manufacturing? Along which characteristics should data quality be evaluated when employing a dataset for trustworthy AI in manufacturing?*
- *(RQb) What are the main AI techniques used to evaluate or improve data reliability in this context?*

To provide a clear and structured narrative, the results are organized into three chronological periods that correspond to major shifts in focus and technological capability. It should be noted that these periods are not rigid or mutually exclusive. Rather, they provide a heuristic structure to highlight predominant trends over time. In practice, overlaps exist: some recent works (e.g., surveys or conceptual reviews) revisit early discussions on the definition of data and data quality, while certain methodological advances anticipated in later stages can already be observed in earlier contributions. The chronological division thus serves as an analytical framework rather than a strict categorisation, enabling a clearer understanding of how different themes have evolved and interacted over the last three decades.

1. 1995–2010: Early efforts to establish shared definitions of *data* and *data quality*, highlighting the absence of a universal standard and the reliance on foundational attributes such as accuracy, completeness, and consistency.
2. 2010–2020: Expansion into the domains of Big Data and the Internet of Things (IoT), with an emphasis on scalability, interoperability, and the integration of new dimensions of data quality into manufacturing systems.
3. 2020–2025: Advanced applications of machine learning (ML), deep learning (DL), and AI, embedding continuous assessment and improvement of data reliability into complex manufacturing environments.

4.1 - Early Period (1995–2010): Foundational Definitions and Initial Approaches

In the earliest years covered by this review, literature predominantly focused on building a conceptual foundation for what would later become the broader discourse on *data quality in manufacturing*. The central concern was to establish clear, operational definitions of *data* and *data quality*, often drawing from parallel domains such as information systems, database management, and software engineering.

The term *data* was generally described as recorded values representing facts, events, or measurements, which could be structured, semi-structured, or unstructured depending on their origin and format (Wang, Strong 1996). In the manufacturing domain, the concept acquires greater specificity, as data were frequently generated by sensors, control systems, and manual entry processes, each with distinct characteristics and potential sources of error. These distinctions were important for understanding the types of quality challenges likely to arise in different manufacturing contexts (Figure 11).

Basis	Data Types	Description
Structure	Structured data	Data with formal schema definition; (e.g., relation tables)
	Unstructured data	Generic sequence of symbols (e.g., video)
	Semi-structured data	Data partly structured or have a descriptive without schema (e.g., XML file)
Change frequency	Stable data	Data impossible to change
	Long-term changing data	Data with very low frequency of change
	Frequently changing data	Dramatically changing data, (e.g., real-time traffic information)
Product	Raw data items	Data that have not been processed
	Information products	Results of manufacturing activities
	Component data items	Semi-processed information
Nature	Federated data	Data from different heterogeneous sources
	Web data	Data from the Web
	High-dimensional data	Big data
	Descriptive data	Consists of many tables with complex interrelationships.
	Longitudinal data	Time series data
	Streaming data	Data generated sequentially at a higher rate in a single source

Figure 11: : Classification for data (Batini et al. 2009)

Data quality in this early period was not yet supported by a universally accepted definition. One of the most influential and enduring conceptualizations was the idea of *fitness for use* (Wang e Strong 1996), which framed *data quality* as a relative and context-dependent property: data is said to be of high quality if they meet the needs of the specific task, decision or process it is intended to support. This perspective underscored that quality requirements are dynamic, shaped by evolving operational contexts, the gradual accumulation of data in repositories, and changing stakeholder expectations.

Within this conceptual frame, the attributes most frequently emphasized were *accuracy*, *completeness* and *consistency*, often complemented by *timeliness* and *relevance* as additional indicators of usability (Wang e Diane M. Strong 1996; Redman 1998). These attributes provided the initial operational foundation for assessing data quality, serving as reference points for both academic research and early industrial applications.

Over time, formal standards began to address the issue more explicitly. The ISO 8000 data quality standard was developed to provide a structured approach for assessing and improving data quality across the product life cycle, from conceptual design to disposal. ISO 8000 defines quality characteristics for data, offers a framework for improvement, and can be applied independently or alongside broader quality management systems. Its structure encompasses *general principles*, *master data quality* (including *syntax*, *semantic encoding*, *provenance*, *accuracy* and *completeness*), *transaction data quality*, and *product data quality* (Figure 12).

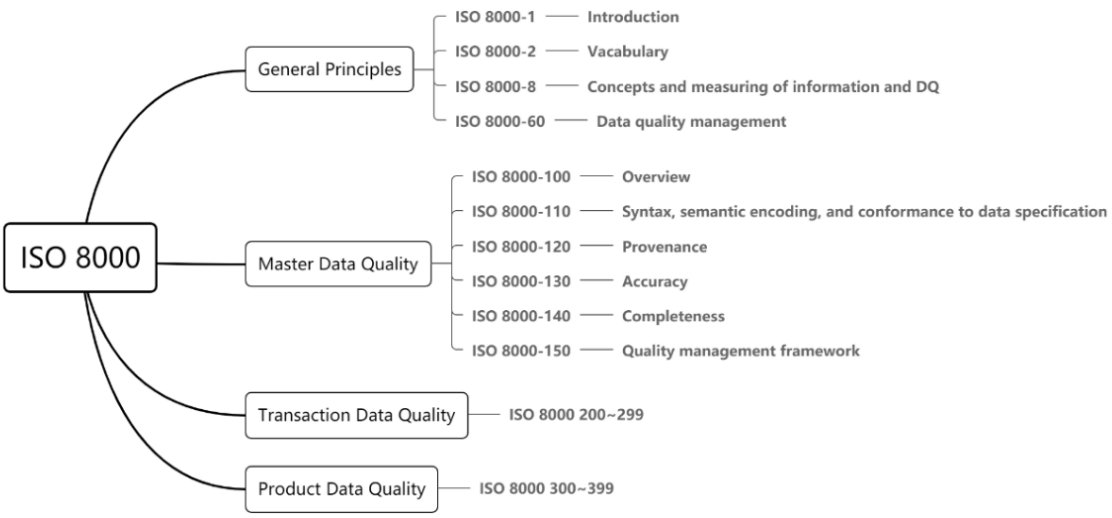

Figure 12: Components of ISO 8000 (Batini et al. 2009)

The conceptual roots of *data quality* also draw from the broader definition of *quality* in ISO 9000, which describes it as the degree to which customer requirements are satisfied (Nikiforova 2020). By extension, *data quality* is understood as the degree to which data meets the requirements of their intended use, reflecting both their inherent properties and their suitability for the context in which they are applied. This definition highlights the inherently relative and dynamic nature of data quality, a property that can change over time as data evolve, accumulate, or are repurposed for new applications.

The focus on inherent dimensions of data quality, *accuracy*, *completeness*, and *consistency*, was a natural reflection of the stage of development. Evaluation methods in this early stage were typically manual or rule-based, with a limited set of metrics applied to discrete datasets. The process followed a linear lifecycle, moving from data generation, often from multiple heterogeneous sources, to acquisition, storage, and finally analysis (Figure 13). At each stage of this lifecycle, there were potential risks of quality degradation, such as errors in collection, transmission delays, incomplete storage, or inaccuracies introduced during processing and visualization. This highlights that data quality is not a static property, but one that can be affected at any point in the data's journey from source to use.

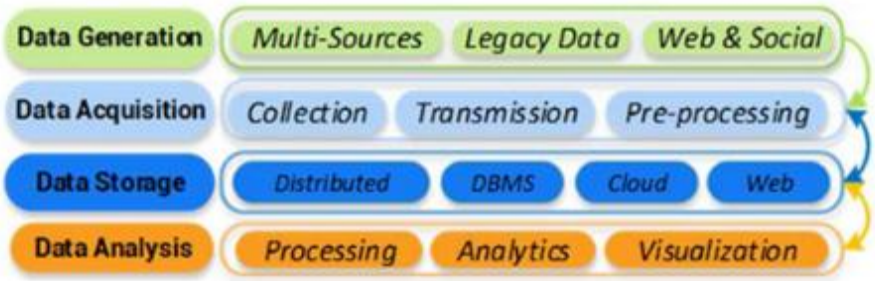

Figure 13: Data lifecycle Value chain (Taleb et al. 2021)

In terms of quality dimensions, early approaches concentrated on intrinsic properties such as *completeness*, *consistency*, *accuracy*, and *timeliness*. Over time, these were complemented by broader categories including contextual dimensions (e.g., *believability*, relevancy, *value-added*, *accessibility*, *reputation*) and representational dimensions (e.g., *interpretability*, *manipulability*) (Figure 14). This expansion reflects the growing recognition that data must not only be correct, but also meaningful, relevant, and usable within the specific operational context of manufacturing.

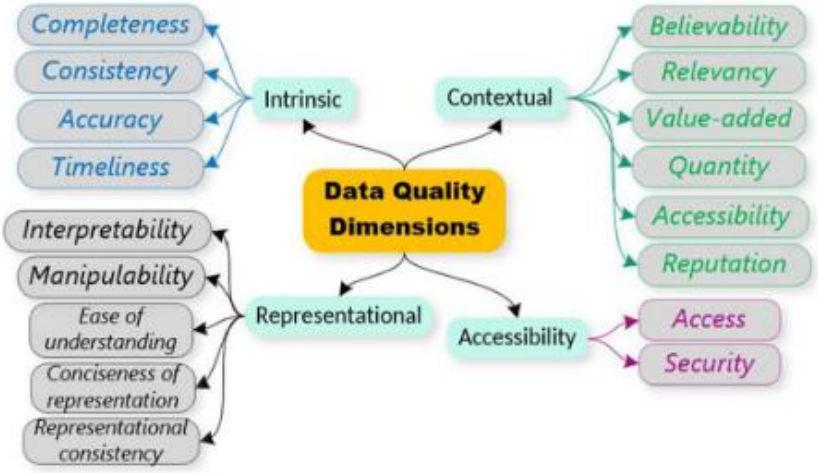

Figure 14: Early conceptual model of Data quality dimensions (Taleb et al. 2021)

Despite the conceptual advances made during this period, several challenges persisted. First, the absence of standardization in defining and measuring data quality continued to hinder the comparability of results and the development of universally applicable frameworks. In addition, the heterogeneity of data sources, ranging from sensor readings and control system logs to manually recorded information, posed significant integration difficulties, often leading to inconsistencies and information loss. Finally, the level of automation in quality assessment remained limited, with few solutions capable of providing real-time evaluation and corrective actions.

4.2 - Intermediate Period (2010–2020): Big Data, IoT, and the Expansion of Data Quality Dimensions

From the early 2000s onwards, literature shifted towards the practical implications of Big Data and the Internet of Things (IoT).

According to IBM, the Internet of Things (IoT) refers to a network of physical devices, vehicles, appliances, and other physical objects embedded with sensors, software, and network connectivity that enable them to collect and share data (IBM 2023). Big Data concerns with massive, complex data sets that traditional data management systems cannot handle. When properly collected, managed and analyzed, Big Data can help organizations discover new insights and make better business decisions (Kosinski 2024).

The rapid advancement of technologies such as social networks, the Internet of Things (IoT), cloud computing, and other digital innovations has introduced the era of Big Data. The exponential growth in data volumes has created substantial value for both enterprises and society at large, while simultaneously raising critical questions about how to manage and exploit these vast resources effectively.

Big Data is commonly characterized by its *four Vs*: *volume*, *variety*, *velocity* and *veracity* (Figure 15) (Zhang et al. 2017).

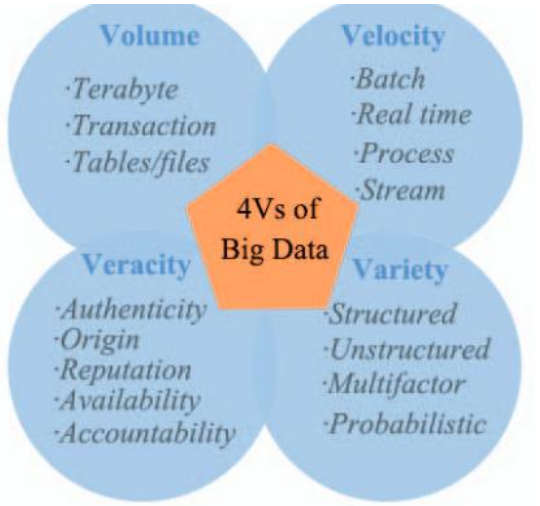

Figure 15: : 4 Vs of Data Quality (Zhang et al. 2021)

Each of these aspects brings its own challenges when it comes to processing and ensuring the quality of data. Large volumes, for instance, call for storage and processing systems that can scale effectively. Variety points to the difficulty of bringing together very different types of information, from structured databases to semi-structured files and unstructured content. Velocity reflects the pressure to deal with data that arrives and changes at high speed, while veracity concerns that not all data can be taken at face value, raising questions about its reliability and trustworthiness (Zhang et al. 2017).

In manufacturing, the effective use of Big Data is contingent on ensuring high data quality and the reliability of its sources. Degradation in data quality can lead to unpredictable consequences, eroding confidence in both the data and its origin. Factors such as the integration of multi-heterogeneous sources and the rapid pace of data generation exacerbate the risk of quality loss, making continuous monitoring and validation essential.

Nevertheless, maintaining Big Data Quality (BDQ) in such environments is inherently costly and resource-intensive, as it often requires substantial computational power and complex pre-processing workflows (Taleb et al. 2021). This reality underscores that data quality management is not a peripheral concern but a prerequisite for the successful application of Big Data techniques in manufacturing, enabling accurate analytics, informed decision-making, and the effective deployment of AI-driven solutions.

The practical implications of the *Four Vs* for data quality management in manufacturing can be illustrated by mapping common issues across the main stages of the Big Data lifecycle, from collection to analysis, together with their primary causes, affected quality dimensions, and potential solutions (Figure 16).

Stages	Issues	Primary Cause	Data Quality	4Vs	Solutions
Collection	**Less data collected and low recall rates**	Network connectivity and dynamics	*Availability*	*Volume Velocity*	Increase collection coverage
	Data sparseness	The interaction between the user and the item is less	*Relevance*	*Volume Variety*	Dimension reduction and processing algorithms
Preprocessing	**Noise data**	Abnormal data and cheating data at the time of collection	*Usability*	*Variety*	Remove noise operation
	Incomplete data	Data distribution is not balanced, the network transmission is unstable	*Reliability*	*Velocity*	Clustering Algorithm
Storage	**Limitations of Storage Technology**	The total amount of data is big and the type is complex	*Usability*	*Volume*	Cloud storage and Tiered storage
	Data timeliness	Response time is long	*Availability*	*Velocity*	Spark platform
Analysis	**Accuracy**	The rule of the data is elusive	*Reliability*	*Veracity*	Processing algorithms and models
	Scalability	High data dimension, high computational complexity	*Usability*	*Volume*	Clustering Algorithm

Figure 16: : Issues, Data Quality, 4V, Solutions (Zhang et al. 2021)

While these challenges are inherent to Big Data environments in general, they become even more pronounced in manufacturing contexts where data are increasingly generated through interconnected devices and systems. This transition leads directly to the domain of the Internet of Things (IoT), whose distributed and heterogeneous nature further amplifies both the opportunities and the complexities of ensuring data quality.

The concept of the Internet of Things (IoT) was first introduced to describe the potential of sensors to connect to the Internet and thereby enable new forms of service provision. It has also been defined more broadly as a network that connects ordinary physical objects, each with an identifiable address, to deliver intelligent services (Batini et al. 2009). In the manufacturing domain, the IoT is best understood as a networked ecosystem of interconnected devices, sensors, machines, and control systems that are capable of collecting, transmitting, and in some cases processing data without direct human intervention (Gubbi et al. 2013). These devices operate across different stages of the production process, from raw material handling to assembly lines and quality control

stations, generating continuous and real-time data streams that form the backbone of modern smart manufacturing.

For example, heterogeneous device specifications, communication protocols, and data formats can lead to interoperability issues and inconsistencies across datasets. Sensor drift, calibration errors, and connectivity disruptions can reduce accuracy and completeness, while the velocity of data flows can hinder effective timeliness control if processing systems cannot keep pace (Batini et al. 2009). Moreover, the distributed architecture of IoT systems demands robust traceability mechanisms to track the provenance and transformation of data across multiple nodes in the network.

This combination of opportunities and challenges has positioned IoT as both a driver and a stress test for AI-based data quality solutions. AI techniques, particularly in anomaly detection, sensor fusion, and real-time quality monitoring, have become essential for managing the complexity of IoT-enabled manufacturing environments.

The adoption of Big Data architectures and IoT technologies in manufacturing brought a significant shift in the way data quality was conceptualized and assessed. While intrinsic dimensions such as *accuracy*, *completeness* and *consistency* remained essential, the new technological landscape required the inclusion of additional attributes that captured the operational and systemic aspects of modern manufacturing data flows.

One of the most prominent was *traceability*, referring to the ability to track each data point back to its source and to reconstruct its transformation across the production chain (Isaja et al. 2023). This capability became crucial in IoT-enabled environments, where multiple heterogeneous devices contribute to the same dataset and where any anomaly must be traced rapidly to its origin to prevent production disruptions.

Equally important was *interoperability*, defined as the continuous integration and exchange of information between different systems, platforms, and devices. In practice, this meant overcoming incompatibility in data formats, communication protocols, and metadata standards, which could otherwise fragment the information landscape and reduce overall reliability.

Timeliness also emerged as a critical dimension, particularly in applications where data is collected and processed in real time or near real time. In such contexts, the value of the data can degrade rapidly if there are delays in acquisition, transmission, or analysis, making time-sensitive quality checks and low-latency data pipelines essential (Mirzaie et al. 2023).

Finally, *data governance* gained relevance as manufacturing systems became more complex and distributed (Sahi et al. 2023). Governance encompasses the policies, procedures, and accountability structures for managing data assets, ensuring not only technical quality but also compliance with standards, security requirements, and ethical considerations.

By integrating these additional dimensions, the assessment of data quality evolved from a narrow, intrinsic focus to a multi-layered framework capable of addressing the complexity of interconnected manufacturing environments. This evolution directly influenced AI-driven quality management, as algorithms increasingly needed to account for contextual, temporal, and systemic factors beyond the traditional scope of data cleaning and validation.

During this period, AI methods for data quality management in manufacturing became more sophisticated and diversified. Feature selection and extraction techniques were increasingly used to refine high-dimensional datasets, ensuring that only the most relevant variables were retained for analysis. Unsupervised anomaly detection methods allowed for the identification of faulty sensor readings and process deviations without requiring exhaustive labelling efforts. Data augmentation strategies were employed to mitigate class imbalance in predictive modelling, while active learning approaches optimized the use of expert labelling by focusing human intervention on the most informative samples (Zhou et al. 2024; Xie et al. 2025). Machine learning (ML) technologies have become substantial in practically all aspects of society and data quality (DQ) is critical for the performance, fairness, robustness, safety, and scalability of ML models. With the large and complex data in data-centric AI, traditional methods like exploratory data analysis (EDA) and cross-validation (CV) face challenges, highlighting the importance of mastering DQ tools.

Despite these advancements, significant challenges persisted. Scalability remained a concern, as even distributed architectures struggled to maintain performance when processing high-volume, high-velocity industrial data streams. Standardization issues continued to limit interoperability and hinder the consistent application of quality metrics across systems. Moreover, achieving real-time assurance of data quality proved difficult, as many AI models lacked the capacity to adapt instantaneously to fluctuations in production conditions.

4.3 - Advanced Period (2020–2025): Machine Learning, Deep Learning, and Integrated AI Systems for Data Quality

The most recent period marks the integration of AI into real-time manufacturing data pipelines, aligned with trustworthy AI principles. Artificial Intelligence (AI) has become one of the primary drivers of digital transformation, with applications that are rapidly expanding across industrial sectors (Oviedo et al. 2024). Manufacturing has been profoundly affected by the integration of AI-based systems, which are now central to predictive analytics, process optimization, and real-time quality monitoring (Sharma et al. 2022). This technological evolution has been accompanied by increasing regulatory and institutional attention, reflected in the development of international standards (e.g., ISO/IEC) aimed at guiding the design, deployment, and assessment of AI solutions, including those that directly affect data quality management.

Within this broader context, Machine Learning (ML) represents one of the most widely used paradigms, providing the foundation for prediction, classification, and anomaly detection tasks in manufacturing (Azimi e Pahl 2025). However, the reliability of ML models is strongly dependent on the quality of the data used during training and validation. Low-quality datasets, characterized by noise, sparsity, or irrelevant attributes, can significantly compromise performance in critical tasks such as defect detection or predictive maintenance. As such, assessing and enhancing dataset quality has become a prerequisite for trustworthy ML applications. This has driven growing emphasis on practices such as data curation (the systematic collection, selection, and

organization of data) and on the development of algorithmic solutions capable of mitigating the effects of low-quality inputs.

Deep Learning (DL), while offering transformative potential through its ability to automatically extract complex patterns from raw data, poses further challenges in industrial environments. DL models are inherently *data-hungry* and require very large, diverse, and reliable datasets to achieve robust performance (Munappy et al. 2022). In real-world manufacturing scenarios, obtaining such high-quality data is not always feasible, as corrupted, incomplete, or biased samples are common. This has raised interest in fairness metrics, data augmentation strategies, and robustness techniques to enable DL models to tolerate imperfections in training data while still delivering dependable results.

The relationship between Artificial Intelligence as a broader application domain and the role of machine learning models within AI systems can be illustrated by considering their structural organization. Figure 17 provides an overview of how training and production data are processed within AI systems, highlighting the continuous interaction between data, models, and outputs.

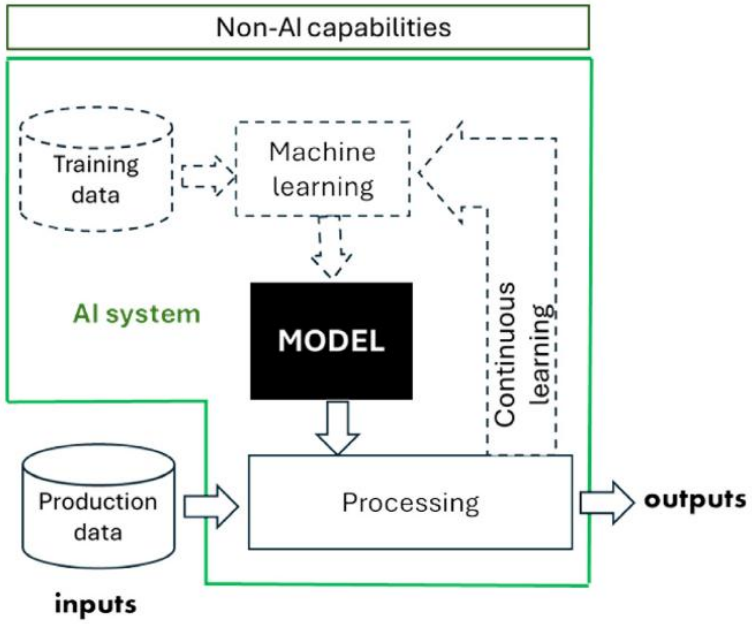

Figure 17: AI application and AI system (Oviedo et al. 2024)

Overall, the most recent period is characterized by the recognition that the success of AI systems in manufacturing depends less on the sophistication of algorithms alone and more on the quality of the data they consume (Majeed e Hwang 2024). Unlike conventional software, where improvements can often be achieved by refining code, AI-based software is inductively derived from data. Consequently, advancing the field requires both rigorous assessment of dataset quality and the development of strategies to safeguard reliability, fairness, and interpretability in AI-driven manufacturing systems.

Artificial Intelligence (AI) has become a pivotal driver of digital transformation in manufacturing, permeating every layer of production, maintenance, and quality control. In this context, ensuring the reliability of AI systems, and, critically, of the data that feeds them, has evolved into a central concern. Recent years have seen the introduction of international standards (notably ISO/IEC standards) that specifically target AI system quality, covering not only processes and products but also the integrity and quality of the data underpinning them (Oviedo et al. 2024).

In contemporary AI-driven manufacturing systems, the spectrum of data quality dimensions extends well beyond the traditional triad of accuracy, completeness, and timeliness. Within data-driven industrial contexts, researchers argue that quality must now be conceived as a multi-layered construct: it should not only reflect correctness but also the ability of data to be meaningful, usable, and trustworthy across various scenarios.

Emerging dimensions such as bias detection and fairness ensure that AI systems do not systematically disadvantage certain outcomes or populations. Semantic accuracy emphasizes that data should faithfully represent real-world phenomena at a level of understanding aligned with human judgment and domain semantics. Cross-domain generalizability, meanwhile, involves ensuring that data collected in one manufacturing context can support AI models deployed across diverse operational environments.

These developments are motivated by evolving industry needs, captured in frameworks like quality-by-design, and the growing recognition that data must maintain its value across changing use cases. Fu et al. highlight this shift, showing how data quality is increasingly framed within socio-technical systems where usability, provenance, and

ongoing value supersede static notions of *correctness* (Fu et al. 2024). In manufacturing, this means moving from verifying data correctness toward ensuring data remain fair, interpretable, and adaptable over time and across applications.

The sophistication of AI techniques applied to manufacturing data quality has likewise grown concurrently with these expanded dimensions. Deep Learning architectures, such as convolutional, recurrent, and transformer models, are now widely used for detecting defects, forecasting maintenance needs, and extracting high-level features from multi-modal sensor streams. These models excel at capturing complex patterns but demand vast amounts of high-quality data.

To mitigate this challenge, practitioners increasingly rely on Transfer Learning: pre-trained models are adapted to new manufacturing environments with limited new training data, enabling faster deployment and improved generalizability. For enhanced robustness, Ensemble and Hybrid Systems combine multiple models, or blend rule-based and learning-based logic, offering better interpretability and error resilience (Fu et al. 2024).

Data scarcity and labelling costs are addressed through Active Learning, guiding human annotation toward the most informative or uncertain instances, thus optimizing expert effort. Additionally, Automated Data Augmentation techniques synthetically enhance data diversity, by adding noise, transformations, or simulated edge cases, to improve model training where real-world samples remain scarce.

Collectively, these techniques shift AI's role from passive analysis to active quality guardianship: models now not only consume data but also help enforce quality standards, detect bias, and adapt to changing operating conditions. This reflects a profound change: in modern manufacturing, AI systems are not just data-driven, but they are data-oriented in the sense that data quality becomes a continuous, integral part of their operation.

Despite these advances, several challenges continue to limit the effectiveness and scalability of AI-based approaches to data quality in manufacturing. One persistent issue is class imbalance, whereby defective cases are underrepresented compared to normal instances. This imbalance hampers the training of reliable predictive models, often

leading to biased outcomes or reduced sensitivity in detecting rare but critical events (Clemente et al. 2023).

A second major obstacle is the high cost of annotation (Kumar et al. 2024). Many AI techniques, particularly in supervised and semi-supervised learning, require large volumes of accurately labelled data. In manufacturing, however, expert labelling is both expensive and time-consuming, and errors in annotation can further compromise model reliability.

Another recurring gap is the absence of benchmark datasets that are standardized and openly available for evaluation and comparison. Without shared references, it is difficult to assess the generalizability of proposed methods or to establish performance baselines across different manufacturing contexts (Nikiforova 2020).

Finally, interoperability gaps persist due to the heterogeneity of manufacturing environments (Oviedo et al. 2024). Differences in data formats, system architectures, and communication protocols create barriers to integrating data from multiple sources, limiting the scope of AI-driven quality management systems.

4.4 - Synthesis and Gaps

The chronological review of the 41 selected papers, spanning the period from 1995 to 2025, reveals a clear trajectory in the way data quality in manufacturing has been conceptualized, evaluated, and enhanced through artificial intelligence. This trajectory can be interpreted as a progressive expansion in both the dimensions of data quality considered relevant (*RQa*) and the AI techniques applied to support them (*RQb*), alongside the persistence of challenges that remain unresolved. Table 2 illustrates an example of 5 documents from the final corpus selected and fully analysed (i.e., 41 documents), structured according to the approach described in Chapter 3 (sub-section 3.6). This table illustrates how the information was organised across application areas, challenges, techniques, and research questions, thereby serving as a representative synthesis of the broader analysis.

In the early period (1995–2010), academic attention was primarily devoted to building a conceptual foundation. *Data* were defined as recorded values of events, facts, or measurements, and the absence of a universal definition of *data quality* led to reliance on intrinsic dimensions such as *accuracy*, *completeness* and *consistency*, with *timeliness* and *relevance* occasionally included. Evaluation methods were largely manual or rule-based, and AI applications were embryonic, limited to statistical methods and simple classification or clustering approaches. The main limitation of this stage lies in the lack of standardization: ISO 8000 only begins to introduce structured principles, and in the inability of available methods to provide real-time or scalable assurance of data quality.

The intermediate period (2010–2020) marked a decisive shift towards Big Data and Internet of Things (IoT). The four Vs, *volume*, *variety*, *velocity* and *veracity*, became central, reflecting the practical challenges of managing large, heterogeneous, and fast-moving data streams in manufacturing environments. These developments brought about a significant expansion in data quality dimensions. While intrinsic properties remained relevant, additional dimensions emerged: *traceability*, ensuring the ability to reconstruct data provenance across distributed IoT networks; *interoperability*, facilitating integration across devices and platforms; *timeliness*, reflecting the critical importance of near real-time data flows; *governance*, encompassing accountability, compliance, and policy frameworks for managing increasingly complex data assets. Correspondingly, AI techniques grew more sophisticated. Feature selection and extraction were used to reduce high-dimensional datasets, unsupervised anomaly detection allowed fault identification without exhaustive labelling and data augmentation and active learning addressed issues of imbalance and labelling costs. Nevertheless, *scalability* and *interoperability* remained problematic, and real-time adaptation of AI methods often failed to meet the needs of rapidly changing industrial contexts.

The advanced period (2020–2025) is characterized by the consolidation of Machine Learning (ML), Deep Learning (DL), and integrated AI systems as central components of manufacturing data quality management. Here, the emphasis shifted from algorithms alone to the quality of the data that fuels them, recognizing that AI software is inductively

derived from data rather than written deterministically. This period introduced new and critical dimensions of data quality: *bias detection* and *fairness*, ensuring that AI systems do not propagate systematic disadvantage; *semantic accuracy*, emphasizing the faithful and meaningful representation of real-world phenomena; *cross-domain generalizability*, enabling models trained in one context to perform reliably in others. Alongside these conceptual advances, AI techniques reached a new level of sophistication. Deep learning architectures such as CNNs, RNNs and transformers became widely used for defect detection and predictive maintenance; transfer learning allowed models to be adapted to new contexts with minimal retraining; ensemble and hybrid systems combined complementary methods for improved robustness; active learning and automated data augmentation alleviated the bottleneck of scarce and costly labelled data. At the same time, the recognition of fairness metrics, robustness techniques and explainability underscored a shift towards data-oriented AI systems, where maintaining quality is an integral and continuous part of operation.

Despite these advances, several persistent gaps remain across all three periods. Class imbalance continues to compromise the reliability of predictive models, as defective cases remain underrepresented in real datasets. The high cost of annotation remains a critical obstacle, especially in supervised learning contexts where expert labelling is indispensable but resource intensive. The absence of standardized benchmark datasets hinders comparability across studies and limits the establishment of universally accepted performance baselines. Finally, interoperability gaps persist due to heterogeneous system architectures, data formats, and communication protocols, which restrict the seamless integration of AI-based quality management systems in diverse manufacturing environments.

This analysis shows that AI applications in manufacturing data quality have evolved from rudimentary preprocessing and classification techniques to highly sophisticated, multi-layered systems capable of addressing complex industrial realities. However, the field is still constrained by structural challenges that impede scalability, generalizability, and standardization. These gaps provide fertile ground for further research and innovation, underscoring the need for collaborative efforts in benchmarking, interoperability frameworks, and the development of data-centric AI methods that explicitly integrate

fairness, adaptability, and robustness as essential dimensions of trustworthy manufacturing systems.

Table 2: Sample of five relevant papers analysed in full text

DOI	Year	Title	Author Keywords	Document Type	Application Area	Problems/Challenges	Techniques/Solutions/Tools	RQ(s)/Aim of the paper	Results
10.1186/s40537-021-00468-0	2021	Big data quality framework: a holistic approach to continuous quality management	Big data quality; Data quality profile; Quality assessment; Quality metrics and scores; Pre-processing	Article	• Big Data environments • Continuous data quality monitoring • Data warehousing and analytics • Cloud-based data processing	• Lack of systematic and scalable approaches to measure and manage data quality in big data contexts. • High volume, variety, and velocity of data make traditional DQ methods inadequate. • Fragmentation of data quality dimensions and responsibilities. • Need for real-time DQ monitoring integrated within processing workflows.	• Proposal of a holistic framework for big data quality (BDQF). • Integration of DQ monitoring into data processing pipelines. • Use of technical and organizational quality dimensions. • Modular architecture enabling continuous assessment, feedback, and correction. • Implementation case based on a real-world big data platform.	• To develop a comprehensive framework for continuous data quality management in big data systems. • To align technical DQ mechanisms with organizational quality governance. • To validate the feasibility of integrated DQ monitoring in practice.	1) The proposed BDQF framework supports automated and continuous DQ monitoring in large-scale data environments. 2) The integration of technical and organizational quality aspects improves traceability and accountability. 3) Experimental results show the framework's effectiveness in identifying and mitigating quality issues in real time. 4) The modular design allows adaptation to different big data architectures and use cases.

10.1145/1541880.1541883	2009	Methodologies for Data Quality Assessment and Improvement	Management; Measurement; Data quality; data quality measurement; data quality assessment; data quality improvement; methodology; information system; quality dimension	Article	• Data warehousing. • Data integration systems. • Database management and governance. • Data cleaning and quality monitoring frameworks.	• Inconsistent, incomplete, and inaccurate data in large datasets. • Lack of standardized procedures for assessing data quality. • Difficulty in reconciling heterogeneous data sources. • Need for continuous monitoring and improvement cycles.	• Definition and formalization of a Data Quality Assessment Methodology (DQAM). • Use of metadata and quality-related information for rule generation. • Integration of user feedback into quality metrics and improvement steps. • Iterative framework including assessment, analysis, improvement, and monitoring phases.	• How can a methodological and structured approach help organizations assess and improve data quality? • What are the key dimensions and procedures necessary for implementing effective data quality management?	1) Presented a formal framework (DQAM) for systematic assessment and improvement of data quality. 2) Emphasized the role of metadata and domain-specific rules in quality evaluation. 3) Demonstrated applicability through use-case discussions and integration strategies. 4) Advocated for feedback-driven, iterative enhancement of data quality over time.

10.3390/data9120151	2024	A Framework for Current and New Data Quality Dimensions: An Overview	data quality; data model; data quality dimensions; data traceability; confidence in data; data metrology; data uncertainty; data structures; big data; IoT	Article	• Data quality assessment and management. • Information systems and databases. •Multidimensional data quality modeling.	• Fragmentation and lack of consensus on definitions of data quality dimensions. • Difficulty in comparing and mapping dimensions across different models • Ambiguity and overlap between dimension definitions. • Need for clarity on how dimensions apply across contexts and domains.	• Comparative literature analysis of existing data quality dimensions. • Development of a unified classification framework (Data Quality Data Model). • Proposal of a meta-model for organizing and categorizing dimensions. • Identification of 49 current and 15 new dimensions and their grouping under broader categories.	• To review, analyze and classify existing and emerging data quality dimensions. • To develop a framework that can consolidate and compare data quality dimensions. • To highlight gaps and overlaps in current dimensional models.	1) Identified and categorized 64 data quality dimensions (49 existing, 15 new). 2) Proposed a unified framework composed of six categories: Intrinsic, Contextual, Representational, Accessibility, Operational, and Organizational. 3) Facilitated comparative analysis and interoperability between DQ models. 4) Provided a basis for future development of tailored DQ assessment tools.

10.1109/ACCESS.2019.2899751	2019	An Overview of Data Quality Frameworks	Data quality assessment; data structures; decision making; information management; quality management	Article	Cross-domain data quality management across diverse business environments, information systems, and data types (structured, semi-structured, and unstructured). Particularly relevant for organizations seeking comprehensive DQ strategies in enterprise data, data warehouses, and Big Data contexts.	•Heterogeneity of data quality requirements across organizations and application domains. •Difficulty selecting appropriate frameworks due to the diversity of existing methodologies. •Complexity in handling different types of data (structured, semi-structured, unstructured) and quality dimensions. •Lack of standardization in defining and applying data quality dimensions and assessment processes. •Inconsistent treatment of improvement costs and decision-making strategies across frameworks.	Comparative survey of 12 general-purpose data quality frameworks that include: •Definition of data quality attributes and dimensions. •Assessment processes (using subjective and/or objective metrics). •Improvement strategies (including root cause analysis, cost-benefit analysis, and decision frameworks). •Classification and comparison of frameworks based on: -Types of data handled -Types of measurements used (e.g., metrics, questionnaires) -Level of detail in improvement steps -Cost considerations and decision models •Proposal of a decision guide to support the selection of suitable data quality frameworks depending on context-specific criteria.	To provide a comprehensive, comparative overview of general-purpose data quality frameworks, enabling organizations to: • Understand core components (definition, assessment, improvement) • Compare available methodologies; • Select the most suitable framework using a structured decision guide.	Identification of 12 general-purpose DQ frameworks, each described in terms of: •Data quality definition and dimensions •Assessment processes and measurement types •Improvement strategies, including cost and decision-making approaches Recognition that accuracy, completeness, and timeliness are the most frequently cited quality dimensions. Emphasis on the need for customization of DQ dimensions based on organizational needs. Highlight of the variation in assessment methods (objective vs. subjective,

									metrics vs. surveys). Presentation of a decision support table to help practitioners choose the most appropriate framework based on factors like data type, organizational needs, and cost awareness.

10.1016/j.jss.2022.111359	2022	Data management for production quality deep learning models: Challenges and solutions	Deep learning; Data management; Production quality DL models; Challenges; Solutions; Validation	Article	• Deep learning in production (deployment-level) environments. • Data management for AI pipelines at scale. • Applications across multiple sectors including manufacturing, automotive, and retail. • MLOps and data-centric AI system development.	• Deep learning models are highly dependent on data quality, yet data processes are often ad hoc or poorly managed. • Lack of standardized practices for managing datasets over the ML lifecycle. • Common pain points include: - Data versioning and traceability - Labeling consistency - Data drift and spurious correlations - Weak supervision and noisy labels. • Difficulty aligning data operations with DevOps/MLOps pipelines.	• Structured a six-stage data management framework: 1. Data acquisition 2. Data cleaning and preparation 3. Data labeling 4. Data versioning 5. Data monitoring and validation 6. Data governance. • Highlighted tools and techniques such as: - Data versioning tools (e.g. DVC) - Active learning and weak supervision frameworks - Continuous monitoring for data and concept drift - Label audits and standardization practices. • Emphasis on aligning ML data lifecycle with software engineering principles (MLOps).	• What are the main data management challenges in deploying production-level deep learning systems? • How can these challenges be addressed with current tools and organizational practices? • Can a structured framework help ensure data quality and traceability across the ML lifecycle?	• Identified six key stages in the data lifecycle critical to production-ready DL systems. • Mapped common issues to each stage, offering actionable practices to address them. • Emphasized the role of standardized data pipelines and governance in improving model reliability. • Showed how poor data handling leads to performance degradation, compliance risks, and scaling issues. • Advocated for data-centric MLOps strategies to ensure consistency and traceability. • Positioned data management as a first-class citizen in AI system development, on par with model architecture.

Chapter 5 – Conclusions

This final chapter brings together the findings of the literature review and provides explicit answers to the research questions formulated in Chapter 1.

The analysis of final *corpus* of 41 papers selected through PRISMA methodology and described step by step in Chapter 3 shows how data quality in manufacturing has been conceptualized, assessed and enhanced through the application of Artificial Intelligence. The results, extracted from the corpus of 41 articles, have been organized chronologically in Chapter 4 to highlight the predominant changes over time. Nevertheless, the periods identified should not be considered as rigid boundaries. Rather, they serve as a heuristic framework that reveals both continuities and turning points in the evolution of concepts, techniques, and challenges.

The review demonstrates that AI applications for manufacturing data quality have evolved from simple preprocessing and validation techniques to highly sophisticated, integrated systems. Initially, AI played a limited role, with rule-based methods and basic statistical checks. Over time, the expansion of Big Data and IoT required scalable and automated approaches, leading to the adoption of machine learning for anomaly detection, feature extraction, and active learning. In the most recent period, deep learning, transfer learning, ensemble models, and hybrid systems have made it possible to embed data quality assurance directly into manufacturing pipelines. Despite this progress, several unresolved challenges persist, including class imbalance, high annotation costs, lack of standardized benchmark datasets, and ongoing interoperability gaps. Together, these findings suggest that AI has become indispensable for managing data reliability, but its full potential remains constrained by structural and methodological limitations.

Across the three decades examined, some dimensions, particularly *accuracy*, *completeness* and *consistency*, have remained central. These intrinsic properties are essential for ensuring that data faithfully reflect manufacturing processes and can support reliable AI-driven decisions. As technologies evolved, additional dimensions became prominent. In the Big Data and IoT era, *traceability*, *interoperability*, *timeliness*,

and *governance* gained importance, reflecting the complexity of distributed and heterogeneous environments. In the most recent period, novel dimensions such as *fairness*, *bias detection*, *semantic accuracy* and *cross-domain generalizability* have emerged, highlighting the alignment of data quality with the broader principles of trustworthy AI (RQa). Overall, the trajectory indicates a progressive broadening of data quality concept, from technical correctness to socio-technical robustness and adaptability. The AI techniques applied to data quality in manufacturing reflect this evolution.

In the early years, applications were limited to simple classification, clustering, and anomaly detection methods. Between 2000 and 2020, machine learning approaches became widespread, including feature selection, unsupervised anomaly detection and data augmentation strategies, often combined with active learning to reduce labelling costs. From 2020 onwards, the field has increasingly relied on deep learning architectures (CNNs, RNNs, transformers), transfer learning for domain adaptation and ensemble or hybrid approaches to improve robustness and interpretability. Moreover, AI has shifted from being a tool for post hoc data cleaning to becoming an integral mechanism for continuous monitoring and quality assurance. These techniques not only enhance data reliability but also reflect the recognition that trustworthy AI depends fundamentally on trustworthy data.

Taken together, the findings reveal a clear trajectory: definitions and conceptual frameworks laid in the late 1990s provided the basis for technical developments during the Big Data and IoT era, which in turn set the stage for today's advanced AI-driven solutions. The dimensions of data quality have expanded from intrinsic attributes to multi-layered constructs that include governance, fairness, and interpretability. AI techniques have moved from simple preprocessing to sophisticated, integrated systems capable of enforcing data quality standards in real time. Nonetheless, the persistence of unresolved issues, such as imbalanced datasets, annotation costs, interoperability, and lack of benchmarks, demonstrates that the field remains incomplete (RQb). Addressing these challenges will be crucial for ensuring that AI in manufacturing can be both technically effective and aligned with the principles of trustworthy AI.

5.1 - Future developments

Across all three periods analysed, a persistent gap was the lack of a common standard for data quality in AI-driven manufacturing, together with the absence of a unified terminology and consistent set of dimensions. This deficiency limited comparability across studies and hindered the development of universally accepted frameworks. Earlier standards partially addressed this issue: ISO 9000 introduced the general concept of *quality* as the degree to which requirements are satisfied, and ISO 8000 extended these principles to data quality through domains such as master, transaction, and product data. Although the ISO 8000 series represented an initial attempt to structure data quality, it was primarily designed for traditional industrial data management and lacked provisions for the complexity of AI- and ML-driven environments, including Big Data and IoT.

However, an important development in the institutionalization of data quality for artificial intelligence and machine learning is represented by the recent publication, in June 2025, of the ISO/IEC 5259 series. This family of standards provides a harmonized set of concepts, characteristics, measures, processes, and governance principles. It offers a comprehensive framework for defining, measuring, managing, and governing data quality in the context of analytics and machine learning. The series is structured into five complementary parts, each addressing a specific level of abstraction, from terminology and metrics to processes and governance.

ISO/IEC 5259-1: Overview, terminology, and examples introduces the fundamental concepts and serves as the entry point to the series. It defines the *data life cycle* as the set of phases covering the entire existence of data, from creation to decommissioning. It distinguishes between roles such as *data originator*, i.e., the party that creates data and may hold rights over them; *data holder*, namely the party with legal control over data use and *data user*, the party authorized to process data under such control. Central to this part is the definition of *data quality* as *the degree to which data satisfy stated and implied needs when used under specified conditions*. Related concepts include *data quality characteristics*, defined as a category of attributes of data that bear on its quality, *data quality model*, namely a defined set of characteristics and relationships that provide a

framework for specifying requirements and evaluating data and *data quality measure*, understood as a variable to which a value is assigned as the result of measurement. This part also provides examples and scenarios showing how deficiencies in quality dimensions can impact the performance of machine learning and analytics.

ISO/IEC 5259-2: Data quality measures specify the *data quality model* to be applied in the context of analytics and ML. It identifies a range of quality characteristics grouped into three categories. *Inherent characteristics* include accuracy, completeness, consistency, credibility, and currentness. *System-dependent characteristics* encompass accessibility, compliance, efficiency, precision, traceability and understandability, as well as availability, portability and recoverability. Finally, *additional characteristics* include auditability, balance, diversity, effectiveness, identifiability, relevance, representativeness, similarity, and timeliness. For each characteristic, the standard provides definitions and guidelines for establishing corresponding *data quality measures*, understood as measurable variables. This part also specifies a *framework for reporting* on data quality, ensuring transparency and comparability across stakeholders.

ISO/IEC 5259-3: Data quality management requirements and guidelines establishes requirements and recommendations for the implementation of a *data quality management system (DQMS)*. It introduces key definitions such as *data quality claim*, namely a statement that data meets a particular quality requirement, and *data quality plan*, a specification of practices, processes, and resources required to achieve stated quality objectives. This part prescribes management principles including the establishment of a *data quality culture*, resource and competence management, auditing and reviewing, and project-specific planning. It further details the management of the *data quality life cycle*, which spans from motivation and specification, through planning, acquisition, preprocessing, augmentation, and provisioning, to decommissioning. Cross-cutting processes include *verification and validation, configuration management, change management,* and *risk management*.

ISO/IEC 5259-4: Data quality process framework provides an operational framework of processes to ensure and improve data quality for ML. It defines concepts central to data preparation and annotation, including *outsourcing* (the use of external organizations for data-related tasks), *stand-off annotation* (annotations kept separate from primary data),

bounding box (a rectangular region enclosing an object of interest), *segmentation* (the separation of objects of interest from context), and *key-point* (a salient point on an object). The framework covers processes for planning, acquisition, preparation (including annotation, labelling, encoding, and de-identification), provisioning, and decommissioning. It provides guidance across different learning paradigms (*supervised, unsupervised, semi-supervised, and reinforcement learning*) and emphasizes the role of annotation and labelling quality in supervised ML. It also specifies the responsibilities of actors in the data ecosystem, including the *data planner, originator, collector, engineer, holder* and *user*.

ISO/IEC 5259-5: Data quality governance framework addresses the governance level, establishing principles and responsibilities for ensuring data quality in organizational and strategic contexts. It emphasizes that governance should ensure the establishment of strategies, policies, and oversight mechanisms to direct and control data quality management. The standard identifies the responsibilities of the *governing body*, which include recognizing the strategic importance of data quality, establishing an enabling environment, formulating strategies and policies, and ensuring oversight. In parallel, the *management* is responsible for implementing these strategies and policies, strengthening internal controls, and integrating risk management mechanisms. This part highlights the importance of accountability, business planning linked to data quality, and the alignment of technical quality practices with organizational objectives.

Taken together, the ISO/IEC 5259 provides a structured and comprehensive reference framework that connects conceptual definitions, measurable characteristics, management processes, operational practices, and governance responsibilities. For the manufacturing sector, this set of standards represents an important step toward harmonizing approaches to data quality in AI and ML, ensuring that technical, organizational, and strategic dimensions are jointly addressed in the pursuit of trustworthy artificial intelligence.

References

Adadi, Amina, e Mohammed Berrada. 2018. «Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)». *IEEE Access* 6: 52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Almeida, Felipe, e Geraldo Xexéo. 2023. «Word Embeddings: A Survey». arXiv:1901.09069. Preprint, arXiv, maggio 2. https://doi.org/10.48550/arXiv.1901.09069.

Alzubaidi, Laith, Aiman Al-Sabaawi, Jinshuai Bai, et al. 2023. «Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements». *Int. J. Intell. Syst.* 2023 (gennaio). https://doi.org/10.1155/2023/4459198.

Andrew Black e Peter van Nederpelt. 2020. «Dimensions of data quality». https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf.

Azimi, Shelernaz, e Claus Pahl. 2025. «Anomaly Analytics in Data-Driven Machine Learning Applications». *International Journal of Data Science and Analytics* 19 (1): 155–80. https://doi.org/10.1007/s41060-024-00593-y.

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, e Andrea Maurino. 2009. «Methodologies for data quality assessment and improvement». *ACM Comput. Surv.* 41 (3): 16:1-16:52. https://doi.org/10.1145/1541880.1541883.

Campbell, Mhairi, Joanne E. McKenzie, Amanda Sowden, et al. 2020. «Synthesis without Meta-Analysis (SWiM) in Systematic Reviews: Reporting Guideline». *BMJ (Clinical Research Ed.)* 368 (gennaio): l6890. https://doi.org/10.1136/bmj.l6890.

Clemente, Fabiana, Gonçalo Martins Ribeiro, Alexandre Quemy, Miriam Seoane Santos, Ricardo Cardoso Pereira, e Alex Barros. 2023. «Ydata-Profiling: Accelerating Data-Centric AI with High-Quality Data». *Neurocomputing* 554 (ottobre): 126585. https://doi.org/10.1016/j.neucom.2023.126585.

Deng, Li. 2018. «Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook [Perspectives]». *IEEE Signal Processing Magazine* 35 (1): 180–177. https://doi.org/10.1109/MSP.2017.2762725.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2019. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». arXiv:1810.04805. Preprint, arXiv, maggio 24. https://doi.org/10.48550/arXiv.1810.04805.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, et al. 2017. «Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks». *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

European Commission, High-Level Expert Group on AI. 2020. «Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment». https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

Fu, Qian, Gemma L. Nicholson, e John M. Easton. 2024. «Understanding Data Quality in a Data-Driven Industry Context: Insights from the Fundamentals». *Journal of Industrial Information Integration* 42 (novembre): 100729. https://doi.org/10.1016/j.jii.2024.100729.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, et al. 2020. «Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?» *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAT* '20, gennaio 27, 325–36. https://doi.org/10.1145/3351095.3372862.

Gubbi, Jayavardhana, Rajkumar Buyya, Slaven Marusic, e Marimuthu Palaniswami. 2013. «Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions». *Future Generation Computer Systems* 29 (7): 1645–60. https://doi.org/10.1016/j.future.2013.01.010.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, e Jian Sun. 2016. «Deep Residual Learning for Image Recognition». *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, giugno, 770–78. https://doi.org/10.1109/CVPR.2016.90.

Hultcrantz, Monica, David Rind, Elie A. Akl, et al. 2017. «The GRADE Working Group Clarifies the Construct of Certainty of Evidence». *Journal of Clinical Epidemiology* 87 (luglio): 4–13. https://doi.org/10.1016/j.jclinepi.2017.05.006.

Hutton, Brian, Ferrán Catalá-López, e David Moher. 2016. «The PRISMA Statement Extension for Systematic Reviews Incorporating Network Meta-Analysis: PRISMA-NMA». *Medicina Clínica (English Edition)* 147 (6): 262–66. https://doi.org/10.1016/j.medcle.2016.10.003.

Isaja, Mauro, Phu Nguyen, Arda Goknil, et al. 2023. «A Blockchain-Based Framework for Trusted Quality Data Sharing towards Zero-Defect Manufacturing». *Computers in Industry* 146 (aprile): 103853. https://doi.org/10.1016/j.compind.2023.103853.

Jess Whittlestone e Stephen Cave. 2019. *Ethical and societal implications of algorithms, data, and artifical intelligence: a roadmap for research*. https://www.researchgate.net/publication/337565648_Ethical_and_societal_implications_of_algorithms_data_and_artificial_intelligence_a_roadmap_for_research.

Jumper, John, Richard Evans, Alexander Pritzel, et al. 2021. «Highly Accurate Protein Structure Prediction with AlphaFold». *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kale, Amruta, Tin Nguyen, Frederick C. Harris Jr., Chenhao Li, Jiyin Zhang, e Xiaogang Ma. 2023. «Provenance documentation to enable explainable and trustworthy AI: A literature review». *Data Intelligence* 5 (1): 139–62. https://doi.org/10.1162/dint_a_00119.

Kumar, Sushant, Sumit Datta, Vishakha Singh, Sanjay Kumar Singh, e Ritesh Sharma. 2024. «Opportunities and Challenges in Data-Centric AI». *IEEE Access* 12: 33173–89. https://doi.org/10.1109/ACCESS.2024.3369417.

Li, Bo, Peng Qi, Bo Liu, et al. 2023. «Trustworthy AI: From Principles to Practices». *ACM Comput. Surv.* 55 (9): 177:1-177:46. https://doi.org/10.1145/3555803.

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, et al. 2009. «The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration». *Journal of Clinical Epidemiology* 62 (10): e1-34. https://doi.org/10.1016/j.jclinepi.2009.06.006.

Liu, Haochen, Yiqi Wang, Wenqi Fan, et al. 2022. «Trustworthy AI: A Computational Perspective». *ACM Trans. Intell. Syst. Technol.* 14 (1): 4:1-4:59. https://doi.org/10.1145/3546872.

Majeed, Abdul, e Seong Oun Hwang. 2024. *Towards Unlocking the Hidden Potentials of the Data-Centric AI Paradigm in the Modern Era*. No. 4. agosto, 4. https://doi.org/10.3390/asi7040054.

McKinsey. 2023. «The state of AI in 2023: Generative AI's breakout year | McKinsey». https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, e Aram Galstyan. 2022. «A Survey on Bias and Fairness in Machine Learning». arXiv:1908.09635. Preprint, arXiv, gennaio 25. https://doi.org/10.48550/arXiv.1908.09635.

Mirzaie, Mostafa, Behshid Behkamal, Mohammad Allahbakhsh, Samad Paydar, e Elisa Bertino. 2023. «State of the Art on Quality Control for Data Streams: A Systematic Literature Review». *Computer Science Review* 48 (maggio): 100554. https://doi.org/10.1016/j.cosrev.2023.100554.

Moher, D., D. J. Cook, S. Eastwood, I. Olkin, D. Rennie, e D. F. Stroup. 1999. «Improving the Quality of Reports of Meta-Analyses of Randomised Controlled Trials: The QUOROM Statement. Quality of Reporting of Meta-Analyses». *Lancet (London, England)* 354 (9193): 1896–900. https://doi.org/10.1016/s0140-6736(99)04149-5.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, e Douglas G Altman. 2009. «Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement». *Open Medicine* 3 (3): e123–30.

Moody's. 2024. «AI in Compliance». https://www.moodys.com/web/en/us/kyc/resources/thought-leadership/ai-in-compliance.html.

Mulrow, C. D. 1987. «The Medical Review Article: State of the Science». *Annals of Internal Medicine* 106 (3): 485–88. https://doi.org/10.7326/0003-4819-106-3-485.

Munappy, Aiswarya Raj, Jan Bosch, Helena Holmström Olsson, Anders Arpteg, e Björn Brinne. 2022. «Data Management for Production Quality Deep Learning Models: Challenges and Solutions». *Journal of Systems and Software* 191 (settembre): 111359. https://doi.org/10.1016/j.jss.2022.111359.

Naseem, Usman, Imran Razzak, Shah Khalid Khan, e Mukesh Prasad. 2020. «A Comprehensive Survey on Word Representation Models: From Classical to State-Of-The-Art Word Representation Language Models». arXiv:2010.15036. Preprint, arXiv, ottobre 28. https://doi.org/10.48550/arXiv.2010.15036.

Nikiforova, Anastasija. 2020. *Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment*. settembre 28. https://doi.org/10.22364/bjmc.2020.8.3.02.

Novack, George. 2020. «Building a One Hot Encoding Layer with Tensorflow». *TDS Archive*, giugno 8. https://medium.com/data-science/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39.

Oviedo, Jesús, Moisés Rodriguez, Andrea Trenta, Dino Cannas, Domenico Natale, e Mario Piattini. 2024. «ISO/IEC Quality Standards for AI Engineering». *Computer Science Review* 54 (novembre): 100681. https://doi.org/10.1016/j.cosrev.2024.100681.

Oxman, A. D., D. J. Cook, e G. H. Guyatt. 1994. «Users' Guides to the Medical Literature. VI. How to Use an Overview. Evidence-Based Medicine Working Group». *JAMA* 272 (17): 1367–71. https://doi.org/10.1001/jama.272.17.1367.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, et al. 2021. «The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews». *BMJ*, marzo 29, n71. https://doi.org/10.1136/bmj.n71.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, et al. 2021. «The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews». *Journal of Clinical Epidemiology* 134 (giugno): 178–89. https://doi.org/10.1016/j.jclinepi.2021.03.001.

Redman, Thomas C. 1998. «The Impact of Poor Data Quality on the Typical Enterprise». *Communications of the ACM* 41 (2): 79–82. https://doi.org/10.1145/269012.269025.

Redmon, Joseph, Santosh Divvala, Ross Girshick, e Ali Farhadi. 2016. «You Only Look Once: Unified, Real-Time Object Detection». *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, giugno, 779–88. https://doi.org/10.1109/CVPR.2016.91.

Richard Y. Wang e Diane M. Strong. 1996. «Beyond accuracy: what data quality means to data consumers: Journal of Management Information Systems». https://dl.acm.org/doi/10.1080/07421222.1996.11518099.

Sacks, H. S., D. Reitman, D. Pagano, e B. Kupelnick. 1996. «Meta-Analysis: An Update». *The Mount Sinai Journal of Medicine, New York* 63 (3–4): 216–24.

Sahi, Louis, Romain Laborde, Mohamed-Ali Kandi, et al. 2023. «Towards Reliable Collaborative Data Processing Ecosystems: Survey on Data Quality Criteria». *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, novembre, 2456–64. https://doi.org/10.1109/TrustCom60117.2023.00345.

Schwabe, Daniel, Katinka Becker, Martin Seyferth, Andreas Klaß, e Tobias Schaeffter. 2024. «The METRIC-Framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review». *Npj Digital Medicine* 7 (1): 203. https://doi.org/10.1038/s41746-024-01196-4.

Scopus. 2025. «Scopus - Analyze search results». https://www.scopus.com/.

Sharma, Manu, Sunil Luthra, Sudhanshu Joshi, e Anil Kumar. 2022. *Implementing Challenges of Artificial Intelligence: Evidence from Public Manufacturing Sector of an Emerging Economy*. ottobre. https://doi.org/10.1016/j.giq.2021.101624.

Silver, David. 2017. «Mastering the game of Go without human knowledge». https://www.nature.com/articles/nature24270.

Sterne, Jonathan A. C., Jelena Savović, Matthew J. Page, et al. 2019. «RoB 2: A Revised Tool for Assessing Risk of Bias in Randomised Trials». *BMJ (Clinical Research Ed.)* 366 (agosto): l4898. https://doi.org/10.1136/bmj.l4898.

Sterne, Jonathan Ac, Miguel A. Hernán, Barnaby C. Reeves, et al. 2016. «ROBINS-I: A Tool for Assessing Risk of Bias in Non-Randomised Studies of Interventions». *BMJ (Clinical Research Ed.)* 355 (ottobre): i4919. https://doi.org/10.1136/bmj.i4919.

Suresh, Harini, e John Guttag. 2021. «A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle». *Equity and Access in Algorithms, Mechanisms, and Optimization*, ottobre 5, 1–9. https://doi.org/10.1145/3465416.3483305.

Swingler, George H., Jimmy Volmink, e John P. A. Ioannidis. 2003. «Number of Published Systematic Reviews and Global Burden of Disease: Database Analysis». *BMJ*

(Clinical Research Ed.) 327 (7423): 1083–84. https://doi.org/10.1136/bmj.327.7423.1083.

Taleb, Ikbal, Mohamed Adel Serhani, Chafik Bouhaddioui, e Rachida Dssouli. 2021. «Big data quality framework: a holistic approach to continuous quality management». *Journal of Big Data* 8 (1): 76. https://doi.org/10.1186/s40537-021-00468-0.

Teoh, Eric R., e David G. Kidd. 2017. «Rage against the Machine? Google's Self-Driving Cars versus Human Drivers». *Journal of Safety Research* 63 (dicembre): 57–60. https://doi.org/10.1016/j.jsr.2017.08.008.

UK Governement. 2023. «Chair's Summary of the AI Safety Summit 2023, Bletchley Park». GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2023. «Attention Is All You Need». arXiv:1706.03762. Preprint, arXiv, agosto 2. https://doi.org/10.48550/arXiv.1706.03762.

Wang e Diane M. Strong. 1996. «Beyond Accuracy: What Data Quality Means to Data Consumers». *Journal of Management Information Systems* 12 (4): 5–33.

Wang, Richard Y., e Diane M. Strong. 1996. «Beyond accuracy: what data quality means to data consumers». *J. Manage. Inf. Syst.* 12 (4): 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Whittlestone J., Cave, S, Dihal, K, Alexandrova, A, e Nyrup, R. 2019. «Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research». https://www.researchgate.net/publication/337565648_Ethical_and_societal_implications_of_algorithms_data_and_artificial_intelligence_a_roadmap_for_research.

Xie, Jiarui, Lijun Sun, e Yaoyao Fiona Zhao. 2025. «On the Data Quality and Imbalance in Machine Learning-Based Design and Manufacturing—A Systematic Review». *Engineering* 45 (febbraio): 105–31. https://doi.org/10.1016/j.eng.2024.04.024.

Yaron. 2019. «Training an AutoEncoder to Generate Text Embeddings». *Yaron Vazana*, settembre 28. https://yaronvazana.com/2019/09/28/training-an-autoencoder-to-generate-text-embeddings/.

Young, Charles, e Richard Horton. 2005. «Putting Clinical Trials into Context». *Lancet (London, England)* 366 (9480): 107–8. https://doi.org/10.1016/S0140-6736(05)66846-8.

Zhang, Pengcheng, Fang Xiong, Jerry Gao, e Jimin Wang. 2017. «Data quality in big data processing: Issues, solutions and open problems». *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, agosto, 1–7. https://doi.org/10.1109/UIC-ATC.2017.8397554.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, e Kai-Wei Chang. 2017. «Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints». arXiv:1707.09457. Preprint, arXiv, luglio 29. https://doi.org/10.48550/arXiv.1707.09457.

Zhou, Yuhan, Fengjiao Tu, Kewei Sha, Junhua Ding, e Haihua Chen. 2024. «A Survey on Data Quality Dimensions and Tools for Machine Learning Invited Paper». luglio. https://doi.org/10.1109/AITest62860.2024.00023.