applied sciences

MDPI

Systematic Review

Big Data Management and Quality Evaluation for the Implementation of AI Technologies in Smart Manufacturing

Alexander E. Hramov [1] and Alexander N. Pisarchik [2,*]

[1] Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Stremyannyy Ln., 36, Moscow 236041, Russia; hramov.ae@rea.ru
[2] Center for Biomedical Technology, Universidad Politecnica de Madrid, Campus Montegancedo, Pozuelo de Alarcón, 28223 Madrid, Spain
* Correspondence: alexander.pisarchik@upm.es

Abstract

This review examines the role of industrial data in enabling artificial intelligence (AI) technologies within the framework of Industry 4.0. Key aspects of industrial data management, including collection, preprocessing, integration, and utilization for training AI models, are analyzed and systematically categorized. Criteria for assessing data quality are defined, covering accuracy, completeness, consistency, and confidentiality, and practical recommendations are proposed for preparing data for effective machine learning and deep learning applications. In addition, current approaches to data management are compared, and methods for evaluating and improving data quality are outlined. Particular attention is given to challenges and limitations in industrial contexts, as well as the prospects for leveraging high-quality data to enhance AI-driven smart manufacturing.

Keywords: big data management; digital twins; prediction and prevention; AI applications; machine learning

Academic Editor: Douglas O'Shaughnessy

Received: 3 October 2025
Revised: 30 October 2025
Accepted: 6 November 2025
Published: 9 November 2025

Citation: Hramov, A.E.; Pisarchik, A.N. Big Data Management and Quality Evaluation for the Implementation of AI Technologies in Smart Manufacturing. *Appl. Sci.* **2025**, *15*, 11905. https://doi.org/10.3390/app152211905

1. Introduction

Modern industry is undergoing a profound transformation driven by the introduction of artificial intelligence (AI) technologies and the digitalization of production processes as part of the Industry 4.0 (I4.0) vision. I4.0, or the fourth industrial revolution, involves the integration of cyber-physical systems, the Internet of Things (IoT) and smart technologies to create 'smart manufacturing' capable of self-optimization and autonomous decision-making [1]. In this context, industrial big data becomes a key resource, and its efficient use becomes the basis for the implementation of AI solutions aimed at increasing productivity, reducing costs, and improving product quality [2].

The shift toward a new production paradigm is driven by increasing global competition and the rise of intelligent industries, compelling the industrial sector to actively adopt innovative solutions [3]. I4.0, implemented in recent years, represents a radical transformation of manufacturing processes, encompassing the development of 'smart factories' and interconnected industrial environments. These systems are founded on key principles such as interoperability, virtualization, decentralization, distributed control, real-time operation, service orientation, modularity, and reduced operational costs [4]. Despite these advancements, traditional centralized control architectures and direct point-to-point device connections are increasingly inadequate for meeting the demands of modern industrial

applications [5]. Consequently, the full realization of the I4.0 vision is regarded as a long-term goal, ultimately resulting in a complex ecosystem that integrates more than 30 distinct technological domains [6].

A key element of I4.0 innovation is the concept of cyber-physical convergence, including the creation of digital twins [7]. Industrial big data management plays a central role in realizing this concept [8]. Technologically advanced devices such as robots, sensor networks, virtual and augmented reality and GPS cameras are already having a significant impact on shaping industrial ecosystems by linking the cyber and physical worlds [9]. Data collected from the physical environment are transmitted to cyberspace for processing which adapts applications and services to the physical context. The results are then returned to the physical world through actuators and robotic systems. A digital twin, which is a virtual model of a physical object, allows simulating its behavior and optimizing production processes [10]. Virtual models analyze the state of physical objects using sensory data. They predict potential changes and subsequently adapt the physical systems based on optimized scenarios. Thus, the digital twin contributes to the creation of a closed cyber-physical system where data management becomes a critical process that short-circuits all production and technological chains [7]. The development of digital twin technologies is closely related to machine learning (ML) methods and complex system theory [11–13] and is determined by the completeness of data collection about the modelled process [14–16].

So, a key aspect of successful implementation of I4.0 technologies is the collection, accumulation, systematization, partitioning, and access to industrial big data. Effective data management is becoming critical to improving productivity, optimizing processes and making informed decisions in production management. However, it should be noted that industrial enterprises face a number of challenges when dealing with data in the context of I4.0 [8]. These include (i) heterogeneity of data coming from different sources (sensors, IoT devices, ERP systems), (ii) problems with their cleaning and integration, as well as (iii) the need to ensure high accuracy and reliability of data to build effective AI models [17]. In addition, AI implementation requires taking into account knowledge from the subject area, in this case the features of industrial processes, which makes this task even more complex and multidisciplinary [18].

The aim of this review is to analyze the peculiarities of data used for implementing AI solutions, to develop an approach to assessing the quality of industrial data and to develop recommendations for their preparation and processing within the framework of the I4.0 concept. This work addresses a gap in the literature by providing a comprehensive, systematic framework that bridges the domains of data management, quality assessment, and AI model readiness specifically for smart manufacturing. The practical significance of the study lies in the possibility of applying the developed recommendations to improve the efficiency of AI implementation in enterprises, which corresponds to the global trends of digitalization and automation in industry.

2. The Role of Industrial Data in Smart Manufacturing Infrastructure

To meet the challenges of cyber-physical convergence under I4.0 and improve the efficiency of digital twins, key technological factors are highlighted. New assembly lines will accelerate the reconfiguration of automated systems, ensuring reliability and short product lifecycles, which are critical to the competitiveness of enterprises [19]. Industrial Internet of Things (IIoT) and cyber-physical systems are revolutionizing business processes, covering the entire cycle—from production to interaction with customers and suppliers. Unlike consumer IoT, IIoT involves the use of powerful devices with advanced data storage and processing capabilities, requiring both local processing and information sharing [20]. The integration of industrial robots reduces costs and increases transparency by facilitating

human–robot interaction, where robots have skills comparable to humans [21]. Wireless sensor and actuator networks (WSANs) provide remote monitoring and control, reducing equipment failures and increasing productivity [22]. Networked control systems (NCSs) eliminate the need for wired connections, simplifying design and reducing costs [23]. New machine-to-machine (M2M) protocols with high data rates, minimal latency, and high reliability are bringing the realization of I4.0 requirements closer [24]. These technologies form the basis for the creation of intelligent production capable of adapting to dynamic market conditions.

Industrial data enables cyber-physical convergence within I4.0 by enabling the creation of digital twins representing physical objects. The natural evolution of data-driven industrial technologies and services leads to the creation of vast amounts of data of varying size and importance. Data serve as a fundamental resource for advancing I4.0 from machine automation to information automation and then to knowledge automation. In addition, data enable fast control cycles for applications such as zero-defect manufacturing, allowing information to be shared between production sites of a single plant operator or between value chains made up of different stakeholders. Indeed, concepts such as shared 'data buses' connecting factory environments have already been identified as the most important enabler of new I4.0 paradigms; for example, the concept of International Data Spaces Association, first introduced in [25]. Over the last few decades, large amounts of data have been generated in industrial environments through the widespread use of NCSs. In the beginning, these large amounts of data were rarely used for detailed analyses; instead, they were only used for routine technical checks and process logging. Later, the realization of the importance of extracting insights from the data took a leading role in I4.0 [26]. This is due to the exponential growth in the number of data sources, both archival and real-time. However, data alone are not useful and data processing processes, including data mining techniques and AI technologies, are needed to utilize them effectively [26,27].

Figure 1 illustrates technologies that enable industrial data and digital services focused on data management and manipulation. Industrial data of varying volume, intensity, and criticality are generated in these technology devices and distributed throughout the industrial and manufacturing ecosystem. This categorization is in line with the general architectural model of industrial automation, commonly known as the industrial automation pyramid [28]. The automation pyramid is an architecture created in 1990 by the International Society of Automation ISA–95 standard [29], which formed the basis of the IEC 62264 standard [30]. It represents a standard for the integration of enterprise management systems that proposes hierarchical levels from the industrial process itself to accounting and business management systems. It has been designed to be applicable to a variety of industries and processes, allowing all components involved in process automation to be represented.

The industrial automation pyramid is divided into five layers. Each layer is characterized by a set of networks and specific requirements (see Figure 1A). The pyramid implies both hierarchical and horizontal relationships. Horizontally, components within the same layer interact, while vertically they connect with subsystems directly above and below. At the bottom of the pyramid are the layer of manufacturing processes and field networks (sensors and actuators) (red), which typically consist of assembly lines, robots, IIoT devices, sensors and actuators. At this ground layer, the main requirements for data transmission are the real-time operation, low latency for data reception/transmission, and low jitter for control applications. The next layer (green) is the control network, which typically consists of controllers and connection servers. Higher levels are the supervision (violet) and production control ones, which consist of operator workstations, engineering and monitoring stations and servers, and much more advanced computation, communication and storage capabilities than the previous levels. At the very top layer is the enterprise

resource planning system (corporative management). In general, the upper layers of the automation pyramid have more relaxed latency constraints and real-time properties than the lower layers. The bottom two layers consist of the operational technology hardware and protocols, which are the core critical infrastructure of the enterprise automation system. All of the layers above consist of information technology hardware and protocols.

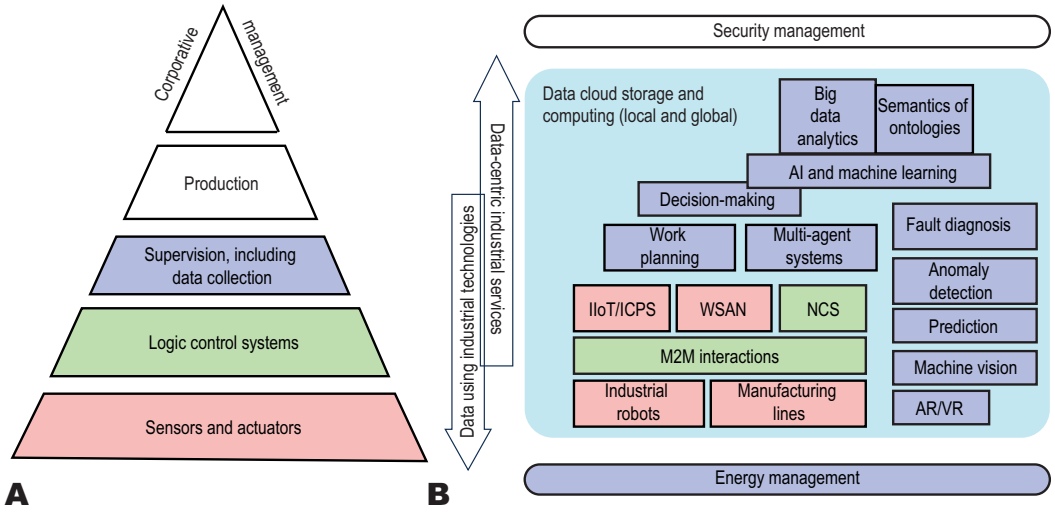

Figure 1. The evolution from traditional industrial automation to the Industry 4.0 paradigm. (**A**) The hierarchical automation pyramid of the Third Industrial Revolution. (**B**) A flexible, service-oriented architecture for smart manufacturing. The color scheme indicates the functional correspondence between the rigid layers in (**A**) and the modern, interconnected technologies in (**B**). The Industry 4.0 model (**B**) is characterized by bidirectional data flows: a bottom–up stream ('Data-centric industrial services') channels data from physical assets (e.g., robots, production lines) through cyber-physical systems (IIoT/ICPS) to enable AI and service applications, and a top–down stream ('Data using industrial technologies') guides the development of new, data-driven functions (e.g., big data analytics, semantic ontologies). Horizontal 'Energy Management' and 'Security Management' layers are cross-cutting enablers, which ensure resource optimization and end-to-end protection across the entire ecosystem. This transition from a rigid hierarchy to an interconnected network facilitates real-time analytics, decentralized decision-making, and autonomous control (adapted from [28,31]).

Smarter industrial data management requires the implementation of digital services across the control layer of the automation pyramid. These services span production control, manufacturing execution, and enterprise resource planning (see Figure 1B). Big data analytics, ML and semantic modeling, facilitate industrial integration and cyber-physical convergence because typical data integration involves large amounts of data, traffic, comparison, and transformation of different data formats [32]. These operations are usually performed in local or global data cloud services that horizontally span industrial installations. Decision-making, job scheduling, and human-in-the-loop approaches are expected to compose hybrid command and control systems with dynamic structure and distributed intelligence, capable of meeting industrial needs and rapid market changes [33]. Augmented reality (AR), virtual reality (VR) services, cameras, and machine vision systems are expected to be able to collect data and mimic the human information processing system to utilize intelligence capabilities and interpret the industrial environment. Prediction and predictive processes, anomaly detection, and fault diagnosis are expected to not only collect data but also support advanced analytics to extract useful insights with high return on investment of such technologies in manufacturing processes and networks [34]. In addition, a sustainable production process is not possible without intelligent energy management and safety solutions, which form two end-to-end services that are present in all control networks at different levels of the automation pyramid [35].

Realizing the high-level control system in Figure 1B depends on two prerequisites: the comprehensive collection and processing of digital industrial data and the training of AI models for deployment across various services and pyramid levels. This foundation enables business intelligence derived from automated data processing. Consequently, this intelligence can be leveraged for production management throughout the entire automation pyramid. This is illustrated in Figure 2, which schematically illustrates the main industrial data flow that we further explore in this article. Let us look at industrial data flows in more detail abd start our consideration with the main sources of industrial data, which should be orientated on when building an intelligent I4.0 infrastructure.

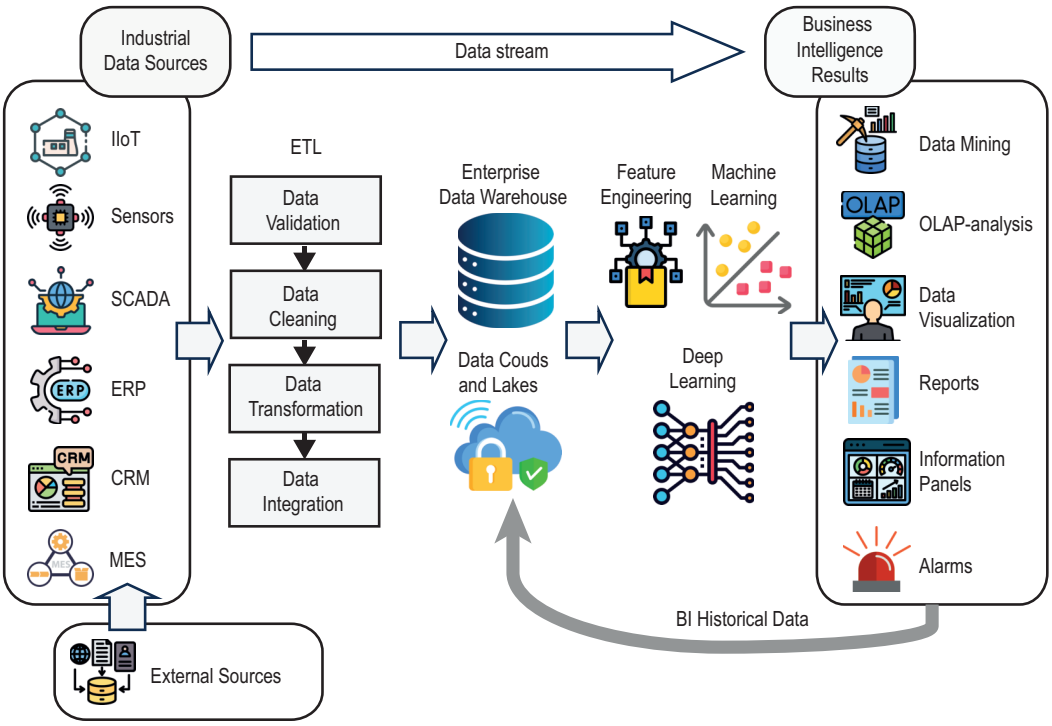

Figure 2. Schematic representation of the industrial data lifecycle in Industry 4.0. The left panel illustrates the main sources of industrial data described in Section 3. The central panel depicts key stages of data handling: primary processing (*left*), storage (*center*), and intelligent processing (*right*). The right panel highlights potential business intelligence outcomes enabled by machine learning and deep learning techniques applied to collected industrial data. Notably, historical process data can be accumulated (grey arrow) and leveraged to continuously refine AI models. This lifecycle underscores that raw data must undergo critical preparation stages (cleaning, integration) before fueling AI models. The closed feedback loop, where historical data refines future models, is essential for creating adaptive and self-improving production systems.

3. Data Sources in Industry

The implementation of AI solutions at industrial enterprises is impossible without the use of data that come from various sources (see Figure 2). These data reflect the state of equipment, parameters of production processes, as well as management and logistics information.

3.1. Sensors and IIoT Devices

Sensors and IIoT devices are key sources of data in industry. They collect information about equipment status, temperature, pressure, vibration, humidity, and other parameters in real time [2]. These data allow monitoring the current state of production processes, predicting possible failures and optimizing equipment performance.

The main features of data from sensors and IoT devices that affect the quality of industrial data are as follows:

- The sampling update rate of the data can be very high ($f_s > 1$ Hz), generating large amounts of information [36].
- Collected data can be both numerical (temperature, pressure) and categorical (equipment status), which fundamentally requires multimodal data processing [37].
- Errors and omissions may occur due to equipment or data transmission failures, which requires data preprocessing, error correction, and gap filling [17].

These features also give rise to the main challenges of utilizing industrial data from IoT sensors. First, the processing of large-scale data in real time requires cloud storage solutions capable of handling massive datasets and supporting efficient partitioning for training intelligent control systems. Second, ensuring the reliability and accuracy of industrial data remains a critical and urgent issue.

The predictive maintenance capabilities of IIoT and AI are demonstrated by several industrial case studies. Siemens, for instance, employs sensors to monitor gas turbine conditions, analyzing vibration, temperature, and pressure data with machine learning. This approach has reduced maintenance costs by 30% and increased equipment uptime [38]. Similarly, the Cenal coal-fired power plant in Turkey used IoT sensor data analyzed by AI to stabilize combustion temperature and optimize soot cleaning, achieving annual savings exceeding USD 700,000 and a 15% reduction in NOx emissions [39]. Beyond predictive maintenance, these technologies enhance quality control, as seen at a BMW factory where convolutional neural networks analyze visual data from IoT cameras to automatically inspect car body quality [40]. Furthermore, companies like Enel leverage IIoT for operational efficiency, using devices to monitor grid load and optimize energy consumption [39].

3.2. Enterprise Resource Planning Systems

Enterprise resource planning (ERP) systems provide data related to enterprise management such as production processes, supply chains, human resources, financial planning and accounting, and product quality [41]. They integrate data from different departments such as finance, production, logistics, and human resources are structured and used to optimize business processes. However, data from ERP systems have their own peculiarities that are important to consider when using them for analysis and decision-making. Let us consider them in more detail:

- Data in ERP systems are stored in a structured manner, usually in relational databases (e.g., SQL). This means that information is organized into tables where each row represents a record and each column represents an attribute (e.g., product name, quantity, price). Importantly, modern ERP systems provide an application programming interface (API) to access the data, allowing integration with other applications such as business intelligence or AI systems. This greatly simplifies data analysis due to the clear structure of data collection and presentation, as well as the ability to use standard database tools (e.g., SQL queries).
- Data are updated less frequently than data from sensors or IIoT devices. For example, inventory data may be updated once a day, while financial reports may be updated once a week or month. Many processes in ERP systems, such as financial reporting, are performed in batch processing, which causes delays in data updates, and data are often manually entered into ERP systems, which also slows down data accumulation and processing. On the one hand, low update frequency simplifies data management, as real-time processing is not required, but on the other hand, it makes it difficult to

make quick decisions. Also, ERP systems have limited applicability for tasks that require real-time data (e.g., production line management).

- ERP systems rarely operate in isolation. They integrate with other enterprise systems such as manufacturing execution system (MES), customer relationship management (CRM), and supply chain management. To integrate ERP with MES or supervisory control and data acquisition (SCADA), middleware such as Apache Kafka or MQTT are often used to provide real-time data transfer [42]. For integration of such systems, the unified standards such as Open Platform Communications Unified Architecture (OPC UA) or electronic data interchange have been developed [43].
- ERP systems are designed to process large volumes of data, making them suitable for large enterprises. However, this requires significant computing resources. By storing historical data, these systems enable trend analysis, which in turn supports long-term strategic decision-making.

The main challenges of using ERP systems in industry are first the differences in data formats between ERP systems and other sources and second the need to synchronize data in real time to make certain operational decisions based on information from ERP systems.

For example, Siemens uses the SAP ERP system to manage production processes at its plants. The system integrates data on inventories, orders, and production capacity to optimize scheduling and reduce downtime [44]. In the automotive company Toyota, ERP is used for supply chain management, providing transparency and control over all stages of supply, from the purchase of raw materials to the delivery of finished products, which minimizes inventories and reduces costs [45].

3.3. SCADA Systems

SCADA systems are used to monitor and control industrial processes, providing real-time data collection, visualization, and analysis [46]. Data from SCADA systems have their unique features, which are important to consider when using them for monitoring and control. Let us consider these features in more detail.

- SCADA systems operate in real time, which means that data from sensors, transducers, and other devices are continuously being collected and processed. This allows operators to react instantly to changes in production processes.
- SCADA systems collect data at a high level of detail, which means there are a large number of parameters for each device or process. For example, for a pump, parameters such as pressure, temperature, vibration, rotational speed, and energy consumption can be monitored. Integrating data from a large number of sensors and devices requires powerful computing resources for processing. To manage such data streams, a 'tagging' system is commonly used, meaning that each parameter in the SCADA system is identified by a unique label ('tag'), allowing data to be easily tracked and analyzed.
- Data in SCADA systems are often redundant, containing duplicate or irrelevant information. This redundancy stems from the large number of data sources and the high frequency of data updates. For instance, pipeline pressure might be measured by two independent sensors, creating duplicates. Furthermore, not all collected data are useful for analysis; ambient temperature readings, for example, may have no impact on the core process but are still logged by default. Consequently, data post-processing is essential to eliminate this redundancy using methods such as averaging, interpolation, and duplicate removal.

A core functionality of SCADA systems is the automated notification of critical events through alarms that activate upon parameter excursions or anomalous conditions, thereby enabling prompt operational response. Nevertheless, leveraging these systems as data sources for broader industrial applications poses two principal challenges: the demand for

high-performance, real-time data processing (a trait shared with IoT architectures) and the imperative of ensuring stringent cybersecurity.

3.4. Other Sources of Industrial Data

In addition to those listed above and most commonly used in manufacturing, industrial companies can use data from:

- MES systems, which provide information about the production process, including data on product quality and lead times [47].
- CRM systems, which contain data about customers, orders and sales, which is useful for demand forecasting [48].
- A variety of external data sources that are not directly industrial: market data, weather, or logistics data that can influence production processes [49].

4. Challenges of Industrial Data Collection, Processing, and Storage at Industrial Facilities

The successful implementation of AI solutions in industrial enterprises depends on robust data management across the entire lifecycle—from collection to analysis. In this context, two widely adopted paradigms for handling data are ETL and ELT. Both approaches comprise the same three stages (extraction, transformation, and loading), but differ in the sequence of operations. In ETL, data undergo transformation during an intermediate preparation phase before being loaded into the target repository, such as an enterprise data warehouse (illustrated in Figure 2). By contrast, ELT first loads raw data directly into the target system (e.g., cloud data warehouses or data lakes), where transformation is performed afterward.

The selection between Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) methodologies represents a critical architectural decision in industrial data management pipelines. Both approaches facilitate data integration from disparate sources, yet differ fundamentally in their execution sequence and operational characteristics, as summarized in Table 1.

Table 1. Comparison between ETL and ELT approaches for industrial data management.

Aspect	ETL (Extract, Transform, Load)	ELT (Extract, Load, Transform)
Processing Sequence	Data transformation occurs before loading into target system	Data transformation occurs after loading into target system
Transformation Location	Separate processing server/staging area	Within target data warehouse/lake
Data Volume Handling	Suitable for moderate volumes of structured data	Optimized for large volumes of structured and unstructured data
Flexibility	Limited flexibility; transformations are predefined	High flexibility; transformations can be modified post-loading
Real-time Processing	Challenging due to preprocessing requirements	More adaptable to real-time and streaming scenarios
Infrastructure Requirements	Requires substantial intermediate processing resources	Demands powerful target system with computational capacity
Data Latency	Higher latency due to staging transformations	Lower latency for raw data availability

Table 1. *Cont.*

Aspect	ETL (Extract, Transform, Load)	ELT (Extract, Load, Transform)
Implementation Complexity	Moderate complexity with well-defined transformation rules	Higher complexity in managing transformations within target system
Cost Considerations	Higher intermediate infrastructure costs	Higher target system and storage costs
Typical Use Cases	Data warehousing, structured business intelligence	Big data analytics, data lakes, exploratory analysis
Industrial Applicability	Mature processes with stable data schemas	Evolving processes requiring analytical flexibility

In ETL, all these operations take place outside the target system, in the preparation phase. In industrial enterprises, where significant amounts of data are generated from multiple sensors and other sources in real time, the use of edge computing technologies can significantly reduce the data transfer load and reduce the amount of information stored in enterprise data warehouses or cloud platforms. For example, data warehouses that support online analytical processing (OLAP) require data to be converted into a SQL-compatible relational format beforehand. However, this approach has a significant drawback: the transformations are performed once, making the ETL process not flexible enough. If there is a need to apply a new type of analysis to already transformed data, a complete redesign of the data processing model may be required.

In contrast to ETL, the ELT method offers more flexibility because the data are loaded into the data warehouse in its original form, where they can be validated, structured, and transformed at any time. This allows for countless transformations of raw data that are stored indefinitely. However, in industrial environments where data volumes are extremely large, storing all data in raw form is often not justified in terms of cost and resources. As a result, the ELT method is not widely used in industry, where more optimized approaches such as ETL using edge computing for data preprocessing are preferred.

Despite these challenges, ELT can become a viable option for medium-sized enterprises under specific conditions. Key enabling factors include (i) access to scalable and cost-effective cloud storage and computing services (e.g., pay-as-you-go models from major cloud providers), which lower the initial infrastructure investment; (ii) evolving or exploratory analytical needs that require the flexibility to reprocess raw data without re-ingesting it; and (iii) the availability of in-house or external expertise in modern data stack technologies (e.g., cloud data warehouses like Snowflake or BigQuery [50]) that are designed for ELT workflows. For such enterprises, ELT can reduce the initial complexity of data pipeline design and accelerate time-to-insight from new data sources.

In industrial contexts, the ETL approach remains predominant due to its maturity and alignment with structured manufacturing data environments. The predefined transformation logic in ETL ensures data quality and consistency, which is crucial for mission-critical manufacturing operations. However, ELT is gaining traction in scenarios requiring rapid ingestion of heterogeneous data sources and when analytical requirements evolve frequently. The choice between these paradigms should consider factors including data characteristics, computational resources, latency tolerance, and the dynamic nature of analytical requirements within the smart manufacturing ecosystem.

In the following section, we focus on the ETL model of data collection, processing, and storage, as it remains the predominant approach in industrial enterprises. Specifically, we examine methods for data acquisition and preparation, preprocessing techniques, and the

application of ML and DL for data analysis, illustrated through the industrial data stream presented in Figure 2.

4.1. Data Collection and Preparation for AI Applications

Data collection in industrial facilities is complicated by a number of factors such as the heterogeneity of data sources, the large volume of data, and the need for real-time operation. This makes it different from, for example, biomedical data collection for AI [51–53] or machine vision applications [54]. The main requirements for industrial data collection that can be further utilized in AI systems include:

- Integration of data from different sources. Data can come from sensors, IIoT devices, ERP and SCADA systems, which requires them to be combined into a unified system [2]. For example, in a chemical plant, data from sensors that monitor pressure and temperature are integrated with data from an ERP system that manages raw material inventory. This allows the production process to be optimized and costs to be reduced.
- Real-time and stream processing. Stream processing technologies such as Apache Flink or Apache Storm [55] are used to process real-time data. For example, in a factory, data from assembly robots are processed in real time to detect defects early in the production process. This minimizes scrap losses and increases product quality. In real-time data processing, edge computing is tried to be used [56]. The latter is an approach to data processing in which computations are performed closer to the data source (at the 'edge' of the network, hence 'edge') rather than in centralized cloud servers. This reduces computational latency, reduces the load on the network as it reduces transmission costs and improves reliability and security by preventing data leakage over the Internet, and as a result, improves system performance, especially in environments where data processing speed is critical. However, edge computing has disadvantages, in particular, computing resources on edge devices are usually limited, and also the distributed network of edge devices requires a complex management and synchronization system.
- Ensuring data quality. During the data acquisition phase, it is important to ensure data accuracy and completeness, which requires the use of calibrated sensors and reliable data transfer protocols. This raises the issue of regularly checking the devices and the industrial data itself for accuracy and errors [57].

The decision between edge and cloud computing architectures represents a critical strategic consideration in industrial data management. Cloud computing refers to the delivery of computing services, including servers, storage, databases, networking, software, and analytics, over the internet ('the cloud') on a pay-as-you-go basis. This paradigm offers centralized resources that can be rapidly provisioned and scaled, making it particularly suitable for applications requiring substantial computational power and storage capacity. In contrast, edge computing brings computation and data storage closer to the location where it is needed, improving response times and saving bandwidth.

The choice between these computing paradigms should be guided by specific operational requirements and constraints rather than technological trends alone [58]. Edge computing is preferred when low latency is critical for applications requiring real-time control, such as robotic assembly lines or safety monitoring systems where milliseconds matter. It is also advantageous in scenarios with bandwidth constraints, such as remote industrial sites with limited network connectivity or high data transmission costs. Furthermore, edge computing becomes essential when data privacy and security are paramount, particularly for sensitive production data that must remain within factory premises due to regulatory or intellectual property concerns. Additionally, it provides operational re-

silience for applications that must continue functioning during network outages or cloud service interruptions.

Conversely, cloud computing is more appropriate when scalable resources are needed for large-scale data analytics, model training, or historical analysis requiring substantial computational power. It excels in scenarios where collaborative analysis is essential for multi-site operations requiring centralized data aggregation and cross-facility insights. Cloud platforms are also ideal for long-term storage and archiving of historical data for regulatory compliance or trend analysis over extended periods. Moreover, they offer advanced AI/ML capabilities for complex model training and inference tasks that benefit from cloud-based AI services and GPU clusters [59].

In practice, most modern industrial implementations adopt a hybrid approach, where edge devices handle time-sensitive preprocessing and immediate control tasks, while the cloud manages resource-intensive analytics, model retraining, and enterprise-wide data integration. This symbiotic relationship ensures optimal performance while maintaining the benefits of centralized intelligence and distributed execution, creating a balanced architecture that addresses both immediate operational needs and long-term analytical requirements.

4.2. Data Preprocessing

Data preprocessing is an important step to prepare data for use in AI models. In the case of industrial data, the main steps include the following procedures [60].

- Data validation and cleaning involves fixing data gaps and filtering out noise in the data. Missing values can be filled in using interpolation or ML techniques, e.g., regression algorithms such as the *k*-nearest neighbor method. For example, reactor temperature data may contain missing values due to sensor failures. Interpolation can recover missing values and ensure data continuity. Data filtration typically uses techniques such as moving average, digital filters, or wavelet transforms that help remove noise and improve data quality. For example, equipment vibration data may contain noise due to external influences. Filtering helps to isolate useful signals to analyze the condition of the equipment.
- Data transformation includes both normalization and standardization. The first is to bring the data to a single normalized range, e.g., $[0, 1]$ or $[-1, 1]$, to eliminate the effects of data scale on the performance of AI models. The latter is particularly relevant when processing data from sensors at different scales and of different physical nature. For example, a manufacturing facility may collect temperature and pressure data inside the plant, which have different scales. Normalization allows them to be used in the same AI model. Standardization involves the conversion of data to a standard normal distribution with zero mean and unit variance. This is also important when building ML models. In the previous example, temperature and pressure data can be standardised for use in clustering algorithms. One of the simplest ways to perform data harmonization is the standard score conversion procedure, which is well known in statistics. If we have a set of measurements $\{x_i | i = 1, \ldots N\}$, where N is the number of measurements, we calculate the standardised score as $z_i = (x_i - \overline{x})/\sigma$, where $\overline{x}$ is the mean and σ is the standard deviation of the set of measurements $\{x_i\}$.

A practical vibration analysis scenario for predictive maintenance demonstrates the utility of filtering and interpolation. In a chemical plant, a centrifugal pump's vibration sensor captures a signal often contaminated with high-frequency noise from electromagnetic interference and adjacent equipment. Here, a moving average or band-pass filter is essential for noise suppression, thereby revealing the underlying vibration signature associated with the pump's rotational components to facilitate fault detection in bearings

or impellers. Furthermore, transient communication glitches that create data gaps can be resolved using linear interpolation between known points. This data reconstruction is vital for maintaining time-series continuity, ensuring the data are suitable for input into machine learning models designed for failure prognosis.

4.3. Data Integration

After data cleaning and transformation, the next stage of data preparation is the process of data integration (see Figure 2), which is one of the key tasks when implementing AI in industrial enterprises. The main problem of industrial data integration is data heterogeneity. Thus, due to the peculiarities of technological processes and equipment, data from different sources may have different formats (CSV, JSON, XML, etc.), sampling frequency, and level of detail. Therefore, ETL processes are used to convert data into a single format. Usually, in specialized cloud platforms (e.g., AWS IoT, Microsoft Azure), edge computing for processing data closer to the source and middleware for data aggregation are used for this purpose [56]. Ensuring the confidentiality and protection of industrial data during transmission and storage is usually performed using encryption and authentication protocols such as OPC UA [61].

4.4. Machine Learning in Industrial Data Analysis

ML plays a key role in analyzing industrial data, enabling the identification of patterns, classification and prediction of processes [26]. Figure 3 shows schematic representation of ML methods applied in various industrial data analysis tasks. ML methods fall into four main categories that find their applications in industrial data processing.

1. Supervised learning is used when all accumulated data are labeled and the desired outcome or goal is known. Then the problem is defined as a type of classification and regression task, such as determining the state of equipment (operational/faulty) or predicting parameters (temperature, pressure) based on collected labeled historical data [62].
2. Unsupervised learning is applied to data which have no labels (tags) and used to find hidden structures in data, such as clustering to detect anomalies or analyzing relationships between parameters of manufacturing processes [63].
3. Reinforcement learning (RL) utilizes a trial and error learning approach to learning by direct interaction with the environment [64]. It does not need supervision or a predefined dataset with/without labels and is applied whenever tasks in dynamic environments requiring real-time decision-making need to be solved, allowing adaptation to changing environmental conditions. Therefore, RL is most effective when applied to tasks requiring interaction with a changing environment, e.g., application in manufacturing process optimization, where the model learns to control parameters (e.g., conveyor speed) to maximize efficiency, robot motion control, or interaction of several linked machines in a serial production line [65].
4. Semi-supervised learning (SSL) is increasingly finding application in industrial data processing tasks when the cost of labeling data samples is expensive or time consuming [66]. The main difference between semi-supervised and fully supervised ML is that the latter can only be trained using fully labeled datasets, while the former uses both labeled and unlabeled data samples during training. For example, determining the type of faulty data samples detected is a difficult labor-intensive task for engineers. As a consequence, most faulty samples turn out to be unlabeled, but they still contain important process information. If these unlabeled samples can be put to good use, the efficiency of the fault classification system can be greatly improved. SSL methods modify or augment the supervised algorithm, the so-called base learner, to

incorporate information from unlabeled examples. The labeled data are used to justify the predictions of the base learner and add structure (e.g., the number of existing classes and the main characteristics of each class) to the learning task. In this case, semi-supervised MLs using both labeled and unlabeled samples yield an improved fault classification model compared to a model that depends only on a small fraction of labeled data samples.

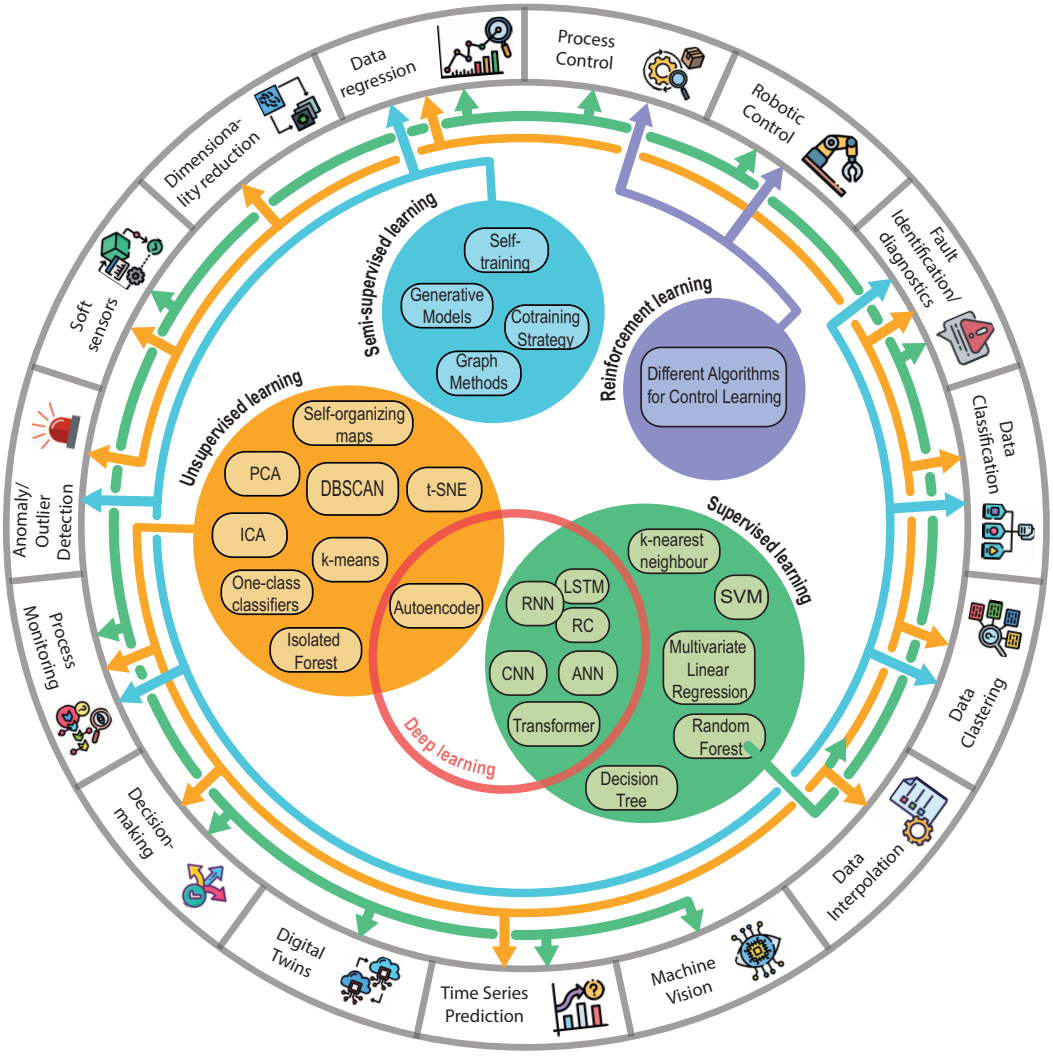

Figure 3. Schematic representation of ML methods applied in various industrial data analysis tasks. The arrows indicate the correspondence between ML methods (see Section 4.4) and their most typical applications in production systems (see Section 5.4). The mapping provides a decision-support framework for selecting AI tools based on problem characteristics: supervised learning for predictive modeling from historical data and reinforcement learning for adaptive control in dynamic environments.

Traditional ML methods require prior feature extraction, the so-called *feature engineering*, including the creation of new features, selection of relevant data, and coding of categorical variables. Applied to industrial data analysis, feature engineering addresses the following tasks.

- *Feature extraction.* Creating new features from existing data, such as calculating derivatives or integrals for time series, can reveal underlying dynamics that are not apparent in the raw signal. For instance, methods inspired by the analysis of complex systems can help identify synchronized patterns or coherent structures within chaotic-looking industrial data [67]. For example, at a wind farm, wind speed data can be converted

into features such as average speed over the last hour, which helps to improve prediction of power generation without creating redundant information.

- *Feature selection or data dimensionality reduction.* Removing redundant or irrelevant features using techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), or Lasso regression. For example, in manufacturing, data on thousands of sensor measured parameters can be reduced to a few key attributes which can be a combination of the original measured quantities.
- *Encoding categorical data* numerically using one-hot encoding or label encoding. For example, in logistics, cargo type data (e.g., "fragile," "dangerous," etc.) can be encoded for use in route optimization models.

However, DL eliminates the need for feature creation by automatically extracting useful insights from raw data. Among the DL techniques, most in demand are:

- *Artificial neural networks* (ANNs) are the foundational architecture of deep learning, consisting of interconnected layers of neurons that can model complex relationships in data. ANNs are applied in a wide range of industries, including retail for demand forecasting, and in logistics for optimizing supply chain operations. Their versatility makes them suitable for both regression and classification tasks across diverse domains.
- *Recurrent neural networks* (RNNs), including long short-term memory (LSTM) and reservoir computing (RC), for time series prediction and digital twin creation. These are particularly useful in industries like finance for stock price forecasting, in energy for predicting electricity demand, and in manufacturing for predictive maintenance.
- *Convolutional neural networks* (CNNs) for visual inspection tasks such as product quality control. CNNs are widely used in the automotive industry for detecting defects in car parts and in retail for automated checkout systems.
- *Autoencoders* (AEs) for automatic feature extraction which is particularly useful in vibration or sound analysis tasks for fault diagnosis. AEs are applied in industries like aerospace for monitoring aircraft engine health and in manufacturing for anomaly detection in machinery.
- *Transformers*, originally developed for natural language processing (NLP), have become a cornerstone in sequence modeling due to their ability to handle long-range dependencies efficiently. Transformers are used in industries for applications such as automated customer support (chatbots) and document summarization in legal and financial sectors. Their self-attention mechanism makes them highly effective in tasks requiring context-aware decision-making.

These techniques enable the creation of digital twins, optimizing production processes and improving equipment reliability. However, the quality of ML models directly depends on the quality and completeness of the data used for training, which requires careful preprocessing and integration of data from different sources. Therefore, in the next section, we focus on the issues of evaluating the quality of industrial data for training AI models.

5. Assessing the Quality of Industrial Data

Data quality is a critical factor for the successful implementation of AI in industry. Poor data quality can lead to incorrect predictions, erroneous decisions and, as a result, significant financial and production losses. Assessing data quality requires context, as it can only be evaluated based on purpose and use. This context is often referred to as 'fitness for use' [68]. The usability of data includes a number of factors that primarily determine its quality [69]. Data quality describes 'the extent to which data is fit for use by data consumers' [70]. To define the various requirements, detailed sets of criteria have been proposed to describe and measure individual aspects of industrial data quality. Here, we summarize these approaches based on the idea of end-to-end integrated analysis of

data quality throughout the entire cycle of data collection, processing and use through the context of specific industrial tasks [71].

The implementation of projects to identify potentially useful patterns in industrial datasets that can be used in I4.0 is usually based on a cross-industry standard process for data mining (CRISP-DM) [72]. Figure 4A shows the six iterative steps of the CRISP-DM process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Due to its success, CRISP-DM is considered one of the most widely used process models in smart manufacturing and serves as an organizational framework for the considerations underlying the formulation of industrial data quality criteria.

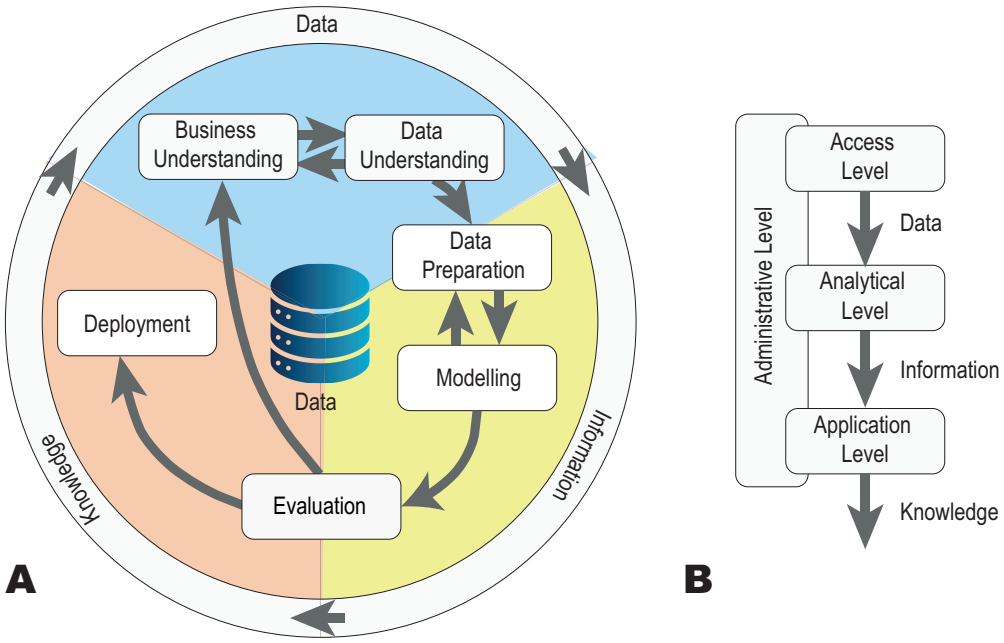

Figure 4. Integrated visualization of the cross-industry standard process for data mining (CRISP-DM) (**A**) and four levels of classification of quality criteria for industrial data analysis (**B**). The arrows in the CRISP-DM diagram (**A**) represent the iterative and non-linear character of the methodology, wherein phase transitions (such as reverting from Modeling to Data Preparation) constitute a fundamental characteristic. This framework synchronizes a four-level data quality assessment (**B**) with the CRISP-DM process (**A**), ensuring data possess both technical fidelity and strategic business value for informed decision-making at administrative and application levels.

When applying quality criteria to industrial data analysis projects, multidimensional sets of criteria should take into account the life-cycle-dependent nature of their target observation [73]. In discussing industrial data quality criteria, we limit our discussion to data, information, and knowledge, which are integrated into the scope of CRISP-DM in Figure 4A. The second step of CRISP-DM, 'Data Understanding', involves the verification of data quality. For this purpose, CRISP-DM proposes exemplary guiding questions and refers to requirements regarding the necessary quality of data and results that have been previously assessed in the business understanding phase. This includes the creation of a data quality report, which becomes the focus for data processing in the next step, 'Data Preparation'. After this point, CRISP-DM considers the data quality as guaranteed and only revises it during the evaluation of the results of the data analysis process [72]. However, given the changing nature of data during the introduction of digital technologies into the production process, the quality criteria require constant monitoring and adjustment throughout the process of analyzing data and building models based on it. This is similar to the concept of total data quality management, which seeks to extend traditional total quality management, which requires the consideration of product quality throughout the

entire lifecycle [74]. In fact, data quality is directly related to the quality of data analysis, which describes the extent to which products based on industrial data are usable by specific customers.

5.1. Steps in Assessing the Quality of Industrial Data

The industrial data quality assessment criteria are grouped for ease of use and take into account the changing perspectives of data handling. As we consider quality assessment criteria focused on industrial applications, they are based on the principles of industrial engineering, which focuses on a systems approach, methods, and problem solving in a continuous improvement manufacturing environment, and include four main steps for data-driven decision-making processes:

1. Data access and provisioning. The first step is to provide access to the necessary data sources, including selecting relevant sources, filling in missing data, and preparing data for analysis. This step corresponds to the initial stages of CRISP-DM—'Business Understanding' and 'Data Understanding' (see Figure 4A).
2. Data analysis. In the second stage, data are analyzed to extract useful information. Various tools and techniques are used, including data science algorithms. This stage covers the 'Data Preparation' and 'Modeling' stages in CRISP-DM.
3. Application of information. The third stage is the utilization of the information obtained in industrial processes. This includes both one-off analyses and long-term monitoring. In CRISP-DM, this corresponds to the 'Deployment' stage, where knowledge of the problem and subject matter knowledge help to extract the necessary knowledge from the data.
4. Process management. The fourth stage covers the management of peripheral processes such as long-term data management, responsibility allocation, security, and ethical use of data. This step has no direct analogue in CRISP-DM, but is critical for industrial applications.

Using these four steps, it is possible to clearly structure the criteria applicable to data quality assessment. Quantitative data quality criteria should be defined at the earliest stage of the project, but if quality problems are later identified, adjustments to the criteria and the data collection and processing procedure can be made at any stage in an iterative manner.

5.2. Level of Assessing the Quality of Industrial Data

Let us consider the criteria for assessing and measuring industrial data quality based on the concept of continuous end-to-end analysis. The data quality criteria are structured into four levels shown in Figure 4B and corresponding to the stages of data analysis presented in Section 5.1.

1. Access level covers the collection of the necessary data according to the defined analysis objectives at the stage of understanding the processes in the industrial system. This level addresses aspects related to the quality of the raw ('raw') data and relevant business processes. The criteria shown in Table 2 support the objectives of collecting and preparing data for analysis.
2. Analytical level refers to the quality of data analysis and is primarily concerned with information. This corresponds to the second and third step of CRISP-DM (see Figure 4A) 'Data Preparation' and 'Modeling'. In this step, data access needs the context of a special use case or problem definition. The context guides the processing steps using at least one special analysis method. The criteria presented in Table 2 assess the quality of the data analysis.

3. Application level concerns the assessment of data quality during the application phase in industrial settings, which is related to the 'Evaluation' and 'Deployment' phases of CRISP-DM (see Figure 4A). The criteria for this level, presented in Table 2, aim to establish high quality of data analysis results. They specifically target the inclusion of personnel not previously involved in the analyses as future users of deployment solutions.
4. Administrative level introduces criteria for assessing the quality of industrial data in terms of being able to administer it effectively. Table 2 shows the proposed criteria that are associated with all phases of CRISP-DM (see Figure 4A).

Using these four steps, it is possible to clearly structure the criteria applicable to data quality assessment. Quantitative data quality criteria should be defined at the earliest stage of the project, but if quality problems are later identified, the criteria and data collection and processing procedures can be adjusted at any stage in an iterative manner.

Table 2. List of criteria for assessing the quality of industrial big data.

Criterion	Description	Measurement/Evaluation
Access level		
Accessibility	The data must be available through defined interfaces for further processing.	Evaluated binary: available/not available.
Relevance	The data must be relevant to the purpose of the analysis.	Evaluated at three levels: insufficient, ideal, excessive.
Timeliness	The data should be available at the right time.	Evaluated binary: timely/not timely.
Uniqueness	The data should be free of technical duplicates and redundancy, which is ensured by basic data integration [see Section 4.3].	Number of duplicate and redundant data identified during analysis using algorithms to identify repeated records.
Consistency	The data should be consistent over time and between different sources which is ensured by basic data integration [see Section 4.3].	Check that data from different systems are not inconsistent. The data are up to date and do not contain time gaps or anomalies. The number of inconsistencies between sources and time periods is assessed.
Validity	The data must conform to established rules (format, value ranges) and must not contain inconsistencies.	Evaluated binary: valid/non-valid.
Analytical level		
Accuracy	The data must match the reference values.	Measured as the proportion of correct values in the total data.
Completeness	The data should be complete, with no omissions.	Measured as the proportion of non-zero values.
Error-free	The data should be free of logical inconsistencies. Defined by the objectives of the analysis.	Evaluated through logical consistency checks, proportion of data passing validation, number of inconsistencies with reference data, automated tests and anomaly analysis.
Value Added	Data should enable the creation of new information useful for analysis.	Evaluated through the cost-benefit ratio of the data.

Table 2. *Cont.*

Criterion	Description	Measurement/Evaluation
Application level		
Cost-effectiveness	Solutions must be economically justifiable.	Evaluated through the ratio of costs to value achieved.
Conciseness of presentation	The results of the analysis should be compact and understandable.	Evaluated through user surveys for ease of comprehension, share of visualized data in the total data volume, number of key indicators in reports compared to the total data volume.
Consistency of presentation	Solutions should be homogeneous and compatible with previous data.	Evaluated through the number of errors or inconsistencies in data structure, the proportion of data validated against data homogeneity metrics.
Interpretability	Results should be presented in understandable terms and units.	Evaluated qualitatively through surveys.
Understandability	Decisions should be easily understandable for operational decision-making.	Evaluated through ad hoc interviews with experts.
Administrative level		
Accessibility	The data must be available through defined interfaces for further processing.	Evaluated binary: available/not available
Relevance	The data must be relevant to the purpose of the analysis.	Evaluated at three levels: insufficient, ideal, excessive
Timeliness	The data should be available at the right time.	Evaluated binary: timely/not timely.
Uniqueness	The data should be free of technical duplicates and redundancy, which is ensured by basic data integration [see Section 4.3].	Number of duplicate and redundant data identified during analysis using algorithms to identify repeated records.
Consistency	The data should be consistent over time and between different sources which is ensured by basic data integration [see Section 4.3].	Check that data from different systems are not inconsistent. The data are up to date and do not contain time gaps or anomalies. The number of inconsistencies between sources and time periods is assessed.
Validity	The data must conform to established rules (format, value ranges) and must not contain inconsistencies.	Evaluated binary: valid/non-valid.

5.3. Developing Recommendations for Preparing Big Data for AI Implementation

The successful deployment of AI models is fundamentally dependent on the quality and structure of the underlying data. This subsection establishes a set of actionable recommendations for the preparation of industrial big data, addressing the critical stages of collection, processing, and integration to ensure robust model training and reliable decision-making.

5.3.1. Industrial Data Collection

A robust data collection framework is the cornerstone of effective AI training in industrial settings. This phase must address three critical requirements to construct a dataset that is comprehensive, timely, and of high fidelity:

- Integration of Heterogeneous Data Sources. Industrial data are inherently multi-source, originating from systems such as SCADA (Supervisory Control and Data Acquisition), ERP (Enterprise Resource Planning), and IoT sensors. Integrating these disparate streams into a unified data platform is essential to form a holistic view of operations, combining equipment status, process parameters, and management data. Technologies like Apache Kafka or MQTT middleware are commonly employed to facilitate this real-time data integration and synchronization.
- Real-time Data Acquisition and Processing. For time-sensitive industrial processes, the ability to acquire and process data in real time is critical. This capability enables rapid response to anomalies, minimizing production scrap and downtime. Implementing stream processing frameworks (e.g., Apache Flink, Apache Storm) allows for continuous data analysis as it is generated. Furthermore, leveraging edge computing architectures processes data closer to source, significantly reducing latency and alleviating network load.
- Proactive Data Quality Assurance. The accuracy and reliability of AI models are directly contingent upon the quality of the input data. To mitigate the risks of erroneous predictions caused by poor data, a proactive approach to quality assurance is mandatory. This involves the regular maintenance and calibration of sensors, the use of robust and reliable data transmission protocols, and the implementation of automated data quality monitoring systems to detect and correct errors and omissions at the point of collection.

5.3.2. Data Processing

Following acquisition, raw industrial data must be cleansed and transformed to rectify errors, inconsistencies, and inaccuracies inherent in collection. This processing stage is critical for constructing a reliable dataset for AI training and primarily involves three core procedures:

- Data Cleansing and Denoizing. Industrial datasets frequently contain noise, outliers, and missing values that can severely degrade model performance. To mitigate this, a suite of signal processing techniques must be applied. This includes using interpolation methods (e.g., linear or spline) to impute missing points and digital filters (e.g., moving average filters or wavelet transforms) to suppress noise and isolate meaningful signals.
- Normalization and Standardization. The heterogeneous nature of industrial sensors results in data with varying scales and units. To ensure stable and efficient model convergence, these data must be brought to a common, dimensionless scale. Normalization (e.g., scaling to a $[0, 1]$ range) and standardization (e.g., scaling to zero mean and unit variance) are essential preprocessing steps, particularly for machine learning algorithms like gradient-based optimizers that are highly sensitive to input feature magnitudes.
- Data Structuring and Feature Engineering. Processed data streams must be structured into a coherent format suitable for model ingestion. This involves aligning time-series data from disparate sources, handling different sampling frequencies, and creating derived features that enhance predictive power. While ETL or ELT pipelines are instrumental in this structuring, the focus at this stage is on the transformational logic: aggregating, windowing, and engineering features to create a curated training dataset. Cloud platforms such as AWS IoT or Microsoft Azure should be used to integrate data from different sources; encryption and authentication protocols such as OPC UA should be used to ensure data security.

5.3.3. Data Governance and Management

Beyond initial processing, sustainable AI implementation requires robust data governance to ensure security, integrity, and long-term usability. This involves establishing policies and systems for two critical areas:

- Data Security and Access Control. Industrial data assets often comprise sensitive operational intelligence and must be protected against unauthorized access and cyber threats. A comprehensive security framework is essential, incorporating encryption (e.g., AES-256) and authentication protocols (e.g., OPC UA) for data in transit and at rest. This must be supplemented with rigorous role-based access control systems and consistently updated security patches to mitigate evolving risks.
- Long-term Data Storage and Lifecycle Management. The continuous refinement of AI models depends on the systematic storage and management of historical data. Implementing scalable, cloud-based data warehouses (e.g., Amazon S3, Google BigQuery, or Azure Data Lake) enables cost-effective long-term archiving and facilitates efficient analysis of large historical datasets. A defined data lifecycle policy ensures that data are retained, archived, and purged according to value, maintaining system performance and relevance for future model retraining.

Adherence to these recommendations for data preparation and management establishes a reliable foundation of high-quality data. This foundation is a critical enabler for developing accurate AI models, which in turn drive significant business value through improved forecasting, operational optimization, and reduction in costs and downtime.

5.3.4. Data Quality Assessment for ML-Based Models

The performance and reliability of ML models are fundamentally constrained by the quality of their training data. To mitigate the risks of erroneous predictions and flawed decision-making, the data quality assessment framework established in Section 5.2 and Table 2 must be actively employed throughout the AI project lifecycle. This entails continuously validating data against these criteria and iteratively refining data collection and processing pipelines based on the results.

In Table 3, we summarize the research findings and present a concrete analytical comparison of ML methods in smart manufacturing. This table systematically matches common industrial tasks with suitable ML paradigms and specific algorithms. Specific results from the application of these technologies are described below, with appropriate references to the original work.

Table 3. Comparison of ML methods for typical industrial data analysis tasks.

Industrial Task	ML Category	Key Methods	Strengths	Notes [Sources]
Predictive Maintenance	Supervised Learning	SVM, Random Forest, ANFIS	High accuracy in failure prediction	Prediction of equipment failures based on historical data from sensors (e.g., vibration, temperature, and other bearing data) [62,75]
	Deep Learning	RNN, LSTM, RC	Automated feature extraction from images; High precision	Sufficient increase in the accuracy of predicting the condition of industrial equipment [76–78]
Quality Control	Deep Learning	CNN, AE	Multivariate time-series forecasting	CNN for defect detections outperformed the traditional computer vision [79–81]

Table 3. *Cont.*

Industrial Task	ML Category	Key Methods	Strengths	Notes [Sources]
Anomaly Detection	Unsupervised Learning	One-Class SVM, Isolation Forest, AE & VAE	Effective with unlabeled data; Identifies novel failure modes	Improving industrial fault detection with one-class deep learning models trained solely on normal data, without needing labeled anomalies [63,82–84]
Process Optimization	Reinforcement Learning	DDPG, TD3, PPO, Q-learning & DQN	Autonomous real-time decision-making; No need for precise physical models; Handles complex state-action spaces	Enables self-improving systems through trial-and-error learning in simulation environments (digital twins); Reduces online computation by 87.7% compared to traditional optimization [85–87]
Soft Sensor	Semi-supervised Learning	Label Propagation, Semi-supervised AE	Reduces need for expensive labeled data; Leverages unlabeled process data	Boosting fault diagnosis accuracy by leveraging unlabeled data to augment scarce labeled examples [66]

5.4. Utilizing ML-Based Techniques

Collected and preprocessed data serve as the foundation for building AI systems using a diverse suite of ML technologies. While Figure 3 provides a high-level overview of key applications in smart manufacturing and Section 4.4 introduces various ML approaches, this section focuses on their practical deployment. Given that the technical specifics of basic ML algorithms in I4.0 are well covered in existing literature [26,65,66,88], including industry-specific reviews [62,89], we instead highlight the strategic mapping of ML paradigms to characteristic industrial problems.

5.4.1. Supervised Learning for Predictive Modeling

Supervised learning is deployed for tasks where the goal is to map inputs to a known output, primarily falling into classification and regression.

- Classification: For tasks such as equipment state determination or part inspection, algorithms like Support Vector Machines (SVMs), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Random Forests are highly effective. Researchers often combine such classifiers to boost their efficiency. A common approach is to integrate them with low-cost sensors, including RGB cameras, accelerometers, and gyroscopes. This synergy, especially when powered by DL algorithms like CNNs, facilitates sophisticated visual monitoring capabilities.
  To enhance the accuracy and robustness of fault detection and diagnosis in an industrial steam turbine, Salahshoor et al. [90] developed a hybrid framework fusing an ANFIS and a SVM via an ordered weighted averaging operator. This data fusion strategy capitalized on the complementary strengths of both classifiers, yielding a system that outperformed either classifier used in isolation. The results demonstrated tangible improvements, including a reduction in diagnosis time for key faults—such as cutting the time for a thermocouple fault from 32 to 13 samples—and the elimination of specific misclassifications, thereby increasing overall diagnostic reliability and reducing the potential for false alarms.
  Other example here is study [91], where a cost-effective IMU sensor (MPU6050) capturing three-axis accelerometer and gyroscope data was used to monitor the condition of a

computer fan. Instead of relying on expensive or specialized hardware, the authors creatively transformed the multi-axis vibration signals into time-frequency spectrograms using the Short-Time Fourier Transform. These spectrograms were then combined into a single RGB image, effectively converting vibration analysis into a visual recognition task. This RGB image was processed using a CNN, which successfully classified the fan's operational state—such as normal, idle, or faulty—with high accuracy. This approach demonstrates how low-cost sensors, when paired with DL techniques like CNNs, can enable sophisticated visual monitoring and fault diagnosis without the need for complex or expensive sensing systems.

- Regression: For predicting continuous variables (e.g., temperature), methods like linear regression and tree-based algorithms are standard. A critical application is in accurate fault diagnosis, where surveys indicate SVM-based algorithms (39% of studies) and ANN-based DL (34% of studies) are most prevalent [26]. The practical implementation and enhancement of these methodologies are well illustrated in recent research on tree-based approaches, which demonstrate both robust predictive capability for continuous parameters and exceptional performance in diagnostic frameworks. The application of tree-based models for forecasting continuous parameters is effectively demonstrated by Tran et al. [92], who employed Regression Trees for one-step-ahead prediction of vibration amplitude in a low methane compressor. Their work shows the viability of tree-based models in time-series forecasting, a critical step for anticipating machine degradation. Furthermore, the flexibility of these models enables their enhancement and integration into sophisticated diagnostic frameworks. For instance, Li et al. [93] developed an improved Decision Tree-based method incorporating virtual sensor-based fault indicators for diagnosing faults in variable refrigerant flow systems. Their hybrid approach, combining physical knowledge with data-driven learning, provided more reliable diagnosis results compared to several other tree-based data-driven models, demonstrating a pathway to augment diagnostic capability. The robustness and high accuracy achievable with tree-based ensembles are further emphasized by Noura et al. [94]. Their bi-phase framework, leveraging an ensemble of tree-based classifiers including Random Forest and XGBoost, achieved perfect accuracy in both detection and diagnosis of faults in a diesel engine system. Notably, their feature importance analysis revealed that optimal performance could be attained with a minimal feature set, underscoring the method's efficiency and interpretability. These findings collectively demonstrate that tree-based algorithms constitute a versatile toolkit, capable of addressing interconnected challenges of continuous state prediction and discrete fault classification with high precision, thereby enriching the ecosystem of machine learning techniques in industry prognostics.

5.4.2. Unsupervised Learning for Data Exploration

Unsupervised learning uncovers hidden structures within unlabeled data. Its primary applications include:

- Clustering: Algorithms such as k-means and DBSCAN are used to group similar data points, identifying natural patterns or operational regimes. These methods are particularly valuable in industrial process monitoring, where they help uncover underlying structures in high-dimensional sensor data without requiring pre-labeled datasets. For instance, Thomas et al. [95] evaluated various clustering techniques, including k-means, DBSCAN, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), and mean shift, combined with dimensionality reduction methods like PCA and ICA. Their study demonstrated that DBSCAN effectively identified fault states in the Tennessee Eastman Process and an industrial separation tower, even when

fault labels were unknown a priori. Similarly, Bagherzade et al. [96] proposed an ensemble clustering approach to detect operational modes in industrial gas turbines. By aggregating multiple partitions from diverse algorithms (e.g., k-means, hierarchical clustering, DBSCAN, and others) and applying a consensus function, their method identified consistent clusters corresponding to distinct operational states such as idle, partial load, and full load. This ensemble strategy improved robustness and enabled the discovery of sub-operational modes, providing a scalable framework for real-time process monitoring and knowledge discovery in complex industrial systems.

The systematic review by Chaudhry et al. [97] provided a comprehensive analysis of unsupervised clustering algorithms, evaluating their strengths and weaknesses for pattern identification in complex data. This overview is particularly valuable for industrial applications, offering guidance on selecting appropriate algorithms like DBSCAN and k-means for process monitoring and operational regime detection in high-dimensional sensor data.

- Anomaly Detection: Techniques like one-class SVM, isolation forest, and AEs are vital for identifying rare events or deviations from normal operation, which is crucial for predictive maintenance [82,84].
- Anomaly Detection: Techniques like one-class SVM, isolation forest, and AEs are vital for identifying rare events or deviations from normal operation, which is crucial for predictive maintenance. These one-class methods are particularly valuable in industrial settings where labeled fault data is scarce, allowing models to be trained exclusively on normal operating data to effectively detect anomalies [82,84].

  For instance, AEs achieve high detection performance with F1-scores up to 93% on industrial datasets like UNSW-NB15, making them suitable for complex pattern recognition in sensor data and images [82]. Variational Autoencoders (VAEs) further enhance this capability by learning probabilistic representations, improving generalization on diverse operational data [84]. In contrast, Isolation Forest offers superior computational efficiency with inference times as low as 1.3 ms per sample, ideal for real-time monitoring applications despite slightly lower accuracy (F1-score: 91%) [82]. One-Class SVM provides a balanced approach with 92% F1-score and moderate latency (2.8 ms), effective for boundary-based anomaly detection in multidimensional sensor data [82]. The choice of method depends on operational constraints: deep learning models (AEs/VAEs) for accuracy-critical applications versus lightweight methods (Isolation iForest) for resource-constrained environments.

5.4.3. Deep Learning for Complex Pattern Recognition

DL excels at automatic feature extraction from high-dimensional, complex data.

- Time-Series Analysis: RNNs, particularly LSTM, and RC networks are standard for analyzing temporal data like sensor streams [98,99]. So, Lei et al. [76] used a novel self-supervised deep LSTM network for industrial temperature prediction in aluminum processes application, which achieved a testing Root Mean Square Error (RMSE) as low as 0.0078 and a prediction accuracy of up to 89.5% in the middle temperature zone, demonstrating effective performance with limited labeled data. In other example, Zhang et al. [78] used a LSTM network optimized with orthogonal experimental design and feature engineering, which reduced the prediction RMSE by up to 97.6% compared to the auto regressive integrated moving average (ARIMA) model on sensor data from an industrial pump. The proposed LSTM-based methods demonstrated not just a small improvement, but a dramatic increase in the accuracy of predicting the condition of industrial equipment on real data, which directly indicates its practical value for predictive maintenance in I4.0.

- Complex Signal Processing: DL is particularly powerful for tasks involving vibration, sound, and image analysis, where raw data contain intricate patterns [100]. For example, in [79], a CNN-based method for online weld defect detection was used with a result of 99.38% mean classification accuracy, outperforming their previous audio-based method (87.16% accuracy). In [81], a Deep CNN with data augmentation was used for weld defect detection, achieving 99.01% accuracy and reducing overfitting on a dataset of 9 680 images.

These capabilities are foundational for creating digital twins—digital representations of physical assets that enable real-time simulation and optimization [101]. Modeling the dynamic and stochastic behavior of such systems often leverages methods from the theory of complex nonlinear systems [102–104].

5.4.4. Semi-Supervised Learning for Industrial Fault Detection and Diagnosis

SSL has emerged as a promising paradigm for industrial fault detection and diagnosis, effectively addressing the challenge of scarce labeled data which is commonplace in real-world industrial settings. By leveraging a small set of labeled instances alongside abundant unlabeled data, various SSL methodologies have demonstrated significant performance improvements.

For instance, the self-training approach has been enhanced with sophisticated confidence measures. Zheng et al. [105] employed a Self-Training framework integrated with a Temporal-Spatial Confidence Measure, enabling more reliable utilization of unlabeled data for process fault diagnosis. Co-training strategies, which utilize multiple classifiers, have also shown considerable efficacy. He et al. [106] implemented a Co-Training method combining a Generative Adversarial Network and a Residual Network for steel surface defect classification, achieving high accuracy with limited labels.

In the realm of feature extraction, DL models have been adapted for SSL. Jiang et al. [107] utilized a Semi-Supervised Dynamic Sparse Stacked Auto-Encoders model for fault classification, effectively capturing essential features from partially labeled datasets. Similarly, in [108], a Semi-Supervised Deep Ladder Network was applied for gear fault diagnosis, enhancing model robustness and diagnostic reliability by fusing information from both data types.

Furthermore, intrinsically semi-supervised methods that incorporate unlabeled data directly into the objective function have been developed. Jia et al. [109] proposed a Dynamic Active Safe Semi-Supervised SVM, incorporating active learning and safety mechanisms to ensure robust fault identification in expensive chemical processes. Generative models have proven particularly powerful for handling complex data scenarios. Xu et al. [110] addressed class imbalance in bearing faults using a Semi-Supervised Conditional Generative Adversarial Network, significantly improving the diagnosis of rare failure modes. Ensemble-based SSL methods like Tri-Training have also been successfully applied; Liu et al. [111] used this approach with multiple base classifiers for milling chatter detection, demonstrating enhanced model stability and noise immunity.

5.4.5. Reinforcement Learning in Industrial Process Optimization

RL has emerged as a transformative technology for industrial process optimization, enabling autonomous systems to learn optimal control policies through trial-and-error interactions with their environment. Unlike traditional optimization methods that require precise first-principles models, RL agents can learn directly from operational data or simulations, making them particularly valuable for complex processes that are difficult to model analytically. In manufacturing, RL algorithms such as Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) have demonstrated remarkable capabilities in op-

timizing transfer lines and flexible manufacturing systems, achieving significant reductions in inventory levels while maintaining high throughput rates. These systems can handle state spaces with dimensions previously considered intractable for conventional optimization techniques, learning to balance production scheduling, maintenance operations, and quality control in dynamic environments [85].

The integration of RL with digital twin technology represents a particularly powerful paradigm for industrial optimization. By training RL agents in high-fidelity virtual replicas of physical processes, manufacturers can safely explore optimal control strategies without risking actual production systems. In plastic injection molding, for instance, Proximal Policy Optimization (PPO) algorithms have been deployed to continuously adjust critical parameters such as mold temperature, fill time, and injection pressure to maintain product quality despite material variations and machine wear. Similarly, in chemical processing, actor–critic architectures have shown 9.6% improvement in annual profit by dynamically optimizing reactor conditions in response to fluctuating commodity prices. These RL-based Real-Time Optimization (RL-RTO) systems reduce computational overhead by up to 87.7% compared to conventional nonlinear programming approaches, while maintaining robust performance across diverse operating conditions [86,87].

For problems with discrete action spaces, Q-learning and its deep variant Deep Q-Networks (DQNs) provide effective solutions for operational decision-making. These value-based methods have been successfully applied in production scheduling and inventory management, where agents learn optimal policies for task sequencing and resource allocation. The combination of discrete Q-learning for high-level decision-making with continuous control algorithms like DDPG for parameter optimization creates comprehensive hierarchical control architectures that can address both strategic and tactical optimization challenges in complex industrial environments.

5.4.6. Specialized Applications: Soft Sensors and Dimensionality Reduction

Two particularly impactful applications of ML in industrial settings are soft sensors and dimensionality reduction.

- Soft Sensors: These are mathematical models that estimate parameters which are difficult to measure directly (e.g., real-time chemical concentration) using data from other, more affordable sensors (e.g., temperature, pressure). Data-driven soft sensors, which build models using regression, ANNs, or SVM to relate input data to a target parameter, are a promising area [112–114]. This is closely related to the problem of *system identification*, where ML also shows significant promise [100,115]. For example, Curreri et al. [116] used transfer learning with RNN and LSTM-based soft sensors, achieving near-optimal performance on a target industrial process while reducing design time by over 100 h compared to full re-training.
- Dimensionality Reduction: Industrial processes often generate vast amounts of data. Dimensionality reduction simplifies models, accelerates training, and improves interpretability by preserving key information while reducing the number of features. Methods like PCA, t-distributed stochastic neighbor embedding (t-SNE), and AEs are used for sensor data analysis, failure prediction, and quality control, as they help extract relevant parameters and reduce noise.

6. Limitations and Prospects for the Use of AI Data in Industry

The introduction of AI into I4.0 offers significant opportunities to optimize processes, improve productivity, and reduce costs. However, despite progress in this area, there are technical and organizational limitations that may slow down or even prevent the successful implementation of AI. At the same time, the prospects for the use of AI data

in industry remain promising, especially given the development of new technologies and methodologies. In this section, we review the key limitations and ways to overcome them, as well as the prospects for the use of data for AI in the context of I4.0.

6.1. Limitations of the Use of Data for AI in Industry and Possible Ways to Overcome Them

One of the key challenges faced by industrial enterprises is data heterogeneity. Data come from different sources such as sensors, IoT devices, ERP, and SCADA systems, and can have different formats, update rates, and levels of detail. This makes it difficult to integrate and utilize them in a single system. Both industry standards for data formats and universal integration tools such as ETL processes can be applied to solve this problem. In strategically important industries such as energy, metals, and engineering, developing industry standards can greatly simplify data integration between companies. For industries with more flexible requirements or low levels of digitalization, the use of ETL processes may be more appropriate as it allows for rapid adaptation to change. Hybrid solutions that combine industry standards for basic data and ETL processes for non-standard sources can also be effective.

Another major challenge is poor data quality, which can lead to incorrect forecasts and decisions, causing significant financial and operational losses. Investments in more accurate sensors and monitoring systems, as well as data preprocessing techniques such as filtering and interpolation, can be used to improve data quality at the data collection stage. In industries where data quality is critical, emphasis should be placed on state-of-the-art equipment. For small and medium-sized enterprises, data preprocessing methods can be a temporary solution to start implementing AI without significant costs. The development of software solutions for data preprocessing can also reduce dependence on imported equipment and stimulate the development of local IT solutions.

Many industrial processes require real-time data processing, which creates additional complexities. Processing large amounts of data in real time requires significant computing resources, which can be difficult for small enterprises. Edge computing allows data to be processed closer to the source of data generation, reducing latency and network load, which is especially important for large enterprises with distributed infrastructure. For SMBs, cloud platforms can be a more affordable solution, offering flexibility and scalability without the need for significant upfront costs. The combination of edge computing for mission-critical processes and cloud platforms for less resource-intensive tasks can be a balanced approach that takes into account both technical and economic aspects.

Thus, overcoming the limitations associated with the use of data for AI in industry requires an integrated approach that includes data standardization, data quality improvement, and the application of modern technologies for real-time processing. These measures can contribute to the successful implementation of AI and unlocking its potential in industry.

6.2. Prospects for the Use of AI Data in Industry

Despite current limitations, the prospects for the use of data for AI in industry remain extremely promising. One of the key prospects is the development of digital twins—virtual models of physical objects that can simulate equipment behavior, predict changes, and optimize manufacturing processes. The question is how to accelerate the adoption of digital twins, given the complexity of creating and maintaining them.

To navigate the inherent complexity and financial constraints of digital twin adoption, small and medium-sized enterprises (SMEs) must prioritize a focused and strategic approach to data management [117–119]. Key initial priorities should include:

- Start with a Critical Asset: Focus on a single, high-value piece of equipment rather than a full production line to demonstrate value and manage scope.
- Ensure Foundational Data Quality: Prioritize the collection of clean, consistent, and time-synchronized data from a few critical sensors over amassing large volumes of unstructured data.
- Establish Robust Data Integration: Build a simple, reliable, and automated pipeline from the asset to a central storage (e.g., a cloud database) to ensure the digital twin receives a live, trustworthy data feed.

This start small, focus on quality strategy allows SMEs to build a scalable foundation without being overwhelmed.

Another important prospect is knowledge automation. AI not only automates processes, but also creates new knowledge based on data analysis, which contributes to a better understanding of manufacturing processes and the development of innovative technologies. However, integrating these techniques into existing processes remains a challenge.

Integration with cloud technologies also opens up new opportunities. Cloud platforms provide powerful tools for storing, processing, and analyzing data, as well as scalability for AI solutions. The main challenge here is ensuring data security and privacy.

The development of IoT and edge computing provides more and more data for real-time analysis, enabling better monitoring and management of production processes. The widespread adoption of such technologies contributes to improving production efficiency and opens up new horizons for industry. Processing multidimensional data in an IIoT network requires coordination across multiple sources. The challenge lies not only in transmitting data but also in ensuring its coherence for decision-making. Fundamental research, such as [120,121], shows that consistent communication between remote network components is achieved through coherence. This principle can be applied to the design of data architecture for digital twins, where various modules (e.g., physical model, wear model, planning system) must be synchronized to form a coherent and accurate picture.

Finally, beyond individual ML models, the next evolutionary step in industrial AI lies in the development of autonomous AI agents [122]. These are sophisticated systems that perceive their environment through data streams, make decisions using AI models, and execute actions to achieve specific manufacturing goals, often operating with a significant degree of autonomy. In the context of Industry 4.0, AI agents can orchestrate complex processes by integrating multiple capabilities. For instance, an agent could continuously monitor sensor data via a CNN-based visual inspection system, analyze temporal patterns using an LSTM for predictive maintenance, and then invoke a reinforcement learning policy to dynamically re-schedule production tasks on a digital twin before a potential failure occurs [123]. This moves the paradigm from isolated AI predictions to closed-loop, intelligent control [124].

The implementation of such agents hinges directly on the foundational themes of this review: high-quality, well-integrated data is the agent's perception; robust, validated ML models form its decision-making core; and secure, reliable data pipelines enable its action. Therefore, the data management and quality assessment frameworks discussed herein are not merely supportive but are essential prerequisites for the deployment of effective and trustworthy AI agents in industrial settings.

7. Conclusions

The integration of AI stands as a cornerstone of the I4.0 paradigm, presenting a transformative potential to enhance manufacturing productivity, optimize processes, and drive down operational costs. However, this review demonstrates that the path to successful AI implementation is critically dependent on overcoming fundamental data-related challenges.

The heterogeneity of industrial data sources, persistent issues with data quality, and the demanding requirements for real-time processing remain significant hurdles.

To navigate this complex landscape, a holistic and strategic approach is essential. This includes the rigorous standardization of data formats and protocols, the adoption of modern data processing frameworks (e.g., for stream processing and edge computing), and targeted investment in sensor infrastructure and connectivity. Furthermore, robust data governance policies encompassing security, lifecycle management, and continuous quality assessment are not ancillary but central to building a reliable data foundation.

Looking forward, the trajectory of industrial AI points toward even greater integration and intelligence. We identify three key frontiers for future development:

- Proliferation of Digital Twins: The evolution from static models to dynamic, self-learning digital twins will enable real-time simulation, predictive what-if analysis, and autonomous optimization of physical assets.
- Rise of Knowledge Automation and Generative AI: Beyond predictive analytics, AI systems will increasingly codify expert knowledge and generate novel process optimizations, shifting their role from decision-support to proactive decision-making.
- Ubiquitous Cloud-Edge Integration: The maturation of hybrid cloud-edge architectures will seamlessly distribute computational load, facilitating scalable AI deployment while ensuring low-latency control for critical operations.

In essence, the future industrial enterprise will be characterized by a tightly coupled, data-driven feedback loop between the physical and digital worlds. While technological advancements continue to be crucial, the ultimate differentiator for competitiveness is an organization's ability to cultivate a data-centric culture and master the entire data lifecycle, from sensor to insight.

Author Contributions: Both authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

AE	Autoencoder
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CRM	Customer Relationship Management
CRISP-DM	Cross-industry Standard Process for Data Mining
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DDPG	Deep Deterministic Policy Gradient
DL	Deep Learning
DQN	Deep Q-Network
ELT	Extract, Load, andTransform
ERP	Enterprise Resource Planning
ETL	Extract, Transform, and Load
I4.0	Industry 4.0
IoT	Internet of Things
IIoT	Industrial Internet of Things
ICA	Independent Component Analysis

LSTM	Long Short-Term Memory
MES	Manufacturing Execution System
ML	Machine Learning
NCS	Networked Control System
OPC UA	Open Platform Communications Unified Architecture
PCA	Principal Component Analysis
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
RL-RTO	RL-Based Real-Time Optimization
RC	Reservoir Computing
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SCADA	Supervisory Control and Data Acquisition
SME	Small and Medium-sized Enterprise
SMB	Small and Medium-sized Business
SSL	Semi-supervised Learning
SVM	Support Vector Machine
TD3	Twin Delayed DDPG
t-SNE	t-distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoder
WSAN	Wireless Sensor and Actuator Network

References

1. Lu, Y. Industry 4.0: A survey on technologies, applications and open research issues. *J. Ind. Inf. Integr.* **2017**, *6*, 1–10. [CrossRef]
2. Lee, J.; Bagheri, B.; Kao, H.A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf. Lett.* **2015**, *3*, 18–23. [CrossRef]
3. Jardim-Goncalves, R.; Romero, D.; Grilo, A. Factories of the future: Challenges and leading innovations in intelligent manufacturing. *Int. J. Comput. Integr. Manuf.* **2017**, *30*, 4–14.
4. Indri, M.; Grau, A.; Ruderman, M. Guest editorial special section on recent trends and developments in industry 4.0 motivated robotic solutions. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1677–1680. [CrossRef]
5. Qiu, J.; Gao, H.; Chow, M.Y. Networked control and industrial applications [special section introduction]. *IEEE Trans. Ind. Electron.* **2015**, *63*, 1203–1206. [CrossRef]
6. Caiado, R.G.G.; Machado, E.; Santos, R.S.; Thomé, A.M.T.; Scavarda, L.F. Sustainable I4. 0 integration and transition to I5. 0 in traditional and digital technological organisations. *Technol. Forecast. Soc. Change* **2024**, *207*, 123582. [CrossRef]
7. Hlophe, M.C.; Maharaj, B.T. From cyber–physical convergence to digital twins: A review on edge computing use case designs. *Appl. Sci.* **2023**, *13*, 13262. [CrossRef]
8. Raptis, T.P.; Passarella, A.; Conti, M. Data management in industry 4.0: State of the art and open challenges. *IEEE Access* **2019**, *7*, 97052–97093. [CrossRef]
9. Conti, M.; Das, S.K.; Bisdikian, C.; Kumar, M.; Ni, L.M.; Passarella, A.; Roussos, G.; Tröster, G.; Tsudik, G.; Zambonelli, F. Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence. *Pervasive Mob. Comput.* **2012**, *8*, 2–21. [CrossRef]
10. Qi, Q.; Tao, F. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access* **2018**, *6*, 3585–3593. [CrossRef]
11. Lai, Y.C. Digital twins of nonlinear dynamical systems: A perspective. *Eur. Phys. J. Spec. Top.* **2024**, *233*, 1391–1399. [CrossRef]
12. Hramov, A.E.; Kulagin, N.; Andreev, A.V.; Pisarchik, A.N. Forecasting coherence resonance in a stochastic Fitzhugh–Nagumo neuron model using reservoir computing. *Chaos Solitons Fractals* **2024**, *178*, 114354. [CrossRef]
13. Kong, L.W.; Weng, Y.; Glaz, B.; Haile, M.; Lai, Y.C. Reservoir computing as digital twins for nonlinear dynamical systems. *Chaos Interdiscip. J. Nonlinear Sci.* **2023**, *33*, 033111. [CrossRef]
14. Andreev, A.V.; Pisarchik, A.N.; Kulagin, N.; Jaimes-Reátegui, R.; Huerta-Cuellar, G.; Badarin, A.A.; Hramov, A.E. Stochastic cloning of dynamical systems with hidden variables. *Phys. Rev. E* **2025**, *112*, 015303. [CrossRef]
15. Human, C.; Basson, A.H.; Kruger, K. A design framework for a system of digital twins and services. *Comput. Ind.* **2023**, *144*, 103796. [CrossRef]

16. Schluse, M.; Rossmann, J. From simulation to experimentable digital twins: Simulation-based development and operation of complex technical systems. In Proceedings of the 2016 IEEE International Symposium on Systems Engineering (ISSE), Edinburgh, UK, 3–5 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–6.
17. Wang, L.; Törngren, M.; Onori, M. Current status and advancement of cyber-physical systems in manufacturing. *J. Manuf. Syst.* **2015**, *37*, 517–527. [CrossRef]
18. Qin, S.J.; Badgwell, T.A. A survey of industrial model predictive control technology. *Control Eng. Pract.* **2003**, *11*, 733–764. [CrossRef]
19. Cupek, R.; Ziebinski, A.; Zonenberg, D.; Drewniak, M. Determination of the machine energy consumption profiles in the mass-customised manufacturing. *Int. J. Comput. Integr. Manuf.* **2018**, *31*, 537–561. [CrossRef]
20. Sisinni, E.; Saifullah, A.; Han, S.; Jennehag, U.; Gidlund, M. Industrial internet of things: Challenges, opportunities, and directions. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4724–4734. [CrossRef]
21. Dean-Leon, E.; Ramirez-Amaro, K.; Bergner, F.; Dianov, I.; Cheng, G. Integration of robotic technologies for rapidly deployable robots. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1691–1700. [CrossRef]
22. Zhu, J.; Zou, Y.; Zheng, B. Physical-layer security and reliability challenges for industrial wireless sensor networks. *IEEE Access* **2017**, *5*, 5313–5320. [CrossRef]
23. Raza, S.; Faheem, M.; Guenes, M. Industrial wireless sensor and actuator networks in industry 4.0: Exploring requirements, protocols, and challenges—A MAC survey. *Int. J. Commun. Syst.* **2019**, *32*, e4074. [CrossRef]
24. Pang, Z.; Luvisotto, M.; Dzung, D. Wireless high-performance communications: The challenges and opportunities of a new target. *IEEE Ind. Electron. Mag.* **2017**, *11*, 20–25. [CrossRef]
25. Otto, B.; Jürjens, J.; Schon, J.; Auer, S.; Menz, N.; Wenzel, S.; Cirullies, J. *Industrial Data Space. Digital Souvereignity Over Data*; Fraunhofer-Gesellschaft: Munich, Germany, 2016.
26. Ge, Z.; Song, Z.; Ding, S.X.; Huang, B. Data mining and analytics in the process industry: The role of machine learning. *IEEE Access* **2017**, *5*, 20590–20616. [CrossRef]
27. Wang, D. Building value in a world of technological change: Data analytics and industry 4.0. *IEEE Eng. Manag. Rev.* **2018**, *46*, 32–33. [CrossRef]
28. Folgado, F.J.; Calderón, D.; González, I.; Calderón, A.J. Review of Industry 4.0 from the perspective of automation and supervision systems: Definitions, architectures and recent trends. *Electronics* **2024**, *13*, 782. [CrossRef]
29. *ISA–95*; Enterprise—Control System Integration. International Society of Automation (ISA): Research Triangle Park, NC, USA, 1995.
30. *IEC 62264*; Enterprise—Control System Integration. International Electrotechnical Commission (IEC): Geneva, Switzerland, 2013.
31. Lucizano, C.; de Andrade, A.A.; Facó, J.F.B.; de Freitas, A.G. Revisiting the Automation Pyramid for the Industry 4.0. In Proceedings of the 2023 15th IEEE International Conference on Industry Applications (INDUSCON), São Bernardo do Campo, Brazil, 22–24 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1195–1198.
32. Jirkovskỳ, V.; Obitko, M.; Mařík, V. Understanding data heterogeneity in the context of cyber-physical systems integration. *IEEE Trans. Ind. Inform.* **2016**, *13*, 660–667. [CrossRef]
33. Trentesaux, D.; Borangiu, T.; Thomas, A. Emerging ICT concepts for smart, safe and sustainable industrial systems. *Comput. Ind.* **2016**, *81*, 1–10. [CrossRef]
34. Lechevalier, D.; Narayanan, A.; Rachuri, S.; Foufou, S. A methodology for the semi-automatic generation of analytical models in manufacturing. *Comput. Ind.* **2018**, *95*, 54–67. [CrossRef]
35. Cardin, O.; Ounnar, F.; Thomas, A.; Trentesaux, D. Future industrial systems: Best practices of the intelligent manufacturing and services systems (IMS2) French Research Group. *IEEE Trans. Ind. Inform.* **2016**, *13*, 704–713. [CrossRef]
36. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **2013**, *29*, 1645–1660. [CrossRef]
37. Tsanousa, A.; Bektsis, E.; Kyriakopoulos, C.; González, A.G.; Leturiondo, U.; Gialampoukidis, I.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. A review of multisensor data fusion solutions in smart manufacturing: Systems and trends. *Sensors* **2022**, *22*, 1734. [CrossRef]
38. de Castro-Cros, M.; Velasco, M.; Angulo, C. Machine-learning-based condition assessment of gas turbines—A review. *Energies* **2021**, *14*, 8468. [CrossRef]
39. Majhi, A.A.K.; Mohanty, S. A comprehensive review on internet of things applications in power systems. *IEEE Internet Things J.* **2024**, *11*, 34896–34923. [CrossRef]
40. Fast, Efficient, Reliable: Artificial Intelligence in BMW Group Production. Retrieved 15 July 2019. Available online: https://www.press.bmwgroup.com/middle–east/article/detail/T0299271EN/fast–efficient–reliable:–artificial–intelligence–in–bmw–group–production?language=en (accessed on 26 October 2025).
41. Shehab, E.; Sharp, M.; Supramaniam, L.; Spedding, T.A. Enterprise resource planning: An integrative review. *Bus. Process Manag. J.* **2004**, *10*, 359–386. [CrossRef]

42. Dingorkar, S.; Kalshetti, S.; Shah, Y.; Lahane, P. Real-Time Data Processing Architectures for IoT Applications: A Comprehensive Review. In Proceedings of the 2024 First International Conference on Technological Innovations and Advance Computing (TIACOMP), Bali, Indonesia, 29–30 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 507–513.
43. Khan, A.A.; Abonyi, J. Information sharing in supply chains–Interoperability in an era of circular economy. *Clean. Logist. Supply Chain* **2022**, *5*, 100074. [CrossRef]
44. Gerig, I. Standardization and Automation as the Basis for Digitalization in Controlling at Siemens Building Technologies. In *The Digitalization of Management Accounting: Use Cases from Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 193–215.
45. Powell, D. ERP systems in lean production: New insights from a review of lean and ERP literature. *Int. J. Oper. Prod. Manag.* **2013**, *33*, 1490–1510. [CrossRef]
46. Pliatsios, D.; Sarigiannidis, P.; Lagkas, T.; Sarigiannidis, A.G. A survey on SCADA systems: Secure protocols, incidents, threats and tactics. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1942–1976. [CrossRef]
47. Shojaeinasab, A.; Charter, T.; Jalayer, M.; Khadivi, M.; Ogunfowora, O.; Raiyani, N.; Yaghoubi, M.; Najjaran, H. Intelligent manufacturing execution systems: A systematic review. *J. Manuf. Syst.* **2022**, *62*, 503–522. [CrossRef]
48. Buttle, F.; Maklan, S. *Customer Relationship Management: Concepts and Technologies*; Routledge: London, UK, 2019.
49. Ikegwu, A.C.; Nweke, H.F.; Anikwe, C.V.; Alo, U.R.; Okonkwo, O.R. Big data analytics for data-driven industry: A review of data sources, tools, challenges, solutions, and research directions. *Clust. Comput.* **2022**, *25*, 3343–3387. [CrossRef]
50. Manjunath, T.; Pushpa, S.; Hegadi, R.S.; Ananya Hathwar, K. A study on big data engineering using cloud data warehouse. In *Data Engineering and Data Science: Concepts and Applications*; Wiley: Hoboken, NJ, USA, 2023; pp. 49–69.
51. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784. [CrossRef] [PubMed]
52. Karpov, O.E.; Pitsik, E.N.; Kurkin, S.A.; Maksimenko, V.A.; Gusev, A.V.; Shusharina, N.N.; Hramov, A.E. Analysis of publication activity and research trends in the field of ai medical applications: Network approach. *Int. J. Environ. Res. Public Health* **2023**, *20*, 5335. [CrossRef]
53. Khorev, V.; Kiselev, A.; Badarin, A.; Antipov, V.; Drapkina, O.; Kurkin, S.; Hramov, A. Review on the use of AI-based methods and tools for treating mental conditions and mental rehabilitation. *Eur. Phys. J. Spec. Top.* **2025**, *234*, 4139–4158. [CrossRef]
54. Steger, C.; Ulrich, M.; Wiedemann, C. *Machine Vision Algorithms and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
55. Davidson, G.P.; Ravindran, D.D. Technical review of apache flink for big data. *Int. J. Aquat. Sci.* **2021**, *12*, 3340–3346.
56. Kumar, R.; Agrawal, N. Analysis of multi-dimensional Industrial IoT (IIoT) data in Edge–Fog–Cloud based architectural frameworks: A survey on current state and research challenges. *J. Ind. Inf. Integr.* **2023**, *35*, 100504. [CrossRef]
57. Goknil, A.; Nguyen, P.; Sen, S.; Politaki, D.; Niavis, H.; Pedersen, K.J.; Suyuthi, A.; Anand, A.; Ziegenbein, A. A Systematic Review of Data Quality in CPS and IoT for Industry 4.0. *ACM Comput. Surv.* **2023**, *55*, 327. [CrossRef]
58. Rahimi, M.; Jafari Navimipour, N.; Hosseinzadeh, M.; Moattar, M.H.; Darwesh, A. Toward the efficient service selection approaches in cloud computing. *Kybernetes* **2022**, *51*, 1388–1412. [CrossRef]
59. Asim, M.; Wang, Y.; Wang, K.; Huang, P.Q. A review on computational intelligence techniques in cloud and edge computing. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 742–763. [CrossRef]
60. Jardine, A.K.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [CrossRef]
61. Kim, W.; Sung, M. Standalone OPC UA wrapper for industrial monitoring and control systems. *IEEE Access* **2018**, *6*, 36557–36570. [CrossRef]
62. Mowbray, M.; Vallerio, M.; Perez-Galvan, C.; Zhang, D.; Chanona, A.D.R.; Navarro-Brull, F.J. Industrial data science–a review of machine learning applications for chemical and process industries. *React. Chem. Eng.* **2022**, *7*, 1471–1509.
63. Frey, C.W. A hybrid unsupervised learning strategy for monitoring complex industrial manufacturing processes. In Proceedings of the IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society, Singapore, 16–19 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
64. Barto, A.G. Reinforcement learning: An introduction. by richard's sutton. *SIAM Rev.* **2021**, *6*, 423.
65. Panzer, M.; Bender, B. Deep reinforcement learning in production systems: A systematic literature review. *Int. J. Prod. Res.* **2022**, *60*, 4316–4341. [CrossRef]
66. Ramírez-Sanz, J.M.; Maestro-Prieto, J.A.; Arnaiz-González, Á.; Bustillo, A. Semi-supervised learning for industrial fault detection and diagnosis: A systemic review. *ISA Trans.* **2023**, *143*, 255–270. [CrossRef]
67. Hramov, A.E.; Koronovskii, A.A.; Kurovskaya, M.K.; Moskalenko, O.I. Synchronization of spectral components and its regularities in chaotic dynamical systems. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2005**, *71*, 056204. [CrossRef]
68. Juran, J.M.; Gryna, F.M. *Quality Planning and Analysis: From Product Development Through Usage*; McGraw-Hill: New York, NY, USA, 1970.

69. Eppler, M.J. *Managing Information Quality: Increasing the Value of Information in Knowledge-Intensive Products and Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
70. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
71. West, N.; Gries, J.; Brockmeier, C.; Göbel, J.C.; Deuse, J. Towards integrated data analysis quality: Criteria for the application of industrial data science. In Proceedings of the 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 10–12 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 131–138.
72. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*; SPSS Inc.: Chicago, IL, USA, 2000; Volume 9, pp. 1–73.
73. North, K.; Kumta, G. *Knowledge Management: Value Creation Through Organizational Learning*; Springer: Berlin/Heidelberg, Germany, 2018.
74. Wang, R.Y. A product perspective on total data quality management. *Commun. ACM* **1998**, *41*, 58–65. [CrossRef]
75. Zhang, W.; Yang, D.; Wang, H. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Syst. J.* **2019**, *13*, 2213–2227. [CrossRef]
76. Lei, Y.; Karimi, H.R.; Chen, X. A novel self-supervised deep LSTM network for industrial temperature prediction in aluminum processes application. *Neurocomputing* **2022**, *502*, 177–185. [CrossRef]
77. Wahid, A.; Breslin, J.G.; Intizar, M.A. Prediction of machine failure in industry 4.0: A hybrid CNN-LSTM framework. *Appl. Sci.* **2022**, *12*, 4221. [CrossRef]
78. Zhang, W.; Guo, W.; Liu, X.; Liu, Y.; Zhou, J.; Li, B.; Lu, Q.; Yang, S. LSTM-based analysis of industrial IoT equipment. *IEEE Access* **2018**, *6*, 23551–23560. [CrossRef]
79. Zhang, Z.; Wen, G.; Chen, S. Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding. *J. Manuf. Process.* **2019**, *45*, 208–216. [CrossRef]
80. Chen, Y.; Wang, J.; Wang, G. Intelligent welding defect detection model on improved r-cnn. *IETE J. Res.* **2023**, *69*, 9235–9244. [CrossRef]
81. Madhav, M.; Ambekar, S.S.; Hudnurkar, M. Weld defect detection with convolutional neural network: An application of deep learning. *Ann. Oper. Res.* **2025**, *350*, 579–602. [CrossRef]
82. Paolini, D.; Dini, P.; Soldaini, E.; Saponara, S. One-Class Anomaly Detection for Industrial Applications: A Comparative Survey and Experimental Study. *Computers* **2025**, *14*, 281. [CrossRef]
83. ismail Hossain, M.; Sanim, S.; Kamruzzaman, S.; Yesmin, M.; Sayem, M.S.; Shufian, A. Anomaly Detection in Industrial Machinery Using Machine Learning and Deep Learning Techniques with Vibration Data for Predictive Maintenance. In Proceedings of the 2025 2nd International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 9–11 February 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
84. Liso, A.; Cardellicchio, A.; Patruno, C.; Nitti, M.; Ardino, P.; Stella, E.; Renò, V. A review of deep learning-based anomaly detection strategies in industry 4.0 focused on application fields, sensing equipment, and algorithms. *IEEE Access* **2024**, *12*, 93911–93923. [CrossRef]
85. Mahadevan, S.; Theocharous, G. Optimizing Production Manufacturing using Reinforcement Learning. In Proceedings of the AAAI Eleventh International FLAIRS Conference, Sanibel Island, FL, USA, 18–20 May 1998.
86. Powell, K.M.; Machalek, D.; Quah, T. Real-time optimization using reinforcement learning. *Comput. Chem. Eng.* **2020**, *143*, 107077. [CrossRef]
87. Khdoudi, A.; Masrour, T.; El Hassani, I.; El Mazgualdi, C. A Deep-Reinforcement-Learning-Based Digital Twin for Manufacturing Process Optimization. *Systems* **2024**, *12*, 38. [CrossRef]
88. Rai, R.; Tiwari, M.K.; Ivanov, D.; Dolgui, A. Machine learning in manufacturing and industry 4.0 applications. *Int. J. Prod. Res.* **2021**, *59*, 4773–4778. [CrossRef]
89. Yildirim, P.; Birant, D.; Alpyildiz, T. Data mining and machine learning in textile industry. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1228. [CrossRef]
90. Salahshoor, K.; Kordestani, M.; Khoshro, M.S. Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers. *Energy* **2010**, *35*, 5472–5482. [CrossRef]
91. Łuczak, D.; Brock, S.; Siembab, K. Cloud Based Fault Diagnosis by Convolutional Neural Network as Time–Frequency RGB Image Recognition of Industrial Machine Vibration with Internet of Things Connectivity. *Sensors* **2023**, *23*, 3755. [CrossRef] [PubMed]
92. Tran, V.T.; Yang, B.S.; Oh, M.S.; Tan, A.C.C. Machine condition prognosis based on regression trees and one-step-ahead prediction. *Mech. Syst. Signal Process.* **2008**, *22*, 1179–1193. [CrossRef]
93. Li, G.; Chen, H.; Hu, Y.; Wang, J.; Guo, Y.; Liu, J.; Li, H.; Huang, R.; Lv, H.; Li, J. An improved decision tree-based fault diagnosis method for practical variable refrigerant flow system using virtual sensor-based fault indicators. *Appl. Therm. Eng.* **2018**, *129*, 1292–1303. [CrossRef]

94. Noura, H.N.; Allal, Z.; Salman, O.; Chahine, K. An optimized tree-based model with feature selection for efficient fault detection and diagnosis in diesel engine systems. *Results Eng.* **2025**, *27*, 106619. [CrossRef]
95. Thomas, M.C.; Zhu, W.; Romagnoli, J.A. Data mining and clustering in chemical process databases for monitoring and knowledge discovery. *J. Process Control* **2018**, *67*, 160–175. [CrossRef]
96. Bagherzade Ghazvini, M.; Sanchez-Marre, M.; Bahlio, E.; Angulo, C. Operational Modes Detection in Industrial Gas Turbines Using an Ensemble of Clustering Methods. *Sensors* **2021**, *21*, 8047. [CrossRef]
97. Chaudhry, M.; Shafi, I.; Mahnoor, M.; Ramírez Vargas, D.L.; Thompson, E.B.; Ashraf, I. A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. *Symmetry* **2023**, *15*, 1679. [CrossRef]
98. Hüsken, M.; Stagge, P. Recurrent neural networks for time series classification. *Neurocomputing* **2003**, *50*, 223–235. [CrossRef]
99. Butcher, J.B.; Verstraeten, D.; Schrauwen, B.; Day, C.R.; Haycock, P.W. Reservoir computing and extreme learning machines for non-linear time-series data analysis. *Neural Netw.* **2013**, *38*, 76–89. [CrossRef]
100. Chiuso, A.; Pillonetto, G. System identification: A machine learning perspective. *Annu. Rev. Control. Robot. Auton. Syst.* **2019**, *2*, 281–304. [CrossRef]
101. Javaid, M.; Haleem, A.; Suman, R. Digital twin applications toward industry 4.0: A review. *Cogn. Robot.* **2023**, *3*, 71–92. [CrossRef]
102. Soleimani, M.; Campean, F.; Neagu, D. Diagnostics and prognostics for complex systems: A review of methods and challenges. *Qual. Reliab. Eng. Int.* **2021**, *37*, 3746–3778. [CrossRef]
103. Cheng, S.; Quilodrán-Casas, C.; Ouala, S.; Farchi, A.; Liu, C.; Tandeo, P.; Fablet, R.; Lucor, D.; Iooss, B.; Brajard, J.; et al. Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1361–1387. [CrossRef]
104. Hramov, A.E.; Kulagin, N.; Pisarchik, A.N.; Andreev, A.V. Strong and weak prediction of stochastic dynamics using reservoir computing. *Chaos Interdiscip. J. Nonlinear Sci.* **2025**, *35*, 033140. [CrossRef]
105. Zheng, S.; Zhao, J. A self-adaptive temporal-spatial self-training algorithm for semisupervised fault diagnosis of industrial processes. *IEEE Trans. Ind. Inform.* **2022**, *18*, 6700–6711. [CrossRef]
106. He, Y.; Song, K.; Dong, H.; Yan, Y. Semi-supervised defect classification of steel surface based on multi-training and generative adversarial network. *Opt. Lasers Eng.* **2019**, *122*, 294–302. [CrossRef]
107. Jiang, L.; Ge, Z.; Song, Z. Semi-supervised fault classification based on dynamic Sparse Stacked auto-encoders model. *Chemom. Intell. Lab. Syst.* **2017**, *168*, 72–83. [CrossRef]
108. Razavi-Far, R.; Hallaji, E.; Farajzadeh-Zanjani, M.; Saif, M.; Kia, S.H.; Henao, H.; Capolino, G.A. Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems. *IEEE Trans. Ind. Electron.* **2019**, *66*, 6331–6342. [CrossRef]
109. Jia, X.; Tian, W.; Li, C.; Yang, X.; Luo, Z.; Wang, H. A dynamic active safe semi-supervised learning framework for fault identification in labeled expensive chemical processes. *Processes* **2020**, *8*, 105. [CrossRef]
110. Xu, M.; Wang, Y. An imbalanced fault diagnosis method for rolling bearing based on semi-supervised conditional generative adversarial network with spectral normalization. *IEEE Access* **2021**, *9*, 27736–27747. [CrossRef]
111. Liu, W.; Wang, P.; You, Y. Ensemble-based semi-supervised learning for milling chatter detection. *Machines* **2022**, *10*, 1013. [CrossRef]
112. Jiang, Y.; Yin, S.; Dong, J.; Kaynak, O. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE Sens. J.* **2020**, *21*, 12868–12881. [CrossRef]
113. Shokry, A.; Vicente, P.; Escudero, G.; Pérez-Moya, M.; Graells, M.; Espuña, A. Data-driven soft-sensors for online monitoring of batch processes with different initial conditions. *Comput. Chem. Eng.* **2018**, *118*, 159–179. [CrossRef]
114. Sun, Q.; Ge, Z. A survey on deep learning for data-driven soft sensors. *IEEE Trans. Ind. Inform.* **2021**, *17*, 5853–5866. [CrossRef]
115. Li, Z.; Andreev, A.; Hramov, A.; Blyuss, O.; Zaikin, A. Novel efficient reservoir computing methodologies for regular and irregular time series classification. *Nonlinear Dyn.* **2025**, *113*, 4045–4062. [CrossRef] [PubMed]
116. Curreri, F.; Patanè, L.; Xibilia, M.G. RNN-and LSTM-based soft sensors transferability for an industrial process. *Sensors* **2021**, *21*, 823. [CrossRef]
117. Yasin, A.; Pang, T.Y.; Cheng, C.T.; Miletic, M. A roadmap to integrate digital twins for small and medium-sized enterprises. *Appl. Sci.* **2021**, *11*, 9479. [CrossRef]
118. Burinskienė, A.; Nalivaikė, J. Digital and sustainable (twin) transformations: A case of SMEs in the European Union. *Sustainability* **2024**, *16*, 1533. [CrossRef]
119. Abolghasem, S.; Carpitella, S.; Mohan, G.T. Digital Twin Implementation in Small and Medium Size Enterprises: A Case Study. In *Analytics Modeling in Reliability and Machine Learning and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 321–341.
120. Pisarchik, A.N.; Maksimenko, V.A.; Andreev, A.V.; Frolov, N.S.; Makarov, V.V.; Zhuravlev, M.O.; Runnova, A.E.; Hramov, A.E. Coherent resonance in the distributed cortical network during sensory information processing. *Sci. Rep.* **2019**, *9*, 18325. [CrossRef]

121. Pisarchik, A.N.; Hramov, A.E. Coherence resonance in neural networks: Theory and experiments. *Phys. Rep.* **2023**, *1000*, 1–57. [CrossRef]
122. Peres, R.S.; Jia, X.; Lee, J.; Sun, K.; Colombo, A.W.; Barata, J. Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. *IEEE Access* **2020**, *8*, 220121–220139. [CrossRef]
123. Canese, L.; Cardarilli, G.C.; Di Nunzio, L.; Fazzolari, R.; Giardino, D.; Re, M.; Spanò, S. Multi-agent reinforcement learning: A review of challenges and applications. *Appl. Sci.* **2021**, *11*, 4948. [CrossRef]
124. Bahrpeyma, F.; Reichelt, D. A review of the applications of multi-agent reinforcement learning in smart factories. *Front. Robot. AI* **2022**, *9*, 1027340. [CrossRef]