applied sciences

MDPI

Article

Ensuring Data Accuracy, Completeness, and Interpretation in Advanced Manufacturing

Nathan Eskue [1,*] **and Amalia Macali** [2]

[1] AI in Manufacturing, Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands

[2] Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands; a.macali@tudelft.nl

* Correspondence: n.d.eskue@tudelft.nl

Abstract

Advanced manufacturing is undergoing a profound transformation, with data quickly becoming its most strategic asset. The industry is pushing toward Industry 4.0 with its sights already on the human-centric Industry 5.0. Manufacturing firms are rapidly integrating AI, IoT, and advanced analytics to enable real-time decision making, predictive maintenance, and full manufacturing lifecycle optimization. However, this data-driven revolution exposes a critical vulnerability: the hidden direct costs and cascading downstream consequences of inaccurate, missing, or corrupt data. This paper provides an in-depth examination of the data quality crisis facing modern manufacturing, exploring its quantifiable impact on cost, safety, and strategic decision making; and identifies the tangible barriers preventing scalable AI in manufacturing today. We investigate how bad data undermines the digital thread, erodes both operational and strategic trust, and stalls the transition to autonomous systems. Supported by recent industry surveys, academic findings, and leading trends, we reveal that most manufacturers suffer from systemic data quality issues, with billions lost annually to inefficiencies, rework, and flawed decisions. Addressing this, the paper evaluates state-of-the-art solutions for real-time data validation, anomaly detection, and predictive imputation. Building upon this, we identify key gaps—including the lack of unified data quality frameworks, integration across legacy/modern systems, and actionable imputation under uncertainty—and propose a roadmap to bridge them. The paper concludes by outlining four research directions that support a seamless, scalable transition toward a trustworthy data foundation in manufacturing. Industry 4.0/5.0 is defined by data, insight, and actionable intelligence: only manufacturers that tame their data chaos will thrive.

Keywords: data quality; Industry 4.0; Industry 5.0; digital thread; manufacturing AI; Industrial IoT (IIoT)

Check for updates

Academic Editor: Wilma Polini

Received: 21 January 2026
Revised: 18 February 2026
Accepted: 24 February 2026
Published: 2 March 2026

1. Introduction

Manufacturing, across industries and geographic regions, has shown an accelerated rate of data-driven sophistication [1] in its processes, workflow, factory optimization, and quality assurance using advanced systems such as the Internet of Things (IOT) [2], and with a growing emphasis on Artificial Intelligence (AI) [3]. While this rapid increase in using data has led to major innovations, it has also created a substantial and underappreciated risk to entire industries. Through the organic evolution driven by open markets, competitive

https://doi.org/10.3390/app16052409

behavior, and adoption of best-in-class practices, advanced manufacturing is in a race to reach Industry 4.0/5.0.

Industry 4.0 has been about making factories intelligent [4], connecting machines and entire workflows, harnessing data for unprecedented analysis, exponentially improving efficiency and flexibility as more processes are connected along the digital thread [5], and in many cases redefining the very business models that have been established over decades. Industry 5.0 is simply an extension of this data-driven evolution, expanding beyond the fully automated factory to merge with a newly focused, human-centric model. Industry 4.0 utilizes a complete system of saturated data extraction, connectivity, advanced analysis and decision making [6], leading to fully automated systems—in many ways removing the traditional roles of humans in manufacturing. Industry 5.0 evolves the system further by bringing back the role of humans, but in ways that benefit from human creativity, design, problem solving, and a focus on sustainability.

Using the context of advanced manufacturing rapidly striving toward Industry 4.0 [7], this paper explores the current state of data creation, collection, validation, and error correction. Specifically: *How can we establish robust data governance that maintains accurate and complete data in our manufacturing processes, given that Industry 4.0's data-driven systems require high quality insight to enable intelligent automation, predictive maintenance, and real-time optimization across the factory?* In answering this question, the paper works to better understand the context and weight of this issue. The cost of bad data was conceptualized as a specific estimate as far back as 1992, in what was called the "1:10:100 rule of data quality." [8] First presented by Geroge Labovitz and Yu Sang Chang, the "rule" (more of a helpful discussion point) stated that it costs a company $1 per record to prevent bad data, $10 per record to correct issues after the data is created, and $100 per record if the bad data is left unchecked. Although the estimates have been updated and debated, the premise creates a starting point for researching, identifying, fixing, and preventing bad data.

We will explore the acceleration and scope of data collection in manufacturing and better understand how data is and will need to be used to achieve Industry 4.0, the cost and consequences of bad/missing data in advanced manufacturing, and how other factors, specifically the phenomenon of growing knowledge gaps in manufacturing due to an aging workforce, can exacerbate the reliance on data to even maintain current processes. After examining the scope of this problem, the paper will discuss current efforts to mitigate the effects of bad/missing data.

After establishing the scale and scope of this data-driven challenge and understanding the current efforts for mitigation, the paper will address the larger issues and gaps that exist today. Specifically, it will address those issues/gaps that will prevent a successful transition into Industry 4.0. The authors provide two innovations to help resolve the research problem. First, a framework for how this problem must be addressed at the industry level in order to be reliable and fully scalable, while simultaneously operating at minimal overhead cost. Second, a novel process that produces effective, uncertainty-aware imputation—which is critical in not only improving the data streams that will be used by advanced manufacturing processes, but adds the high-value uncertainty context that is necessary to improve the human-in-the-loop effectiveness. The paper concludes with four proposed future research directions—each framed with a research question, objectives, and justification—aimed at filling critical gaps in ensuring data integrity required for Industry 4.0/5.0.

2. Data Proliferation in Advanced Manufacturing

Today's manufacturing centers are awash in data, although what seems to be a saturation of data sources continues to be redefined as new technologies emerge and provide novel, more focused, and higher definition data streams. IoT has been a major contributor

https://doi.org/10.3390/app16052409

to this trend [9], with a dramatic decrease in sensor cost combined with the industrialization of wireless, low power connectivity architectures [10]. Connected machines, a greater ability to stream high bandwidth data, and the decreasing costs to store and manage this data all contribute to a nearly exponential rise in data collection. This all, however, is a symptom of the underlying cause of increased data creation/collection. The reason behind this rise in data, collecting it wherever it can be found, is our growing ability to utilize it for decision making.

Advances in computing, most notably with AI but also in areas of modeling, simulation, and more robust traditional statistical analysis, have driven a digital revolution. Across industries, companies are discovering that data once considered irrelevant can now be analyzed to generate valuable insight. With the proper application of AI, this data collection/process/analysis cycle has the potential for real-time insight, which can improve the actionable intelligence provided to the human in the loop for rapid decision making. In cases that are time-critical with high confidence of the data insight, the human can be removed in favor of a self-contained autonomous decision making process.

The data collected, especially as more and varied sensor types are implemented, come in diverse forms [11]. Data can be structured streams (e.g., time-series sensor readings, equipment logs), but can also be unstructured or semi-structured data (e.g., machine vision images, maintenance notes). As AI becomes more prolific in manufacturing applications, particularly those on the edge, the AI insights themselves become data streams that flow into more system-wide analyses, helping to optimize or make critical decisions across a factory. This data being collected, especially with the many potential types of data, can create new strains on systems that were not designed to process so much information. A single production line in an advanced manufacturing plant could have thousands of sensors continuously providing data such as temperatures, pressures, speeds, vibrations and more, each updating in milliseconds. In parallel, the same production line could employ a high resolution video feed, infrared cameras, complex vibrometer readings, and more. All of this data must be reliably captured, aligned with the same timestamp, collected and stored in such a way that is labeled for rapid analysis, and must be protected from loss, damage, theft, or sabotage.

These massive data feeds exist because we are developing more and more ways to use the data. For example, the production line mentioned above could use these data streams to improve the quality and speed or reduce the cost of critical smart manufacturing techniques [12]. This type of manufacturing process is likely monitored and analyzed to ensure it is operating as intended, whether with AI or other data-driven tools. However, with real-time insight [13] and embedded controls, an advanced AI could evolve process monitoring to automated process control, predicting necessary tweaks in the process to ensure a “first time right” result for the material produced. The data can be used to ensure the machines themselves are operating at an optimum level, using predictive maintenance to deliver a “just in time” approach to repairs and maintenance in a way that minimizes cost while managing the risk of unplanned downtime. The data could be used to improve how the material is processed, ensuring that the quality standards are met [14]. Finally, this data can be connected to other processes along the production line and factory itself, creating a true digital thread that paves the way for data-driven optimization at the factory level [15]. To gain this level of advanced simulation and optimization [16], the data necessary becomes exponentially larger, as does the storage and processing power required to produce these insights.

The insights we are able to now gain from these data streams have the capability to vastly improve efficiency, reduce costs, improve quality, and optimize the entire workflow. As analytical applications are continuously developed, companies are mining their current processes and machines to extract data that already exists, but was not collected because it did not have a purpose [17]. Called “dark data [18,19],” this is one more element pointing to the rise in data volumes being collected, and the challenges faced by the wide variation in data type and quality. While dark data can be difficult to retrieve not just due to physical connection access and the reliability of aging sensors, there are novel approaches that can utilize AI on legacy data that is even encrypted for IP protection. In the healthcare system, a parallel application is Homomorphic Encryption (HE), which is used to gain insights on privacy-protected data without decryption, both protecting privacy while still gaining process insights [20].

3. From Volume to Criticality: Data as a Strategic Asset

As more and more tools have evolved to enable advanced insights and decision making, manufactures have placed more value on the potential data that can be collected [21]. Combined with the abundance of sensors and data handling infrastructure, the issue is no longer a lack of data. Instead, the challenge for manufacturers is now the ability to convert massive quantities of raw data into *trusted and usable data* [22]. Raw data by itself has no value. Creating and collecting this raw data has a growing cost. The paradigm change in data-driven insights across the industry can encourage the blind collection of as much data as possible, but if this data is siloed and ungoverned, it creates both cost and risk without gaining any return.

Industry 4.0 requires that companies avoid a “data rich, information poor” paradox by investing heavily in the full data process. To better align this focus, leading manufacturers now regard high quality data as a strategic asset on par with other critical business elements including physical inventory and capital investments [23]. By implementing a focused governing data quality policy ensuring accuracy, consistency, and context-rich information [24], there is significant investment toward the conversion of mass raw data into actionable intelligence [25].

An additional consequence of this newfound treatment of data is that its use is no longer confined to operational efficiency, but is key to strategic decision making [26]. As the digital thread is further connected and fortified, these data-driven insights can not only provide insight to the breadth of the manufacturing process, but they can look forward and backward in time to gain a better understanding of past events, and provide invaluable insight as leaders make decisions on product designs, new markets, technology investments, and supply chain reinforcement.

Original Equipment Manufacturers (OEMs) equipped with a robust digital thread of connected data, especially if using AI-based insights [27], can even move beyond the manufacturing process in their strategic decision making. Aligning manufacturing data and insights with field performance data (sensors on the products themselves), the OEM can create new insights that inform design changes and process optimization [28]. This brings the level of strategic value even higher as it closes the lifecycle loop to create a positive feedback loop of knowledge [29], driving the company closer toward not just Industry 4.0, but a sustainable-focused Industry 5.0 [30]. That said, OEMs must also take caution as they study manufacturing data. They have a responsibility to not expose (intentionally or unintentionally) proprietary data, processes, or insights from manufacturers, which could potentially occur if they adjust their products to include these insights without proper obfuscation. In addition, they must also conduct the extensive experimentation and validation needed when taking specific insights and generalizing them for a product.

Two key forces emerge from these insights. The first is that as data as a resource has become more mission-critical not just for manufacturing but for long-term strategy, companies are relying on data more while rapidly reducing tolerance for bad data. While this tolerance has dropped, it is not always clear what actions can be taken to ensure data that is of a guaranteed high quality. The second force that has emerged is the need for all manufacturers, big and small, no matter the industry, to utilize their data for both operational and strategic decision making. If companies do not evolve quickly toward Industry 4.0, they will soon lose their competitive edge. Even large companies could face extinction from more nimble companies that utilize their data-driven insights well, reducing costs, improving quality, reducing cycle time, and even closing the loop to directly improve product design.

4. The High Stakes of Data Dropouts and Errors

Having established the dramatic increase in data within manufacturing, along with the significant value placed on data for both operational and strategic decision making, it is important to understand the consequences of using data that is incomplete, corrupted, or misinterpreted.

4.1. Pervasive Data Quality Issues

Data quality issues are extremely common among manufacturers. A 2024 survey [31] of 500 manufacturing leaders—representing industry verticals including electronics, automotive, healthcare, energy, aerospace, and defense—revealed that 98% of manufactures are struggling with data issues in their operations. The key issues respondents reported were significant data streams that had incomplete/missing data; data that was outdated and could not be used for relevant decision making; and data that was proven to be inaccurate, creating either bad decision making or an erosion of trust with all data. A total of 40% of respondents believed that their companies are falling behind competitively specifically because they are failing to adopt automation, for which high quality data is critical. In summary, the survey indicates that reaching Industry 4.0, or even advanced manufacturing capability, is greatly hindered by systemic data issues, and that this problem could contribute directly to companies losing any competitive advantage.

4.2. Decision Making Gaps

Another survey of 750 business and tech leaders [32] showed a concerning gap in the ability to rely upon data for decision making. A total of 58% of leaders stated that *most or all* of their key business decisions are based on inaccurate or inconsistent data. The implications of this are clear: strategic decisions, investments, and attempts at optimization are made using intelligence that is either incomplete or potentially in conflict with ground truth. A total of 61% of respondents said that no one in their organization fully comprehends the data used and does not know how to access it. A full 51% believe their organizations must significantly increase current data management investments in order to achieve business goals. The results from this survey provide strong implications that the manufacturing industry will not achieve Industry 4.0/5.0 without a foundational change in how data is captured, managed, and used. In fact, the gulf between companies with strong data practices and those without will grow rapidly in the near future, because only those companies with high quality data can fully implement advanced tools such as AI and the digital thread in their organizations [33].

4.3. Cost of Poor Data Quality

It is difficult to account for all the direct and indirect results from poor data quality, although an industry study [34] estimated that, on average, bad data costs organizations $12.9 million per year. This estimate factors elements such as overproduction, higher defect rates, delays in production and shipments, compliance penalty risk, missed quality standards, risk of product recalls, and an overall loss in the ability to compete effectively.

Another report [35] estimated that scrap and rework typically cost manufacturers up to 2.2% of annual revenue on average, with even more costs that cascade due to a reduction in OEE (Overall Equipment Effectiveness). A different method to estimate the cost of poor data [36] reviewed the data records and their downstream effects, using industry data to estimate a cost of approximately $10 per bad data record to correct, but $100 per bad data record if left unchecked. This illustrates the order of magnitude higher costs due to cascading problems, such as process inefficiencies, manual workarounds, and bad decisions. Companies that allow bad data put themselves at a quantifiable disadvantage, creating major inefficiency in the form of cost increases and delays, along with larger market risks including poor quality, customer dissatisfaction, and regulatory penalties.

4.4. Strategic Cascading Effects

In addition to compounding downstream costs, bad data can prevent a company from taking advantage of advances in AI [37] and other advanced analytical tools. A McKinsey report [38] indicates that common data issues such as missing sensor data, a poor data architecture, and miscalibrated data collection components can destroy any benefit from AI or other continuous improvement efforts. Case studies showed that many high-value, high-potential AI projects that had been vetted for strong Return on Investment (ROI) were delayed for years (some permanently) due to poor data. If this data is not meticulously controlled, ongoing projects can result in not just ineffective, but harmful results [39]. Finally, these cascading effects due to bad data can impact entire production lines due to false alarms, the masking of actual equipment failures, and misled decisions that amplify risk to the operation and potential safety of personnel [40].

5. Demographic Shifts and the Knowledge Gap

The importance of high quality data has been demonstrated by examining the cost of poor quality data, cascading operational costs, misinformed decision making, and the prevention of leveraging strategic tools such as AI. However, there is another global force that illustrates the importance of high quality data (and the consequences of bad data). A 2017 systematic review of aging worker populations showed that in the US, 18.6% of the adult working population was 65 years or older, with the percentage expecting to increase 0.6% yearly over the next decade [41]. At the same time, the COVID-19 pandemic provided a unique window into what this knowledge loss might look like over time, with significant critical knowledge missing as experienced workers were disproportionately absent [42]. While there have been a growing number of efforts in this area to either capture this critical knowledge [43,44], transfer it to younger workers [45], or even use tools like generative AI to help supplement the gaps [46], these solutions cannot be effective without good data. Moreover, with a growing risk of a smaller experienced workforce, companies are already forced to do more with less, working “smarter, not harder” by better utilizing data as actionable intelligence [47]. These data-driven insights must not only be able to provide workers with accurate information, but they must also be of a high enough quality to drive interpretability and trustworthiness to aid workers who have less experience and intuition. With fewer people—and a much lower experience level on average—expected to maintain the same efficiency, quality, and strong decision making, manufacturers must implement force multiplier tools such as automation, AI, and real-time insights [48]. This process does not happen overnight, and it is clear that investments for data quality improvements must be made immediately so that advanced data-driven tools can be effectively incorporated into manufacturing operations.

https://doi.org/10.3390/app16052409

6. Ensuring Data Integrity: State-of-the-Art Methods and Research

Addressing the significant challenges outlined in the previous sections requires a multifaceted approach to improving manufacturing data. The key processes required are those that accurately detect, correct, and prevent data quality issues in real time. A growing amount of academic research is focusing on improving data fidelity for manufacturing data generally, and IoT data in particular. This section surveys current state-of-the-art methods for (a) detecting data dropouts and anomalies in real time, (b) identifying and handling corrupted data, and (c) predicting or imputing missing values to maintain continuous data streams.

6.1. Detecting Data Dropouts and Anomalies in Real-Time

Real-time monitoring and profiling of data streams is foundational to identifying data issues as they occur. One foundational strategy is to perform this data validation/profiling at the source, preferably on an edge device that is physically located near the sensors and other data stream origins. Effective, reliable, and automated methods are required to ensure that data dropouts are caught immediately (within milliseconds) [49]. When anomaly detection is also used, data can be checked against its expected ranges, patterns, or continuity. Another family of methods developed uses either redundant sensing or multi-sensor fusion to identify data dropouts [50,51]. Multiple sensors can measure identical, related, or offset parameters in order to compare values and identify any significant differences. Other techniques, depending on the application, can compare sensor data to physics-based measurements, which allows maximum independence between the measurements [52]. This can help if sensor data drops out completely, or if there is a sensor error causing a significantly different output. The sensor fusion data can even be used to create a virtual sensor [53] in order to create a more independent calculation that should align with the other measurements and alert to missing/anomalous data if not. This architecture has been expanded beyond an individual sensor feed and into a larger system [54], creating a digital twin that can measure not only individual feeds but the complex expected interactions between subsystems, data feeds, and physics-based simulations.

As with many analysis-driven problems, AI methodologies are being used to innovate the ability to not only detect dropouts [55], but to find anomalies. As an example, autoencoders and generative models can learn an internal representation of normal sensor behavior; when new data is passed through the model, a high reconstruction error can indicate an anomaly. Likewise, other AI models such as deep feedforward neural networks, convolutional neural networks (CNNs), and hybrid models can successfully detect dropouts and anomalies [56–58]. In order to first train an AI model for this purpose, the use of data quality assessment (DQA) [49,59] is a required step to ensure the data to be used for training is high quality.

State-of-the-art practices for dropout and anomaly detection have increased, and involve techniques such as combining multi-layered monitoring with a mix of rules-based and AI-based methods. The purpose of these tools is to catch issues immediately, before they affect production. If a dropout or anomaly is detected, the system can trigger mitigating actions (alarm if there is a human in the loop), switching to backup/redundant components, and isolating/flagging the suspected data so it will not be used in real-time analysis (or downstream decisions).

Despite the progress in this area, significant challenges remain in balancing false alarms with missed errors, efficiently adapting high quality models to new contexts, and creating intelligent mitigation systems that are reliable when used in real time. Research continues to refine these detectors for greater robustness and generality, as discussed further in the Section 7.

6.2. Identifying Corrupted or Erroneous Data

Although data dropout/anomaly detection has its own challenges, the ability to detect data that has been corrupted can be significantly more difficult. Sensor data noise, biases, drift, or even tampering from a bad actor can introduce errors that are not simply missing or significantly off, but are still incorrect. Identifying data quality issues that are much more subtle than dropout/anomalies [60] is an active area of research; it has seen significant improvement with the evolution of AI-based techniques, but large gaps in generally effective practices still remain. Common data errors in sensor data feeds can include outliers, drifting measurements over time, values stuck at a constant reading, and inconsistent readings across duplicate sensors [61,62]. Establishing data quality dimensions and metrics to systematically evaluate data feeds is critical, with methods assessing attributes such as accuracy, precision, bias, consistency, and timeliness. Accuracy, for example, could be assessed if a ground truth measurement is available for comparison. Various context-based logical comparisons could be established (e.g., measuring how much product has been placed in a box and comparing it to the maximum volume possible). In these cases, however, there is a clear weakness: while these can be effectively implemented for an individual process, these methods are unique to the process and cannot be generalized [63].

The use of Statistical Process Control (SPC) [64] on the data itself is an interesting and often effective method to identify subtle errors. Manufacturers use SPC charts to monitor the quality of products/processes, and this same technique can be used to measure the data stream itself, treating it as its own process. Control charts tuned to the expected parameters of the sensor can alert to anomalies, bias, or drift. While mentioned in the previous section, the tools used to identify anomalies have potential overlap with the process of identifying more subtle differences.

Although tools that effectively measure subtle errors will likely also flag large anomalies, an anomaly detection method will rarely be effective for subtle errors.

AI again is making strong improvements in this area, using a wide range of pattern recognition models and algorithms [65] that can effectively spot when data “just looks wrong” compared to historical patterns. To aid with generalizing AI models and making them easier to use on a variety of processes, a promising trend in research is to develop robust anomaly classifiers that can distinguish between process anomalies and data anomalies, and to combine domain knowledge with data-driven models for this. One approach is hybrid modeling: using physical models to set bounds and ML models to detect subtle patterns within those bounds.

A growing concern for Industry 4.0 is that as the use of data-driven decision making becomes more valuable and powerful, it alters the profile of risk management for the organization. Many of the tools mentioned are meant to handle incidental data errors, but this heavy reliance on data can also create a major security risk for the organization. Specifically, a bad actor could perform a “false data injection” attack [66–68], introducing specific data into a feed that will predictably cause the process to fail. This can take the form of introducing false positives (resulting in line down alerts, wasted time and material, etc.), but could also introduce more subtle altered data that causes a process to shift without tripping an error alert, causing the process to potentially run the material in a production process, create issues but cause the quality control to miss them, or even adjust the process so that it eventually damages the manufacturing machinery.

The risk of false data injection attacks should be especially concerning to manufacturers, simply because we have not evolved enough toward Industry 4.0 for this to be on our threat radar. Leaders of manufacturers are well equipped to handle established threats to security, but as the role of data shifts to a central position, it will gradually become a company’s most valuable asset, as well as its biggest risk. There has been progress made to both protect data from theft through many different cryptographic strategies, but also to ensure that data has not been modified since it was created at the sensor level. Blockchain has been a novel consideration for this challenge [69], providing a hash of a data package to place on-chain, ensuring that the data's authenticity can be validated indefinitely by the owner. This subject holds significant potential within Manufacturing 4.0, and will be the focus of additional research by the authors in the near future.

6.3. Predicting and Imputing Missing Data to Preserve Completeness

Despite the use of advanced tools aimed at prevention, some data loss in manufacturing is inevitable. Because of this, a critical capability is the ability to predict or impute missing values (and to do so in real time), such that operations can continue with minimal disruption. Data imputation [70–72] is well established in the field of statistics, but the high-frequency, multivariate, time-sensitive nature of manufacturing data demands advanced approaches. Traditional imputation methods such as mean substitution, forward-fill, or linear interpolation can be effective for simple sensor measurements and have the added benefit of simplicity and speed [73]. These methods can work adequately for short gaps or data that changes slowly. However, these methods quickly lose effectiveness when dealing with complex sensor feeds, and are especially ill equipped to handle multivariate context. The multivariate relationships between data feeds can act as a major strength, however, and tools suited to manage these complex connections can effectively predict what the missing value should be.

State-of-the-art solutions leverage machine learning and deep learning for imputation [74,75]. Of particular effectiveness are those models that treat imputation as a prediction problem. Regression or time-series forecasting [76,77] models work especially well for many data types. More recently, deep learning models have shown impressive results in accurately reconstructing missing data from complex sensor data streams. One such example is the application of self-attention-based deep learning for time-series imputation [78].

Multi-variate imputation [79] is especially important for ensuring high quality data in manufacturing. Instead of imputing each variable independently, these models leverage correlations between variables. For example, if a pressure reading is missing, this type of model can examine related variables such as temperatures or flow rates to infer what the pressure likely was.

In process manufacturing applications where certain sensors might report only occasionally, methods like matrix completion or tensor factorization [80,81] have been applied to infer the likely data value in between readings. For this technique, a machine's dataset is treated as a matrix (sensors × time). Algorithms fill the missing data values by finding low-rank structures.

7. Research Trends, Gaps, and Implications

This discussion summarizes key trends discovered in the extensive survey of this field, highlighting areas of promise for manufacturers. We will then identify particular areas of concern where significant gaps still exist in our ability to ensure high quality data at scale. Finally, we will consolidate these findings into a framework that must be achieved in the manufacturing industry in order to reach Industry 4.0, which can fully utilize data to its maximum potential, building a strong digital thread and a foundation of AI-driven insights at all levels of the manufacturing cycle.

https://doi.org/10.3390/app16052409

7.1. Trends

7.1.1. Growing Research Emphasis on Data Quality

In the last five years, there has been an exponential increase in publications that explore data quality issues in manufacturing. While this was more of a niche focus for manufacturing operations, the subject is now considered a foundational component of smart manufacturing and Industry 4.0 across all industries.

7.1.2. Shift from Reactive to Proactive Data Management

Traditional quality control has typically taken a reactive approach, largely due to the limitations in real-time data analysis and actionable intelligence. However, with the increase in AI and edge computing, there is a growing capability toward proactive data quality management [82]. Because of the clear benefits of proactive data management, a growing number of manufacturers are adopting governance frameworks and automated systems to continually ensure data integrity throughout the data lifecycle [83]. There are a variety of approaches within this effort, from integrating quality checks at data generation, using governance architecture to remove data silo effects, and investing time, money, and expert personnel into data health. The investment is more easily justified due to increasing evidence of a clear ROI for proactive data management.

7.1.3. AI and Hybrid Techniques at the Forefront

A significant trend is the growth of innovative methods that blend the domain knowledge of a manufacturing process with AI techniques to handle data issues. Hybrid physics/ML approaches for anomaly detection are especially useful as the physics models are well established, and using ML with a ground truth ballast ensures more accurate and reliable results [84]. AI models for imputation as well as data augmentation are rapidly improving and increasing in variety as more attention is focused on this area [85].

7.1.4. Real-Time and Edge Computing Focus

The drive for real-time responsiveness has encouraged more and more solutions to be implemented on the edge [86]. With edge device capability increasing to handle more and more complex computational tasks (including AI), there is a continuously improving trend toward real-time data validation, compression, and even local ML inference on or near the machines [87]. This improves latency, bandwidth usage, and reliability.

7.2. Gaps and Challenges

7.2.1. Unified Data Quality Frameworks

Manufacturing companies often lack a holistic framework that covers data quality expectation across all levels. Standards can provide general data quality guidelines [88], but as of yet there is no widely adopted, manufacturing-specific standard that integrates with production systems. The focus of research projects tends to limit its scope to a narrow slice (sensors, or databases, or specific processes). However, a unified framework [89] that provides guidelines on measuring, monitoring, and improving data quality continuously across the manufacturing process is still an open challenge. Even having common definitions of what constitutes data redundancy, how to quantify trust in data, etc., have not been settled across the manufacturing industry. Developing this framework could help the industry benchmark and certify its data processes (similar to how ISO 9001 [90] certifies quality management systems).

7.2.2. Interoperability and Heterogeneity

For the manufacturing industry, there is a wide range of equipment and tooling that combines modern systems with legacy hardware. Ensuring data accuracy and consistency across this range of heterogeneity is difficult [91]. Moreover, the integration of legacy systems into a data quality standard [92] remains a wide a gap, and highlights the need for a robust, flexible, and wide-reaching framework that can accommodate AI and century-old sensors alike. Techniques to retrofit or infer data quality from limited outputs are needed, such as guidelines for supplementing legacy systems with modern, external sensors to act as reliable proxy data collection. Data format and semantic inconsistencies are another challenge to be solved [93], with a vast array of machinery that spans not only the globe, but many decades. Different machines label and structure data differently, which will lead to misinterpretation if not aligned.

7.2.3. Real-Time Assurance vs. Volume

Maintaining high data quality in real time as data scales up is non-trivial. This problem is exponential as more and more sensors are used, but the relationships between data streams are also analyzed [94]. Even for single data streams that require large amounts of data, that data might be generated faster than it can be properly analyzed. There is currently a gap in ultra-fast data validation that is important today for these large data processes, but will continue to be critical as data collection continues to scale up [95]. Solving this might require novel hardware solutions or smarter data reduction that does not lose critical information. The balance between data volume and quality is tricky, because collecting less data can make quality management easier, but might forfeit insight [96]. On the other extreme, collecting all data creates a "needle in the haystack" problem for quality monitoring. There are promising techniques like selective sampling or adaptive fidelity (variable data rates based on process state) that could help alleviate this gap, but further research is needed.

7.2.4. Human Factors and Data Culture

Technology alone is not enough to solve the larger problem. Adding to this gap is the organizational and human aspects of data quality. Across manufacturing fields, many operators and engineers are not yet fully trained in data literacy [97], and as a result they might not know how to interpret a data quality metric, or understand what to do when a data anomaly alert appears. As mentioned earlier, this is exacerbated by the growing number of older workers retiring, taking that intuition through experience with them. Developing intuitive tools and training for the workforce to fully understand data quality is crucial, and remains a large gap in both research and industry [98]. Moreover, there is a strong need to establish a "data quality culture" where everyone (from machine operators to C-level executives) will value and promote data integrity [99]. This is a major culture shift for not just one generation of workers, but most/all the generations working together in a manufacturing organization.

7.2.5. Missing Data Prediction Confidence and Use Cases

While there has been progress in advanced imputation, a gap remains in quantifying confidence and in deciding what automated actions can be taken on imputed data. If a system predicts a critical value in a high-impact process, should the control system act on it when the consequences of a false positive or negative could be devastating? Under what conditions is that acceptable? Research must explore frameworks for decision making under data uncertainty, with the ability to create a consistent model that acts predictably, yet is purposely flexible to ensure that the risk management element is applied proportionately to the severity of the consequences [100].

7.3. Illustrative Example: Output Data Dropout

This section presents both a controlled illustrative example and a realistic industrial case study to evaluate the proposed uncertainty-aware framework for missing output reconstruction. The objective is twofold: first, to analyze the behavior of deterministic and probabilistic models in a transparent synthetic environment; second, to validate their applicability in a complex manufacturing-like process characterized by nonlinear dynamics and strong multivariate dependencies.

Rather than treating missing data as a mere preprocessing inconvenience, this study frames output dropout as a critical data quality issue capable of compromising downstream analytics, digital twins, and AI-driven decision systems. Reconstruction methods must therefore provide not only accurate estimates but also a quantification of prediction reliability.

7.3.1. Synthetic Benchmark: Extended Ishigami Dataset

A synthetic multivariate dataset derived from an extended version of the Ishigami function [101] was employed to create a controlled validation environment. The Ishigami function (Y_1) is widely adopted in uncertainty quantification due to its strong nonlinearity, non-monotonic response, and significant interaction effects among variables.

Five independent inputs ($\mathrm{X}_1, \ldots, \mathrm{X}_5$) were sampled uniformly in the interval $[-\pi, \pi]$, generating three outputs:

$$\mathrm{Y}_1 = \sin(\mathrm{X}_1) + \alpha \sin^2(\mathrm{X}_2) \; + \; b\, X_3^4 \sin(\mathrm{X}_1)$$

$$\mathrm{Y}_2 = \sin(\mathrm{X}_4)\cos(\mathrm{X}_5) + 0.5\mathrm{X}_1$$

$$\mathrm{Y}_3 = \exp(-X_2^2) \; + \mathrm{X}_3 \sin(\mathrm{X}_4)$$

where ($\alpha = 7$) and ($b = 0.1$). A total of 1000 samples were generated. Y1 follows the canonical Ishigami function, while Y2 and Y3 are additional nonlinear outputs introduced to form a multi-output benchmark. Although synthetic, the dataset reproduces key properties commonly observed in industrial sensing systems, including nonlinear relationships, heterogeneous outputs, and cross-variable dependencies.

To emulate realistic data degradation, 20% of the output values were randomly removed, mimicking scenarios such as temporary sensor failures, communication losses, or synchronization issues that disrupt the digital thread.

Two models were evaluated:

- A deterministic artificial neural network (ANN);
- A probabilistic Deep Kernel Learning (DKL) model combining neural feature extraction with Gaussian Process regression.

A schematic comparison between the deterministic ANN and the DKL architecture is illustrated in Figure 1.

7.3.2. Model Implementation and Training Configuration

The architecture and training configuration of the ANN and DKL models are summarized in Table 1. The models were implemented in PyTorch (Python 3.13.2). For the ANN model, a fully connected feedforward architecture was adopted with four linear layers, input → 32 → 64 → 64 → output, using ReLU activations after each hidden layer.

The ANN was trained using the Adam optimizer (learning rate = 0.01) for 500 epochs, minimizing the Mean Squared Error (MSE). To simulate output data dropout, 10% of the target values were randomly masked. Only samples with complete targets were used for training, while samples with missing targets were used for reconstruction evaluation. Input and output variables were standardized using zero-mean, unit-variance normalization (StandardScaler), and an 80/20 train–test split was adopted. For both datasets, the hyperparameters were selected based on validation performance using a hold-out validation set.

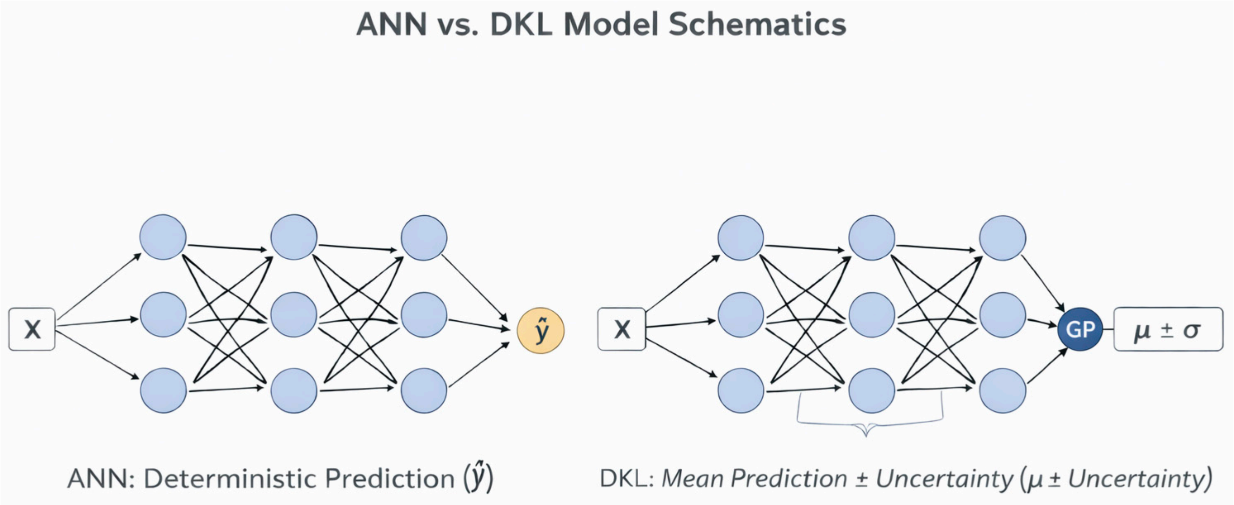
ANN vs. DKL Model Schematics
X
ŷ
X
GP
μ ± σ
ANN: Deterministic Prediction (ŷ)
DKL: *Mean Prediction ± Uncertainty* (*μ ± Uncertainty*)

Figure 1. Schematic comparison between the deterministic ANN architecture and the Deep Kernel Learning (DKL) model. The ANN produces a point estimate ($\hat{y}$), while the DKL combines the same neural architecture with a Gaussian Process (GP) head to provide mean prediction and predictive uncertainty ($\mu \pm \sigma$).

Table 1. ANN and DKL architecture and training configuration adopted for the Ishigami dataset.

Parameter	ANN	DKL
Architecture	32-64-64	64-64 + GP
Optimizer	Adam	Adam
Learning Rate	0.01	3×10^{-2}
Epochs	500	500
Kernel	-	Matérn 1.5
Inducing	-	200
Mask Rate	10%	10%

While the ANN produces fast point predictions (Figure 2), the DKL framework (Figure 3) additionally provides predictive uncertainty, enabling reliability-aware reconstruction.

Both models successfully identify abnormal outputs; however, the probabilistic formulation consistently achieves higher reconstruction accuracy while simultaneously quantifying prediction confidence. This capability transforms imputation from a purely numerical task into a reliability-aware decision mechanism.

7.3.3. Industrial Case Study: Tennessee Eastman Process

While the synthetic benchmark provides a controlled validation environment, the industrial case study demonstrates the practical applicability of the proposed framework under realistic operating conditions. For the Extended Tennessee Eastman Process (TEP) dataset, the same reconstruction protocol was adopted. Input variables were standardized to zero mean and unit variance, and an 80/20 train–test split was applied. The variable xmeas_17 was selected as reconstruction target. Output data dropout was emulated by randomly masking 10% of the target values in the evaluation set. The DKL model was implemented in GPyTorch (Python 3.13.2) using a neural feature extractor followed by a sparse variational Gaussian Process head. The GP head employed a Matérn kernel and Gaussian likelihood, and variational inference was optimized via the ELBO objective using inducing points. The specific DKL hyperparameters (learning rate, epochs, and inducing point count) are reported in Table 2 with the implementation code used for the TEP experiments.

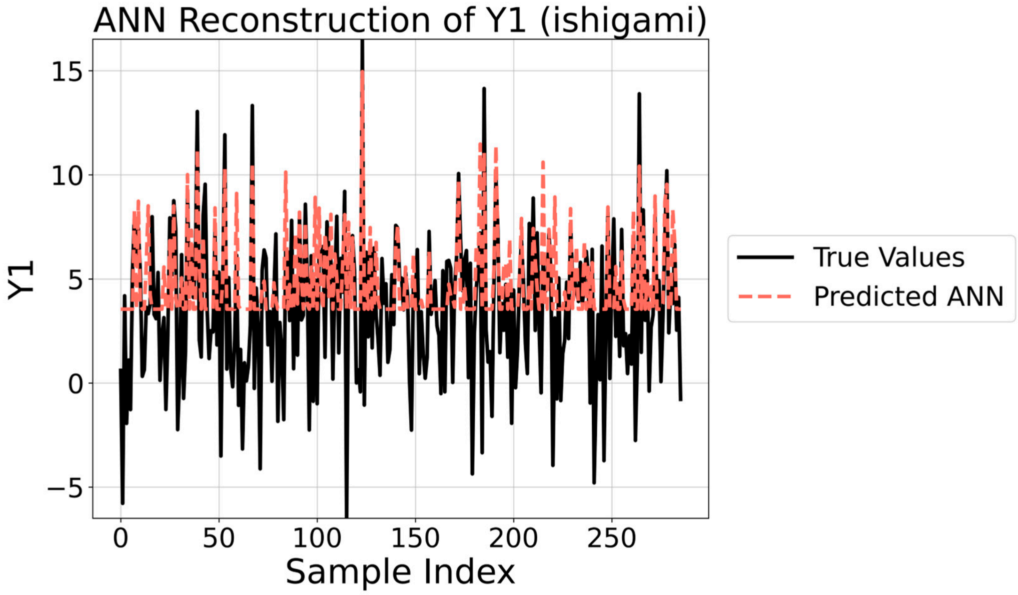

Figure 2. Reconstruction of Y1 (canonical Ishigami function output) using the deterministic ANN model. The nonlinear trend is captured; however, the model provides only a point prediction without uncertainty estimation.

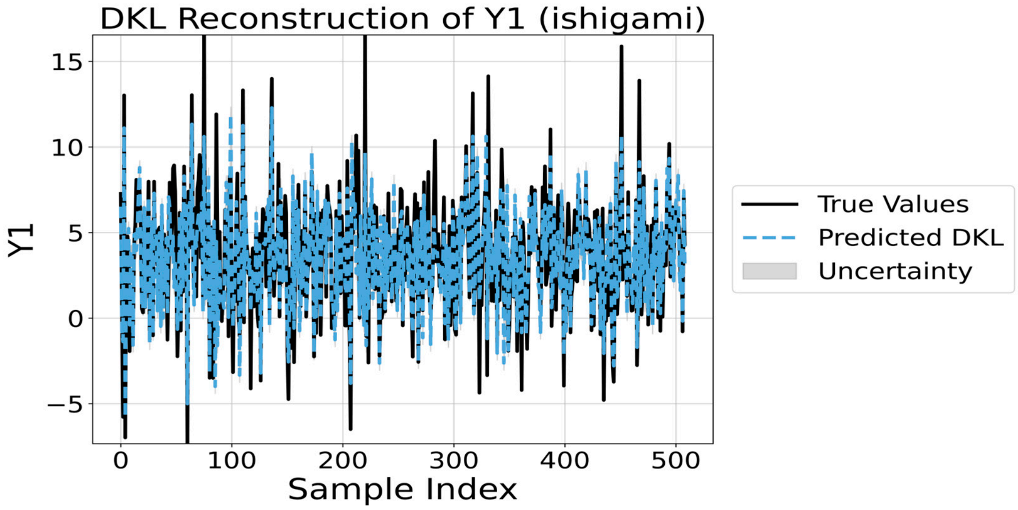

Figure 3. Probabilistic reconstruction of Y1 (canonical Ishigami function output) using the DKL model. The shaded region represents the predictive uncertainty (μ ± σ).

The extended Tennessee Eastman Process (TEP) dataset introduced by Reinartz et al. [102] was adopted as an industrial benchmark. The TEP simulates a complex chemical production plant characterized by nonlinear dynamics, strong cross-variable correlations, and multiple operating regimes. Each observation contains 55 process variables, closely resembling modern sensorized manufacturing environments where reliable data streams are essential for automated decision making.

Table 2. ANN and DKL configuration and training parameters used for the output reconstruction experiments.

Parameter	ANN	DKL
Architecture	128-64-1	64-32 + GP
Optimizer	Adam	Adam
Learning Rate	1×10^{-3}	3×10^{-3}
Epochs	400	300
Kernel	-	Matérn 1.5
Inducing	-	400
Mask Rate	10%	10%

The process variable *xmeas_17* was selected as the reconstruction target due to its structured temporal behavior and measurable dependency on other process variables. To emulate realistic data quality degradation, 20% of the target values were randomly removed, reflecting practical scenarios such as sensor outages, communication failures, or temporary acquisition faults.

Figure 4 illustrates the reconstruction obtained using the deterministic ANN model. The network successfully captures the global trajectory of the sensor signal; however, visible deviations remain, particularly in regions characterized by sharper transitions. Quantitatively, the ANN achieves an RMSE of approximately 0.27 and an R^2 score close to 0.83, indicating good yet imperfect reconstruction capability. These residual errors highlight a fundamental limitation of deterministic models, which provide point estimates without expressing prediction confidence.

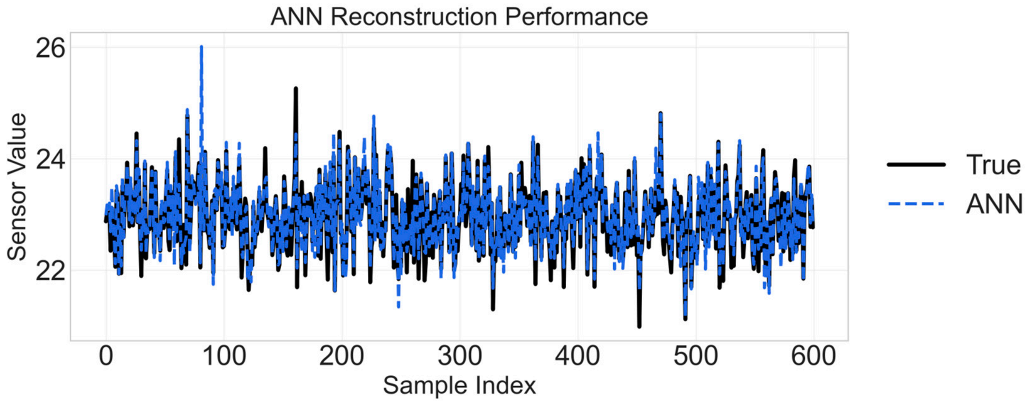

Figure 4. Reconstruction of the xmeas_17 sensor using the deterministic ANN model. While the global trend is captured, noticeable deviations from the ground truth remain in regions of higher signal variability.

Similarly, the coefficient of determination increases from approximately 0.83 for the ANN to nearly 0.87 for the DKL (Figure 5), confirming an improved ability to explain the variance of the process signal. This result indicates that the probabilistic formulation enhances both local reconstruction accuracy and the global representation of system dynamics.

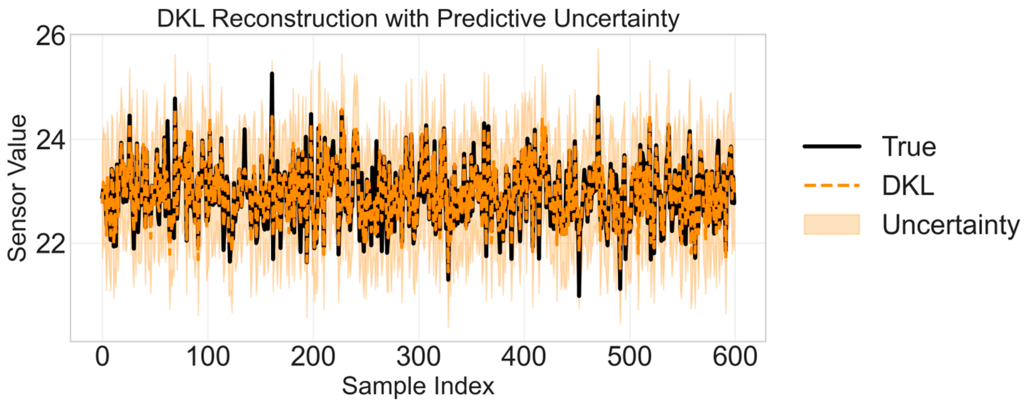

Figure 5. Probabilistic reconstruction produced by the DKL model. The shaded region represents the uncertainty quantification, illustrating stable uncertainty estimates and strong alignment with the ground truth signal.

A direct comparison of the performance metrics is reported in Figure 6. The consistent improvement across both RMSE and R^2 suggests that incorporating uncertainty into the learning process yields more reliable reconstructions without sacrificing the predictive stability.

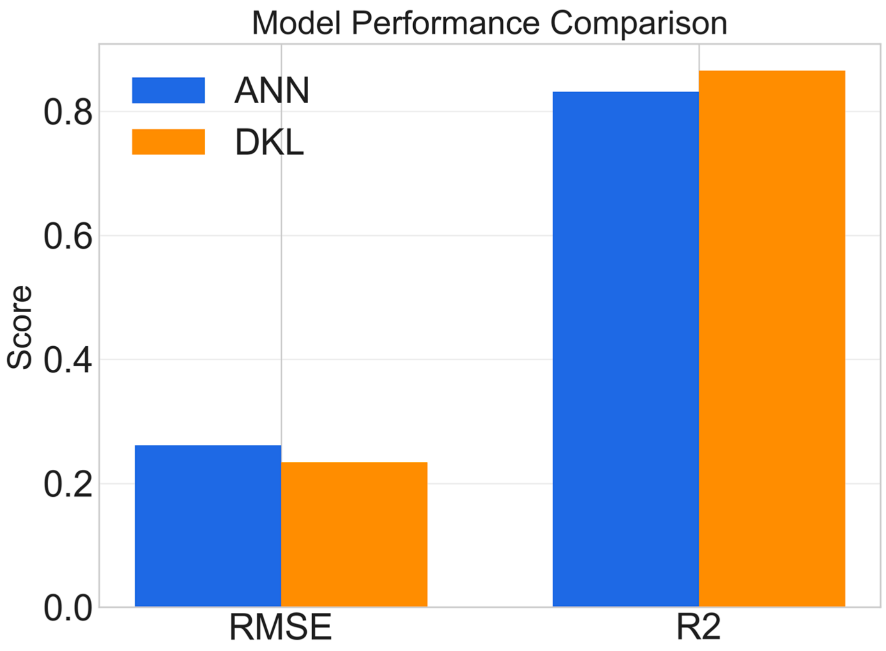

Figure 6. Performance comparison between ANN and DKL models. The probabilistic approach achieves lower reconstruction error and higher explanatory power.

The predicted-versus-true relationship shown in Figure 7 further supports these findings. DKL predictions remain tightly concentrated around the ideal diagonal, whereas ANN outputs exhibit a broader dispersion pattern. This reduced spread indicates improved calibration and stronger agreement between predicted and observed values.

Beyond accuracy, uncertainty plays a central role in assessing reconstruction reliability. The median coefficient of variation (CoV) is approximately 0.023, indicating that predictive uncertainty corresponds to only about 2–3% of the reconstructed signal magnitude. Such a low relative uncertainty suggests that the model operates in a highly confident regime across most operating conditions.

Importantly, uncertainty bands expand near sharper signal transitions, demonstrating adaptive behavior: the model expresses higher uncertainty precisely where the reconstruction task becomes more challenging. This property is particularly desirable in industrial environments, where distinguishing between confident and uncertain estimates is essential for risk-aware decision making.

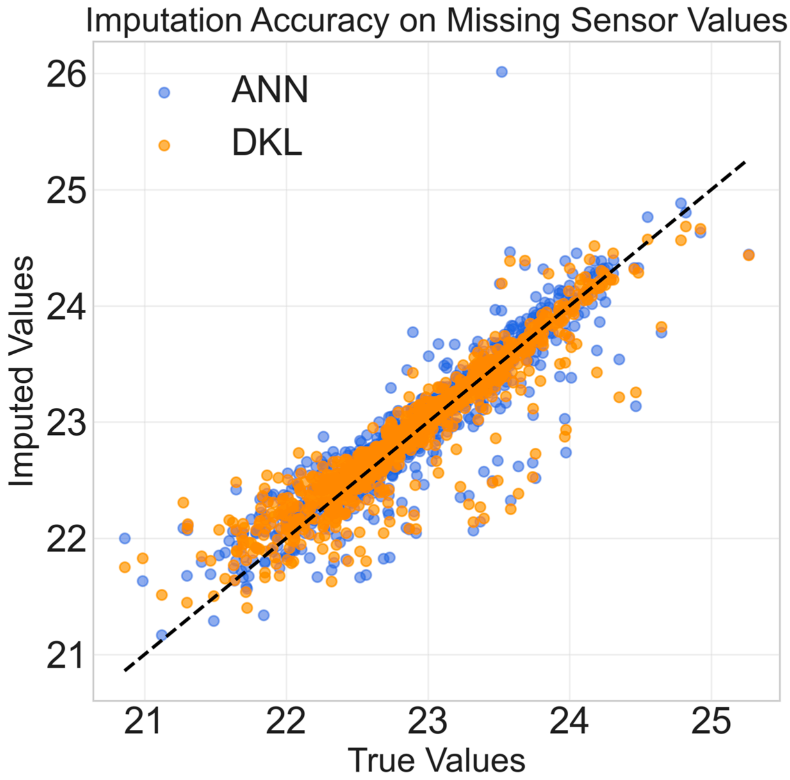

Figure 7. Predicted versus true values for the imputed samples. The tighter clustering of DKL predictions around the diagonal indicates improved reconstruction fidelity.

Notably, the alignment between reduced reconstruction error and low predictive uncertainty indicates that the probabilistic model is not only more accurate but also better calibrated.

Overall, the industrial experiment confirms the trends observed in the synthetic benchmark: uncertainty-aware models improve reconstruction accuracy while simultaneously providing actionable information regarding prediction confidence. This dual capability is critical for maintaining trustworthy data pipelines and supporting reliable automation in smart manufacturing systems.

The consistency between synthetic and industrial results strengthens the generalizability of the proposed framework and highlights its potential for deployment in real-world industrial scenarios.

From an operational perspective, the availability of both reconstructed values and associated uncertainty estimates enables more informed decision making within industrial data pipelines. Rather than blindly replacing missing sensor readings, maintenance systems could prioritize inspections when uncertainty exceeds predefined thresholds, while highly confident reconstructions could be safely reintegrated into analytics workflows.

For instance, in predictive maintenance scenarios, low-uncertainty imputations may allow continuous monitoring without interrupting production, whereas high-uncertainty estimates could trigger fallback strategies such as sensor redundancy checks or manual validation. This distinction transforms missing value handling from a purely data-centric task into a risk-aware operational process.

More broadly, embedding uncertainty-aware reconstruction into digital infrastructures contributes to preserving the integrity of the digital thread, ultimately supporting safer automation and more trustworthy AI deployment in smart manufacturing environments.

7.4. Implications and Recommended Framework

Given these observations of the trends and gaps of ensuring high quality data for manufacturing, we recommend concrete guidelines that will help to build the framework/standards needed. These guidelines are critical in moving the industry toward better data quality control, moving the whole of manufacturing closer to the ideals of Industry 4.0/5.0.

First, the framework should be developed in a way that supports the uniquely wide range of manufacturing applications. The level of detail should help to promote consistency, but be mindful of the global variation in manufacturing practices and standards. To address the challenge stated in Section 7.2.1 (Unified Data Quality Frameworks), the framework should begin with relevant existing standards within the targeted industry, then develop a guided path for data engineers, process owners, and manufacturing technicians to first identify the company-level problems to be solved (increased revenue, reduced cost, faster throughput, etc.) that the process being examined contributes toward. To address the challenge stated in Section 7.2.3. (Real-Time Assurance vs. Volume), the team should then examine the specific metrics that are most critical to monitor for the process and explore (through experimentation, user experience, hybrid modeling, etc.) which driving data streams contribute to the process metric's performance. Once identified, the data streams should be labeled and stored in a way that a data analyst unfamiliar with the process can find, understand, and effectively use the data. Examples of this framework for a given data stream could include the sensor used, process point where the sensor is placed, units of measurement and timing, historical accuracy, logs of data failures, etc.). In summary, the data selection and standards should be driven by both the top-level metrics and process-level metrics. This provides discrete guidance on priority, the amount of effort needed, and the specific level of detail required to map and connect a given data stream. The approach should be driven by "minimum effective effort", which ensures that the data will be collected and framed in a way that solves the needed problems, without committing excess resources to this already intensive effort.

Furthermore, the framework should be developed heavily toward the support of a real-world manufacturing environment. Rather than the sterile, controlled laboratory, the framework should be structured specifically to support manufacturing processes that have operated for five or more decades, with equipment that has been operating that length of time alongside modern IoT devices and AI process monitoring systems.

Second, the framework should be both modular and scalable. In testing how adaptive it must be, we need to plan for the framework to support even quantum computing applications in the factory, with computing capabilities able to handle a real-time analysis of multivariate data. While further research is required, it seems most beneficial to develop the specific framework for data quality at the sensor/instrument/machine level. This ensures that the data integrity can be performed as soon as the data is created, providing the maximum opportunity for data errors of any type to be detected and mitigated.

To address the challenge stated in Section 7.2.4. (Human Factors and Data Culture), the framework should emphasize the modular level of sensor/instrument/machine not just for scalability, but to better connect to the humans who will use this data. Ensuring that the data is tied to metrics at the process level will empower the humans who are directly managing that process. It increases their ability to understand the data and its connection to the process causes and effects. This creates a much shorter chain of explainability, has a higher chance of including context, and ensures less skepticism by the operator or process owner, even if they have not been specifically trained in data literacy. The effect of this modular encapsulation of process, data, insights, and understanding has potential benefits as the data framework expands upward. This is because validated data be connected with other processes and factory-level insights, but can also be accompanied by the human insights and acceptance from those sensor-/instrument-/machine-level modules.

This is also naturally efficient because only data certified as high quality is transmitted to higher levels in the system architecture. This framework should be able to operate at that single sensor, single unit level, but seamlessly integrate with other modular calculations for data quality so that multivariate data quality can also take place at the lowest possible level (e.g., an edge device that collects data from a group of sensors, and uses both individual sensor data along with sensor fusion to properly determine data quality).

Third, there should be particular emphasis on the risk management elements for this data quality framework. The data collection and models verifying quality should include an embedded confidence assessment, providing rich insight into the reliability of any data captured or imputation. A firm, quantifiable understanding of confidence can be weighed against the process being monitored and its criticality (e.g., if it has significant consequences for false positives or false negatives). This can create a consistent but highly flexible framework to ensure that the level of risk appetite for a given process intervention is well understood and appropriate. The need for confidence levels and an understanding of uncertainty is also critical due to these models likely involving AI-driven decision making. While AI can be very accurate, it must be balanced with a clear understanding of confidence in the AI model's output.

The example shown in Section 7.3 directly addresses this challenge, illustrating how risk management can be better informed through a framework that includes an uncertainty-aware assessment. This can be a critical improvement when placing greater trust on data-driven insights, and can also serve as a partial gap when compensating for the knowledge gaps left by retiring experts described in Section 5. While gaining that level of insight from decades of experience may be impossible if not thoroughly documented while experts are still among the workforce, providing a layer of uncertainty-aware context to data insights can enhance the decision making of operators who are trained but less experienced.

Fourth, the framework should create as little overhead as possible. This is important to ensure the framework can be indefinitely scalable, but has practical implications as well. The verification that collected data is complete, accurate, and interpretable is necessary. However, that does not mean that it is value-added. It adds no additional value to the manufacturing process, even though it is required. As such, the process must be as lean as possible to minimize the overhead burden on organizations. Only this framework will ensure that large enterprises and small job shops can actively use the framework without undue burden. An aspect of this should be the preference for intelligent software-based validation over the use of redundant sensors. While there may be high risk implementations where redundant sensors are necessary, it should be understood that there is always a cost and added risk to increasing the number of sensors simply for the sake of data quality verification. It can certainly help in some cases where the overhead and added complexity are well worth the benefit created. For most implementations, however, there is the risk of creating the Two Generals' Problem [103], which illustrates how the redundancy used for validation means that the redundant element must also be validated, effectively adding overhead with no added benefit.

In order to address the challenge discussed in Section 7.2.2. (Interoperability and Heterogeneity), it is critical to know how much effort should be spent in the process of accessing, validating, and standardizing data feeds from legacy equipment with limited data output capabilities. Addressing this challenge involves the other recommended framework elements described in this section: creating a unified framework, ensuring the framework is modular/scalable, and incorporating risk assessment through uncertainty-aware methods. While the process will vary by machine/sensor/data, a team working to answer the question of incorporating legacy sensor data should consider the following elements. At this point the team should have detailed insight into the top-level problems the company is working to solve, and how the process in question directly contributes to affecting those result metrics (e.g., revenue, costs, throughput). The team should also know what process metrics must be tracked, and what level of accuracy, confidence levels, and timeliness these metrics require. Based on this, the team will be able to explore the sensor data required within the process, assessing what data is currently available. In the case of potentially available legacy sensor data, the team can conduct a fairly accurate assessment of the cost/benefit comparison. This comparison weighs the overall benefit from acquiring/validating legacy sensor data vs. installing a new proxy sensor. Firm requirements such as accuracy, accessibility, and reliability will already be known at this time, allowing this cost/benefit decision to be well informed by the team.

8. Conclusions

The rise in data-centric tools such as advanced analytics, AI, IoT, vastly improved computation capabilities, and an emerging digital thread in leading manufacturing firms has significantly elevated the role and value of data in the manufacturing industry. In this exploration of this topic, we successfully validated that the amount of data collected and used for critical decision making has increased dramatically over the last several decades, and the rate of data collection continues to accelerate as AI-driven tools require more and more data to be effective. We have explored the level at which business leaders are relying upon data to make not just operational but strategic decisions—even though these same leaders simultaneously admit that their companies have major data shortcomings, with bad/incomplete data informing decisions on a regular basis. A number of different studies attempted to quantify the cost of this bad data, showing a multifaceted problem that identifies both direct and indirect (though not insignificant) costs. We studied research and applications for solving this problem, focusing on identifying missing data and corrupt data and predicting missing values. After exploring the relevant trends of research and development in this area, we identified critical gaps that must be addressed before the manufacturing industry can approach a functional Industry 4.0/5.0. Finally, we proposed a list of critical guidelines that should be followed as the industry builds toward a unifying framework for data quality assurance in manufacturing.

The need for reliable, accurate, complete, and interpretable data cannot be overstated. Manufacturing organizations are still relatively immature in creating strong digital threads that can fully leverage connected, high quality data. AI applications are quickly growing, however, and can make use of quality data at a specific process level today. That said, the gap between today and an industry with consistently high quality data is extensive, and is at risk of growing larger, not smaller, as more and more data collection is implemented without a strong data quality framework. There are tangible, quantitative costs to poor data quality, measured in the millions of dollars for the average manufacturing company. There are opportunity costs that are incurred as poor data quality companies lose out on rapidly evolving AI, optimization insights that span the entire operation, and protection against a knowledge drain from retired experts and experience the risk of strategic blunders by using inaccurate information; these costs are difficult to quantify, but will likely be realized in the near future as data quality serves as a key differentiator between successful manufacturers and companies that are forced to exit the market.

Future Research Needed

To this end, the industry and research institutes globally must continue to pursue efficient, effective, and flexible standards to ensure that data quality is verified/mitigated at its source. There are four recommended research questions that can help propel the industry toward Industry 4.0/5.0:

- Can we develop a unified framework that monitors and maintains data quality across heterogeneous manufacturing systems (legacy and modern equipment, different data types) in a scalable way?
- How can we design manufacturing control systems that not only detect data anomalies/dropouts in real time, but can automatically react (through fail-safes or corrective actions) to prevent quality or safety incidents?
- How can we incorporate domain expert knowledge into AI systems to create a consistent interpretation of manufacturing data (enabling a smaller and less-experienced workforce to use effective decision making)?
- What untapped insights can be gained by analyzing the "dark data" in manufacturing, and how can we reliably integrate those insights into process optimization without creating data quality issues?

Author Contributions: Conceptualization, N.E. and A.M.; methodology, N.E. and A.M.; software, A.M.; validation, N.E. and A.M.; formal analysis, N.E. and A.M.; investigation, N.E. and A.M.; resources, N.E. and A.M.; data curation, A.M.; writing—original draft preparation, N.E. and A.M.; writing—review and editing, N.E. and A.M.; visualization, N.E. and A.M.; supervision, N.E.; project administration, N.E.; funding acquisition, N.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data sets used for this study are available at the repositories cited in the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Xu, K.; Li, Y.; Liu, C.; Liu, X.; Hao, X.; Gao, J.; Maropoulos, P.G. Advanced Data Collection and Analysis in Data-Driven Manufacturing Process. *Chin. J. Mech. Eng.* **2020**, *33*, 43. [CrossRef]
2. Hu, C.; Sun, Z.; Li, C.; Zhang, Y.; Xing, C. Survey of Time Series Data Generation in IoT. *Sensors* **2023**, *23*, 6976. [CrossRef]
3. Sharp, M. NIST Explores AI-Enhanced Monitoring in Manufacturing Processes. *NIST*. 2024. Available online: https://www.nist.gov/blogs/manufacturing-innovation-blog/nist-explores-ai-enhanced-monitoring-manufacturing-processes (accessed on 30 December 2025).
4. Bonnard, R.; Arantes, M.D.S.; Lorbieski, R.; Vieira, K.M.M.; Nunes, M.C. Big data/analytics platform for Industry 4.0 implementation in advanced manufacturing context. *Int. J. Adv. Manuf. Technol.* **2021**, *117*, 1959–1973. [CrossRef]
5. Eskue, N. Digital Thread Roadmap for Manufacturing and Health Monitoring the Life Cycle of Composite Aerospace Components. *Aerospace* **2023**, *10*, 146. [CrossRef]
6. Byabazaire, J.; O'Hare, G.; Delaney, D. Using Trust as a Measure to Derive Data Quality in Data Shared IoT Deployments. In Proceedings of the 2020 29th International Conference on Computer Communications and Networks (ICCCN), Honolulu, HI, USA, 3–6 August 2020; pp. 1–9.
7. Sony, M.; Naik, S. Key ingredients for evaluating Industry 4.0 readiness for organizations: A literature review. *Benchmarking Int. J.* **2019**, *27*, 2213–2232. [CrossRef]
8. Labovitz, G.H.; Chang, Y.S.; Rosansky, V. *Making Quality Work: A Leadership Guide for the Results-Driven Manager*; Harper Business: New York, NY, USA, 1993; ISBN 9780887305825.

9. Khang, A.; Rath, K.C.; Satapathy, S.K.; Kumar, A.; Das, S.R.; Panda, M.R. Enabling the Future of Manufacturing: Integration of Robotics and IoT to Smart Factory Infrastructure in Industry 4.0. In *Handbook of Research on AI-Based Technologies and Applications in the Era of the Metaverse*; IGI Global Scientific Publishing: London, UK, 2023; pp. 25–50.
10. Hu, Y.; Jia, Q.; Yao, Y.; Lee, Y.; Lee, M.; Wang, C.; Zhou, X.; Xie, R.; Yu, F.R. Industrial Internet of Things Intelligence Empowering Smart Manufacturing: A Literature Review. *IEEE Internet Things J.* **2024**, *11*, 19143–19167. [CrossRef]
11. Bi, Z.; Jin, Y.; Maropoulos, P.; Zhang, W.-J.; Wang, L. Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). *Int. J. Prod. Res.* **2023**, *61*, 4004–4021. [CrossRef]
12. Wang, J.; Xu, C.; Zhang, J.; Zhong, R. Big data analytics for intelligent manufacturing systems: A review. *J. Manuf. Syst.* **2022**, *62*, 738–752. [CrossRef]
13. Bejani, M.; Appello, D.; Mauri, M.; Todaro, S.; Mariani, S. AI-Assisted Framework for Real-Time Monitoring and Management of Probe Cards in Electrical Wafer Sort Applications. In Proceedings of the 2025 IEEE European Test Symposium (ETS), Tallinn, Estonia, 26–30 May 2025; pp. 1–4.
14. Ruiz-Cárcel, C.; Cao, Y.; Mba, D.; Lao, L.; Samuel, R.T. Statistical process monitoring of a multiphase flow facility. *Control Eng. Pract.* **2015**, *42*, 74–88. [CrossRef]
15. Singh, H.A.; Almangour, B.P. *Handbook of Smart Manufacturing: Forecasting the Future of Industry 4.0*; CRC Press: Boca Raton, FL, USA, 2023.
16. Preuveneers, D.; Ilie-Zudor, E. The intelligent industry of the future: A survey on emerging trends, research challenges and opportunities in Industry 4.0. *J. Ambient. Intell. Smart Environ.* **2017**, *9*, 287–298. [CrossRef]
17. A Conceptual Framework to Liberate Open Data from Dark Data—ProQuest. Available online: https://www.proquest.com/openview/9a2b8a1f8a0c360d458c0e4337a2e3c0/1?pq-origsite=gscholar&cbl=2026366&diss=y (accessed on 11 December 2025).
18. Roman, D.; Prodan, R.; Nikolov, N.; Soylu, A.; Matskin, M.; Marrella, A.; Kimovski, D.; Elvesæter, B.; Simonet-Boulogne, A.; Ledakis, G.; et al. Big Data Pipelines on the Computing Continuum: Tapping the Dark Data. *Computer* **2022**, *55*, 74–84. [CrossRef]
19. Corallo, A.; Crespino, A.M.; Vecchio, V.D.; Lazoi, M.; Marra, M. Understanding and Defining Dark Data for the Manufacturing Industry. *IEEE Trans. Eng. Manag.* **2023**, *70*, 700–712. [CrossRef]
20. Williamson, S.M.; Prybutok, V. Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Appl. Sci.* **2024**, *14*, 675. [CrossRef]
21. Babu, M.M.; Rahman, M.; Alam, A.; Dey, B.L. Exploring big data-driven innovation in the manufacturing sector: Evidence from UK firms. *Ann. Oper. Res.* **2024**, *333*, 689–716. [CrossRef]
22. Eichenseer, P.; Winkler, H. A data-oriented shopfloor management in the production context: A systematic literature review. *Int. J. Adv. Manuf. Technol.* **2024**, *134*, 4071–4097. [CrossRef]
23. Corban, T. Data as a Strategic Asset. EBSCOhost. 2021. Available online: https://openurl.ebsco.com/contentitem/gcd:149599922?sid=ebsco:plink:crawler&id=ebsco:gcd:149599922 (accessed on 18 February 2026).
24. Bejani, M.; Appello, D.; Mauri, M.; Missaglia, E.; Mariani, S. Digital Twin-Assisted Optimal Sensor Placement for Real-Time Monitoring of Probe Cards in EWS Applications. In Proceedings of the 2025 26th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), Utrecht, The Netherlands, 6–9 April 2025; pp. 1–6.
25. Lodhi, S.K.; Gill, A.Y.; Hussain, I. AI-Powered Innovations in Contemporary Manufacturing Procedures: An Extensive Analysis. *Int. J. Multidiscip. Sci. Arts* **2024**, *3*, 15–25. [CrossRef]
26. Lorenz, R.; Kraus, M.; Wolf, H.; Feuerriegel, S.; Netland, T.H. Selecting advanced analytics in manufacturing: A decision support model. *Prod. Plan. Control* **2024**, *35*, 711–724. [CrossRef]
27. Rakholia, R.; Suárez-Cetrulo, A.L.; Singh, M.; Carbajo, R.S. Advancing Manufacturing Through Artificial Intelligence: Current Landscape, Perspectives, Best Practices, Challenges, and Future Direction. *IEEE Access* **2024**, *12*, 131621–131637. [CrossRef]
28. Shahin, M.; Maghanaki, M.; Hosseinzadeh, A.; Chen, F.F. Improving operations through a lean AI paradigm: A view to an AI-aided lean manufacturing via versatile convolutional neural network. *Int. J. Adv. Manuf. Technol.* **2024**, *133*, 5343–5419. [CrossRef]
29. Damtew, A.W. Application of Digital Thread Impacts on Sustainable Manufacturing and Smart Production Systems. *Res. Sq.* **2024**, *preprint*. [CrossRef]
30. Xiang, W.; Yu, K.; Han, F.; Fang, L.; He, D.; Han, Q.-L. Advanced Manufacturing in Industry 5.0: A Survey of Key Enabling Technologies and Future Trends. *IEEE Trans. Ind. Inform.* **2024**, *20*, 1055–1068. [CrossRef]
31. Data Challenges Affect Nearly all Manufacturers. *Aerospace Manufacturing and Design*, 30 April 2024. Available online: https://www.aerospacemanufacturinganddesign.com/article/data-challenges-affect-nearly-all-manufacturers (accessed on 12 December 2025).
32. New SoftServe Study: Big Decisions Made with Bad Data. 2025. Available online: https://www.softserveinc.com/en-us/news/bad-data-makes-bad-decisions (accessed on 12 December 2025).

33. Rangineni, S.; Bhanushali, A.; Suryadevara, M.; Venkata, S.; Peddireddy, K. A Review on Enhancing Data Quality for Optimal Data Analytics Performance. *Int. J. Comput. Sci. Eng.* **2023**, *11*, 51–58. [CrossRef]
34. Data Governance in Manufacturing: Your Complete 2024 Guide. Available online: https://atlan.com/data-governance-in-manufacturing/ (accessed on 12 December 2025).
35. How Scrap Rework Affect Cost of Quality and OEE. Available online: https://www.ease.io/blog/scrap-rework-affect-cost-of-quality-and-oee/ (accessed on 12 December 2025).
36. KashTech LLC. The Cost of Poor Data Quality and How It Impacts Manufacturing 4.0. 2025. Available online: https://www.kashtechllc.com/blog/data-analytics/the-cost-of-poor-data-quality-and-how-it-impacts-manufacturing-4-0/ (accessed on 12 December 2025).
37. Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; Tabona, O. A survey on missing data in machine learning. *J. Big Data* **2021**, *8*, 140. [CrossRef] [PubMed]
38. Improving AI Data Quality in Manufacturing | McKinsey. Available online: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/clearing-data-quality-roadblocks-unlocking-ai-in-manufacturing (accessed on 12 December 2025).
39. Cooper, R.G. Why AI Projects Fail: Lessons from New Product Development. *IEEE Eng. Manag. Rev.* **2024**, *52*, 15–21. [CrossRef]
40. Goel, P.; Datta, A.; Mannan, M.S. Industrial alarm systems: Challenges and opportunities. *J. Loss Prev. Process Ind.* **2017**, *50*, 23–36. [CrossRef]
41. Calzavara, M.; Battini, D.; Bogataj, D.; Sgarbossa, F.; Zennaro, I. Ageing workforce management in manufacturing systems: State of the art and future research agenda. *Int. J. Prod. Res.* **2020**, *58*, 729–747. [CrossRef]
42. Jennex, M.; Durcikova, A.; Ilvonen, I.; Babb, J. Assessing and Mitigating the Risk of Critical Knowledge Loss in Organizations: Insights from COVID-19 and the Great Resignation. In Proceedings of the Hawaii International Conference on System Sciences 2024 (HICSS-57), Honolulu, HI, USA, 3–6 January 2024.
43. Summerscales, J. Harvesting tacit knowledge for composites workforce development. *Compos. Part A Appl. Sci. Manuf.* **2024**, *185*, 108357. [CrossRef]
44. Krumsiek, K.J. Retention of Knowledge From "Baby Boomers" Prior to Leaving the Workforce. In *Effective Human Resources Management in the Multigenerational Workplace*; IGI Global Scientific Publishing: London, UK, 2024; pp. 23–50.
45. Kuyken, K.; Costanza, D. Because Work is Changing: A New Paradigm for Intergenerational Workplace Knowledge Sharing. *J. Intergener. Relatsh.* **2025**, *23*, 91–107. [CrossRef]
46. Automate. Closing the Manufacturing Skills Gap: Generative AI as a Workforce Solution. Available online: https://www.automate.org/ai/industry-insights/genai-can-help-close-skills-gap (accessed on 11 December 2025).
47. Ghasemaghaei, M. Does data analytics use improve firm decision making quality? The role of knowledge sharing and data analytics competency. *Decis. Support Syst.* **2019**, *120*, 14–24. [CrossRef]
48. Whig, P.; Madavarapu, J.B.; Yathiraju, N.; Thatikonda, R. Managing Knowledge in the Era of Industry 4.0. In *Knowledge Management and Industry Revolution 4.0.*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2024; pp. 239–273.
49. Peixoto, T.; Oliveira, B.; Oliveira, Ó.; Ribeiro, F. Data Quality Assessment in Smart Manufacturing: A Review. *Systems* **2025**, *13*, 243. [CrossRef]
50. Qian, H.; Wang, M.; Zhu, M.; Wang, H. A Review of Multi-Sensor Fusion in Autonomous Driving. *Sensors* **2025**, *25*, 6033. [CrossRef]
51. Lin, T.; Ren, Z.; Zhu, L.; Zhu, Y.; Feng, K.; Ding, W. A Systematic Review of Multi-Sensor Information Fusion for Equipment Fault Diagnosis. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 3507848. [CrossRef]
52. Kasilingam, S.; Yang, R.; Singh, S.K.; Farahani, M.A.; Rai, R.; Wuest, T. Physics-based and data-driven hybrid modeling in manufacturing: A review. *Prod. Manuf. Res.* **2024**, *12*, 2305358. [CrossRef]
53. Sahoo, S. Sensor Fusion and Virtual Sensor Design for Enhanced Multi-Sensor Data Accuracy in Autonomous Systems. *Int. J. Smart Sustain. Intell. Comput.* **2024**, *1*, 21–39. [CrossRef]
54. Krempl, G.; Puolamäki, K.; Miliou, I. *Advances in Intelligent Data Analysis XXIII: 23rd International Symposium on Intelligent Data Analysis, IDA 2025, Konstanz, Germany, May 7–9, 2025, Proceedings*; Springer Nature: Cham, Switzerland, 2025.
55. Guo, H.; He, K.; Luo, Y.; Chang, Y. Physics-Informed Neural Networks for robust thermal comfort prediction: Overcoming data quality limitations through physiological constraints. *Build. Environ.* **2025**, *285*, 113588. [CrossRef]
56. Liso, A.; Cardellicchio, A.; Patruno, C.; Nitti, M.; Ardino, P.; Stella, E. 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]
57. AlKalbani, W.; AlAzri, A.; Ahmad, M.N. Data Quality Assessment Framework for Predictive Maintenance (PdM) in Industrial AI: Review. In *Advances in Visual Informatics*; Badioze Zaman, H., Robinson, P., Smeaton, A.F., Shih, T.K., Jørgensen, B.N., Huajing, Z., Xiaoping, L., Mohamad Ali, N., Mat Surin, E.S., Eds.; Springer Nature: Singapore, 2026; pp. 35–49.
58. Raeiszadeh, M.; Ebrahimzadeh, A.; Glitho, R.H.; Eker, J.; Mini, R.A.F. Real-Time Adaptive Anomaly Detection in Industrial IoT Environments. *IEEE Trans. Netw. Serv. Manag.* **2024**, *21*, 6839–6856. [CrossRef]

59. Xie, J.; Sun, L.; Zhao, Y. On the Data Quality and Imbalance in Machine Learning-based Design and Manufacturing—A Systematic Review. *Engineering* **2024**, *45*, 105–131. [CrossRef]
60. Shamsa, M.; Lerner, D. Defect Mechanisms Responsible for Silent Data Errors. In Proceedings of the 2024 IEEE International Reliability Physics Symposium (IRPS), Grapevine, TX, USA, 14–18 April 2024; pp. 1–5.
61. Rodríguez, M.; Tobón, D.P.; Múnera, D. A framework for anomaly classification in Industrial Internet of Things systems. *Internet Things* **2025**, *29*, 101446. [CrossRef]
62. Rong, Z.; Pang, R.; Xu, B.; Zhou, Y. Dam safety monitoring data anomaly recognition using multiple-point model with local outlier factor. *Autom. Constr.* **2024**, *159*, 105290. [CrossRef]
63. Pemula, L.; Zhang, D.; Dabeer, O. Robust AD: A Real World Benchmark Dataset for Robustness in Industrial Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Nashville, TN, USA, 11–15 June 2025; pp. 4086–4096.
64. Oakland, J.; Oakland, R. *Statistical Process Control and Data Analytics*, 8th ed.; Routledge: London, UK, 2024.
65. Lee, Y.; Park, C.; Kim, N.; Ahn, J.; Jeong, J. LSTM-Autoencoder Based Anomaly Detection Using Vibration Data of Wind Turbines. *Sensors* **2024**, *24*, 2833. [CrossRef]
66. Chang, Z.; Wu, J.; Liang, H.; Wang, Y.; Wang, Y.; Xiong, X. A Review of Power System False Data Attack Detection Technology Based on Big Data. *Information* **2024**, *15*, 439. [CrossRef]
67. Lin, W.-T.; Chen, G.; Zhou, X. Privacy-preserving federated learning for detecting false data injection attacks on power system. *Electr. Power Syst. Res.* **2024**, *229*, 110150. [CrossRef]
68. Oshnoei, S.; Aghamohammadi, M.R.; Khooban, M.H. Smart Frequency Control of Cyber-Physical Power System Under False Data Injection Attacks. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2024**, *71*, 5582–5595. [CrossRef]
69. Miguel, P.; Massimo, V. Towards Trusted Data on Decentralized IoT Applications: Integrating Blockchain in Constrained Devices. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020. [CrossRef]
70. Jäger, S.; Allhorn, A.; Bießmann, F. A Benchmark for Data Imputation Methods. *Front. Big Data* **2021**, *4*, 693674. [CrossRef] [PubMed]
71. Jiang, D.; Yang, H.; Cao, H.; Xu, D. A missing data imputation method for industrial soft sensor modeling. *J. Process Control* **2025**, *152*, 103485. [CrossRef]
72. Peng, D.; Zou, M.; Liu, C.; Lu, J. RESI: A Region-Splitting Imputation method for different types of missing data. *Expert Syst. Appl.* **2021**, *168*, 114425. [CrossRef]
73. Bertsimas, D.; Delarue, A.; Pauphilet, J. Adaptive optimization for prediction with missing data. *Mach. Learn.* **2025**, *114*, 124. [CrossRef]
74. Pan, X.; Wang, H.; Lei, M.; Ju, T.; Bai, L. A method for filling missing values in multivariate sequence bidirectional recurrent neural networks based on feature correlations. *J. Comput. Sci.* **2024**, *83*, 102472. [CrossRef]
75. Sun, Y.; Li, J.; Xu, Y.; Zhang, T.; Wang, X. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Syst. Appl.* **2023**, *227*, 120201. [CrossRef]
76. Yuan, X.; Zhang, J.; Wang, K.; Wang, Y.; Yang, C.; Gui, W.; Shen, F.; Ye, L. Missing Data Imputation for Industrial Time Series with Adaptive Median Iteration Based on Generative Adversarial Networks. *IEEE Sens. J.* **2024**, *24*, 35081–35091. [CrossRef]
77. Zhang, K.; Yang, Q.; Li, C.; Sun, X.; Chen, J. Missing Data Recovery Methods on Multivariate Time Series in IoT: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2025**, *28*, 149–180. [CrossRef]
78. Bejani, M.; Mauri, M.; Mariani, S. Self-Attention-Based Deep Learning for Missing Sensor Data Imputation in Real-Time Probe Card Monitoring. *Sensors* **2025**, *25*, 7194. [CrossRef]
79. Zhang, Y.; Zhou, B.; Cai, X.; Guo, W.; Ding, X.; Yuan, X. Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Inf. Sci.* **2021**, *551*, 67–82. [CrossRef]
80. Zhang, X.; Sun, X.; Xia, L.; Tao, S.; Xiang, S. A Matrix Completion Method for Imputing Missing Values of Process Data. *Processes* **2024**, *12*, 659. [CrossRef]
81. Jeong, C.; Byon, E.; He, F.; Fang, X. Tensor-based statistical learning methods for diagnosing product quality defects in multistage manufacturing processes. *IISE Trans.* **2025**, *57*, 706–723. [CrossRef]
82. Shah, K.; Gami, S.; Trehan, A. An Intelligent Approach to Data Quality Management AI-Powered Quality Monitoring in Analytics. *Int. J. Adv. Res. Sci. Commun. Technol.* **2024**, *4*, 2581–9429. [CrossRef]
83. Gami, S.; Shah, K.; Katru, C.R.; Nagarajan, S.K.S. Interactive Data Quality Dashboard: Integrating Real-Time Monitoring with Predictive Analytics for Proactive Data Management. *Int. J. Comput. Sci. Eng.* **2024**, *12*, 5. Available online: https://www.researchgate.net/publication/387631619_Interactive_Data_Quality_Dashboard_Integrating_Real-Time_Monitoring_with_Predictive_Analytics_for_Proactive_Data_Management (accessed on 23 February 2026). [CrossRef]
84. Kuppurajuy, S.Y.; Anand, G.; Choudhury, A. Data Augmentation Techniques for Building Robust AI Models in Enterprise Applications. *Int. J. Innov. Res. Eng. Manag.* **2025**, *12*, 69–74. [CrossRef]

85. Ekwaro-Osire, H.; Ponugupati, S.L.; Al Noman, A.; Bode, D.; Thoben, K.-D. Data augmentation for numerical data from manufacturing processes: An overview of techniques and assessment of when which techniques work. *Ind. Artif. Intell.* **2025**, *3*, 1. [CrossRef]
86. Krishnamurthi, R.; Kumar, A.; Gopinathan, D.; Nayyar, A.; Qureshi, B. An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques. *Sensors* **2020**, *20*, 6076. [CrossRef]
87. Sharma, M.; Tomar, A.; Hazra, A. Edge Computing for Industry 5.0: Fundamental, Applications, and Research Challenges. *IEEE Internet Things J.* **2024**, *11*, 19070–19093. [CrossRef]
88. Miller, R.; Chan, S.H.M.; Whelan, H.; Gregório, J. A Comparison of Data Quality Frameworks: A Review. *Big Data Cogn. Comput.* **2025**, *9*, 93. [CrossRef]
89. Miller, R.; Whelan, H.; Chrubasik, M.; Whittaker, D.; Duncan, P.; Gregório, J. A Framework for Current and New Data Quality Dimensions: An Overview. *Data* **2024**, *9*, 151. [CrossRef]
90. ISO 9001:2015. ISO. Available online: https://www.iso.org/standard/62085.html (accessed on 30 December 2025).
91. Muñoz, M.; Torres, M.; Gil, J.D.; Guzmán, J.L. An Internet of Things platform for heterogeneous data integration: Methodology and application examples. *J. Netw. Comput. Appl.* **2025**, *240*, 104197. [CrossRef]
92. Sadeghi, M.; Carenini, A.; Corcho, O.; Rossi, M.; Santoro, R.; Vogelsang, A. Interoperability of heterogeneous Systems of Systems: From requirements to a reference architecture. *J. Supercomput.* **2024**, *80*, 8954–8987. [CrossRef]
93. Fatemi, N.; Fattahi, J. Adaptable Semantic Interoperability in Heterogeneous Smart Grids Using Large Language Models. *Energy Rep.* **2025**, *14*, 5774–5789. [CrossRef]
94. Zhang, H.; Jia, X.; Chen, C. Deep Learning-Based Real-Time Data Quality Assessment and Anomaly Detection for Large-Scale Distributed Data Streams. *Int. J. Med. All Body Health Res.* **2025**, *6*, 1–11. [CrossRef]
95. Immadisetty, A. Mastering Data Platform Design: Industry-Agnostic Patterns for Scale. *Int. J. Res. Comput. Appl. Inf. Technol.* **2021**, *7*, 2259–2270. [CrossRef]
96. Manduva, V.C. Scalable AI: Leveraging Cloud and Edge Computing for Real-Time Analytics. *Int. J. Sci. Res. Manag.* **2024**, *12*, 1788–1813. Available online: https://ijsrm.net/index.php/ijsrm/article/view/5819/3801 (accessed on 23 February 2026).
97. Pothier, W.; Condon, P. Data Literacy Skills: Industry Perspectives and Professional Practice. *Portal Libr. Acad.* **2025**, *25*, 271–298. [CrossRef]
98. Terras, M.; Jones, V.; Osborne, N.; Speed, C. (Eds.) *Data-Driven Innovation in the Creative Industries*, 1st ed.; Routledge: London, UK, 2024. [CrossRef]
99. Lefebvre, H.; Krasikov, P.; Flourac, G.; Legner, C. Toward Cross-company Value Generation from Data: Investigating the Role of Data Sharing Communities. *Pre-ICIS FRAIS* **2022**, *2022*, 3.
100. Gao, J.; Lu, Y.; Ashrafi, N.; Domingo, I.; Alaei, K.; Pishgar, M. Prediction of sepsis mortality in ICU patients using machine learning methods. *BMC Med. Inf. Decis. Mak.* **2024**, *24*, 228. [CrossRef]
101. Iooss, B.; Ribatet, M.; Marrel, A. Global Sensitivity Analysis of Stochastic Computer Models with joint metamodels. *arXiv* **2009**, arXiv:0802.0443. [CrossRef]
102. Reinartz, C.; Kulahci, M.; Ravn, O. An extended Tennessee Eastman simulation dataset for fault-detection and decision support systems. *Comput. Chem. Eng.* **2021**, *149*, 107281. [CrossRef]
103. Danny, D. The Byzantine generals strike again. *J. Algorithms* **1982**, *3*, 14–30. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/app16052409