Data issues in industrial AI systems

A meta-review and research strategy

Li, Xuejiao; Cheng, Yang; Møller, Charles; Lee, Jay

Published in:
Computers in Industry

DOI (link to publication from Publisher):
10.1016/j.compind.2025.104361

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Computers in Industry 173 (2025) 104361

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Data issues in industrial AI systems: A meta-review and research strategy

Xuejiao Li [a,b,*], Yang Cheng [a,c,*], Charles Møller [d,1], Jay Lee [b,2]

[a] Department of Materials and Production, Aalborg University, Aalborg, Denmark
[b] Industrial Artificial Intelligence Center, Department of Mechanical Engineering, University of Maryland, MD, USA
[c] Department of Economics and Business Administration in Hungarian Language, Faculty of Economics and Business Administration, Babes-Bolyai University, Cluj-Napoca, Romania
[d] Department of Mechanical and Production Engineering, Aarhus University, Denmark

ARTICLE INFO

Keywords:
Data Issue
Data Lifecycle
Data-Centric AI
Industrial AI
Machine Learning

ABSTRACT

In the era of Industry 4.0, artificial intelligence (AI) is assumed to play an increasingly pivotal role within industrial systems. Despite the recent trend within various industries to adopt AI, the actual adoption of AI is not as developed as perceived. A significant factor contributing to this lag is the data issues in AI implementation. How to address these data issues stands as a significant concern confronting both industry and academia. Thus, this study conducts a comprehensive meta-review of data issues and corresponding methods in industrial AI. Eighty-two data issues are identified and categorized into seven stages of the data lifecycle. To supplement the existing research that focuses more on data issues arising in historical data, this study subsequently discusses the management of real-time sensor data and expert domain knowledge. Meanwhile, it proposes a model-aware data preparation approach, which integrates the data characteristics with specific AI model requirements to enhance data usability and algorithm alignment. This approach is further integrated into a conceptual framework that combines managerial and technical perspectives for systematically resolving data issues. The framework provides actionable insights and a systematic method for AI practitioners and industrial system developers to anticipate and address data-related challenges. Finally, the study highlights future research directions. This study advances the existing body of knowledge, supports a seamless transition from traditional model-centric AI to data-centric AI, and offers practical guidelines for professionals navigating the complexities of achieving data excellence in industrial AI applications.

1. Introduction

In the era of Industry 4.0, artificial intelligence (AI) is assumed an increasingly pivotal role within industrial systems. Consequently, the discussion on industrial AI draws increasing attention. Industrial AI can be defined as a systematic discipline to enable engineers to systematically develop and deploy AI algorithms with repeating and consistent successes (Lee et al., 2019). It focuses on the development, validation, deployment, and maintenance of AI solutions for industrial applications with sustainable performance (Lee et al., 2018).

AI algorithms cannot stand alone. When discussing AI, the specific requirements of the industrial application and the available data are inevitable. The relationship between data, AI, and industrial applications is symbiotic: data serves as the foundation upon which AI algorithms operate within industrial systems. Through AI, industrial applications can harness the power of data to make data-driven decisions. In turn, industrial applications generate vast amounts of data that fuel further AI development and refinement. This cyclical relationship enables continuous improvement and innovation in industrial processes, leading to increased efficiency, productivity, and competitiveness.

The progression from data to AI and industrial applications is further depicted in Fig. 1. This process begins with data collection, storage, quality assurance, and analysis to support AI implementation. AI models

* Correspondence to: Department of Materials and Production, Aalborg University, Fibigerstræde 16, Aalborg 9220, Denmark.
E-mail addresses: xueli@mp.aau.dk, lixuejiao326@gmail.com (X. Li), cy@business.aau.dk, cy@mp.aau.dk (Y. Cheng), charles@mpe.au.dk (C. Møller), leejay@umd.edu (J. Lee).
[1] Department of Mechanical and Production Engineering, Aarhus University, Katrinebjergvej 89, Aarhus N, 8200, Denmark
[2] Clark Distinguished Chair, Director of Industrial Artificial Intelligence Center, Department of Mechanical Engineering, Maryland University, 2181 Glenn L. Martin Hall, College Park, MD, USA

https://doi.org/10.1016/j.compind.2025.104361
Received 29 May 2024; Received in revised form 15 April 2025; Accepted 31 August 2025
Available online 8 September 2025

are then introduced through feature extraction, model selection, training, validation, and evaluation. Once deployed, AI supports industrial applications such as process optimization, predictive maintenance, quality control, collaborative robotics, and workforce training (Peres et al., 2020). AI adoption spans sectors including embedded AI, resilient manufacturing, smart health initiatives, predictive energy systems, and AI-driven education (Lee et al., 2018). Throughout this workflow, systematic documentation of data processes and model deployment fosters iterative improvements, creating a closed-loop system that enhances industrial AI integration.

However, despite the critical role of data in AI implementation, data issues remain a significant bottleneck, hindering progress in industrial AI applications. Research has shown that up to 80 % of the time spent on AI projects is dedicated to data preparation and cleaning (Stonebraker et al., 2019; Zhang et al., 2023). Problems such as insufficient data, imbalanced datasets, and data fusion challenges significantly impact the performance of AI models. For example, data insufficiency limits AI's reliability. In healthcare, AI models for rare diseases struggle due to small, fragmented datasets, leading to unreliable diagnoses (Miotto et al., 2018). Similarly, in manufacturing, AI-driven quality control relies on labeled defect data, but the rarity of certain defects hampers model accuracy (Ghahramani et al., 2020). Imbalanced datasets further distort AI decision-making. In fraud detection, AI often favors the majority class, leading to poor identification of fraudulent transactions (Dal Pozzolo et al., 2015). In workplace safety, AI models trained on limited accident data may fail to detect hazardous conditions, reducing their preventive effectiveness (Shah et al., 2024). Data fusion challenges also limit AI's effectiveness. In smart factories, integrating IoT sensor data, ERP systems, and supply chain records is complex due to inconsistencies in formats and timestamps (Krishnamurthi et al., 2020). These challenges highlight the critical importance of addressing data issues to unlock AI's full potential. As industries become increasingly data-driven, how to address data issues to ensure the usability and usefulness of data stands as a significant concern.

In academia, in the conventional Model-Centric AI (MCAI) logic, researchers and developers typically focus on finding more effective models to enhance AI performance, while leaving the data largely unchanged (Zha et al., 2023). In contrast, a new research stream known as Data-Centric AI (DCAI) has emerged recently. DCAI is the discipline of systematically engineering the data needed to successfully build an AI system (Strickland, 2022). The core idea of DCAI is to pay ample attention to the data, while benefiting from the pre-trained or already developed AI models as much as possible. It involves the engineering and monitoring of data throughout the lifecycle of an AI project, addressing any potential issues that may arise (Majeed and Hwang, 2024). Our perspective aligns with DCAI, underscoring the critical importance of data in the AI projects.

We have noticed that while numerous studies address data-related topics in AI projects, a significant research gap, to the best of our knowledge, remains: no study has thoroughly examined the landscape of data issues and corresponding solutions within industrial systems. Existing literature often has two major limitations that hinder a comprehensive understanding of data issues in industrial AI. First, the existing literature often has a limited scope, with studies concentrating on specific domains or narrow technical aspects. Many studies focus on particular industries such as healthcare or manufacturing, without considering the broader, cross-industry implications of data challenges in AI applications. Additionally, technical research tends to address individual data-related problems—such as missing values, outlier detection, or feature selection—without integrating these issues into a larger framework that reflects the end-to-end AI lifecycle. This fragmented approach prevents a holistic understanding of how multiple data challenges interact and compound throughout the AI implementation process. Second, there is a lack of specificity regarding data issues across the AI lifecycle, as many studies adopt high-level conceptual frameworks without systematically identifying and analyzing the concrete data challenges encountered in real-world AI projects. While theoretical discussions on data quality, data governance, and AI ethics are abundant, they often remain abstract and do not provide empirical insights into how these issues manifest in industrial AI applications. As a result, existing frameworks do not offer actionable solutions that align with the dynamic and evolving nature of data-driven AI systems. In addition, the current frameworks lack lifecycle-based analyses of data issues, failing to account for how data issues accumulate throughout the entire AI lifecycle.

This study aims to bridge these gaps, by systematically categorizing and analyzing data challenges encountered throughout the AI project lifecycle in industrial systems. To achieve this, this study conducts a meta-review of existing literature reviews on data issues and corresponding methods in the context of AI implementation in industrial systems. The decision to undertake a meta-review is driven by the observation that while numerous literature reviews discuss data-related

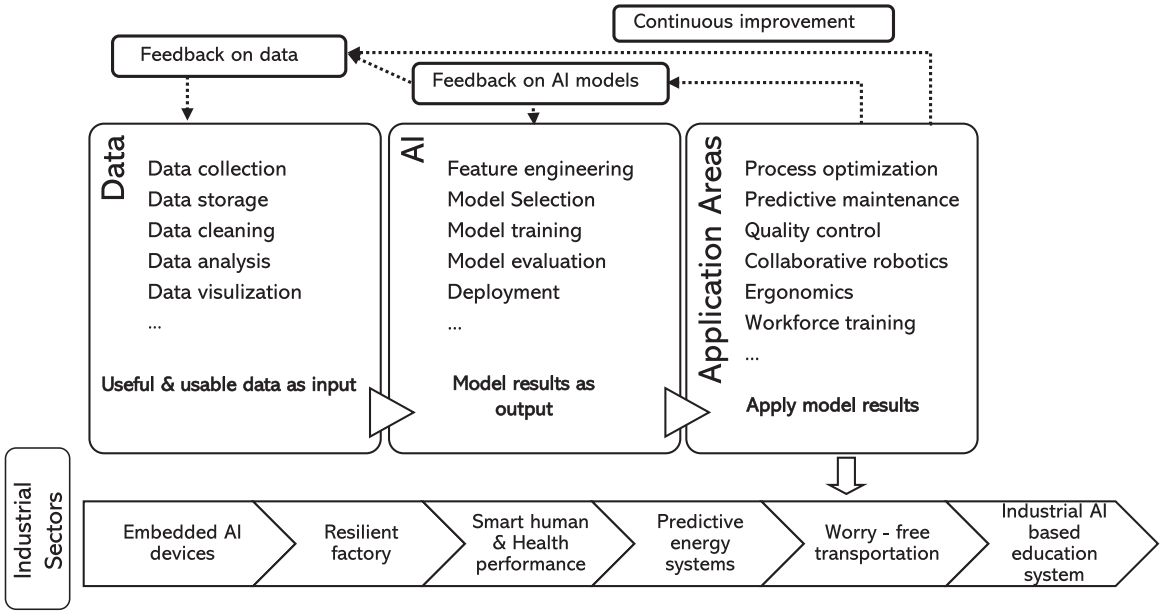


Fig. 1. Closed loop of data, AI, and industrial applications (developed by authors).

challenges, they often do so in a fragmented manner, focusing on specific technical aspects or industry domains without providing a comprehensive, cross-disciplinary synthesis. By consolidating insights from existing reviews, our study seeks to offer a holistic perspective on data challenges and their solutions. Through this approach, we systematically identify data issues reported in the literature and classify them according to data lifecycle stages, enabling a structured resolution of these challenges. Building upon these findings, we propose a framework for comprehensively managing data issues in industrial AI, ensuring that AI implementations are better aligned with real-world industrial needs.

In pursuit of this objective, the study seeks to address the following research questions:

RQ1. What are the specific data issues encountered during the implementation of AI in industrial systems?

RQ2. What methods are associated with addressing these issues?

RQ3. How can these data issues be systematically resolved?

By answering these questions, this study contributes to both academic discourse and practical applications across four key dimensions. From an academic perspective, first, this study provides a structured taxonomy of data issues across the AI data lifecycle, linking these issues to their corresponding resolution methods. This categorization deepens the theoretical understanding of how data-related challenges evolve and impact AI performance in industrial settings. By offering a comprehensive and structured overview, the study serves as a valuable reference for researchers working on AI data management, data governance, and quality control, as well as paving the way for future investigations into DCAI. Second, this study systematically reviews and maps existing methods used to mitigate data challenges in AI projects. This contribution enhances the theoretical foundation of data engineering, bridging the gap between abstract concepts and applied techniques. This mapping also identifies areas where current solutions are insufficient, providing direction for future research. Third, this study introduces a conceptual framework for systematically resolving data issues in AI implementation, integrating perspectives from model-aware data preparation, management strategies, and technical methodologies. This framework offers a foundation for future studies by highlighting key areas where data issues impact AI performance. It also identifies gaps that require further exploration, encouraging interdisciplinary research on data quality, governance, and AI reliability. Finally, from a practical standpoint, this study offers actionable insights for AI practitioners and industrial system developers, equipping them with a systematic approach for identifying and mitigating data challenges. The proposed framework serves as a strategic tool for organizations seeking to improve the efficiency and reliability of AI deployment. By helping industries anticipate and address data-related obstacles, the study not only enhances AI adoption but also reduces operational inefficiencies, improves decision-making accuracy, and supports scalable AI integration into industrial processes.

The rest of this study is organized as follows. Section 2 investigates and compares the research that aligns with similar interests with our research. Section 3 introduces the methodology of how we collect, extract, and analyze data. Section 4 presents the results of the literature analysis. Section 5 provides discussions and proposes a data management framework for systematically handling data issues in industrial AI. Section 6 identifies the future research directions on data research in industrial AI. Finally, conclusions, implication and limitations are described in Section 7.

2. Related work

2.1. Analysis of previous literature reviews

Systematic literature reviews are widely recognized as one of the most rigorous forms of scientific evidence. They offer a robust methodology for identifying and synthesizing existing research (Mallett et al., 2012; Matricciani et al., 2019). A substantial body of literature discusses the role of data in AI within industrial systems, with many of these publications being review articles. Among these, some align with our interest in examining data as the primary focus of research. While these reviews provide valuable insights, they differ in research focus and scope from our study.

Some reviews address data quality and data-related challenges in AI, but they are typically confined to specific domains, such as additive manufacturing (AM) (Zhang et al., 2023b) or healthcare (Baloch et al., 2023). These reviews provide valuable insights into how big data and AI technologies are leveraged within their respective fields. While these domain-specific studies enhance understanding within particular industries, a key limitation of these studies is that they often address data challenges in isolation, without considering how similar or overlapping issues manifest across different industrial sectors. As a result, existing literature lacks a comprehensive, cross-industry synthesis of data issues, which is essential for developing generalizable methodologies applicable across various AI-driven industrial environments.

Meanwhile, other reviews discuss data quality and data science methods across multiple industrial sectors but fail to specifically identify, categorize, or analyze data issues in AI-driven industrial systems. For instance, Madhikermi et al. (2016) provide a comprehensive review of expert maintenance systems that incorporate Multi-Criteria Decision Making (MCDM) techniques and data quality frameworks. Their discussion of data quality dimensions, such as believability, completeness, and timeliness, offers valuable insights into the theoretical aspects of data quality management. However, their study remains at a conceptual level, without examining the concrete data challenges that arise in real-world AI applications, such as missing data or inconsistencies in industrial datasets. Similarly, Arruda et al. (2023) conduct a systematic literature review on the application of data science methods and tools across various industrial sectors. While their work effectively summarizes industrial segments and the methods and tools used, these methods serve broader purposes rather than addressing specific data issues. Furthermore, the consideration of data quality in their study is limited to the quantity and origins of datasets, which overlooks critical aspects of data quality, such as accuracy, consistency, bias, and representativeness. While these studies acknowledge the importance of high-quality data, they do not explore the practical obstacles that industries face.

Finally, several reviews focus on specific types of techniques rather than offering a comprehensive analysis of approaches for addressing data challenges throughout the AI lifecycle. For instance, Zhang and Gao (2021) focus on data curation techniques, summarizing key methods for extracting information from noisy, incomplete, insufficient, and unannotated data. While their study provides useful technical insights, it remains narrowly focusing on data preprocessing and does not address how these challenges affect AI performance across different stages of implementation. Similarly, Chander and Kumaravelan (2022) explore outlier detection in Wireless Sensor Networks (WSNs), examining methods for identifying anomalous sensor readings. However, their work is limited to a single data issue (i.e., outliers) within a specific application domain (i.e., WSNs), failing to account for broader industrial AI challenges. Likewise, Wang et al. (2022) review feature engineering techniques for data-driven building energy prediction, focusing on feature selection, construction, and extraction to optimize AI model performance. While providing a detailed analysis of feature engineering, their discussion remains limited to a specific technique, without addressing how feature-related challenges interact with broader data issues. In short, a fundamental limitation of these studies is their focus on isolated techniques rather than the full spectrum of methods that can be applied in AI-driven industrial systems. The methods they discuss represent only a small subset of the broader range of AI technologies used in industrial applications. Many real-world AI systems require integrated solutions that combine multiple data processing, governance, and quality assurance techniques to ensure robust and scalable AI

deployment.

2.2. Synthesis of previous reviews and identified gaps

As highlighted previously, while various reviews touch upon data-related topics in AI applications, they exhibit two key limitations: limited scope and lack of specificity on data issues. Regarding limited scope, many studies focus on isolated domains or narrow technical topics, such as additive manufacturing, healthcare, or predictive maintenance, without considering cross-industry data challenges that are essential for scalable and transferable AI solutions. Others examine specific techniques, such as data curation, outlier detection, or feature engineering, but do not explore how these techniques interact within the broader AI implementation process. As a result, existing studies provide fragmented insights, preventing a holistic understanding of how multiple data challenges emerge, interact, and evolve throughout AI-driven industrial systems. Regarding lack of specificity, some reviews take a high-level conceptual approach, discussing data quality dimensions (e.g., believability, completeness, timeliness) or general data science methodologies, but they do not systematically analyze concrete data issues that arise throughout AI projects. While some studies acknowledge the importance of high-quality data, they do not explore the practical obstacles industries face in ensuring data reliability, integration, and governance (Zhang et al., 2023b; Baloch et al., 2023; Madhikermi et al., 2016; Zhang and Gao, 2021; Chander and Kumaravelan, 2022; Wang et al., 2022).

We also observe that the current state of the art lacks lifecycle-based analyses of data issues. As previously mentioned, some studies focus primarily on data preprocessing or specific AI model training challenges, but they overlook how data issues accumulate and evolve throughout the entire AI lifecycle—from data collection and integration to application and long-term maintenance. Without this comprehensive lifecycle perspective, industries struggle to implement proactive data governance strategies, resulting in suboptimal AI performance, unreliable predictions, and inefficiencies in industrial AI applications.

This study aims to bridge these gaps by conducting a meta-review, a rigorous "review of reviews" approach that systematically evaluates existing reviews, offering a comprehensive assessment of academic work within a specific field (Ryan et al., 2009). Meta-reviews are increasingly used in cases where conducting a systematic review becomes impractical due to the sheer volume of literature or the existence of multiple reviews with diverse findings. This method is particularly suitable for our study, given the abundance of literature reviews that touch on data issues without offering a holistic examination. Notably, no meta-review has yet addressed data issues in industrial AI systems.

Addressing the gaps identified above, this study differs from previous reviews in terms of several aspects. First, it develops a structured taxonomy ensuring that data issues are systematically classified rather than addressed in isolation. Second, it maps existing methods to address data issues, offering a consolidated and structured reference for AI practitioners and researchers. Third, it introduces a comprehensive framework that integrates both management and technical perspectives to provide actionable solutions for resolving data issues at different AI lifecycle stages. By providing a holistic, cross-domain perspective on data issues in industrial AI, this study contributes both academically and practically. It not only enables researchers to build upon a structured taxonomy of data issues, but also equips industry professionals with practical strategies to anticipate and mitigate data-related risks. In doing so, more efficient, scalable, and robust AI implementations in industrial environments can be ensured.

3. Methodology

3.1. Study identification, screening, and eligibility

The meta review followed the guidelines specified in the PRISMA statement (Mallett et al., 2012). Fig. 2 displays the PRISMA flow chart, illustrating the various phases of the study.

Initially, a keyword string was formulated with a focus on data research related to the implementation of AI within industrial systems. The construction of the keyword string involved ensuring the inclusion of at least one element from each aspect listed in Table 1, employing a combination of OR and AND operators. The three aspects encompass data issues, industrial system, and AI. Regarding the first two aspects, the elements in Table 1 represent synonyms or antonyms associated with these aspects. As for AI, the elements in Table 1 constitute the core concepts associated with AI, sourced from reference (Peres et al., 2020).

Subsequently, this search string was tailored for application across the three electronic databases central to this study: Web of Science, Scopus, and ScienceDirect. The search encompassed academic research that fulfilled the following criteria: (1) Only Peer Reviewed publications; (2) published between 2014 and August 2024 (recent 10 years); (3) Document type: Review; (4) containing at least one term from each group in the abstract, title, or keywords; and (5) composed in the English language. Following this, the obtained records were consolidated, and duplicates were removed afterwards.

After the identification phase, two authors conducted the first screening process independently based on the exclusion criteria for the screening phase listed in Table 2. Any discrepancies were resolved through discussion.

Finally, the remaining articles underwent a detailed analysis of their full text, based on the exclusion criteria for the eligibility phase listed in Table 3. No further restriction was applied.

3.2. Data extraction

For each eligible article included in the study, two types of data were extracted. First, basic information about the publication was gathered, encompassing (1) source title, (2) publication title, (3) authors, (4) keywords, (5) abstract, (6) year, (7) subject area, and (8) country/territory.

Second, the data was dedicated to address the research questions. For RQ1 "*What are the specific data issues encountered during the implementation of AI in industrial systems*", the data extracted from the eligible publications included: 1) specific AI technologies; 2) domain with the industrial systems; 3) data issues. For RQ2 "*What methods are associated with addressing these issues*", the data extracted from the eligible publications was: 4) methods for addressing data issues. For RQ3 "*How can these data issues be systematically resolved*", the data extracted from the eligible publications was: 5) data lifecycle stage of the data issues. Nevertheless, when it came to point 5), extracting data from the text in the publications might not always be straightforward. Hence, distinguishing and interpretation during the data analysis phase was essential to discern the corresponding data lifecycle stage.

3.3. Data analysis

We used data lifecycle theory for data analysis in this study. The data lifecycle theory provides a systematic approach to understanding, managing, and improving the progression of data throughout its lifecycle. When applied to analyzing data issues, this theory offers a structured framework for comprehensively examining data from inception to disposal. By systematically addressing data issues within this framework, we were able to even establish clear guidelines for data management and governance.

The data lifecycle encompasses a sequence of phases covering its entire useful lifespan, typically including various stages. More studies explore the phases of the data lifecycle. For example, an influential article by Tao et al. (2018) delineates seven distinct phases within the manufacturing data lifecycle, comprising data sources, data collection, data storage, data processing, data visualization, data transmission, and data application. Another study (Wing, 2019) defines the data lifecycle

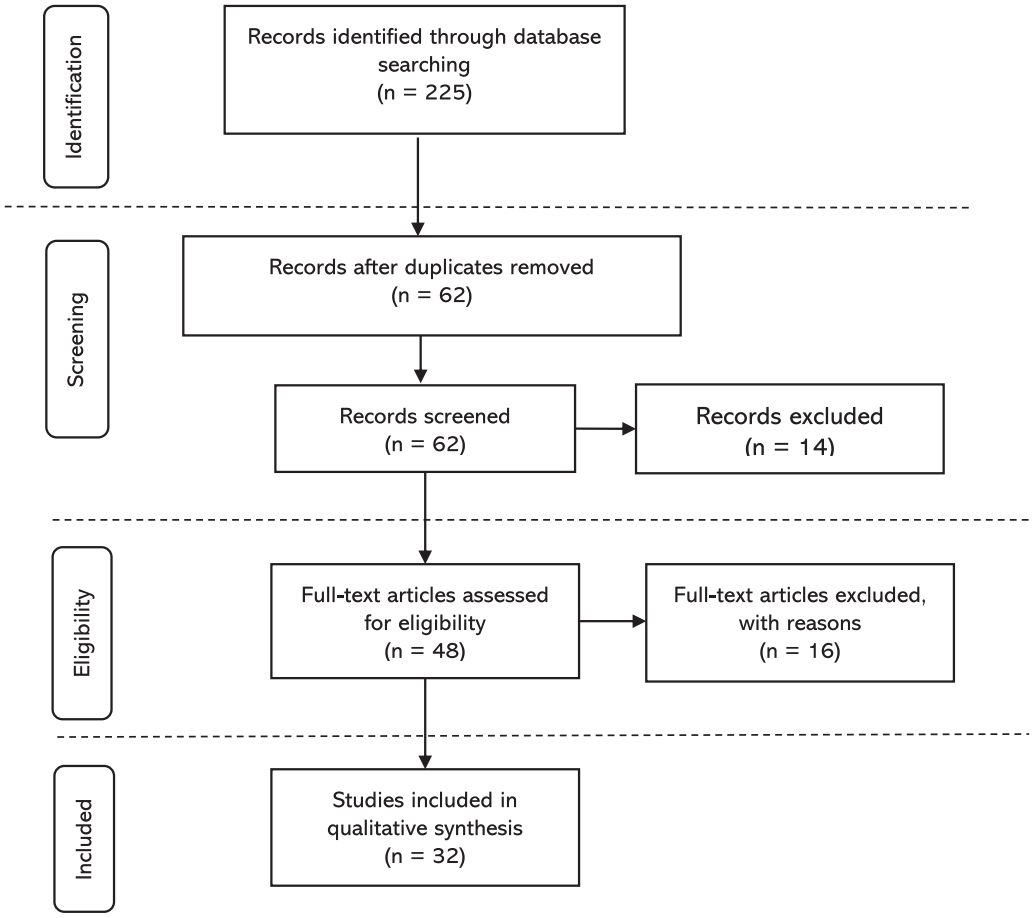


Fig. 2. PRISMA flowchart of study inclusions and exclusions for the systematic literature review (Adopted from Moher et al., 2009).

Table 1
The aspects that the subject contains and the search terms.

Aspect 1 (Data issue)	Aspect 2 (Industrial System)	Aspect 3 (AI)
Data issue	Industr*	AI
Data error	Manufactur*	Artificial intelligence
Data problem	Craft	Machine Learning
Data quality	Sector	Deep Learning
	Commerce	Data science
		Predictive analytics

Table 2
Exclusion criteria for the screening phase.

No.	Screening criteria
1	The record must be in English.
2	The record must include at least the Title, Year, Source, Abstract, and DOI.
3	The record must be a systematic review.
4	The abstract must discuss data in the context of AI implementation. Any discrepancies were resolved through discussion.

Table 3
Exclusion criteria for the eligibility phase.

No.	Screening criteria
1	The record is not relevant to data quality or data issues.
2	The record is not relevant to AI.
3	The domain of the record is beyond the industrial system.
4	Full text is not available.

as involving generation, collection, processing, storage, management, analysis, visualization, and interpretation of data. Ashmore et al. (2021) and Paleyes et al. (2022) specifically emphasize the significance of data management in the implementation of ML. They highlight that data management plays a crucial role in acquiring the data essential for synthesizing ML models. This process encompasses data collection, data preprocessing, data augmentation, and data analysis. However, scholars have not yet reached a consistent consensus regarding the precise phases that constitute the data lifecycle. In this study, with a specific emphasis on AI-related data, we synthesized existing research while specifically addressing the data pipeline in methodologies such as ML. Therefore, we emphasized phases including *data source and collection, data access and storage, data integration and interoperation, data pre-processing, data processing,* and *data security and privacy* in the data lifecycle. Additionally, we observed that apart from data itself, there are also data issues associated with the technologies used in implementing AI. Hence, we have incorporated *AI Technology adoption* as an additional category within our analytical framework.

4. Descriptive analysis of reviewed literature

This section provides the results of data analysis, which include the statistical description of the selected articles and the distribution of AI technologies and industrial domains.

4.1. Statistical description of selected articles

Understanding the landscape of selected articles is paramount to gaining insights into the prevailing trends, thematic concentrations, and scholarly contributions. This section provides a statistical overview,

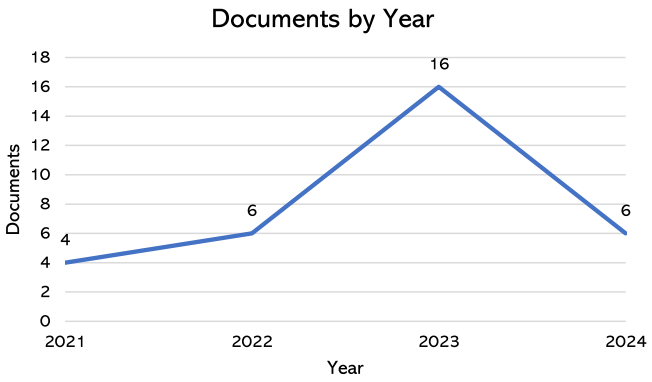


Fig. 3. Documents by year.

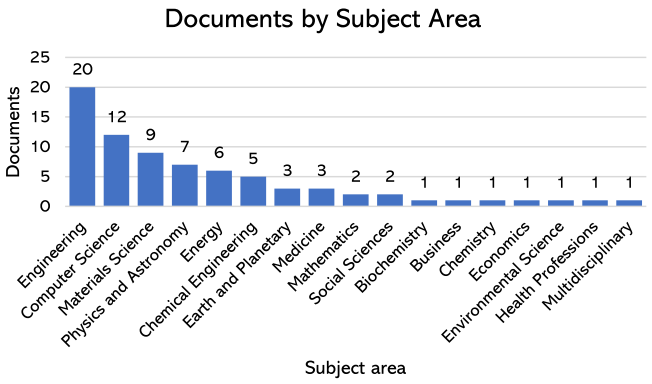


Fig. 4. Documents by Subject Area.

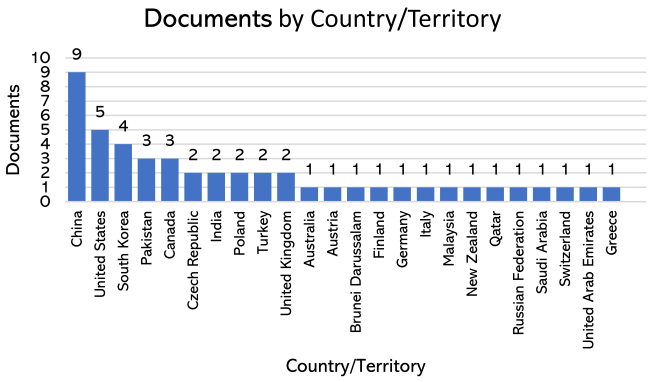


Fig. 5. Documents by Country/Territory.

employing Figs. 3, 4, and 5 to offer an examination of the selected literature.

Fig. 3 traces the distribution of articles across different publication years, shedding light on the evolving patterns of research activity. Based on the articles we collected, the peak of interest in the domain appeared to be in 2023. However, we can only gain a full picture of the number of publications for 2024 in early 2025. Therefore, there is a possibility that the number of publications in 2024 exceeds that of 2023, indicating an increasing trend.

Fig. 4 categorizes articles by subject area, revealing the thematic diversity and concentration. The primary emphasis of most articles lies within the areas of Engineering, Computer Science, and Material Science.

Fig. 5 categorizes articles by country/territory, showcasing the diversity and concentration in geographical distribution. The primary contributors to selected data issue reviews are observed to be China, the United States, and South Korea.

4.2. Distribution of AI technologies and industrial domains

The analysis of 32 articles reveals a comprehensive distribution of AI technologies, with the majority focusing on ML, including both unsupervised and supervised methods (Roy et al., 2022) and underscoring ML as a core component of AI. Among the articles, deep learning (DL), a subfield of ML, is mentioned 10 times and leverages neural networks for tasks like image recognition and natural language processing.

Two articles specifically mention Artificial Neural Networks (ANNs) (Akhtar et al., 2023; Strielkowski et al., 2023a), which are inspired by biological neural networks and used across various ML tasks. Natural Language Processing (NLP) is also highlighted in two references (Baloch et al., 2023; Papadimitriou et al., 2024) for its critical role in text analysis.

Some articles delve into specific techniques, such as Transformer architectures (Sengupta, 2023). Other techniques mentioned in the articles include traditional ML algorithms like Support Vector Machines (SVMs), Decision Trees, and Random Forests (Strielkowski et al., 2023), and Evolutionary Algorithms (EA) (Papadimitriou et al., 2024). The rise of Generative Adversarial Networks (GANs) and deep architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Papadimitriou et al., 2024) reflect their importance in generating realistic data and processing sequential information.

Further specialized techniques mentioned include Feature Engineering for improving model performance (Wang et al., 2022) and Transfer Learning (Zhu et al., 2022), which use models from one task to start new tasks, Contrastive Learning, a self-supervised learning technique, and Incremental Learning (Zhu et al., 2022), which allows models to adapt as new data arrives.

Last, AutoML (Wen and Li, 2022), which automates the ML pipeline, and Physics-Based Machine Learning (PBML) (Yüce et al., 2022) are also discussed, indicating their growing role in enhancing AI model development. In four instances (Wilson et al., 2021; Chuo et al., 2022; Aldoseri et al., 2023; Majeed and Hwang, 2024), AI technologies are referenced generally, demonstrating their broad, interdisciplinary applications.

Regarding domains in the industrial systems, the selected articles, as outlined in Table 4, cover a broad spectrum of domains within industrial systems. The most prominent areas include energy and power systems (Liu et al., 2023; Akhtar et al., 2023; Strielkowski et al., 2023a, 2023b; Yüce et al., 2022; Wang et al., 2022; Zhu et al., 2022), healthcare (Cellina et al., 2023; Sengupta, 2023; Baloch et al., 2023; Wong et al., 2023; Roy et al., 2022; Wilson et al., 2021), manufacturing (Zhang et al., 2023b; Xu et al., 2022; Tripathi et al., 2021; Zhang and Gao, 2021), and the materials industry (Ma et al., 2023; Guo et al., 2023; Papadimitriou et al., 2024). This diversity underscores AI's wide applicability across various industrial sectors.

5. Data issues and corresponding methods

As mentioned before, this study aims to provide a comprehensive overview of existing reviews on data issues and corresponding methods in industrial AI implementation. 82 data issues are identified and categorized into various stages of the data lifecycle, along with corresponding methods for addressing these issues. The data lifecycle stages include data source and collection, data access and storage, data integration and interoperation, data pre-processing, data processing, data security and privacy, and AI Technologies adoption. The findings are presented in Tables 5 to 11 and reported below.

5.1. Data source and collection stage

In the data source and collection stage, a total of 16 data issues have been identified, which can be classified into three types: *insufficient data,*

Table 4
AI technologies and Industrial domains of references.

Domain		Specific AI technology	Reference
AM		ML	Zhang et al., (2023b)
Blast furnace ironmaking		ML	Shi et al., (2023)
Building energy prediction		ML, feature engineering	Wang et al., (2022)
Energy industry		ANNs, ML	Akhtar et al., (2023)
Integrated Energy Systems		Transfer Learning, Contrastive Learning, Incremental Learning	Zhu et al., (2022)
Geothermal energy industry (above-ground geothermal operations)		ML	Abrasaldo et al., (2024)
Healthcare	Digital human twins	DL	Cellina et al., (2023)
	Dermatology	DL, Transformer architectures	Sengupta, (2023)
	Healthcare and biomedical research	SML	Roy et al., (2022)
		ML	Wong et al., (2023)
	-	NLP, ML, DL	Baloch et al., (2023)
	-	AI technology in general	Wilson et al., (2021)
Machining operations		AI technology in general	Chuo et al., (2022)
Maritime industry (maritime operations and maintenance)		ML	Durlik et al., (2023)
Material industry (material microscopic image analysis)		DL	Ma et al., (2023)
Materials science and engineering		ML	Guo et al., (2023)
		ML, DL, NLP, EA, GANs, CNNs, RNNs	Papadimitriou et al., (2024)
Metrology or precision measurement		ML	Zhang et al., (2023a)
Mining industry (optimal utilization of minerals)		DL	Liu et al., (2023b)
Oil and gas industry	Corrosion monitor	ML	Khalaf et al., (2024)
	Oil recovery	ML	Du et al., (2024)
Power systems sector	-	ML, SVMs, decision trees, random forests; ANNs	Strielkowski et al., (2023a)
	Electric power systems	DL	Strielkowski et al., (2023b)
Production and Manufacturing	-	ML	Tripathi et al., (2021)
	Smart manufacturing	DL	Xu et al., (2022)
		DL	Zhang and Gao, (2021)
Radiation oncology		ML	Field et al., (2021)
Spatial decision support systems		AutoML	Wen and Li, (2022)
Wind energy infrastructure		ML, PBML	Yüce et al., (2022)
Wind power - prediction, forecasting		ML, DL	Liu et al., (2023a)
General		AI technology in general	Aldoseri et al., (2023)
		AI technology in general	Majeed and Hwang, (2024)

excessive data, and overly diverse data. *Insufficient data* implies a lack of necessary information. For instance, challenges in collecting sufficient, accurate, and high-quality data are evident (Khalaf et al., 2024; Sengupta, 2023; Majeed and Hwang, 2024), along with limited dataset (Du et al., 2024) and limitations in data availability (Khalaf et al., 2024; Ma et al., 2023; Wong et al., 2023; Akhtar et al., 2023; Zhu et al., 2022; Wilson et al., 2021; Papadimitriou et al., 2024). Du et al. (2024) also report that some datasets are generated solely through simulations due to the challenges in acquiring authentic data from experimental or field sources. The corresponding methods vary and can be categorized into both management approaches and technical methods. Regarding management approaches, the suggestions include fostering collaborative efforts among developers, startups, and academia for data collection and technical methodologies. Regarding technical methods, it is suggested to involve constructing large, anonymized, representative, and regularly updated open-source datasets, creating comprehensive benchmarking datasets, using data augmentation techniques (e.g., generative algorithms) to artificially produce representative data, and employing transfer learning and domain adaptation techniques. Specifically, Papadimitriou et al. (2024) recommend the integration of physical laws into ML models, a concept known as Physics-Informed Neural Network (PINN). They indicate that by incorporating physical laws, PINNs can better generalize and make reliable predictions even in regions of the input space where data might be scarce or unavailable. Du et al. (2024) summarize that when faced with restricted data, researchers often use single shot learning strategies, wherein models are pre-trained on similar datasets and subsequently refined through experience. Last, Majeed and Hwang (2024) propose a Data-Centric AI paradigm that encompasses a series of data-tailored actions, and it is expected to address the insufficient data problem.

Excessive data refers to a situation where data is immense in volume, high in velocity, and diverse in variety (Durlik et al., 2023; Aldoseri et al., 2023). To address this, industry-wide initiatives promoting data standardization can be advocated (Durlik et al., 2023). Additionally, establishing clear data requirements, including the identification of data types, sources, and quantities, is recommended (Aldoseri et al., 2023).

The challenge of *overly diverse data sources* arises when dealing with varying data formats from different sources (Khalaf et al., 2024), coupled with a lack of standardization (Strielkowski et al., 2023a). While specific methods for addressing overly diverse data sources are not explicitly outlined in the references, one potential approach could involve establishing standards for data sources and collection processes. Furthermore, regarding diverse data sources, it is important to recognize that data can come from a variety of origins, including historical storage, real-time sensor data driven by IoT developments, and human interactions or knowledge. Managing data from these heterogeneous sources is crucial. This topic will be discussed in more detail in Section 5.1.1.

Table 5 presents the data issues and corresponding methods identified at the data source and collection stage, as outlined below.

5.2. Data access and storage stage

In the Data access and storage stage, two data issues have been identified, labeled as *"no access" and "cost issue." No access* refers to the inability to utilize large datasets due to legal or regulatory constraints (Liu et al., 2023). The recommended method for addressing this issue involves applying for the necessary permissions and licenses to access and use the data. On the other hand, *cost issues* indicate challenges in storing data while meeting scalability and performance requirements within budget constraints (Aldoseri et al., 2023). Proposed methods to mitigate this challenge include maintaining existing data storage and retrieval systems and employing technologies such as nonvolatile memory or distributed storage systems for storing data. Table 6 presents the data issues and corresponding methods identified at the data access and storage stage, as outlined below.

5.3. Data integration and interoperation stage

Within the data integration and interoperation stage, four data issues have been identified and characterized as *"system issue" and "data fusion issue"*. *System issues* entail the lack of integration in existing systems

Table 5
Data issues and corresponding methods in the data source and collection stage.

Data issue type	Specific issue	Method	AI technology	Domain
Insufficient data	1. Difficult to collect sufficient and accurate data (Khalaf et al., 2024)	/	ML	Oil and gas industry – maintenance
	2. Lack of historical data (Khalaf et al., 2024)	/	ML	Oil and gas industry – maintenance
	3. Difficult to find high-quality, diverse, and representative data (Sengupta, 2023)	Developers, startups, and academia collaborate on data collection; Build large open-source, anonymized, representative, and regularly updated datasets; Develop generative algorithms to artificially create representative images	Transformer architectures, DL	Healthcare
	4. Limited availability of annotated data for training DL models (Ma et al., 2023)	Image Augmentation	DL	Material industry, material microscopic image analysis
	5. Limited data availability and generalizability (Wong et al., 2023)	Comprehensive benchmarking datasets for antigen selection and vaccine efficacy	ML	Biomedical research and healthcare
	6. Lack of data (Akhtar et al., 2023)	Data augmentation techniques, transfer learning and domain adaptation techniques	ML, ANNs	Energy industry
	7. Limited data (Zhu et al., 2022)	Transfer Learning, involves leveraging knowledge from a correlated domain to improve outcomes in the target domain; Generative Models: such as GANs to generate synthetic but realistic time-series data, addressing the challenge of limited data on hand	Transfer Learning, Contrastive Learning, GANs, Incremental Learning	Integrated Energy Systems
	8. Not all events are captured in the medical record (Wilson et al., 2021)	Data cleaning, artefact removal	/	Healthcare
	9. Data being scarce or unavailable (Papadimitriou et al., 2024)	The integration of physical laws into ML models - PINNs	ML	Materials science and engineering
	10. Limited dataset, data scarcity (Du et al., 2024)	Deploying data augmentation techniques	ML	Oil and gas industry
	11. Database was generated by simulation, difficult to obtain authentic data from experimental and/or field sources (Du et al., 2024)	Use single-shot learning strategies, wherein models are pre-trained on similar datasets and subsequently refined through experience	ML	Oil and gas industry
	12. Owing to constrained budgets for data collection or a lack of expertise in handling datasets, datasets either do not exist or cannot be curated (Majeed and Hwang, 2024)	Data centric-AI paradigm encompassing a series of data-tailored actions	/	/
Excessive data	1.Sheer volume and variety of data, immense in volume, high in velocity, and diverse in variety (Durlik et al., 2023)	Champion industry-wide data standardization initiatives	ML	Maritime industry, maritime operations and maintenance
	2. Data deluge, the diversity of data sources, and the need for representative samples (Aldoseri et al., 2023)	Data requirements for identifying data types, sources, and quantities	ML	/
Overly diverse data sources	1. Varying data formats from different sources (Khalaf et al., 2024)	/	ML	Oil and gas industry – maintenance
	2. Lack of standardization across different data sources (Strielkowski et al., 2023a)	/	ML	Power systems sector

Table 6
Data issues and corresponding methods in the data access and storage stage.

Data issue type	Specific issue	Method	AI technology	Domain
No access	1. Not able to use large datasets due to legal or regulatory constraints (Liu et al., 2023)	Necessary permissions and licenses to access and use data	ML, DL	Wind power - prediction, forecasting
Cost issue	1. Difficult to meet scalability, performance requirement within budget on data storage (Aldoseri et al., 2023)	Maintaining data storage and retrieval systems; Nonvolatile memory; Distributed storage systems	ML	/

(Khalaf et al., 2024; Wilson et al., 2021), coupled with a deficiency in standardization and interoperation between systems (Durlik et al., 2023). When efforts are made to integrate systems, challenges such as compatibility issues and resistance to change from traditional practices may arise (Khalaf et al., 2024). In some instances, unintegrated systems are still reliant on paper records (Wilson et al., 2021). The methods for addressing these issues can be broadly categorized into management and technical approaches. For management methods, initiatives, such as catalyzing joint development and uniting stakeholders, advocating for collaborative data sharing platforms for cross-verification, anticipating challenges in data encoding and extraction, and fostering collaboration with IT teams, are recommended. On the technical front, it is suggested to employ advanced data integration techniques, integrate versatile middleware solutions for system harmonization, prepare for extensive data wrangling to align datasets from various sources, and accurately define project data.

Regarding the *data fusion issues*, Zhang et al. (2023a) suggest the presence of potential data quality issues without providing specific details. Gaussian Processes (GPs) algorithms are proposed as a viable solution. GPs, a type of non-parametric model widely used in machine learning for regression and probabilistic classification, create a distribution over functions. This unique characteristic allows them to adeptly capture complex relationships between input and output variables. Yüce et al. (2022) discuss the difficulties in quantifying confidence in fused data and suggest the adoption of a Bayesian framework to construct more precise surrogate models. Table 7 presents the data issues and corresponding methods identified at the data integration and interoperation stage, as outlined below.

Table 7
Data issues and corresponding methods in the data integration and interoperation stage.

Data issue type	Specific issue	Method	AI technology	Domain
System issue	1.Integration with Existing Systems: Compatibility issues, resistance to change from traditional practices (Khalaf et al., 2024)	Develop advanced data integration techniques	ML	Oil and gas industry – maintenance
	2. Lack of standardization and interoperation between systems (Durlik et al., 2023)	Integrate versatile middleware solutions for system harmonization; Catalyze joint development initiatives, uniting stakeholders; Advocate for collaborative data sharing platforms for cross-verification	ML	Maritime industry, maritime operations and maintenance
	3. IT systems are not integrated and are often supported by paper records (Wilson et al., 2021)	Be prepared for extensive data wrangling to align datasets from various sources; Anticipating data encoding and extraction challenges; Accurately defining project data; Collaboration with IT teams	/	Healthcare
Data fusion issue	1. Data quality issues in data fusion (Zhang et al., 2023a)	GPs algorithms	ML	Metrology, precision measure
	2. Quantify the confidence in the fused data (Yüce et al., 2022)	Using a Bayesian framework to construct more accurate surrogate models	ML	Wind energy infrastructure

5.4. Data pre-processing stage

Most of data issues are concentrated in the data pre-processing stage. A total of 41 issues have been identified, grouped into 13 types. These 13 types can be further divided into two groups. The first group pertains to data quality issues, where data quality is assessed based on criteria such as accuracy, reliability, consistency, completeness, and relevance for a specific purpose. Within this group, we identify issues like *incompleteness, imbalance, inaccuracy, outliers, and mislabeling*. The second group focuses on data nature issues, referring to the inherent characteristics or properties of data. Data nature is influenced by factors such as structure, type, format, and context. In this context, our study identifies data nature issues such as *high dimensionality, lack of structure, mixed-frequency data, non-stationarity, non-Gaussian distribution, nonlinearity, non-heterogeneity, and heteroscedasticity*.

5.4.1. Data quality issue

The *incompleteness of data*, often referred to as missing, fragmented, or insufficient data, may arise due to unstandardized data collection or data loss during storage or transfer. To tackle this issue, Liu et al. (2023) recommend leveraging a multitude of meteorological data prediction sources and employing data assimilation techniques. Shi et al. (2023) suggest that for a small amount of missing data (<5 %), it can be safely disregarded, while for a significant amount of missing data (>60 %), deletion of the dataset without utilization is advised. For intermittently short-term data loss, interpolation techniques can be applied. In cases of continuous long-term data loss, they also recommend analyzing the correlation between missing data variables and other complete variables, followed by filling in the missing data. Furthermore, Chuo et al. (2022) propose the use of data imputation methods, involving the replacement of missing or incomplete data values with estimated or predicted values. In situations where there are insufficient samples in the collected dataset, they suggest considering the application of additional sensors or generating virtual data using GANs. Tripathi et al. (2021) also acknowledge the appropriateness of deletion when necessary. Alternatively, imputation techniques, such as replacing with average/most frequent values, multiple imputations, expectation maximization algorithm, and classification/regression trees, can be employed. Zhang and Gao (2021) emphasize the importance of data imputation as well. Moreover, they suggest for time series data, recurrent neural networks can be applied; and for image convolutional data, neural networks or hybrid approaches are recommended. While data augmentation serves to tackle data incompleteness, it is not without challenges. Generated images through data augmentation may exhibit high similarity, and an overreliance on this technique may introduce the multicollinearity problem (Liu et al., 2023). Strategies to address these issues include acquiring more image data, regulating the extent of data augmentation, and leveraging emerging data collection technologies such as crawlers, big data, and image data augmentation based on GANs.

Data imbalance is a very common issue in the context of ML and data analysis (Abrasaldo et al., 2024; Aldoseri et al., 2023; Strielkowski et al., 2023b; Chuo et al., 2022; Zhang and Gao, 2021). It refers to a situation where the distribution of classes in a dataset is not uniform. To address this issue, two types of approaches are suggested. First, at the data level, Abrasaldo et al. (2024) propose employing data resampling techniques, such as random under-sampling. Additionally, Aldoseri et al. (2023) emphasize the importance of meticulous attention to data collection and pre-processing. Chuo et al. (2022) advocate for the use of under/over-sampling methods, while Zhang and Gao (2021) suggest data interpolation, exemplified by the Synthetic Minority Over-sampling Technique. Another approach involves the use of generative models, such as the Generative Model, a type of machine learning model designed to produce new data instances resembling a given training dataset, or GANs. GANs consist of two neural networks – a generator that creates synthetic data samples and a discriminator that evaluates the authenticity of a given sample. Second, at the algorithm level, Aldoseri et al. (2023) underscore the significance of understanding the principles of algorithmic fairness, fairness metrics, explainable AI, and interdisciplinary collaboration. Strielkowski et al. (2023b) posit that DL algorithms may need to generalize more effectively to underrepresented classes. Furthermore, Chuo et al. (2022) suggest adjusting the weight values of minor classes during the training process as a means to address data imbalance effectively.

Data inaccuracy is another issue. In dealing with inaccurate values, Chuo et al. (2022) recommend several strategies, including reducing sampling frequency, filling empty sections using interpolation methods, and ensuring synchronization in the data processing stage when dealing with different sampling rates. Addressing data noise, Zhang and Gao (2021) suggest employing data denoising techniques. Specifically, for projection-based noise, the recommendation is to utilize local geometric projection. For frequency-based noise, employing methods, such as empirical model decomposition or wavelet transform, is suggested. In the case of noise-assisted noise, they advise using stochastic resonance, while for data-driven/hybrid noise, the proposed approach involves utilizing generative prior and unrolled optimization. Majeed and Hwang (2024) highlight that recent advancements in sophisticated pipelines address noisy data and data scarcity issues, significantly improving ML accuracy with limited data. These include tools like DataPerf, which tackles data fragmentation and evaluates data quality in real-world applications (Mazumder et al., 2024), the open-source platform YMIR (Huang et al., 2021), which uses APIs to optimize data for computer vision tasks, and dcbench, which is designed to enhance the quality of training data (Eyuboglu et al., 2022).

Data outliers are observations or data points that deviate significantly

Table 8
Data issues and corresponding methods in the data pre-processing stage.

Data issue type	Specific issue	Method	AI technology	Domain
Incompleteness (data quality issue)	1. Data incomplete (Liu et al., 2023)	Considering as many predictions meteorological data sources as possible; using data assimilation techniques	ML, DL	Wind power - prediction, forecasting
	2. Missing entries in heterogeneous data sets (Roy et al., 2022)	/	Supervised machine learning	Healthcare and biomedical research
	3. Missing data, 1) Small amount of missing data ($< 5\ \%$) or significant amount of missing data (>60 %), 2) Intermittently short-term loss of data, 3) Continuously long-term loss of data (Shi et al., 2023)	1) Ignore or delete, 2) Interpolation techniques, 3) analyzing the correlation between the missing data variables and other complete variables and then filling in the missing data	ML	Blast furnace ironmaking
	4. Fragmented data (Chuo et al., 2022)	Data imputation methods	Various	Machining operations
	5. Insufficient samples of the collected data set (Chuo et al., 2022)	Applying more sensors or generating virtual data using GAN	Various	Machining operations
	6. Incomplete and Missing data (Tripathi et al., 2021)	Deletion; replacement with average/most frequent values, multiple imputations, and expectation maximization algorithm, classification/regression trees for imputation	ML	Production and Manufacturing
	7. Missing or incomplete data (Zhang and Gao, 2021)	Data imputation; For time series data: Recurrent neural network; For image convolutional: neural network, hybrid approach	DL	Smart manufacturing
	8. The generated images by data augmentation may have high similarity; excessive reliance on data augmentation may lead to the multicollinearity problem (Liu et al., 2023)	Gather more image data; control the level of data augmentation; use emerging data collection technologies such as crawler, big data, and GAN-based image data augmentation	DL	Mining industry, sorting and optimal utilization of minerals
Imbalance (data quality issue)	1. Imbalanced input dataset (Abrasaldo et al., 2024)	Data resampling. e.g., random under sampling	ML	Geothermal energy industry
	2. Data bias and unfairness (Aldoseri et al., 2023)	Pay attention to data collection and pre-processing, Algorithmic Fairness, Fairness Metrics, Explainable AI, Interdisciplinary Collaboration	ML	/
	3. Unbalanced datasets (Strielkowski et al., 2023b)	DL algorithms may need to generalize better to underrepresented classes	DL	Electric power systems
	4. Significant difference in the amount of data between classes (Chuo et al., 2022)	Under/over-sampling of data (approaches at data level) or adjusting the weight values of minor classes in training process (approach at algorithm level)	Various	Machining operations
	5. Data imbalance (Zhang and Gao, 2021)	Data interpolation: Synthetic minority over-sampling technique; Generative model Variational autoencoder; GANs	DL	Smart manufacturing
Inaccuracy (data quality issue)	1. Inaccurate values measured and intermittent sampling delay (Chuo et al., 2022)	Reducing the sampling frequency, filling the empty section using interpolation methods, and synchronization in the data processing stage for different sampling rates	Various	Machining operations
	2. Data noise (Zhang and Gao, 2021)	Data denoising. Projection-based: Local geometric projection; Frequency-based: Empirical model decomposition, Wavelet transform; Noise-assisted: Stochastic resonance; Data-driven/hybrid: Generative prior, Unrolled optimization.	DL	Smart manufacturing
	3. Noisy data (Majeed and Hwang, 2024)	Data-enhanced pipelines or instruments, such as DataPerf tool, YMIR	/	/
Outlier (data quality issue)	1.Manual input error or sensing device failure; Abnormal furnace state (Shi et al., 2023)	Statistical methods such as box plot and threesigma; ML methods such as clustering and isolated forests; Process methods such as operating guidelines	ML	Blast furnace (BF) ironmaking
	2. Data outliers (Zhang and Gao, 2021)	Outlier detection. Data-level: Autoencoder; Model-level: Probabilistic neural network, Temperature scaling, Input perturbation	DL	Smart manufacturing
Mislabeling (data quality issue)	1. Low label accuracy, mislabeling errors (Zhang et al., 2023b)	Finite element method model	ML	Additive manufacturing
	2. Inefficient interpretability of datasets (Zhang and Gao, 2021)	Data annotation. Image annotation: Fully convolutional network, U-Net, Mask region-based CNN; Natural language processing: Word embedding, Transformer, BERT	DL	Smart manufacturing
	3. Burden of manual data annotation (Aldoseri et al., 2023)	Active learning, weak supervision, transfer learning	ML	/
	4. Time consuming manual labeling process (Zhang et al., 2023b)	Annotation tool	ML	Additive manufacturing
	5. Requiring large amount of labeled data (Papadimitriou et al., 2024)	Active learning	ML	Materials science and engineering
High dimensionality (data nature issue)	1. Curse of dimensionality (Sengupta, 2023)	/	Transformer architectures, DL	Healthcare
	2. High-dimensional and complex nature of the data (Akhtar et al., 2023)	Feature selection techniques; dimensionality reduction techniques	ML, ANNs	Energy industry
	3. Spatial dimension of load data, load data varies over time and space (Akhtar et al., 2023)	Spatial–temporal models; clustering and spatial interpolation techniques	ML, ANNs	Energy industry

(*continued on next page*)

Table 8 (*continued*)

Data issue type	Specific issue	Method	AI technology	Domain
	4. The data dimension is large, but the sample size is very small (Roy et al., 2022)	/	SML	Healthcare and biomedical research
	5. High dimensionality (Tripathi et al., 2021)	Dimensionality reduction (e.g., PCA) and feature selection methods	ML	Production and manufacturing
	6. Materials data is often high dimensional, causing issues related to overfitting and the curse of dimensionality (Papadimitriou et al., 2024)	Dimensionality reduction	ML	Materials science and engineering
Lack of structure (data nature issue)	1. Big data lacks defined structure (Baloch et al., 2023)	Information extraction tech, and advanced analytics like NLP, ML, and DL	NLP, ML, and DL	Healthcare
Mixed frequency data (data nature issue)	1. 1) High-frequency data are converted to low-frequency ones, 2) Low-frequency data are converted to high-frequency ones, 3) Mixed-frequency data model (Shi et al., 2023)	1) Average or sum or latest value, 2) Copy, 3) Mixed data sampling and mixed frequency vector autoregression	ML	Blast furnace ironmaking
Non-stationary (data nature issue)	1. Dynamic nature of load data, non-stationary data (Akhtar et al., 2023)	Time varying models, such as autoregressive moving average models; DL models, such as DNNs	ML, ANNs	Energy industry
	2. Non-established and non-stationarity of data (Roy et al., 2022)	/	SML	Healthcare and biomedical research
	3. Non-Stationary Data Distribution (Field et al., 2021)	Attention or Memory Mechanisms can be embedded into the model development process to handle changes over time and variations in different geographical locations	ML	Radiation oncology
Non-Gaussian distribution (data nature issue)	1. Non-Gaussian and heavy-tailed nature of the load data (Akhtar et al., 2023)	Robust STLF models; distributional STLF models	ML, ANNs	Energy industry
Nonlinearity (data nature issue)	1. Nonlinear and non-monotonic relationships between the load data, weather data, and other predictors (Akhtar et al., 2023)	Non-parametric models, such as decision trees or random forests; kernel-based models, such as kernel regression or support vector machines	ML, ANNs	Energy industry
	2. Presence of highly correlated variables; Large number of variables (Guo et al., 2023)	Evaluating performance against a gold standard, analyzing performance based on experimental groups, matching ML classifiers, and testing against expert-labeled test sets	ML	Materials science and engineering
	3. Massiveness, nonlinearity, and high dimensionality of data (Guo et al., 2023)	Combining the statistical data-driven methods with the ML based methods	ML	Wind energy Infrastructure
	4. Data often captures complex, non-linear relationships among variables (Papadimitriou et al., 2024)	Advanced data integration, data mining, and ML techniques	ML	Materials science and engineering
Non-heterogeneity (data nature issue)	1. Non-heterogenous data across different models (Yüce et al., 2022)	Connecting previously unconnected data sources	ML	Wind energy infrastructure
Heteroscedasticity (data nature issue)	1. Data Heteroscedasticity (Tripathi et al., 2021)	Residual analysis, statistical tests, and alternative models like weighted least squares	ML	Production and Manufacturing

Table 9
Data issues and corresponding methods in the data processing stage.

Data issue type	Specific issue	Methods	AI technology	Domain
Lack of efficiency	1. Lack of efficient data processing and analysis capabilities, and lack of reliable communication systems (Khalaf et al., 2024)	Edge Computing and IoT	ML	Oil and gas industry – maintenance
	2. Processing large datasets requiring enormous computing resources to (Aldoseri et al., 2023)	Specialized hardware such as GPUs and TPUs: Model compression, pruning, and quantization; Transfer learning	ML	/
	3. Dealing with dataset sizes to balance performance vs. acceptable running times (Wen and Li, 2022)	Ensuring diverse, representative, and reliable data through resource distribution	AutoML	Spatial decision support systems
	4. Using longer data sets (Yüce et al., 2022)	Develop models with large degrees of freedom	ML	Wind energy infrastructure
Complicated procedures	1. Complicated testing procedures specifically related to using ML for vaccine development (Wong et al., 2023)	/	ML	Biomedical research and healthcare
Model Overfitting	1. Models becoming overly complex and lacking generalizability to new data (Strielkowski et al., 2023a)	/	ML	Power systems sector
	2. Insufficien training data and the absence of well-defined stopping criteria during training caused model overfitting (Du et al., 2024)	Adjust the model's structure, including weight modification	ML	Oil and gas industry

from the overall pattern or distribution of the dataset. In the context of Blast Furnace ironmaking, Shi et al. (2023) suggest that outliers may arise from manual input errors, sensing device failures, or abnormal furnace states. They propose employing statistical methods, such as box plots and three-sigma analysis, for outlier identification. Additionally, ML methods like clustering and isolated forests can be utilized. From a management perspective, process-oriented approaches, such as implementing operating guidelines, are recommended. For outlier detection, Zhang and Gao (2021) advocate the use of Autoencoder at the data level. They also suggest that at the model level, techniques such as Probabilistic Neural Network, Temperature Scaling, and Input Perturbation can be applied. These methods enhance the ability to identify and manage outliers in the Blast Furnace ironmaking process.

Mislabeling is also a common issue. ML algorithms sometimes require large amounts of labeled data (Papadimitriou et al., 2024). Low label accuracy and mislabeling errors can cause inefficient interpretability of

Table 10
Data issues and corresponding methods in the data security and privacy.

Data issue type	Specific issue	Methods	AI technology	Domain
Information leakage	1. Compromise on data security leading to information leakage (Durlik et al., 2023)	Prioritize data encryption; Mandate regular security audits and penetration testing; Establish robust access controls and rigorous authentication protocols	ML	Maritime industry, maritime operations and maintenance
	2. Inadvertent release of identifiable information, or identification through deductive disclosure (Wilson et al., 2021)	Bring data scientists and data processing capabilities into the organization; Build a trusted research environment	/	Healthcare
Information misuse	1. Unauthorized access, breaches, and misuse of sensitive information (Cellina et al., 2023)	Implementing robust data privacy and security measures, including encryption, access restrictions, and compliance with data protection regulations	DL	Healthcare
	2. Breaches (Baloch et al., 2023)	Unsynchronized sensor data analytics model	NLP, ML, DL	Healthcare
Attacks	1. Inference Attacks, model inversion or membership inference attacks (Aldoseri et al., 2023)	Federated training; DP	ML	/
	2. Adversarial Attacks, Data Poisoning, Model and data tampering (Aldoseri et al., 2023)	Adversarial training techniques; Monitoring and anomaly detection techniques; Compliance with Data Protection Regulations	ML	/

datasets. In terms of label accuracy, Zhang et al. (2023b) indicate that the experimental results (such as mechanical properties) and computational results (such as thermal distribution) obtained from the finite element method model are more promising than manual labeling. Zhang and Gao (2021) introduce various techniques for data annotation. Specifically, in the realm of image annotation, the suggested methods include Fully Convolutional Network, U-Net, and Mask Region-Based CNN. In the domain of NLP, the recommended techniques involve Word Embedding and Bidirectional Encoder Representations from Transformers. Regarding manual data annotation, it can pose a challenge due to its significant time and energy consumption. Aldoseri et al. (2023) recommend techniques such as active learning, weak supervision, and transfer learning to alleviate this burden. Among these techniques, active learning can be used to select the most informative examples to label, based on their potential to improve the performance of ML models (Papadimitriou et al., 2024). Furthermore, Zhang et al. (2023b) propose the utilization of annotation tools.

5.4.2. Data nature issue

The challenge of *high dimensionality*, often referred to as the "curse of dimensionality" (Sengupta, 2023), is a prevalent issue in machine learning and AI. It arises from the complexity and high dimensionality of data, leading to the utilization of numerous predictors and potentially resulting in overfitting and diminished accuracy (Akhtar et al., 2023; Papadimitriou et al., 2024). What compounds the problem is the scenario where the data dimension is large, yet the sample size is very small (Roy et al., 2022). To mitigate these challenges, the recommended approaches include the use of feature selection techniques and dimensionality reduction methods, such as PCA (Akhtar et al., 2023; Tripathi et al., 2021). Specifically for addressing the spatial dimension of load data and the temporal and spatial variability in load data, the suggestions involve employing spatial–temporal models, clustering, and spatial interpolation techniques (Akhtar et al., 2023). These strategies aim to enhance the handling of high-dimensional and complex data, contributing to more effective and accurate machine learning outcomes.

Lack of structure is another significant data nature issue, which can be addressed through Information Extraction (IE) technologies and advanced analytics, such as NLP, ML, and DL (Baloch et al., 2023).

The complexity introduced by *mixed-frequency data* is also an issue, which makes future data processing and model development challenging. To tackle this issue, the suggestions include converting high-frequency data to low-frequency by using methods like averaging, summing, or selecting the latest value; converting low-frequency data to high-frequency through copying; and developing mixed-frequency data models using mixed data sampling and mixed frequency vector autoregression (Shi et al., 2023).

Certain statistical characteristics of data can pose challenges as well. The first characteristic is *non-stationarity*, which refers to variations in the statistical properties of time series data over time (Akhtar et al., 2023; Roy et al., 2022; Field et al., 2021). To address this issue, the methods, such as employing time-varying models like autoregressive moving average models or utilizing DL models such as DNNs, have been suggested (Akhtar et al., 2023). Additionally, Field et al. (2021) propose incorporating attention or memory mechanisms into the model development process to handle temporal changes and geographical variations effectively.

Akhtar et al. (2023) also note the second characteristic, i.e., *non-Gaussian and heavy-tailed* nature of load data. They further recommend the use of robust Short-Term Load Forecasting (STLF) models and distributional STLF models.

Addressing *nonlinearity* is identified as the third challenge (Akhtar et al., 2023; Guo et al., 2023; Papadimitriou et al., 2024). The proposed methods include the application of non-parametric models like decision trees or random forests, kernel-based models such as kernel regression or support vector machines (Akhtar et al., 2023), and other advanced data integration, data mining, and ML techniques (Papadimitriou et al., 2024). Evaluation strategies involve assessing performance against a gold standard, analyzing performance based on experimental groups, ML classifiers, testing against expert-labeled test sets, and integrating statistical data-driven methods with ML-based approaches (Guo et al., 2023).

Non-heterogeneous data across various models is mentioned as the fourth challenge, which indicates a lack of diversity or variability among the components. Yüce et al., (2022) suggest connecting previously unconnected data sources as the solution for this challenge.

The last issue of *data heteroscedasticity* is highlighted in Tripathi et al. (2021), signifying a statistical phenomenon where the variability of the dependent variable is not constant across different levels or values of an independent variable. To address this concern, the authors recommend employing residual analysis, statistical tests, and alternative models

Table 11
Data issues and corresponding methods in AI technologies adoption stage.

Data issue type	Specific issue	Methods	AI technology	Domain
Cost	1. Implementing AI technologies, installing the required sensors and infrastructure for data collection can be costly (Khalaf et al., 2024)	Cost-Benefit Analysis	ML	Oil and gas industry – maintenance
	2. Implementing advanced data analytics can be costly, including hardware, software, infrastructure, hiring or training staff (Durlik et al., 2023)	/	ML	Maritime industry, maritime operations and maintenance
Reusability	1. Challenges on updating existing models with new data, while maintaining performance consistency and achieving reproducible solutions (Wen and Li, 2022)	/	AutoML	Spatial decision support systems
Interpretability	1. Challenges on understanding why models perform better or worse, and explaining the reasons behind certain model actions (Wen and Li, 2022)	/	AutoML	Spatial decision support systems
	2. Doubts on Deep Neural Networks Reliability (Xu et al., 2022)	Concept Drift Detection, Uncertainty Estimation, Out of Distribution Detection	DL	Smart manufacturing
	3. Tradeoff between model accuracy and interpretability (Papadimitriou et al., 2024)	/	ML	Materials science and engineering

such as weighted least squares. Table 8 presents the data issues and corresponding methods identified at the data pre-processing stage, as outlined below.

5.5. Data processing stage

In the data processing stage, seven data issues have been identified, falling into the categories of *lack of efficiency, complicated procedures*, and *model overfitting*. Addressing the *lack of efficiency*, Khalaf et al. (2024) point out the absence of efficient data processing and analysis capabilities, along with a deficiency in reliable communication systems. This can be mitigated through the implementation of Edge Computing and IoT. Edge Computing and IoT contribute to reducing latency and facilitating real-time data processing and analysis. Furthermore, Aldoseri et al. (2023) emphasize the challenge of processing large datasets that demand enormous computing resources. To address this, specialized hardware like GPUs and TPUs can be employed to accelerate AI training and inference. Techniques such as model compression, pruning, and quantization are suggested for optimizing AI models, while transfer learning proves beneficial in reducing the amount of required training data and enhancing the efficiency of the training process. A challenge highlighted by Wen and Li (2022) pertains to dealing with dataset sizes to strike a balance between performance and acceptable running times. Ensuring diverse, representative, and reliable data through resource distribution becomes imperative in overcoming this challenge. Additionally, Yüce et al. (2022) underscore the necessity, at times, to use longer datasets. Therefore, the development of models with a large degree of freedom becomes crucial in such scenarios.

In terms of *complicated procedures*, it is merely indicated in Wong et al. (2023), which note the existence of complicated testing procedures, specifically related to using ML for vaccine development.

Concerning *model overfitting*, Strielkowski et al. (2023a) highlight the risk of models becoming overly complex and lacking generalizability to new data. Du et al. (2024) also indicate that overfitting is a significant concern in ML applications, often resulting from insufficient training data and the absence of clearly defined stopping criteria during model training. They believe that potential solutions include modifying the model architecture, such as adjusting weights, although such modifications may introduce additional complexity and potentially limit the generalization beyond the specific dataset. Table 9 presents the data issues and corresponding methods identified at the data processing stage, as outlined below.

5.6. Data security and privacy stage

In the realm of data security and privacy, six issues are identified, focusing on *information leakage, information misuse*, and *attacks*. Addressing *information leakage*, Durlik et al. (2023) highlight the risk of compromising data security leading to the leakage of information. To mitigate this, the following suggestions are proposed: prioritizing data encryption, mandating regular security audits and penetration testing, and establishing robust access controls along with rigorous authentication protocols. Additionally, Wilson et al. (2021) point out the risk of inadvertent release of identifiable information or identification through deductive disclosure. From a management perspective, they recommend bringing data scientists and data processing capabilities into the organization and building a trusted research environment.

Addressing *information misuse*, unauthorized access, breaches, and the improper use of sensitive information become concerns. Thus, it is imperative to implement robust data privacy and security measures, encompassing encryption, access restrictions, and compliance with data protection regulations (Cellina et al., 2023). Baloch et al. (2023) also reference breaches and recommend the utilization of an unsynchronized sensor data analytics model as a potential solution.

Regarding *attacks*, the presence of inference attacks, model inversion, or membership inference attacks raises concerns. Solutions such as Federated training and Differential Privacy (DP) are suggested as protective measures (Aldoseri et al., 2023). Federated training involves training ML models across decentralized devices or servers, ensuring privacy by keeping raw data local and not shared. DP is a privacy and security concept aiming to safeguard individuals' sensitive information while still allowing valuable insights from aggregated data. This is achieved by introducing noise to the data and preserving statistical properties. Attacks also encompass threats like adversarial attacks, data poisoning, and the tampering of both models and data. Addressing these threats requires the implementation of adversarial training techniques to bolster the robustness of machine learning models and anomaly detection techniques to identify abnormal patterns or deviations from expected behavior. In addition, from a management perspective, compliance with Data Protection Regulations is imperative (Aldoseri et al., 2023). Table 10 presents the data issues and corresponding methods identified in data security and privacy, as outlined below.

5.7. AI technologies adoption stage

When contemplating the adoption of AI technologies, several challenges come to light. Six specific issues have been identified, which can be categorized as *cost issue, reusability issue, and interpretability issue*. In terms of *cost*, Khalaf et al. (2024) argue that installing the required sensors and infrastructure for data collection can be costly, thus a Cost-Benefit Analysis should be conducted first. Durlik et al. (2023) further underscore the financial burden associated with implementing

advanced data analytics, encompassing expenses related to hardware, software, infrastructure, and staff training or hiring.

In terms of *reusability*, Wen and Li (2022) highlight challenges tied to updating existing models with new data while ensuring performance consistency and reproducibility, without providing specific solutions.

On the front of *interpretability*, Wen and Li (2022) also raise challenges related to understanding why models perform as they do and explaining the reasons behind specific model actions. Additionally, Xu et al. (2022) express doubts about the reliability of Deep Neural Networks and recommend employing Concept Drift Detection, Uncertainty Estimation, and Out-of-Distribution Detection when deploying models in real-world and dynamic environments to ensure their reliability. Concept Drift Detection aims to identify when the statistical properties of the target variable change, which allows for timely model adaptation. Uncertainty estimation involves quantifying the uncertainty linked to predictions made by ML models. Out-of-distribution detection focuses on identifying instances where input data significantly deviates from the distribution on which the model is trained. Papadimitriou et al. (2024) also point out that ML models are capable of delivering highly precise predictions, but their intricate architecture often acts as a "black box''. Understanding the 'why' and 'how' behind predictions is essential for gaining insights into material properties and behaviors. Table 11 presents the data issues and corresponding methods identified in AI technologies adoption stage, as outlined below.

6. Discussion

This section discusses the extended critical points based on the literature analysis results and presents a data lifecycle management framework for handling data issues in industrial AI. This section aims to answer the RQ3: How can these data issues be systematically resolved?

6.1. Manage data from different sources

As noted in Section 4.3, it is crucial to recognize that data can originate from various sources, including historical storage, real-time sensors, and human interactions or expertise, etc. While most of data issues discussed in previous sections focus on historical data, this section will explore two critically important data sources that have less frequently addressed: real-time sensor data and expert domain knowledge.

6.1.1. Real-time data from sensors

In the data issues we identified, most of the data consists of historical information that has already been collected and stored in systems. However, with the advancement of IoT technology in industrial settings, real-time data is gaining increasing attention. Real-time data poses unique challenges for AI applications, as it requires processing and analysis with minimal delays or near-instantaneous responses. Key challenges associated with real-time data in AI include managing high data volume and velocity, ensuring low latency and fast response times, addressing issues related to real-time data quality and noise, scalability and resource constraints, data synchronization, adapting models in real time, and providing real-time analytics and visualization (Aldoseri et al., 2023).

For example, in the domain of computer vision, Song et al. (2015) introduce an RGB-D benchmark suite aimed at advancing the state of the art in major scene understanding tasks. Their dataset, captured by four different sensors, contains 10,335 RGB-D images, all of which are densely annotated with 146,617 2D polygons and 64,595 3D bounding boxes. Cho and Kang (2021) further augment the dataset to incorporate the necessary data for color images, gaze data with corresponding depth information, and object labels. They also generate virtual gaze data from bounding boxes and depth images and apply this to real-time, gaze-aware attentive object detection. In their work, they identify several challenges, including missing data in certain ranges, the time-consuming nature of manual annotation for 3D objects and scene categories, and the issue of cross-sensor bias. Since real-world data often comes from different sensors, it is crucial for algorithms to generalize across them. Their research demonstrates the existence of sensor bias and emphasizes the need for developing RGB-D algorithms with improved sensor generalization capabilities.

In the manufacturing process domain, Bastani et al. (2016) proposes an online sparse estimation-based classification (OSEC) approach for real-time monitoring in advanced manufacturing processes using heterogeneous sensor data. They apply this method to a semiconductor chemical mechanical planarization (CMP) process for polishing copper wafers to a specular finish. The data in their study is acquired from a two-axis wireless vibration sensor at a sampling rate of 685 Hz. The authors highlight the complexity of the vibration data in the CMP process and the difficulty of detecting subtle process drifts using traditional statistical and data mining methods. Their OSEC approach proves effective in addressing these challenges.

From a more theoretical point of view, Aldoseri et al. (2023) emphasize the importance of timeliness as a critical dimension of AI data quality. Timeliness refers to the extent to which data is current and relevant to the present context. They argue that AI systems require timely data to adapt to dynamic environments and provide accurate predictions. Using outdated data can lead to poor performance and, in some cases, harmful consequences, as the AI system may fail to account for recent changes in the underlying phenomena. To address these challenges, they propose edge computing as a solution. By processing data on edge devices, such as smartphones, IoT devices, or edge servers, edge computing reduces the amount of data that needs to be transmitted to the cloud or a centralized data center. This approach enables real-time data processing, reduces latency, and conserves bandwidth.

6.1.2. Data from experts' domain knowledge

Human-generated data, especially expert domain knowledge, is playing an increasingly important role alongside machine- or sensor-generated data. Expert knowledge is essential for decision-making, as it provides contextual understanding, improving the accuracy and relevance of decisions. In qualitative research, expert knowledge has been widely used in Delphi studies (Yousuf et al., 2007; Laupichler et al., 2023) and MCDM methods, such as Fuzzy AHP (Mehrparvar et al., 2024a, 2024b). In recent years, expert domain knowledge has also become integral to AI projects, enhancing the accuracy and relevance of AI models by aligning them with real-world complexities and industry-specific requirements. Expert insights help refine data inputs, validate model outputs, and address nuanced challenges, which ultimately improve AI reliability.

However, using expert knowledge raises concerns about subjectivity and bias, as expert opinions may reflect individual or sector-specific perspectives that are not universally applicable. It thus might lead to biased training data or skewed model outcomes in AI projects. Inconsistencies between different experts' knowledge can also complicate data integration and model development. Additionally, capturing and formalizing expert knowledge into structured data is challenging, limiting its full potential in AI systems. To address these issues, it is suggested that standardization techniques can be employed to minimize subjectivity and bias, while incorporating multiple experts' perspectives could help balance individual biases. Utilizing formal knowledge representation methods, such as ontologies, may aid in structuring expert insights and enhancing data integration and consistency. Additionally, regular validation of AI models with expert feedback could ensure their continued alignment with evolving domain knowledge.

6.2. Strategy for addressing cost issue in data storage

In section 4.3, two issues pertain to the data access and storage stage. For data access, the key question is how tight access control should be, balancing the trade-off between data democracy and data security.

Regarding data storage, the primary concern is cost. The solutions highlighted in the literature primarily focus on technical approaches, such as applying non-volatile memory and utilizing distributed storage systems (Aldoseri et al., 2023). However, from a management perspective, it is equally important to develop strategies not only for determining what data should be collected and stored but also for planning when to archive or delete data at the end of its lifecycle. This ensures efficient resource utilization and long-term sustainability; and requires implementing a well-defined data lifecycle management strategy, which ensures that data is stored and maintained efficiently throughout its useful life. Once the data is no longer needed for active use, organizations should establish criteria for archiving or permanently deleting it. Archiving allows data to be preserved for future reference or regulatory compliance while freeing up valuable storage resources. On the other hand, deletion removes outdated or redundant data to reduce storage costs and minimize the risk of unnecessary data exposure or breaches. Effective data lifecycle management not only reduces operational costs but also ensures that data governance policies are upheld.

6.3. Data quality issues vs. Data nature issues

In section 4.3, 41 issues are related to the data preprocessing stage, which is arguably the most critical phase in the data lifecycle for industrial AI. This stage is where most of the challenges arise and where they are expected to be resolved. As previously mentioned, these issues can be categorized into two types: data quality issues and data nature issues. For data quality, there is a long history and wealth of research on data quality management that addresses its definitions (Wang and Strong, 1996; Abraham et al., 2019), key dimensions (Wang and Strong, 1996; Hazen et al., 2014), evaluation methods (Kontokostas et al., 2014; Pipino et al., 2002). Another concept, data integrity, is also closely related to data quality but focuses on different aspects of managing data. Data quality refers to the overall utility of data based on various characteristics. The main goal of data quality management is to ensure that the data is fit for its intended uses. Data integrity, on the other hand, is specifically concerned with the accuracy and consistency of data over its lifecycle (Chen et al., 2012). In other words, it is the degree to which data is reliable and trusted to be correct. Compromised data integrity can result in AI systems making decisions based on corrupt or inconsistent data, which leads to unreliable or flawed outcomes (Aldoseri et al., 2023). In essence, data integrity can be seen as a fundamental aspect of data quality. As noted, extensive research has been conducted on data quality and its related concepts to address data quality issues. However, less attention has comparatively been given to data nature issues. These issues are particularly critical in the context of AI, where the nature of the data significantly impacts model selection and performance. To address this gap, it is important to consider both the characteristics of the data and the specific AI models being used, in order to ensure a more effective alignment between data and model requirements. This approach will be discussed in detail in Section 5.1.4.

6.4. Model-aware data preparation

In AI projects, data and models are inherently interconnected. Enhancing model performance requires careful examination of the data, while improving data quality necessitates consideration of the characteristics of the models that will be applied. Some research refers to this as "model-aware cleaning" (Majeed and Hwang, 2024). In our study, we adopt the term "model-aware data preparation," as it encompasses not only data cleaning but also other stages of the data lifecycle, particularly data collection. In data science, being model-aware involves understanding the assumptions, strengths, and limitations of the models being used, which directly influences how data is processed, interpreted, and visualized.

In this section, we propose a model-aware data preparation approach that considers both the characteristics of the data and the AI models being utilized. Since most of the articles we reviewed focus on ML, our analysis primarily addresses the data requirements of different ML algorithms. According to Sarker (2021), ML algorithms can be classified into four main categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning, as illustrated in Fig. 6.

6.4.1. Supervised learning

Supervised learning utilizes labeled training data and a set of training examples to deduce a function. The primary tasks in supervised learning are classification, which segregates the data, and regression, which fits the data. It is important to have a well-structured and representative labeled dataset. Both the quality and quantity of the data significantly impact the model's performance and its ability to generalize to new and unseen data.

The key elements of the data requirement for classification in supervised learning include:1) Input Features (X): these are the features used for making predictions or classifications; 2) Output Labels (Y): each input sample is associated with a corresponding output label, distinguishing between classes; 3) Labeled Dataset: a dataset exists, containing input-output pairs used for training the model; 4) Sufficient Data Variety: the dataset should encompass diverse scenarios relevant to the problem, aiding generalization; 5) Adequate Data Size: the dataset should be sufficiently large to capture underlying patterns without being excessively large, which could increase computational costs; and 6) Balanced Classes: having a balanced distribution of examples across different classes is beneficial for training unbiased models.

In contrast to classification, where the target variable signifies discrete classes, regression involves a continuous target variable that represents a numerical value. The objective in regression is to predict a continuous output rather than categorizing samples into specific classes. Despite this distinction, the data requirements for regression are generally similar to those of classification algorithms.

6.4.2. Unsupervised learning

Unsupervised learning involves the analysis of unlabeled datasets, operating without explicit supervision or labeled output. In unsupervised learning, the system tries to learn the patterns, relationships, or structures inherent in the data without being given specific target values to predict. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc.

In clustering tasks, success depends on the inherent structure of the data and the appropriateness of the chosen algorithm for the given dataset. The key elements of the data requirement for clustering include: 1) Feature Set (X): the dataset should consist of a set of features that describe the characteristics of each data point; 2) Homogeneous Data Distribution: the dataset should ideally represent a somewhat homogeneous distribution of data, where clusters can be naturally identified; 3) Sufficient Data Size: having a sufficiently large dataset is beneficial. However, the appropriate size of the dataset depends on the complexity of the data and the clustering algorithm used.

Density estimation in unsupervised learning involves estimating the probability density function of a dataset, providing insights into the underlying distribution of the data. The data requirements for density estimation include: 1) Unlabeled Data; 2) Continuous Data; 3) Representative Sample: the dataset should be a representative sample of the population or system being modeled and the quality and accuracy of the density estimation depend on how well the dataset captures the true underlying distribution; 4) Sufficient Data Size: smaller datasets may lead to higher uncertainty in estimating the density, especially in areas with fewer observed data points; and 5) Appropriate Feature Representation: the features used for density estimation should be appropriate for capturing the essential characteristics of the data distribution.

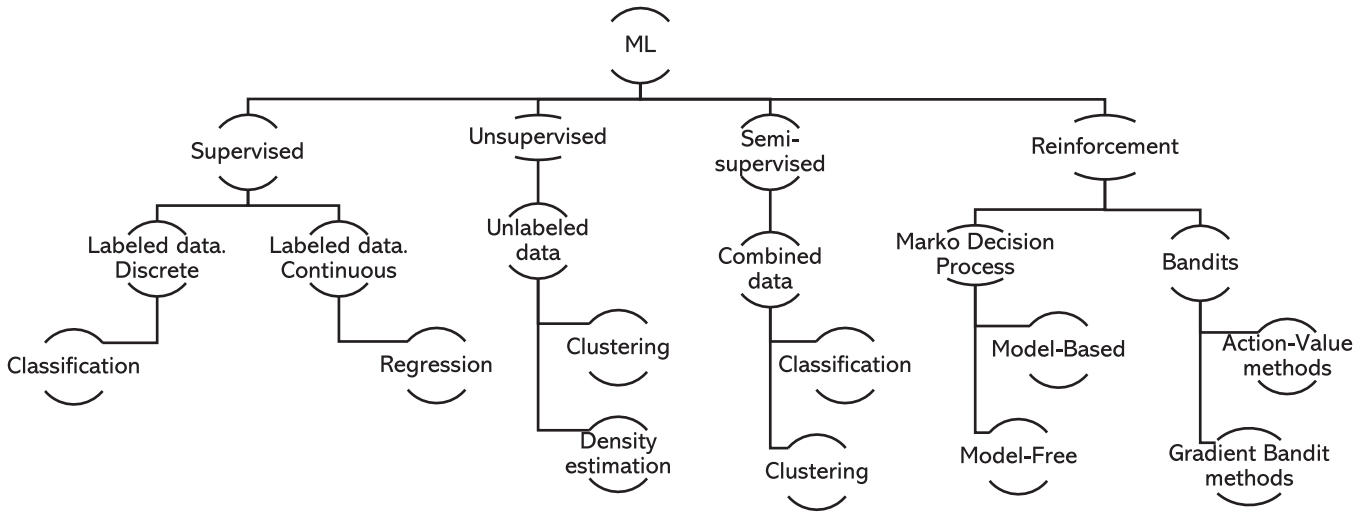


Fig. 6. Categories of machine learning algorithms (adapted from Sarker, 2021; Zhang and Yu, 2020).

6.4.3. *Semi-supervised learning*

Semi-supervised learning can be defined as a hybridization of the above-mentioned supervised and unsupervised methods. It involves training a model using a combination of labeled and unlabeled data. This type of learning can be particularly useful when obtaining labeled data is costly or time-consuming. Both the labeled and unlabeled data should be representative samples of the underlying distribution. The distribution of labeled and unlabeled examples should be balanced appropriately based on the available resources and the goals of the semi-supervised learning task. The labeled and unlabeled data should be relevant to the specific task at hand. The semi-supervised learning approach is most effective when the labeled and unlabeled examples share similarities, allowing the model to leverage the unlabeled data for improved performance on the labeled task.

6.4.4. *Reinforcement learning*

Reinforcement learning is a type of ML where an agent learns to make decisions by interacting with an environment. The data requirements for reinforcement learning are distinct from those of supervised and unsupervised learning. It requires agent and environment interaction, environment dynamics, state representation, action space, reward signal, trajectories or episodes. In addition, the data should reflect a balance between exploration and exploitation.

6.4.5. *DL*

As DL gains increasing attention, it is noteworthy to emphasize the distinctive data requirements associated with DL. DL is a subfield of ML, and it encompasses various types of learning paradigms, including supervised, unsupervised, and semi-supervised learning. DL models often require large and diverse datasets to learn complex patterns and generalizations. Table 12 summarizes the different data requirements for various ML algorithms.

Considering the diverse data requirements outlined in Table 12, we can only evaluate data usability and anticipate potential data issues based on the specific needs of different AI algorithms, allowing for strategic data collection and preprocessing in advance. For example, small datasets are usually not suitable for DL algorithms, so it is essential to assess the availability of large datasets before initiating a DL project. This proactive approach ensures better preparedness, helping to avoid unnecessary costs and time delays in AI project development.

Table 12
Different Data requirements in various ML algorithms.

ML algorithms	Task	Data requirement
Supervised learning	Classification	Input features (X); Output labels (Y); Labeled dataset; Sufficient data variety; Adequate data size; Balanced classes
	Regression	Similar with classification, yet the target variable is continuous and represents a numerical value
Unsupervised learning	Clustering	Unlabeled dataset, Feature set (X); Homogeneous data distribution; Sufficient data size
	Density estimation	Unlabeled data; Continuous data; Representative sample; Sufficient data size; Appropriate feature representation
Semi-supervised learning	Classification, Clustering	Labeled data; Unlabeled data; Representative sample; Balance between labeled and unlabeled Data; Task relevance
Reinforcement learning	Reward or penalty	Agent and environment interaction; Environment dynamics; State representation; Action space; Reward signal, Trajectories or episodes; Balance between exploration and exploitation
DL		Large and diverse datasets

6.5. *Data management framework for systematically handling data issues in industrial AI*

Building on the findings, from a data-centric point of view, this section introduces a comprehensive data management framework, specifically designed to systematically address data issues in industrial AI. Through the synthesis of the analyses, the framework is structured into three integral components aimed at addressing the complex nature of data challenges in industrial AI projects.

The first component involves data usability evaluation which comprises three key steps. Every AI project originates from a specific problem or operational challenge that needs to be addressed. Therefore, before delving into data analysis or model development, the first step is to fully understand the problem at hand by asking fundamental questions such as: *What is the business or operational problem? What is the purpose of using AI? What task needs to be accomplished?* This step ensures a clear understanding of the AI project's objectives, which is essential for determining the types of data required and how to effectively use them. By reflecting on these questions, project teams can better identify the appropriate AI algorithms and gain insight into the necessary data characteristics. Once the problem is understood, the second step is

preliminary data exploration. This involves thoroughly analyzing existing data to comprehend its nature, including type, features, distribution, patterns, size, and complexity. This exploration enables an initial assessment of whether the available data can adequately address the identified problem. If gaps are found, strategies for additional data collection must be considered. The third step emphasizes model-aware data preparation, which is crucial for aligning the data with the specific AI models being employed. As discussed in Section 5.1.4, a model-aware approach should be used to evaluate data usability, ensuring that the data meets the requirements of the selected algorithms. This involves asking key questions such as: *What AI algorithms are we using? What are the data requirements for these algorithms? What potential data issues might arise with these algorithms?* Anticipating these challenges allows for strategic data collection and preprocess, ensuring both data quality and usability throughout the project. This structured approach minimizes project risks, optimizes resource allocation, and enhances the effectiveness of industrial AI applications. Once data usability is confirmed—ensuring alignment with both the operational objectives and the technical requirements of the AI models—the framework progresses to the next component to activate the full data lifecycle.

As outlined in the second component of the proposed framework, the 82 identified data issues are categorized into 29 types and further classified based on the stages of the data lifecycle. This categorization offers a comprehensive map, providing clarity on when and where specific issues may arise. As noted in Section 3.3, data lifecycle theory is used in this study because it provides a systematic approach to understand and manage data throughout its lifecycle. This theory offers a structured framework for analyzing data issues from inception to disposal, enabling the development of clear guidelines for data management and governance. The reason of further categorizing individual data issues into different types is that specific data issues are numerous. They often appear trivial and vary across AI projects, so addressing each issue individually would be time-consuming and inefficient. Meanwhile, upon reviewing these individual data issues, we observe that many of them exhibit patterns and a high degree of repeatability. For instance, several studies refer to the problem of data imbalance using different terminology, such as "imbalanced input dataset" (Abrasaldo et al., 2024), "data bias and unfairness" (Aldoseri et al., 2023), "unbalanced datasets" (Strielkowski et al., 2023b), "significant difference in the amount of data between classes" (Chuo et al., 2022), and "data imbalance" (Zhang and Gao, 2021). Despite the varying terms, they all refer to the same fundamental issue. Therefore, by categorizing and managing these recurring data issues instead of treating them as individual cases, we can streamline the process of resolving them. Moreover, common data issues, such as data incompleteness, inaccuracy, or mislabeling, frequently arise across different AI projects. Accumulating knowledge and solutions for different types of data issues allows us to reuse and transfer these strategies to other projects, thereby avoiding repetitive work and improving efficiency. This is the reason that we emphasize the need for a structured approach to categorize and manage data issues within our framework. This structured classification acts as a roadmap, revealing specific data issues of different types at various stages, thereby enabling targeted resolution strategies, and ultimately improving the overall data management in industrial AI.

The final component of the framework presents the methods for addressing data issues at various stages of the data lifecycle, integrating both managerial and technical perspectives. It is important to note that it is not feasible to encompass all methods within a single chart. Therefore, we provide selected examples of the methods shown in Table 5 to Table 11, which address data issues at each stage of the data lifecycle. For example, for some data issues in the data source and collection stage, managerial solutions include fostering collaboration in data collection, establishing clear data requirements, and promoting standardization, while technical solutions focus on using data augmentation techniques to address data insufficiency or imbalance. For some issues in the data access and storage stage, managerial strategies may involve securing licenses for data access and usage, while technical solutions could include building distributed storage systems or utilizing NVM technologies for enhanced capacity and performance. For some issues in the data integration and interoperation stage, managerial strategies can promote joint development initiatives among stakeholders to facilitate seamless integration, while technical strategies may involve deploying advanced data integration techniques. For some issues in the data preprocessing stage, cooperation between IT and business teams is key from a managerial perspective, whereas technical solutions include applying data imputation, cleaning techniques, annotation tools, and feature engineering to improve data quality. For some issues in the data processing stage, a managerial approach might involve conducting cost-benefit analyses on system updates or replacements, while technical strategies focus on leveraging specialized hardware and flexible models. For some issues in the data security and privacy stage, managerial strategies emphasize building trusted research environments and ensuring compliance with data protection regulations, while technical solutions may include federated learning, differential privacy, and adversarial training techniques to enhance security. Last, for some issues in the AI technology adoption stage, a cost-benefit analysis of AI implementation from a managerial perspective is essential, while technical approaches should focus on detecting concept drift, estimating uncertainty, and managing out-of-distribution data to ensure the robustness and effectiveness of AI systems.

One of the key advantages of the final component of the framework is its recognition of the interdependence between management and technology. By integrating both managerial and technical perspectives, the framework emphasizes the importance of collaboration between management and technical teams. This collaborative approach is crucial for achieving sustainable improvements in industrial AI projects, as it ensures that strategic objectives align with operational needs. Management teams provide the oversight, direction, and regulations necessary for data handling, while technical teams offer the expertise required to address specific data issues and implement advanced solutions. This synergy between management and technology not only enhances the overall effectiveness of data handling processes, but also fosters a more cohesive and responsive approach to addressing challenges throughout the data lifecycle. The whole framework is presented in Fig. 7.

7. Future research on data management in industrial AI

As highlighted earlier, data is essential for making actual progress in AI advancement. It is imperative for future research to prioritize the discussion of data-related topics. To guide future research topics, seven pivotal points can be considered.

1. Assessing the effectiveness of existing methods. The methods discussed in this paper have shown the potential to address specific data issues; however, their effectiveness needs thorough evaluation. This can be accomplished by implementing these methods in real-world scenarios or conducting case studies on projects that have employed these techniques. By analyzing and comparing the outcomes, researchers and practitioners can identify areas for enhancement and ensure that the solutions remain relevant and effective.

2. Developing model-aware data preparation. Future studies should delve deeper into model-aware data preparation techniques. This involves not only refining data cleaning processes but also developing strategies for effective data collection and preprocessing that align with the specific requirements of various AI models. Research could explore how different AI algorithms impact data needs and how to optimize data for the diverse algorithms.

3. Advancing data usability assessment. Future research should focus on advancing methods for assessing data usability. This includes developing frameworks for evaluating data quality, identifying potential issues, and ensuring that data aligns with both operational goals and technical requirements. Research could also explore the development of standardized metrics and tools for data usability assessment.

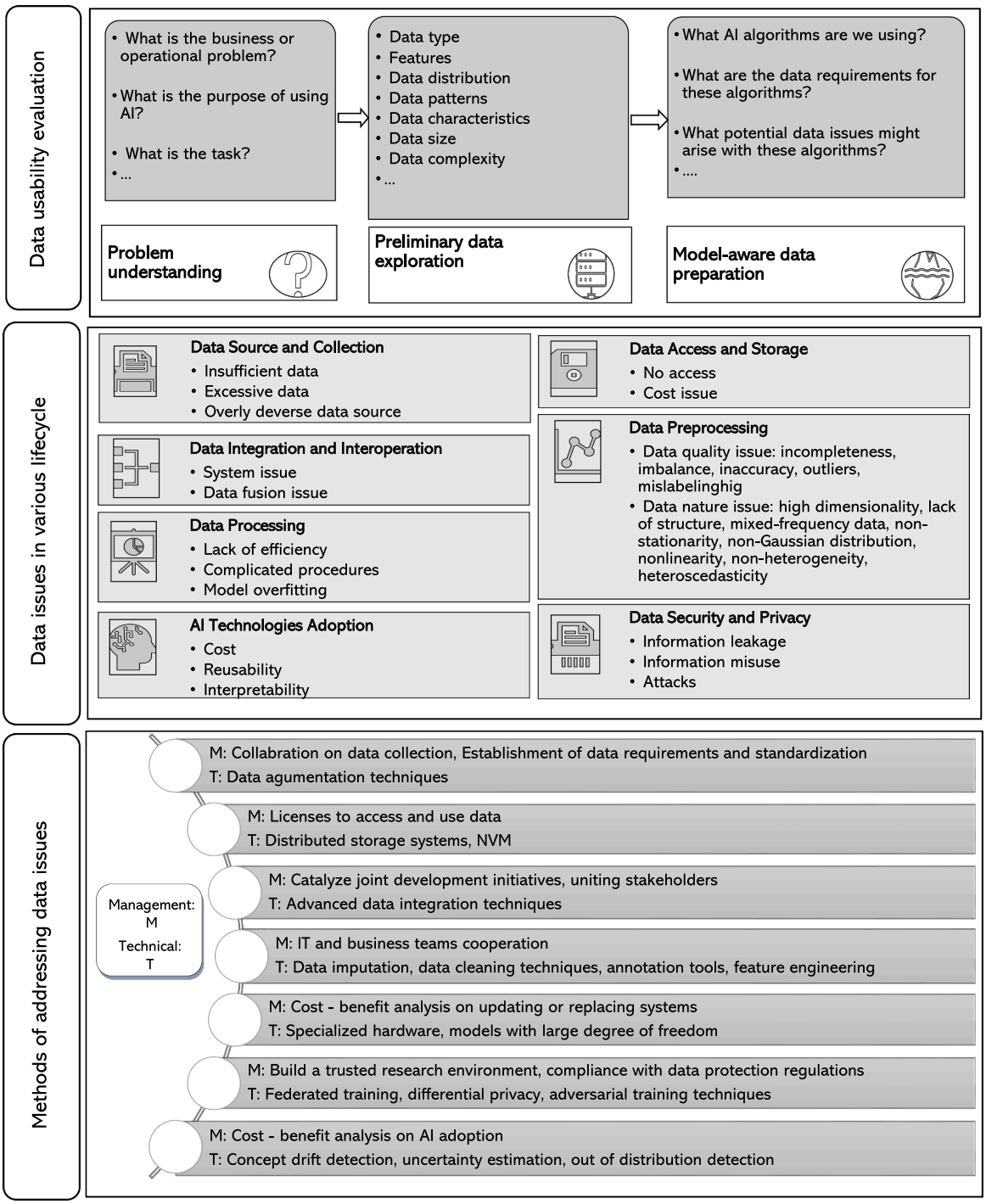


Fig. 7. Data management framework for systematically handling data issues in industrial AI.

4. Establishing collaborative management and technical approaches. It is crucial to investigate how to better integrate managerial and technical methods for addressing data issues. Research should explore models for fostering collaboration between management and technical teams to manage data more effectively.

5. Developing longitudinal data lifecycle management. Research should focus on the long-term management of data throughout its lifecycle. This includes strategies for data archiving, deletion, and continuous evaluation to ensure that data remains relevant and useful over time. Developing frameworks for managing the entire data lifecycle will support sustainable AI practices and improve overall data management.

6. Translating knowledge on solving data issues into industry experience. While the literature has highlighted numerous data issues and corresponding methods, there is a need to propel the industry towards confronting these data issues pragmatically. This requires developing and continuously practicing problem-solving skills to translate knowledge into practical experience. Given the privacy, confidentiality, and complexity of industry data, obtaining real data from specific companies is difficult. Therefore, leveraging open data sources, such as participating in initiatives like the PHM Data Challenge competition (Jia et al., 2018; Su and Lee, 2023), presents a valuable avenue for honing problem-solving skills.

7. Achieving a more profound integration of AI and industry. Looking ahead, beyond addressing data issues, the future should also focus on

expanding the application of AI. This necessitates a dual perspective: from the AI standpoint, exploring potential applications in various fields; and from the industry perspective, contemplating how to prepare industry comprehensively to pave the way for a more profound and widespread AI integration. For instance, the introduction of ChatGPT has highlighted the potential of large language models (LLMs) in showcasing artificial general intelligence. However, in industry, where there is a need for domain-specific knowledge, LLMs may not be ideal due to their training on general knowledge. Therefore, exploring the development of Industrial Large Knowledge Models (ILKMs) tailored for industrial systems could be a promising avenue for future research (Lee and Su, 2024). A data forum can also be established to provide both researchers and practitioners with a platform to convene and exchange insights, best practices, and innovative ideas within the realms of data science, data analytics, data management and DCAI.

8. Conclusions, implications, and limitations

8.1. Conclusions

This study conducts a comprehensive meta-review of data issues and corresponding methods in the implementation of industrial AI. It identifies and categorizes 82 data issues according to the stages of the data lifecycle: *data source and collection, data access and storage, data integration and interoperation, data preprocessing, data processing, data security and privacy, and AI technologies adoption.* By doing so, this study answers two research questions: *What are the specific data issues encountered during the implementation of AI in industrial systems?* and *What methods are associated with addressing these issues?* Subsequently, this study discusses not only historical data but also the management of data from other critical sources, particularly real-time data from sensors and data derived from expert domain knowledge. Afterwards, it proposes a model-aware data preparation approach, analyzing the data requirements of various ML algorithms to guide data preparation and usability evaluation for AI models. Synthesizing all the analyses, this study introduces a data management framework to answer the research question: *How can these data issues be systematically resolved?* Finally, the study outlines seven future research directions for addressing data issues, advancing data management, and integrating AI and industry.

8.2. Implications

The discussions and findings of this study have significant implications for both academic knowledge and practical applications in industrial AI.

Academically, it fills a notable gap in the literature by creating a clear taxonomy of data issues across the data lifecycle stages, as well as linking these issues to their corresponding resolution methods. This taxonomy not only maps existing issues and solutions but also guides the development of new tools and techniques for engineering and monitoring data. By adopting the lifecycle perspective, it also deepens the theoretical understanding of how data-related challenges evolve and impact AI performance in industrial settings. In addition, the discussion on real-time sensor data and expert domain knowledge highlights the need for advanced data management beyond traditional historical data. Moreover, by proposing a model-aware data preparation approach, this study illustrates that aligning data characteristics with specific AI algorithms improves model usability and effectiveness. Last, this study introduces a conceptual framework that integrates model-aware data preparation, managerial perspectives, and technical methodologies, providing a systematic methodology for DCAI research. This framework not only structures future research on data governance and quality management, but also highlights underexplored challenges, which accordingly encourages interdisciplinary studies between AI, data science, and industrial applications. By offering a foundational structure, the framework contributes to the development of new theories and methodologies for AI data lifecycle management.

Practically, this study provides valuable insights for AI practitioners and industrial system developers by offering a comprehensive framework for addressing data issues in AI projects. The proposed framework serves as a strategic tool that enables practitioners to anticipate and mitigate data-related obstacles, enhance AI adoption and scalability, improve decision-making accuracy, and support long-term AI lifecycle management. By equipping industries with practical strategies for improving data quality and governance, this study not only enhances AI-driven decision-making and automation, but also reduces operational inefficiencies, ensuring that AI applications deliver consistent and valuable insights over time.

In summary, this study offers actionable insights for both researchers and practitioners when implementing industrial AI, and advocates for a systematic and integrated approach to data management that incorporates both strategic and operational considerations. The proposed framework and future research directions provide a valuable foundation to advance the field and improve the deployment and performance of industrial AI systems.

8.3. Limitations

The study acknowledges its limitations, as the literature gathered through the meta-review method may not encompass all data issues and corresponding methods. The analysis and categorization of data issues heavily rely on the availability and quality of literature sources, potentially introducing biases or gaps that may have influenced the identification and categorization process. Moreover, despite efforts to systematically analyze data issues and methods, the study is inherently subject to limitations associated with the meta-review methodology, including selection bias, publication bias, and the potential for researcher interpretation. Scholars, who are interested in conducting a more exhaustive exploration, can consider employing alternative methods such as NLP technique to automate the extraction and analysis of data-related information from a vast array of textual sources.

CRediT authorship contribution statement

Yang Cheng: Writing – review & editing, Project administration. **Xuejiao Li:** Writing – original draft, Investigation, Methodology, Data curation. **Jay Lee:** Methodology, Conceptualization, Supervision. **Charles Møller:** Writing – review & editing, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work, the first author used ChatGPT to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this article.

Acknowledgements

This study is supported by MADE FAST (MADE – Flexible, Agile, and Sustainable production enabled by Talented employees) Project (470100), Double Thousand Plan of Jiangxi Province, and Strengthening the digitalization of businesses in Eastern Europe – a micro and macro-level approach" project funded by the European Union – NextGenerationEU and the Romanian Government, under the National Recovery and Resilience Plan for Romania, contract no. 760036/23.05.2023, cod PNRR-C9-I8-CF 198/28.11.2022, through the Romanian Ministry of

Research, Innovation and Digitalization, within Component 9, Investment I8.

Data availability

No data was used for the research described in the article.

References

Abraham, R., Schneider, J., Vom Brocke, J., 2019. Data governance: a conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. 49, 424–438.

Abrasaldo, P.M.B., Zarrouk, S.J., Kempa-Liehr, A.W., 2024. A systematic review of data analytics applications in above-ground geothermal energy operations. Renew. Sustain. Energy Rev. 189, 113998.

Akhtar, S., Shahzad, S., Zaheer, A., Ullah, H.S., Kilic, H., Gono, R., Leonowicz, Z., 2023. Short-Term load forecasting models: a review of challenges, progress, and the road ahead. Energies 16 (10), 4060.

Aldoseri, A., Al-Khalifa, K.N., Hamouda, A.M., 2023. Re-Thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Appl. Sci. 13 (12), 7082.

Arruda, H.M., Bavaresco, R.S., Kunst, R., Bugs, E.F., Pesenti, G.C., Barbosa, J.L.V., 2023. Data science methods and tools for industry 4.0: a systematic literature review and taxonomy. Sensors 23 (11), 5010.

Ashmore, R., Calinescu, R., Paterson, C., 2021. Assuring the machine learning lifecycle: desiderata, methods, and challenges. ACM Comput. Surv. (CSUR) 54 (5), 1–39.

Baloch, L., Bazai, S.U., Marjan, S., Aftab, F., Aslam, S., Neo, T.K., Amphawan, A., 2023. A review of big data trends and challenges in healthcare. Int. J. Technol. 14 (6), 1320–1333.

Bastani, K., Rao, P.K., Kong, Z., 2016. An online sparse estimation-based classification approach for real-time monitoring in advanced manufacturing processes from heterogeneous sensor data. IIE Trans. 48 (7), 579–598.

Cellina, M., Cè, M., Alì, M., Irmici, G., Ibba, S., Caloro, E., Papa, S., 2023. Digital twins: the new frontier for personalized Medicine? Appl. Sci. 13 (13), 7940.

Chander, B., Kumaravelan, G., 2022. Outlier detection strategies for WSNs: a survey. J. King Saud. Univ. Comput. Inf. Sci. 34 (8), 5684–5707.

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact. MIS Q. 1165–1188.

Cho, D.Y., Kang, M.K., 2021. Human gaze-aware attentive object detection for ambient intelligence. Eng. Appl. Artif. Intell. 106, 104471.

Chuo, Y.S., Lee, J.W., Mun, C.H., Noh, I.W., Rezvani, S., Kim, D.C., Park, S.S., 2022. Artificial intelligence enabled smart machining and machine tools. J. Mech. Sci. Technol. 36 (1), 1–23.

Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G., 2015. Calibrating probability with undersampling for unbalanced classification (December). In 2015 IEEE Symposium Series on Computational Intelligence. IEEE, pp. 159–166 (December).

Du, X., Salasakar, S., Thakur, G., 2024. A comprehensive summary of the application of machine learning techniques for CO2-Enhanced oil recovery projects. Mach. Learn. Knowl. Extr. 6 (2), 917–943.

Durlik, I., Miller, T., Cembrowska-Lech, D., Krzemińska, A., Złoczowska, E., Nowak, A., 2023. Navigating the sea of data: a comprehensive review on data analysis in maritime IoT applications. Appl. Sci. 13 (17), 9742.

Eyuboglu, S., Karlaš, B., Ré, C., Zhang, C., & Zou, J. (2022, June). dcbench: A benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning* (pp. 1-4).

Field, M., Hardcastle, N., Jameson, M., Aherne, N., Holloway, L., 2021. Machine learning applications in radiation oncology. Phys. Imaging Radiat. Oncol. 19, 13–24.

Ghahramani, M., Qiao, Y., Zhou, M.C., O'Hagan, A., Sweeney, J., 2020. AI-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA J. Autom. Sin. 7 (4), 1026–1037.

Guo, Y., Sun, M., Zhang, W., Wang, L., 2023. Machine learning in enhancing corrosion resistance of magnesium alloys: a comprehensive review. Metals 13 (10), 1790.

Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. 154, 72–80.

Huang, P.X., Hu, W., Brendel, W., Chandraker, M., Li, L.J., & Wang, X. (2021). Ymir: A rapid data-centric development platform for vision applications. *arXiv preprint arXiv: 2111.10046*.

Jia, X., Huang, B., Feng, J., Cai, H., Lee, J., 2018. A review of PHM data competitions from 2008 to 2017: methodologies and analytics (November). Proc. Annu. Conf. Progn. Health Manag. Soc. 1–10.

Khalaf, A.H., Xiao, Y., Xu, N., Wu, B., Li, H., Lin, B., Tang, J., 2024. Emerging AI technologies for corrosion monitoring in oil and gas industry: a comprehensive review. Eng. Fail. Anal. 155, 107735.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A., 2014. Test-driven evaluation of linked data quality (April). Proc. 23rd Int. Conf. World Wide Web 747–758.

Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., Qureshi, B., 2020. An overview of IoT sensor data processing, fusion, and analysis techniques. Sensors 20 (21), 6076.

Laupichler, M.C., Aster, A., Raupach, T., 2023. Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. Comput. Educ. Artif. Intell. 4, 100126.

Lee, J., Davari, H., Singh, J., Pandhare, V., 2018. Industrial artificial intelligence for industry 4.0-based manufacturing systems. Manuf. Lett. 18, 20–23.

Lee, J., Singh, J., & Azamfar, M. (2019). Industrial artificial intelligence. *arXiv preprint arXiv:1908.02150*.

Lee, J., Su, H., 2024. A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing. Int. J. AI Mater. Des. 3681, 20.

Liu, Y., Wang, Y., Wang, Q., Zhang, K., Qiang, W., Wen, Q.H., 2023a. Recent advances in data-driven prediction for wind power. Front. Energy Res. 11, 1204343.

Liu, Y., Wang, X., Zhang, Z., Deng, F., 2023b. A review of deep leaning in image classification for mineral exploration. Miner. Eng. 204, 108433.

Ma, J., Hu, C., Zhou, P., Jin, F., Wang, X., Huang, H., 2023. Review of image augmentation used in deep Learning-Based material microscopic image segmentation. Appl. Sci. 13 (11), 6478.

Madhikermi, M., Kubler, S., Robert, J., Buda, A., Främling, K., 2016. Data quality assessment of maintenance reporting procedures. Expert Syst. Appl. 63, 145–164.

Majeed, A., Hwang, S.O., 2024. A Data-Centric AI paradigm for Socio-Industrial and global challenges. Electronics 13 (11), 2156.

Mallett, R., Hagen-Zanker, J., Slater, R., Duvendack, M., 2012. The benefits and challenges of using systematic reviews in international development research. J. Dev. Eff. 4 (3), 445–455.

Matricciani, L., Paquet, C., Galland, B., Short, M., Olds, T., 2019. Children's sleep and health: a meta-review. Sleep. Med. Rev. 46, 136–150.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., ... & Janapa Reddi, V. (2024). Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, *36*.

Mehrparvar, M., Majak, J., Karjust, K., 2024a. A comparative analysis of fuzzy AHP and fuzzy VIKOR methods for prioritization of the risk criteria of an autonomous vehicle system. Proc. Est. Acad. Sci. 73 (2), 116–123.

Mehrparvar, M., Majak, J., Karjust, K., 2024b. Effect of aggregation methods in fuzzy technique for prioritization of criteria of automated vehicle system (January). In: In AIP Conference Proceedings, 2989. AIP Publishing.

Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2018. Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinforma. 19 (6), 1236–1246.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann. Intern. Med. 151 (4), 264–269.

Paleyes, A., Urma, R.G., Lawrence, N.D., 2022. Challenges in deploying machine learning: a survey of case studies. ACM Comput. Surv. 55 (6), 1–29.

Papadimitriou, I., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I., 2024. AI methods in materials design, discovery and manufacturing: a review. Comput. Mater. Sci. 235, 112793.

Peres, R.S., Jia, X., Lee, J., Sun, K., Colombo, A.W., Barata, J., 2020. Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. IEEE Access 8, 220121–220139.

Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data quality assessment. Commun. ACM 45 (4), 211–218.

Roy, S., Meena, T., Lim, S.J., 2022. Demystifying supervised learning in healthcare 4.0: a new reality of transforming diagnostic Medicine. Diagnostics 12 (10), 2549.

Ryan, R.E., Kaufman, C.A., Hill, S.J., 2009. Building blocks for meta-synthesis: data integration tables for summarising, mapping, and synthesising evidence on interventions for communicating with health consumers. BMC Med. Res. Methodol. 9, 1–11.

Sarker, I.H., 2021. Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. 2 (3), 160.

Sengupta, D., 2023. Artificial intelligence in diagnostic dermatology: challenges and the way forward. Indian Dermatol. Online J. 14 (6), 782–787.

Shah, I.A., Mishra, S., 2024. Artificial intelligence in advancing occupational health and safety: an encapsulation of developments. J. Occup. Health 66 (1), uiad017.

Shi, Q., Tang, J., Chu, M., 2023. Key issues and progress of industrial big data-based intelligent blast furnace ironmaking technology. Int. J. Miner. Metall. Mater. 30 (9), 1651–1666.

Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: a rgb-d scene understanding benchmark suite. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 567–576.

Stonebraker, M., El Kindi, Rezig, 2019. Machine learning and big data: what is important? IEEE Data Eng. Bull. 42 (4), 3–7.

Strickland, E., 2022. Andrew ng, ai minimalist: the machine-learning pioneer says small is the new big. IEEE Spectr. 59 (4), 22–50.

Strielkowski, S., Adeel, M., Iqbal, M., Namoun, A., Tufail, A., Kim, K.H., 2023b. Deep learning methods utilization in electric power systems. Energy Rep. 10, 2138–2151.

Strielkowski, W., Vlasov, A., Selivanov, K., Muraviev, K., Shakhnov, V., 2023a. Prospects and challenges of the machine learning and Data-Driven methods for the predictive analysis of power systems: a review. Energies 16 (10), 4025.

Su, H., & Lee, J. (2023). Review of Machine Learning Approaches for Diagnostics and Prognostics of Industrial Systems Using Industrial Open Source Data. *arXiv preprint arXiv:2312.16810*.

Tao, F., Qi, Q., Liu, A., Kusiak, A., 2018. Data-driven smart manufacturing. J. Manuf. Syst. 48, 157–169.

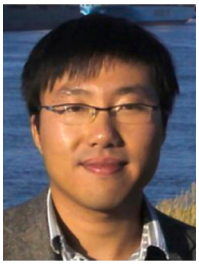
Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., Emmert-Streib, F., 2021. Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. Front. Artif. Intell. 4, 576892.

Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. 12 (4), 5–33.

Wang, Z., Xia, L., Yuan, H., Srinivasan, R.S., Song, X., 2022. Principles, research status, and prospects of feature engineering for data-driven building energy prediction: a comprehensive review. J. Build. Eng., 105028

Wen, R., Li, S., 2022. Spatial decision support systems with automated machine learning: a review. ISPRS Int. J. GeoInf. 12 (1), 12.
Wilson, A., Saeed, H., Pringle, C., Eleftheriou, I., Bromiley, P.A., Brass, A., 2021. Artificial intelligence projects in healthcare: 10 practical tips for success in a clinical environment. BMJ Health Care Inform. 28 (1).
Wing, J.M., 2019. The data life cycle. Harv. Data Sci. Rev. 1 (1), 6.
Wong, F., de la Fuente-Nunez, C., Collins, J.J., 2023. Leveraging artificial intelligence in the fight against infectious diseases. Science 381 (6654), 164–170.
Xu, J., Kovatsch, M., Mattern, D., Mazza, F., Harasic, M., Paschke, A., Lucia, S., 2022. A review on ai for smart manufacturing: deep learning challenges and solutions. Appl. Sci. 12 (16), 8239.
Yousuf, M.I., 2007. Using expertsopinions through delphi technique. Pract. Assess. Res. Eval. 12 (1).
Yüce, C., Gecgel, O., Doğan, O., Dabetwar, S., Yanik, Y., Kalay, O.C., Ekwaro-Osire, S., 2022. Prognostics and health management of wind energy infrastructure systems. ASCEASME J. Risk Uncertain. Eng. Syst. Part B Mech. Eng. 8 (2), 020801.
Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., & Hu, X.(2023a). Data-centric ai: Perspectives and challenges. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)(pp. 945-948). Society for Industrial and Applied Mathematics.
Zhang, Z.M., Catalucci, S., Thompson, A., Leach, R., Piano, S., 2023a. Applications of data fusion in optical coordinate metrology: a review. Int. J. Adv. Manuf. Technol. 124 (5-6), 1341–1356.
Zhang, J., Gao, R.X., 2021. Deep learning-driven data curation and model interpretation for smart manufacturing. Chin. J. Mech. Eng. 34, 1–21.
Zhang, Y., Safdar, M., Xie, J., Li, J., Sage, M., Zhao, Y.F., 2023b. A systematic review on data of additive manufacturing for machine learning applications: the data quality, type, preprocessing, and management. J. Intell. Manuf. 34 (8), 3305–3340.
Zhang, H., Yu, T., 2020. Taxonomy of reinforcement learning algorithms. Deep Reinf. Learn. Fundam. Res. Appl. 125–133.
Zhu, J., Dong, H., Zheng, W., Li, S., Huang, Y., Xi, L., 2022. Review and prospect of data-driven techniques for load forecasting in integrated energy systems. Appl. Energy 321, 119269.

Miss Xuejiao Li is a Ph.D. candidate at Aalborg University, Denmark, specializing in data governance and data-centric AI within the context of operations and supply chain. She has received a Master of Management Science and Engineering from Chinese Academy of Sciences, China. Her research focuses on developing frameworks and methods that address companies' data challenges, building effective data governance strategies and embracing digital and AI solutions. Her goal is to help organizations unlock the full potential of their data assets, hence driving improved performance and efficiency in digitalization and AI implementation. With a strong foundation in data management, AI, and analytics, she bridges the gaps between technical innovation and practical application.



Dr. Yang Cheng is currently an associate professor in Department of Materials and Production, Aalborg University, Denmark. He received a Master of Management Science and Engineering from Beihang University, China and a PhD from Aalborg University. He has an extensive research and consulting experience in business process reengineering, manufacturing strategy, global manufacturing network management and knowledge transfer. In these fields, he has authored numerous articles. He is the associate editor of Production Planning and Control and Journal of Manufacturing Technology Management.

Prof. Charles Møller holds a Ph.D. in industrial engineering from Aalborg University, Denmark, where he used to be a full professor in Enterprise Systems and Process Innovation. Currently, he is a full professor at the Department of Mechanical and Production Engineering, Aarhus University, Denmark. His research interests include ERP/MES systems, IT/OT integration, virtual factories, and smart production. He is also a Principal Investigator at the Manufacturing Academy of Denmark.

Prof. Jay Lee is Clark Distinguished Professor and Director of Industrial AI Center in the Mechanical Engineering Dept. of the Univ. of Maryland College Park. His current research focuses on developing non-traditional machine learning including transfer learning, domain adaptation, similarity-based machine learning, stream-of-x machine learning, as well as industrial large knowledge model (ILKM), etc. In addition, he is leading Data Foundry, which consists of over 100 diversified industrial datasets including semiconductor manufacturing, jet engines, wind turbine, EVs, high speed train, machine tools, robots, medical TBI, etc. for machine learning research. These datasets are used to rapidly develop and validate Industrial AI system with scalable and systematic approaches.