frontiers | Frontiers in Manufacturing Technology

TYPE Review
PUBLISHED 13 February 2024
DOI 10.3389/fmtec.2024.1320166

Check for updates

OPEN ACCESS

EDITED BY
Kunpeng Zhu,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Xin Lin,
Wuhan University of Science and Technology, China
Xueliang Zhou,
Hubei University of Automotive Technology, China

*CORRESPONDENCE
Jaime Camelio,
jcamelio@uga.edu

RECEIVED 11 October 2023
ACCEPTED 15 January 2024
PUBLISHED 13 February 2024

CITATION
Buggineni V, Chen C and Camelio J (2024), Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility. *Front. Manuf. Technol.* 4:1320166.
doi: 10.3389/fmtec.2024.1320166

Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility

Vishnupriya Buggineni[1], Cheng Chen[2] and Jaime Camelio[2]*

[1]Institute of Artificial Intelligence, University of Georgia, Athens, GA, United States, [2]College of Engineering, University of Georgia, Athens, GA, United States

Addressing the challenges of data scarcity and privacy, synthetic data generation offers an innovative solution that advances manufacturing assembly operations and data analytics. Serving as a viable alternative, it enables manufacturers to leverage a broader and more diverse range of machine learning models by incorporating the creation of artificial data points for training and evaluation. Current methods lack generalizable framework for researchers to follow and solve these issues. The development of synthetic data sets, however, can make up for missing samples and enable researchers to understand existing issues within the manufacturing process and create data-driven tools for reducing manufacturing costs. This paper systematically reviews both discrete and continuous manufacturing process data types with their applicable synthetic generation techniques. The proposed framework entails four main stages: Data collection, pre-processing, synthetic data generation, and evaluation. To validate the framework's efficacy, a case study leveraging synthetic data enabled an exploration of complex defect classification challenges in the packaging process. The results show enhanced prediction accuracy and provide a detailed comparative analysis of various synthetic data strategies. This paper concludes by highlighting our framework's transformative potential for researchers, educators, and practitioners and provides scalable guidance to solve the data challenges in the current manufacturing sector.

KEYWORDS

synthetic data generation, machine learning models, data quality and bias, industry 4.0, data quality assessment, defect classification, quality control

1 Introduction

The convergence of digitization, automation, and the Internet of Things (IoT) under the umbrella of Industry 4.0 continues to influence the evolution of modern manufacturing. Having the right data at the right time is still a significant challenge for manufacturers seeking to enhance their decision-making processes (Whitley, 2022). The global synthetic data generation market, valued at $168.9 million between 2021 and 2022, is anticipated to reach $3.5 billion by 2031, reflecting an impressive growth rate of 35.8% (Shrawanty, 2022; Research, 2023). Recognizing this remarkable opportunity and necessity, various industries have integrated data-driven applications, propelling foundational research towards synthetic data. Many studies, for instance, utilize synthetic data to enhance manufacturing processes, including process monitoring (Fecker et al., 2013), quality inspection (Nguyen et al., 2022), production scheduling (Andres et al., 2021), and process optimization (Apornak et al., 2021). Despite considerable progress in enhancing production efficiency, reducing waste, and improving product quality, notable challenges remain in generating synthetic data for the development and validation of innovative assembly models and algorithms, particularly with real-world production data (Tao et al., 2018). Therefore, there is a significant need for the development of novel, data-driven applications, and an expansion of the utilization of artificial intelligence in manufacturing processes.

Having a large, accurate, and reliable data repository is crucial for scaling up the production system or conducting large-scale simulations. However, it often leads to challenges such as data scarcity, costs, quality, and proprietary constraints (Gao et al., 2020). Generating synthetic data for manufacturing assembly systems proves challenging due to the complexity, realism, size, and cost considerations of these systems (Gao et al., 2020; Mubarak et al., 2020). Creating a synthetic data generator can address these difficulties by providing flexible solutions for generating test instances (Mubarak et al., 2020). In addition, the process of generating synthetic data demands a thorough understanding of process mechanics and physics, along with substantial computational resources. The use of incomplete or inaccurate data can introduce bias, degrade model performance, and impede both collaboration and innovation. Nevertheless, synthetic data, offering publicly accessible variations in manufacturing assembly processes (Tao et al., 2018; Mubarak et al., 2020), emerges as a potent solution to these obstacles, accurately reflecting authentic production patterns.

Enhancing the accuracy and utility of synthetic data in manufacturing assembly processes entails addressing various challenges. Current models have difficulty representing real-world production processes accurately. Furthermore, there is a significant knowledge gap regarding high-quality synthetic data generation, despite its critical role in producing more realistic data. The absence of a standardized method for the validation and testing of synthetic data also constitutes a barrier, inhibiting the comparison of different synthetic data generation methods. A final key challenge lies in the need for more concentrated research into identifying and mitigating biases in synthetic data generation, given its significant potential to boost data accuracy and reliability. These multifaceted issues inspire a set of research questions:

- What are the most effective methods for generating synthetic data in production systems?
- How can an effective framework be designed for the production, validation, and testing of synthetic data sets to support Industry 4.0 practices?
- How can the proposed framework be implemented through a case study to assess the effectiveness of synthetic data across various performance trade-offs?

By addressing these research questions, this paper begins by examining the existing literature on synthetic data generation in manufacturing assembly systems and presents a framework for generating synthetic data that simulates a variety of production scenarios. The proposed framework provides a wide array of production scenarios, proving beneficial for researchers, educators, and industry practitioners. This work enhances the design and evaluation of synthetic data generation, and aids in assembly systems transformation through data-driven approaches, thereby providing substantial benefits to researchers, educators, and industry practitioners.

The rest of the paper is organized as follows: Section 2 presents a systematic review of synthetic data generation, encapsulating the current state-of-the-art approaches. Section 5 outlines a comprehensive framework for generating synthetic data,

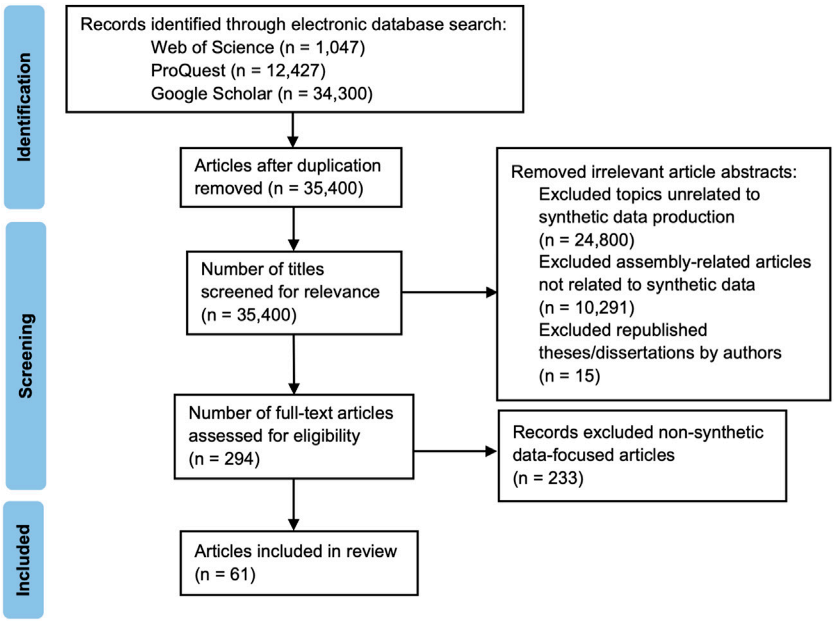

Identification
Screening
Included
Records identified through electronic database search:
Web of Science (n = 1,047)
ProQuest (n = 12,427)
Google Scholar (n = 34,300)
Articles after duplication removed (n = 35,400)
Number of titles screened for relevance (n = 35,400)
Removed irrelevant article abstracts:
Excluded topics unrelated to synthetic data production (n = 24,800)
Excluded assembly-related articles not related to synthetic data (n = 10,291)
Excluded republished theses/dissertations by authors (n = 15)
Number of full-text articles assessed for eligibility (n = 294)
Records excluded non-synthetic data-focused articles (n = 233)
Articles included in review (n = 61)

FIGURE 1
Systematic review search process.

TABLE 1 Literature on synthetic data generation.

	Type of data	Synthetic data generation method	Application of manufacturing
Discrete	Binary	ADASYN Han et al. (2019), Random Sampling Kim et al. (2020), Digital Twin Simulation Ademujimi and Prabhu (2022), GMM Fecker et al. (2013), SMOTE Syafrudin et al. (2018)	Quality Control Han et al. (2019); Kim et al. (2020); Ademujimi and Prabhu (2022), Syafrudin et al. (2018); Fecker et al. (2013)
	2D coordinates	GANs Sibona and Indri (2021)	Human Robot Collaboration Sibona and Indri (2021)
	Point-cloud	IPS cable simulation and Blender Nguyen et al. (2022)	Quality Control (Automated optical inspection) Nguyen et al. (2022)
	Multi-class	Simufact additive software Biczó et al. (2021)	Quality Control Biczó et al. (2021)
	Time-series	By varying parameters Laxman et al. (2007); Andres et al. (2021), GANs Malekzadeh et al. (2017), Promodel-PC simulation Bikes et al. (1994), Simul8 simulation Guner et al. (2016), SIMIO simulation Sisca et al. (2015), Fiasché et al. (2016), Hidden Markov Models Mubarak et al. (2020), Simpy Library Martin et al. (2020), Virtual Factory Prototype Jain et al. (2018), Taguchi simulation Apornak et al. (2021), and Wiener Process, Gaussian Noise and by varying standard deviation Cai et al. (2021)	Quality Control Laxman et al. (2007), Production Scheduling Andres et al. (2021), Activity Recognition Malekzadeh et al. (2017), Component Delivery Bikes et al. (1994), Preventive Maintenance Guner et al. (2016), Production Planning Sisca et al. (2015); Fiasché et al. (2016), Pipe-Spooling Mubarak et al. (2020), Quality Control Martin et al. (2020), Cycle-Time estimation Jain et al. (2018), Process Optimization Apornak et al. (2021), Stream Processing Cai et al. (2021)
Continuous	Image	Unity 3D and Revit software Zheng et al. (2020), Unity 3D and CAD models Lai et al. (2020); Kohtala and Steinert (2021), Domain Randomization Grappiolo et al. (2021), Geometric transformations de la Rosa et al. (2022), By varying levels of environmental noise Sikora et al. (2021), Through rotating and modifying the colors of the images Singh et al. (2020), GANs Qian et al. (2022), and Blender software and Domain randomization Ameperosa and Bhounsule (2020), Blender software Maliks and Kadikis (2021)	Process Optimization Zheng et al. (2020) Increasing Productivity Lai et al. (2020), Quality Control de la Rosa et al. (2022); Sikora et al. (2021); Ameperosa and Bhounsule (2020); Maliks and Kadikis (2021), Production Monitoring Kohtala and Steinert (2021), Operator Guidance Grappiolo et al. (2021), Braille Display Singh et al. (2020), Industrial Internet of Things Qian et al. (2022)
	3D image	Ksim9 Rio-Torto et al. (2021), Unreal Engine4 Outón et al. (2021)	Quality Control Rio-Torto et al. (2021), Autonomous industrial mobile manipulator Outón et al. (2021)
	Video	GANs da Silva et al. (2021)	Defect detection da Silva et al. (2021)

elaborating on data types and applicable synthetic data techniques. Section 6 showcases a case study utilizing the proposed framework and analyzed results. Lastly, Section 7 provides an overview of potential future research directions in synthetic data generation within the manufacturing domain.

2 Systematic review

Synthetic data generation plays a vital role in manufacturing, fostering innovation and process optimization while safeguarding data privacy. It facilitates diverse scenario simulations, empowering predictive maintenance, and bolstering quality control. Importantly, this approach trains machine learning models without risking sensitive information and remedies real-world data's limitations. Consequently, synthetic data generation enhances efficiency, reduces costs, and bolsters competitiveness—key components for Industry 4.0 success.

2.1 Literature search and selection process

The purpose of this section is to narrow down the related literature regarding synthetic data generation in assembly production, as depicted in Figure 1. Various databases (e.g., Web of Science and Google Scholar) were utilized as a resource for discovering relevant papers. Initially, the keywords "Synthetic Data" AND "Generation" AND ("Manufacturing" OR "Production") were used, yielding 35,400 articles. Refining the search to focus specifically on manufacturing, unrelated articles were eliminated. This yielded 10,600 articles primarily about manufacturing, a field that employs diverse techniques and equipment for product creation, including phases like design, engineering, prototyping, and testing (Blau et al., 1976). By excluding keywords not associated with production—a comprehensive term encompassing all activities related to goods or services creation—the results were further refined to articles specific to the research topic. This process led to a greater understanding of synthetic data applications in assembly processes. Incorporating the keyword "Assembly Line" refined the search to 309 relevant articles. After assessing titles for duplication and screening abstracts for their relevance, a total of 61 pivotal papers were identified. These were divided into two categories: 33 papers on synthetic data generation and 28 on the application of synthetic data in research. The first category revolves around the creation of simulated data sets mirroring real-world data, while the second category exploits synthetic data for experimental and analytical purposes, offering an insightful framework to comprehend the diverse applications of synthetic data in research.

2.2 Synthetic data generation methods

Upon reviewing 61 papers, the studies, organized in Table 1, can be classified into two primary categories. Two distinct topics for synthetic data generation are covered in these papers: physical simulations and numerical modeling.

2.2.1 Physical simulation

Physical simulations, gaining increasing popularity in the manufacturing and assembly domain, leverage mathematical models to replicate real-world phenomena. This provides a platform for researchers to generate synthetic data that faithfully mirrors the system under study. By operating simulations under varied parameters and configurations, researchers can generate synthetic data conducive to the testing and refinement of assembly processes, ultimately enhancing efficiency and productivity.

Recent years have seen an increasing use of physical simulations to generate synthetic data for manufacturing and assembly applications. These include the use of digital twin models for fault diagnostics (Ademujimi and Prabhu, 2022) and ProModel-PC simulation software for assembly system analysis (Bikes et al., 1994). Data analytics approaches have also been compared using simulation-based synthetic data (Jain et al., 2018), and machine learning methods have been employed to predict additive manufacturing process distortion (Biczó et al., 2021).

Synthetic data has found diverse applications such as multispectral data classification in plastic bottle sorting (Maliks and Kadikis, 2021), and autonomous navigation in unstructured industrial environments using Unreal Engine 4 (Outón et al., 2021). Virtual environments like ViTroVo have been employed for *in vitro* assembly search (Grappiolo et al., 2021) and virtual prototyping has been utilized for detecting modules in modular integrated construction (Zheng et al., 2020).

Various research has used synthetic data generated from discrete-event simulation models to develop hybrid models for aggregate production planning (Sisca et al., 2015) and solve flexible flow-shop scheduling problems (Apornak et al., 2021). The same technique has also helped develop decision support systems for plant-level maintenance (Guner et al., 2016).

In deep learning applications, synthetic data has been used to train convolutional neural networks for quality control (Sikora et al., 2021) and enable deep learning in automotive wiring harness manufacturing (Nguyen et al., 2022). It has also been used in a hierarchical approach for automatic quality inspection in the automotive industry (Rio-Torto et al., 2021) and to train deep learning models in a smart augmented reality instructional system for mechanical assembly (Lai et al., 2020).

While the use of physical simulations for generating synthetic data and optimizing assembly processes requires significant expertise and resources, the approach can lead to considerable improvements in manufacturing and assembly industries.

2.2.2 Numerical modeling

The recent upsurge in generating synthetic data to facilitate machine learning applications in industrial settings represents a promising development. The synthetic data, created by simulating real-world conditions on assembly systems via data-driven methodologies, enables training and evaluation of machine learning models, optimizes their performance, and mitigates the time and cost involved in securing labeled data.

Studies have showcased synthetic data generation's efficacy in enhancing machine learning models' performance by balancing data sets and enriching training data (Han et al., 2019; Kim et al., 2020). Synthetic data sets have also been used to evaluate algorithm and model performance (Laxman et al., 2007; Andres et al., 2021). To complement limited real-world data for model training, a Generative Adversarial Network (GAN) is utilized for synthetic data generation, improving defect detection in models (da Silva et al., 2021).

Synthetic data can optimize human-robot collaborative systems (Sibona and Indri, 2021) by training models to anticipate collision probability and optimize robot motion. Furthermore, synthetic data generation can economize memory usage for stream join operations (Cai et al., 2021), support process improvement and optimization efforts in various industries (Martin et al., 2020), and create balanced data sets for detecting abnormal events in assembly operations (Syafrudin et al., 2018).

Overall, numerical models for generating synthetic data have become instrumental across various research domains. These techniques provide a holistic understanding of complex systems and facilitate training and assessment of machine learning algorithms. Despite their distinct advantages and challenges, both physical simulations and data-driven algorithms are pivotal in generating synthetic data.

3 Data challenges

This study, synthesizing insights from physical simulations and numerical modeling literature, identifies four key challenges in synthetic data generation for manufacturing: Firstly, data scarcity and high costs of collection and labeling, a significant hurdle for small to medium-sized enterprises. Secondly, quality control issues, particularly in ensuring realism and adherence to standards in synthetic data. Thirdly, the imperative of safeguarding sensitive manufacturing data, highlighting the need for secure data connectivity. Lastly, the high cost of collecting substantial real data often becomes a barrier for researchers implementing data-driven models. These challenges, crucial in the context of generating, storing, and analyzing unstructured data, such as 2D and 3D images, underline the importance of addressing them for the successful development and deployment of robust, data-driven industrial applications.

3.1 Data scarcity

The challenge of data scarcity significantly hampers the generation of accurate synthetic data for assembly processes (Sibona and Indri, 2021; Ademujimi and Prabhu, 2022). The problem is often caused by a lack of data on specific production scenarios, unique worker behaviors, or a lack of quality labeled data required to construct supervised learning models. Solutions to this issue may include identifying and prioritizing critical data needs, installing additional sensors, fostering collaboration with other organizations, or using unsupervised learning techniques or transfer learning methods for labeling. By addressing data scarcity, the accuracy of synthetic data can be enhanced to better reflect actual assembly processes, thereby understanding manufacturing operations.

3.2 Data quality

Maintaining data quality is imperative to produce dependable synthetic data for assembly systems, as this data is instrumental in training machine learning models. Factors influencing source data quality include incomplete data, data errors, biases, and data normalization impact (Tayi and Ballou, 1998). Mismanagement of these factors can result in inaccuracies in synthetic data, potentially leading to process control challenges and diminished efficiency. Hence, it is valuable to utilize precise, comprehensive, and representative source data, possibly requiring data cleansing, standardization, and augmentation techniques. Moreover, machine learning models should be resilient to errors and biases to yield accurate and reliable synthetic data. The inherent biases in the data used for training models can significantly influence the outcomes. When data reflects certain predispositions, the resulting models may perpetuate and amplify these biases, leading to skewed or unfair results. It is crucial to recognize and address these biases to ensure the models developed are fair and unbiased.

3.3 Proprietary data

The complexity of creating synthetic data to augment manufacturing processes often increases when data is spread across various departments or companies (Esposito et al., 2016). Entities may hesitate to disclose proprietary information or be constrained by ethical and legal boundaries. To navigate these obstacles, it is beneficial to establish explicit guidelines on data ownership, access, and usage, possibly through data-sharing agreements or ethical directives. Employing methods like data anonymization or differential privacy can help preserve privacy while maintaining the synthetic data quality. Tackling these issues is essential for successfully generating synthetic data that accurately reflects real assembly processes.

3.4 Data costs

The generation of synthetic data in manufacturing can be a significant financial undertaking, requiring extensive computational resources and specialized expertise. This process involves the development of a large data repository through the use of advanced simulation algorithms and high-performance computing, both of which come with substantial operational costs (Koren et al., 1999). Particularly for smaller manufacturing organizations with constrained budgets, these costs may be prohibitive. Furthermore, the financial implications can escalate due to the requirement to store, manage, and process the vast quantities of data produced.

In sum, the data used in model training encompasses a variety of characteristics and patterns. Recognizing these hidden patterns in real data enables the creation of high-quality synthetic data, assisting researchers in developing more data-driven models for targeted analysis and model advancement.

4 Synthetic data generation in assembly: data types and opportunities

To harness the unique characteristics of assembly systems for streamlined manufacturing processes, an exhaustive overview of synthetic data generation is provided in Table 2. This resource enables stakeholders to systematically assess and compare a range of strategies and opportunities. The table summarizes examples, methods, and potential avenues for growth, supported by a comprehensive classification of the collected data types.

4.1 Multidimensional importance of synthetic data

This section reviews the applications of synthetic data in experimental research across various disciplines. As synthetic data

TABLE 2 Examples, techniques, and opportunities for synthetic data generation across various data types in assembly applications.

	Type of data	Example in the assembly station	Synthetic data generation method	Opportunities
Discrete	Binary	Presence of the operator, Station is active or not	Random sampling, SMOTE, ADASYN, ROS, SLSMOTE, BLSMOTE, GMMs	Assessing assembly line uptimes, downtimes, and process reliability
	Image	Inventory images, Operator's image, Final product image	GANs, VAEs, StyleGAN, DBNs, CAE, PixelCNN, Generative Flow Networks	Inventory component identification, operator identity verification, and quality control
	Point cloud	Human positions, Robotic arm equipped with machine vision camera	PointNetGAN, DCGANs, VAE-GAN, VAEs, Autoencoders	Object detection, human-robot collaboration for obstacle detection, and assembly phase classification
Continuous	Time-series	Assembly time, Production count, Biomedical data (EEG data, Wristband data)	TTSGANs, Hidden Markov Model, ARIMA, LSTM, RNN, VAE, CGAN, DCGAN, VAE, GMM, RNNs	Cycle time acquisition, production efficiency tracking, correlation assessment with various data types
	3D Image (RGB + D)	Final product images	GANs, VAEs, CNNs	Quantify the number of assembled parts by the operator and assess quality deficiencies
	Video	Operators' interactions	GANs, RVAE, CVAE	Operator identity verification, operator count determination in the assembly process, operator's emotional state, and object detection

continues to gain prominence across various domains, it delivers significant advantages such as cost reduction, increased accessibility, and reinforced privacy protection, thereby underscoring its pivotal role in contemporary research paradigms. Its utilization is evident in the discovery of causal relationships in manufacturing processes (Marazopoulou et al., 2016), predictive maintenance training in digital twins (Mihai et al., 2021), automotive manufacturing applications (Luckow et al., 2018), anomaly detection (Shetve et al., 2021), and production scheduling optimization (Georgiadis et al., 2022).

Its role in the realm of big data analytics for future smart factories has been highlighted (Gao et al., 2020), along with its potential for improving quality and efficiency in the casting industry (Sun et al., 2021). Other studies focus on the use of synthetic data for scene text recognition (Zhang et al., 2021), performance enhancement of industrial systems (Bécue et al., 2020), object recognition and tracking (Godil et al., 2013), and harmonizing digital twins (Cimino et al., 2021).

The value of synthetic data in exploring CAD models (Ramanujan and Bernstein, 2018), training digital twins (Mihai et al., 2022), and empirical evaluation of optimization algorithms (Rardin and Uzsoy, 2001) has been underscored. Its significance is evident in industrial machine learning applications (Bertolini et al., 2021) and in testing algorithms' ability to discover injective episodes (Achar et al., 2012).

Further, synthetic data proves crucial in the creation of digital twins (Thelen et al., 2022) and implementing deep learning techniques in smart manufacturing (Xu et al., 2022). It is also beneficial in the simulation of blockchain-based digital twins (Suhail et al., 2022) and improving hardware trust and assurance (Botero et al., 2021).

Researchers also propose the use of GANs to generate synthetic data for the Industrial Internet of Things (Qian et al., 2022) and infrastructure maintenance (Mahmoodian et al., 2022). Additionally, synthetic data aids in statistical quality control procedures (Flores et al., 2021), machine creativity (van Doorn et al., 2020), and factor misallocation analysis in firms (Asturias and Rossbach, 2023).

Several studies successfully trained models using existing synthetic data sets or simulated data, negating the need for custom synthetic data generation. The absence of sufficient training data for the synthetic data generation methods used by these researchers raises a number of concerns about the potential quality and validity of the produced data. The prevalent lack of transparency, compounded by potential data quality concerns, creates a widespread need for a standardized framework for synthetic data generation and evaluation.

5 Data generation framework

This section introduces a framework designed for generating synthetic data in assembly processes, crucial for analysis and prediction tasks. Figure 2 provides a visual representation of this framework. The framework details several techniques at each step, assisting researchers in creating high-quality synthetic data for their projects. The process initiates with 1) data collection from embedded sensors in assembly systems, followed by thorough data cleaning to correct errors. This step ensures the integrity and reliability of the initial dataset. Next, 2) data preprocessing techniques are applied to minimize noise and outliers, thereby enhancing the real data's quality. An

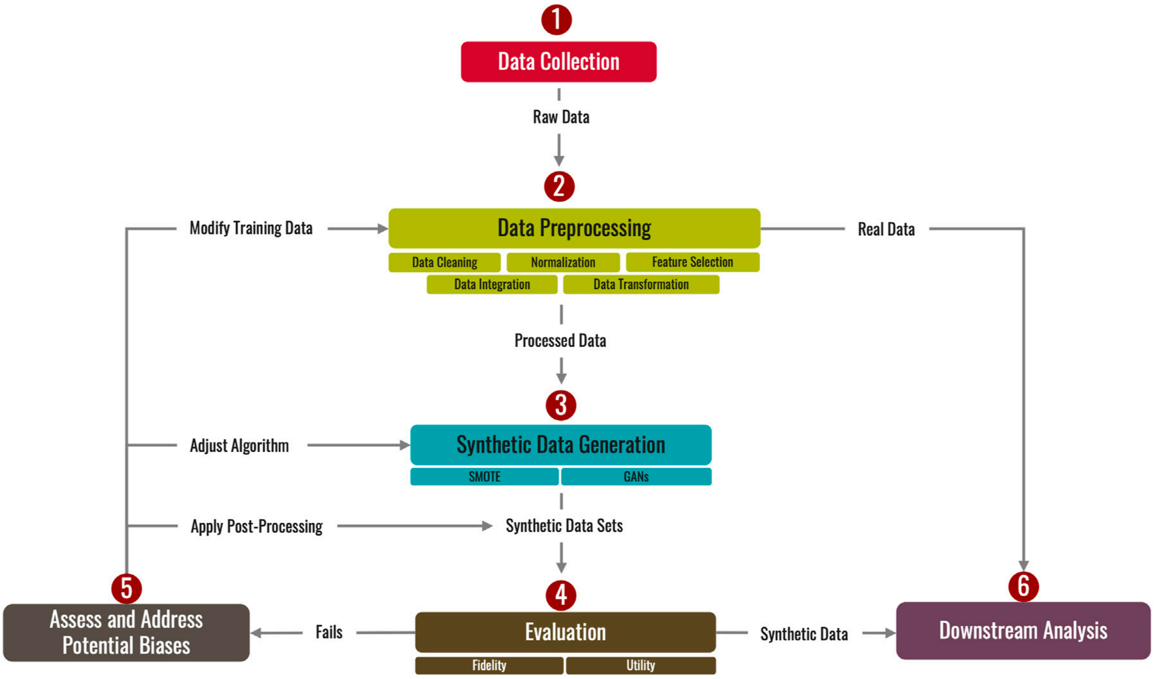


FIGURE 2
Synthetic data generation framework.

example of this is the exploratory data analysis phase, where raw data undergoes cleansing to remove duplicate entries, ensuring dataset uniqueness and accuracy. Subsequently, 3) depending on the target features' types, various synthetic data generation techniques are selected and assessed. For instance, the Generative Adversarial Network (GAN) model is a popular choice in the literature for generating synthetic data points. 4) The next phase involves evaluating the synthetic data's quality, focusing on its realism and similarity to the original data. For example, synthetic images produced by a GAN model are assessed against real images using the Structural Similarity Index (SSI), which quantifies visual similarity. Moreover, 5) the framework includes a validation step to address potential data biases and fairness issues, which are common in many datasets. The aim here is to counteract the amplification of biases or unfairness in the collected dataset. This is particularly important as overrepresentation of certain features or patterns due to training data imbalances can lead to discrepancies in model performance. Such biases can result in misclassifications or inaccurate predictions in real-world applications. Finally, 6) the last step involves preparing and amalgamating both synthetic and real data for downstream analysis. For instance, blending synthetic and real images enriches the dataset, enabling the development of machine learning models with optimal performance. These models benefit from exposure to a broader range of scenarios and conditions, which enhances their generalization capabilities and interpretation of new, unseen data. Iterating through these steps, this framework effectively addresses the challenges in dataset generation for various assembly systems and outlines strategies to enhance the quality of synthetic data.

5.1 Step 1: data collection

The development of synthetic data accurately mirroring real-world assembly data requires careful data source selection, drawing from various inputs such as sensors and cameras. To generate high-quality synthetic data, which closely emulates the original dataset, precision and reliability of the gathered data must be ensured. Considerations such as data collection duration and frequency, along with strategic sensor and equipment placement, significantly impact the accuracy and dependability of the data. Rigorous evaluation of these factors is crucial to yield synthetic data that faithfully represents the original, as overlooking these aspects may introduce errors with potentially detrimental effects on the synthetic data output (Luckow et al., 2018).

5.2 Step 2: data pre-processing

Following data collection in the synthetic data generation framework for assembly systems, the subsequent step is pre-processing. This phase refines and formats the collected data, ensuring its suitability for synthesis, which is pivotal for maintaining data precision and authenticity. Pre-processing methodologies vary based on the nature of the data and typically encompass data cleaning, transformation, normalization, feature selection, and integration. These processes handle the removal of missing or flawed data, conversion into a usable format, standardization for comparative purposes, relevancy-based feature selection, and integration of data from various sources into a comprehensive format for analysis. Applying suitable pre-processing techniques, tailored to the data type and research goals, improves the quality of the synthetic data, ensuring its accuracy and reliability.

5.3 Step 3: synthetic data generation

Leveraging data obtained from assembly systems can furnish useful insights into the efficiency and performance of the manufacturing process. Various data types such as binary data, point-cloud data, biomedical data, image data, time-series data, 3D image data, and video data can be gathered and scrutinized to refine the assembly process.

5.4 Step 4: evaluation

Evaluating the quality and relevance of synthetic data for assembly processes is crucial before its use in analysis or model development to ensure it can reliably replace the original data. This evaluation largely concentrates on two factors: fidelity, indicating how closely the synthetic data resembles the original, and utility, demonstrating the effectiveness of the synthetic data in intended applications.

5.4.1 Fidelity

Fidelity, in the context of synthetic data, refers to the degree to which the synthetic data replicates or mirrors the original, actual data. This is a critical factor, particularly in machine learning applications, where synthetic data is often used to augment training sets or generate additional data sets that are statistically similar to the original ones (Figueira and Vaz, 2022). The importance of fidelity lies in its influence on the performance of models. If synthetic data does not accurately reflect the characteristics of the actual data, models trained on such synthetic data might underperform when tested against the original data. Thus, ensuring high fidelity in synthetic data is essential to maintain model accuracy and robustness. As a practical matter, several techniques may be combined, such as visual examinations, statistical tests, or domain-specific metrics, to assess the authenticity of synthetic data.

When assessing synthetic data generation, the concepts of fidelity and reliability are intricately interconnected. Synthetic data reliability can be assessed through various methods. One approach uses statistical tests, such as Kolmogorov-Smirnov and Anderson-Darling (Santos et al., 2021), to compare synthetic and real data properties, with matching distributions suggesting synthetic data dependability. Another method involves training models with synthetic data and evaluating them using original data, often done in fields like speech recognition or image classification using domain-specific metrics. Lastly, metrics like the F1 score and ROC-AUC score can measure the accuracy of downstream tasks by comparing the performance of models trained on synthetic data with those trained on the original data (Hand and Till, 2001). Equation (1) is employed to determine the Area Under the Curve (AUC) of a classifier specifically designed for binary classification scenarios.

$$\hat{A} = \frac{S_0 - n_0 (n_0 + 1)/2}{n_0 n_1} \tag{1}$$

where n_0 and n_1 indicate the number of positive and negative samples, respectively. $S_0 = \sum r_i$, where r_i is the ist positive example in the ranked list. High congruence in scores indicates trustworthy synthetic data.

5.4.2 Utility

Assessing the utility of synthetic data, its applicability and effectiveness in a specific context, is critical for its successful deployment in applications like training machine learning models (Khan et al., 2022). Various methods can be used to gauge synthetic data's utility, including:

- Similarity metrics: These allow comparisons between synthetic and original data distributions using metrics like mean squared error or correlation coefficient.
- Classification accuracy: Training a classifier on original data and testing it on synthetic data can validate the utility of the synthetic data, provided the accuracy levels are similar.
- Regression error: Analogously, training a regression model on original data and testing it on synthetic data allows for a measure of utility—if the error levels are comparable, the synthetic data's utility is affirmed.
- Clustering: Similar groupings in synthetic and original data following clustering techniques signify comparable data structures and relationships.
- Visual inspection: A subjective assessment of synthetic data can establish whether it retains key features and characteristics of the original data.

The utility of synthetic data can be evaluated using similarity metrics, classification accuracy, regression error, clustering, and visual inspection to ensure its adequacy for the intended purpose.

5.5 Step 5: assessing and addressing potential biases

In addition to fidelity and utility, mitigating potential biases is also crucial for enhancing the quality of synthetic data.

- Modify training data and re-generate synthetic data: By modifying the training data used for synthetic data generation and subsequently re-generating the data, this strategy can help reduce biases present in the original training data, which may have been transferred to the synthetic data. For instance, if synthetic data is biased towards a certain group, broadening the range of examples in the training data can help to reduce this bias.
- Adjust the synthetic data generation algorithm: This strategy involves adjusting the algorithm that generates synthetic data to better align with the target outcome, then re-generating the data. If the synthetic data's fidelity is insufficient, you can fine-tune the algorithm to better mirror the original data. If the utility of the synthetic data is lacking, enhancing the algorithm to incorporate more essential features or variables can increase model accuracy.
- Apply post-processing on synthetic data: This strategy involves modifying the synthetic data post-generation to address any identified issues. Techniques such as data smoothing or imputation can improve the quality, utility, and confidentiality of the synthetic data if its fidelity is found to be inadequate.

5.6 Step 6: downstream analysis

Generating high-quality synthetic data through an iterative evaluation process allows for the leveraging of both synthetic and real datasets in downstream analysis. Incorporating varying ratios of these data types enables the development of more comprehensive machine learning models, expanding exploration of the solution space. Such integration supports the study of foundation models that demand large training datasets, ultimately yielding more accurate and robust results. Increased utilization of synthetic data also mitigates the need for labor-intensive tasks and costly manual data collection processes.

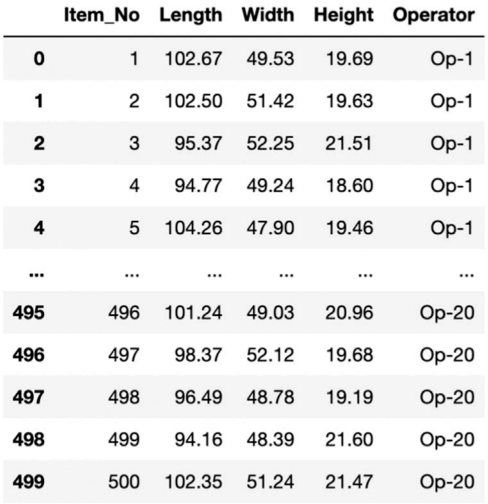
6 Case study

This section presents a use-case study employing high-quality synthetic data, generated through the iterative evaluation process shown in Figure 2, and combines this data with real datasets for subsequent analysis. The integration of synthetic and real data facilitates the training of more comprehensive machine learning models, expanding the scope of predictive capabilities. This combined approach is particularly beneficial for large models that require extensive training data sets, leading to improved accuracy and robustness in the results. Moreover, the increased utilization of synthetic data can alleviate the burden of labor-intensive tasks, thereby reducing the necessity for costly and time-consuming manual data collection processes.

6.1 Dataset

In this proposed case study, the Parts Manufacturing Industry Dataset[1] sourced from Kaggle was utilized. The dataset encompasses data on 500 parts manufactured by 20 operators within a designated timeframe, featuring parameters such as length, width, height, and operator ID, as illustrated in Figure 3. The normal variables of length, width, and height define the dimensions of manufacturing parts designed for packaging purposes. This dataset exemplifies our synthetic data generation framework's ability to classify dimensions of parts for potential anomalies.

[1] https://www.kaggle.com/datasets/gabrielsantello/parts-manufacturing-industry-dataset



	Item_No	Length	Width	Height	Operator
0	1	102.67	49.53	19.69	Op-1
1	2	102.50	51.42	19.63	Op-1
2	3	95.37	52.25	21.51	Op-1
3	4	94.77	49.24	18.60	Op-1
4	5	104.26	47.90	19.46	Op-1
...	...	...	...	...	...
495	496	101.24	49.03	20.96	Op-20
496	497	98.37	52.12	19.68	Op-20
497	498	96.49	48.78	19.19	Op-20
498	499	94.16	48.39	21.60	Op-20
499	500	102.35	51.24	21.47	Op-20

FIGURE 3
Samples of manufacturing parts with varying features.

6.2 Data pre-processing

We conduct an exploratory data analysis to streamline the manufacturing dataset by eliminating unnecessary features and identifying outliers for a subsequent classification task. The initial step involves discarding irrelevant attributes such as "Item_No" and "Operator". We define outliers as data points that fall outside the 95% confidence interval, under the presumption that they significantly deviate from standard manufacturing tolerances. Following this, we categorize each data point as either "perfect" or "defective", transforming our task into a binary classification problem. Figure 4 illustrates the distribution of dimensions, where parts deviating by more than two standard deviations from the mean are marked as defective. This approach effectively pinpoints anomalies and generates labels for further classification analysis.

The predominant class is classified as "perfect," comprising 480 samples (96%), while the 'defective' minority class includes 20 samples (4%). O rectify this, synthetic data generation was employed to even out the class distribution to 50–50. This approach involves understanding the distribution of the minority "defective" class distribution and generating additional samples to enhance the robustness of subsequent analysis.

6.3 Synthetic data generation

Following Figure 2's subsequent step for synthetic data generation, this section employs the techniques outlined in Table 2 for binary classification. Methods like SMOTE, ADASYN, ROS, BLSMOTE, SLSMOTE, and GMM generate synthetic "defective" samples, with SMOTE and ADASYN interpolating to augment the minority class—ADASYN particularly for complex cases. ROSE leveraged diverse resampling methodologies to generate synthetic samples for the minority class. BLSMOTE and SLSMOTE targeted the creation of synthetic samples for borderline instances between minority and majority classes; BLSMOTE concentrated on instances nearer to the majority class, while SLSMOTE focused on those closer to the minority class. Lastly, GMM created synthetic data by fitting the original data to a mixture of Gaussian distributions and sampling from these distributions. Through these methods, a balanced dataset was achieved via synthetic data generation. For this evaluation, fidelity and utility are used to identify the optimal method for generating synthetic data for the manufacturing dataset.

6.4 Fidelity evaluation

To assess the fidelity of various synthetic data generation techniques, accuracy, F-1 score, and ROC-AUC score were employed as performance metrics, as illustrated in Table 3.

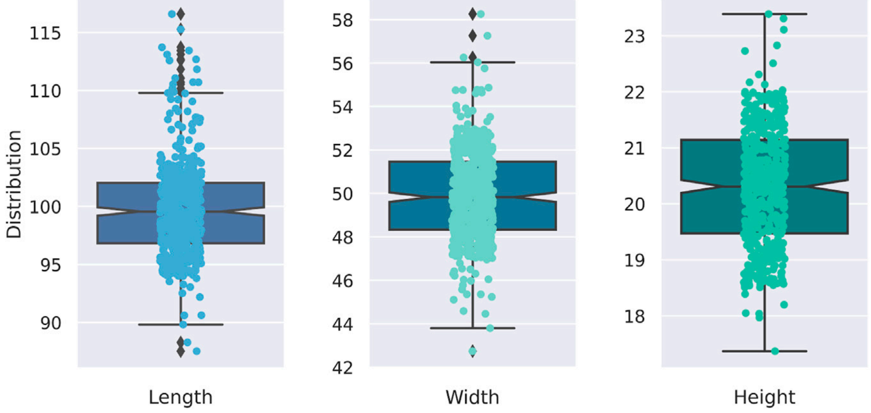


FIGURE 4
Box plot distributions of manufacturing feature dimensions.

TABLE 3 Fidelity comparison.

	Train and test on real data			Train and test on synthetic data		
	Accuracy	F-1 score	ROC-AUC score	Accuracy	F-1 score	ROC-AUC score
SMOTE	0.97	0.98	0.75	0.91	0.95	0.72
ADASYN	0.97	0.98	0.75	0.91	0.95	0.72
ROS	0.97	0.98	0.75	0.91	0.95	0.72
BLSMOTE	0.97	0.98	0.75	0.93	0.97	0.81
SLSMOTE	0.97	0.98	0.75	0.91	0.95	0.72
GMM	**0.97**	**0.98**	**0.75**	**0.95**	**0.95**	**0.96**

TABLE 4 Utility comparison.

	Train and test on real data	Train on hybrid and test on real data	Train and test on hybrid data
Accuracy	0.98	**0.98**	0.94
Precision	0.97	**0.98**	0.92
Recall	0.50	**0.97**	0.96
F1-score	0.67	**0.97**	0.94
ROC-AUC score	0.75	**0.95**	0.94

To conduct a comparative evaluation, a logistic regression model was utilized in two distinct scenarios, with consideration given to its general assumptions. Initially, the model was trained and validated on real data, serving as a baseline. Subsequently, it was trained on synthetic data and tested on real data, to demonstrate the model's performance under these conditions. The best model performance is highlighted in bold.

Our findings indicated that the GMM outperformed other synthetic data generation techniques, indicating high-quality synthetic data that closely resembled real data. GMM's advantage lies in its flexibility to model intricate and non-linear data distributions by combining multiple Gaussian components. Unlike other techniques that rely on simple heuristics or oversampling approaches, GMM leverages a probabilistic modeling framework to learn the data distribution and generate synthetic samples accordingly. This allows GMM to generate data that better represents the original data set, leading to improved performance in terms of accuracy, F-1 score, and ROC-AUC score.

6.5 Utility evaluation

To evaluate the utility of GMM-generated synthetic data for enhancing model precision, a random forest model with stratified 10-fold cross-validation was employed. Three case scenarios were conducted: 1) training and testing solely on real data sets, 2) training on hybrid data sets and testing on real data, and 3) training and testing on hybrid data sets. Model performance was assessed using metrics such as $accuracy$, $precision$, $recall$, $F1 - score$, and $ROC - AUC$ score, as detailed in Table 4. Equations (2)-(5) describes these metrics as follows:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5}$$

where TP (true positive) denotes the number of correctly classified positive instances, TN (true negative) represents the number of correctly classified negative instances, FP (false positive) is the count of negative instances incorrectly classified as positive, and FN (false negative) indicates the number of positive instances that are misclassified as negative.

The results showed that the random forest model, when trained and tested on real data, achieved high accuracy and precision but lower recall, indicating missed positive cases. Training on a hybrid dataset and testing on real data, the model excelled in accuracy, precision, recall, F1-score, and ROC AUC score, showcasing effective generalization and minority class identification. The third scenario, trained and tested on hybrid data, also displayed high performance across all metrics. These outcomes are anticipated as the inclusion of synthetic data, generated from a GMM, adds more information to train the model, naturally enhancing performance. To maximize the benefits, it is advisable to experiment with different ratios of synthetic to real data when training for the most effective model performance.

7 Conclusion

This study presents a systematic review of synthetic data generation techniques used in manufacturing systems, delineating data challenges within the manufacturing domain and corresponding data-driven techniques for creating synthetic data for both discrete and continuous data types. Based on these findings, a framework for generating synthetic data in machine learning applications is proposed, which includes mitigation strategies for data generation, quality evaluation, and bias reduction. The practicality and effectiveness of this framework is tested through a case study that explores its application to imbalanced manufacturing part data sets. A performance comparison, based on fidelity and utility tests, illustrates the impact of incorporating synthetic data alongside real data. Results indicate that models trained on a hybrid data set and tested on real data outperform those trained solely on real data. This study, therefore, expands opportunities for researchers to devise more data-driven approaches in manufacturing applications.

Future work involves testing the framework in more complex manufacturing scenarios, exploring novel synthetic data techniques, and validating the framework with real data from manufacturing assembly systems. Additionally, investigating the potential of combining synthetic and real data to enhance machine learning model performance in assembly applications could significantly benefit manufacturing productivity and product quality.

Author contributions

VB: Data curation, Formal Analysis, Investigation, Validation, Visualization, Writing–original draft. CC: Formal Analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. JC: Conceptualization, Methodology, Project administration, Supervision, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Achar, A., Laxman, S., Viswanathan, R., and Sastry, P. (2012). Discovering injective episodes with general partial orders. *Data Min. Knowl. Discov.* 25, 67–108. doi:10.1007/s10618-011-0233-y

Ademujimi, T., and Prabhu, V. (2022). Digital twin for training bayesian networks for fault diagnostics of manufacturing systems. *Sensors* 22, 1430. doi:10.3390/s22041430

Ameperosa, E., and Bhounsule, P. A. (2020). Domain randomization using deep neural networks for estimating positions of bolts. *J. Comput. Inf. Sci. Eng.* 20, 051006. doi:10.1115/1.4047074

Andres, B., Guzman, E., and Poler, R. (2021). Corrigendum to "a novel milp model for the production, lot sizing, and scheduling of automotive plastic components on parallel flexible injection machines with setup common operators". *Complexity* 2021, 1–17. doi:10.1155/2021/9850964

Apornak, A., Raissi, S., and Pourhassan, M. R. (2021). Solving flexible flow-shop problem using a hybrid multi criteria taguchi based computer simulation model and dea approach. *J. Industrial Syst. Eng.* 13, 264–276.

Asturias, J., and Rossbach, J. (2023). Grouped variation in factor shares: an application to misallocation. *Int. Econ. Rev.* 64, 325–360. doi:10.1111/iere.12605

Bécue, A., Maia, E., Feeken, L., Borchers, P., and Praça, I. (2020). A new concept of digital twin supporting optimization and resilience of factories of the future. *Appl. Sci.* 10, 4482. doi:10.3390/app10134482

Bertolini, M., Mezzogori, D., Neroni, M., and Zammori, F. (2021). Machine learning for industrial applications: a comprehensive literature review. *Expert Syst. Appl.* 175, 114820. doi:10.1016/j.eswa.2021.114820

Biczó, Z., Felde, I., and Szénási, S. (2021). "Distorsion prediction of additive manufacturing process using machine learning methods," in 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 19-21 May 2021, 000249–000252.

Bikes, A., Williams, G., and O'connor, R. (1994). "Assembly systems sensitivity to component delivery: a logistics study using simulation," in Fourth International Conference on Factory 2000-Advanced Factory Automation (IET), 3-5 October, 1994, 638–644.

Blau, P. M., Falbe, C. M., McKinley, W., and Tracy, P. K. (1976). Technology and organization in manufacturing. *Adm. Sci. Q.* 21, 20–40. doi:10.2307/2391876

Botero, U. J., Wilson, R., Lu, H., Rahman, M. T., Mallaiyan, M. A., Ganji, F., et al. (2021). Hardware trust and assurance through reverse engineering: a tutorial and outlook from image analysis and machine learning perspectives. *ACM J. Emerg. Technol. Comput. Syst. (JETC)* 17, 1–53. doi:10.1145/3464959

Cai, W., Bernstein, P. A., Wu, W., and Chandramouli, B. (2021). Optimization of threshold functions over streams. *Proc. VLDB Endow.* 14, 878–889. doi:10.14778/3447689.3447693

Cimino, C., Ferretti, G., and Leva, A. (2021). Harmonising and integrating the digital twins multiverse: a paradigm and a toolset proposal. *Comput. Industry* 132, 103501. doi:10.1016/j.compind.2021.103501

da Silva, L. A., dos Santos, E. M., Araújo, L., Freire, N. S., Vasconcelos, M., Giusti, R., et al. (2021). Spatio-temporal deep learning-based methods for defect detection: an industrial application study case. *Appl. Sci.* 11, 10861. doi:10.3390/app112210861

de la Rosa, F. L., Gómez-Sirvent, J. L., Sánchez-Reolid, R., Morales, R., and Fernández-Caballero, A. (2022). Geometric transformation-based data augmentation on defect classification of segmented images of semiconductor materials using a resnet50 convolutional neural network. *Expert Syst. Appl.* 206, 117731. doi:10.1016/j.eswa.2022.117731

Esposito, C., Castiglione, A., Martini, B., and Choo, K.-K. R. (2016). Cloud manufacturing: security, privacy, and forensic concerns. *IEEE Cloud Comput.* 3, 16–22. doi:10.1109/mcc.2016.79

Fecker, D., Märgner, V., and Fingscheidt, T. (2013). Density-induced oversampling for highly imbalanced datasets. *Image Process. Mach. Vis. Appl. VI* 8661, 211–221. doi:10.1117/12.2003973

Fiasché, M., Ripamonti, G., Sisca, F. G., Taisch, M., and Tavola, G. (2016). "A novel hybrid fuzzy multi-objective linear programming method of aggregate production planning," in *Advances in neural networks: computational intelligence for ICT* (Berlin, Germany: Springer), 489–501.

Figueira, A., and Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and gans. *Mathematics* 10, 2733. doi:10.3390/math10152733

Flores, M., Fernández-Casal, R., Naya, S., and Tarrío-Saavedra, J. (2021). Statistical quality control with the qcr package. *R J.* 13, 194–217. doi:10.32614/rj-2021-034

Gao, R. X., Wang, L., Helu, M., and Teti, R. (2020). Big data analytics for smart factories of the future. *CIRP Ann.* 69, 668–692. doi:10.1016/j.cirp.2020.05.002

Georgiadis, K., Nizamis, A., Vafeiadis, T., Ioannidis, D., and Tzovaras, D. (2022). Production scheduling optimization enabled by digital cognitive platform. *Procedia Comput. Sci.* 204, 424–431. doi:10.1016/j.procs.2022.08.052

Godil, A., Eastman, R., and Hong, T. (2013). *Ground truth systems for object recognition and tracking.* Gaithersburg, MA, USA: National Institute of Standards and Technology.

Grappiolo, C., Pruim, R., Faeth, M., and de Heer, P. (2021). Vitrovo: *in vitro* assembly search for *in vivo* adaptive operator guidance: an artificial intelligence framework for highly customised manufacturing. *Int. J. Adv. Manuf. Technol.* 117, 3873–3893. doi:10.1007/s00170-021-07824-7

Guner, H. U., Chinnam, R. B., and Murat, A. (2016). Simulation platform for anticipative plant-level maintenance decision support system. *Int. J. Prod. Res.* 54, 1785–1803. doi:10.1080/00207543.2015.1064179

Han, S., Choi, H.-J., Choi, S.-K., and Oh, J.-S. (2019). Fault diagnosis of planetary gear carrier packs: a class imbalance and multiclass classification problem. *Int. J. Precis. Eng. Manuf.* 20, 167–179. doi:10.1007/s12541-019-00082-4

Hand, D. J., and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.* 45, 171–186. doi:10.1023/a:1010920819831

Jain, S., Narayanan, A., and Lee, Y.-T. T. (2018). "Comparison of data analytics approaches using simulation," in 2018 Winter Simulation Conference (WSC) (IEEE), Gothenburg Sweden, December 9 - 12, 2018.

Khan, M. S. N., Reje, N., and Buchegger, S. (2022). Utility assessment of synthetic data generation methods. arXiv preprint arXiv:2211.14428

Kim, S. W., Lee, Y. G., Tama, B. A., and Lee, S. (2020). Reliability-enhanced camera lens module classification using semi-supervised regression method. *Appl. Sci.* 10, 3832. doi:10.3390/app10113832

Kohtala, S., and Steinert, M. (2021). Leveraging synthetic data from cad models for training object detection models–a vr industry application case. *Procedia CIRP* 100, 714–719. doi:10.1016/j.procir.2021.05.092

Koren, Y., Heisel, U., Jovane, F., Moriwaki, T., Pritschow, G., Ulsoy, G., et al. (1999). Reconfigurable manufacturing systems. *CIRP Ann.* 48, 527–540. doi:10.1016/s0007-8506(07)63232-6

Lai, Z.-H., Tao, W., Leu, M. C., and Yin, Z. (2020). Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. *J. Manuf. Syst.* 55, 69–81. doi:10.1016/j.jmsy.2020.02.010

Laxman, S., Sastry, P., and Unnikrishnan, K. (2007). Discovering frequent generalized episodes when events persist for different durations. *IEEE Trans. Knowl. Data Eng.* 19, 1188–1201. doi:10.1109/tkde.2007.1055

Luckow, A., Kennedy, K., Ziolkowski, M., Djerekarov, E., Cook, M., Duffy, E., et al. (2018). "Artificial intelligence and deep learning applications for automotive manufacturing," in 2018 IEEE International Conference on Big Data (Big Data) (IEEE), Seattle, WA, USA, Dec. 10 2018 to Dec. 13 2018, 3144–3152.

Mahmoodian, M., Shahrivar, F., Setunge, S., and Mazaheri, S. (2022). Development of digital twin for intelligent maintenance of civil infrastructure. *Sustainability* 14, 8664. doi:10.3390/su14148664

Malekzadeh, M., Clegg, R. G., and Haddadi, H. (2017). Replacement autoencoder: a privacy-preserving algorithm for sensory data analysis. arXiv preprint arXiv:1710.06564

Maliks, R., and Kadikis, R. (2021). "Multispectral data classification with deep cnn for plastic bottle sorting," in 2021 6th International Conference on Mechanical Engineering and Robotics Research (ICMERR) (IEEE), Krakow, Poland, December 11-13, 2021, 58–65.

Marazopoulou, K., Ghosh, R., Lade, P., and Jensen, D. (2016). Causal discovery for manufacturing domains. *arXiv preprint arXiv:1605.04056*

Martin, N., Depaire, B., Caris, A., and Schepers, D. (2020). Retrieving the resource availability calendars of a process from an event log. *Inf. Syst.* 88, 101463. doi:10.1016/j.is.2019.101463

Mihai, S., Davis, W., Hung, D. V., Trestian, R., Karamanoglu, M., Barn, B., et al. (2021). "A digital twin framework for predictive maintenance in industry 4.0," in 2020 International Conference on High Performance Computing and Simulation, Barcelona, Spain, December 10-14, 2020.

Mihai, S., Yaqoob, M., Hung, D. V., Davis, W., Towakel, P., Raza, M., et al. (2022). Digital twins: a survey on enabling technologies, challenges, trends and future prospects. *IEEE Commun. Surv. Tutorials* 24, 2255–2291. doi:10.1109/comst.2022.3208773

Mubarak, A.-A., Mohamed, Y., and Bouferguene, A. (2020). Application of industrial pipelines data generator in the experimental analysis: pipe spooling optimization problem definition, formulation, and testing. *Adv. Eng. Inf.* 43, 101007. doi:10.1016/j.aei.2019.101007

Nguyen, H. G., Habiboglu, R., and Franke, J. (2022). Enabling deep learning using synthetic data: a case study for the automotive wiring harness manufacturing. *Procedia CIRP* 107, 1263–1268. doi:10.1016/j.procir.2022.05.142

Outón, J. L., Merino, I., Villaverde, I., Ibarguren, A., Herrero, H., Daelman, P., et al. (2021). A real application of an autonomous industrial mobile manipulator within industrial context. *Electronics* 10, 1276. doi:10.3390/electronics10111276

Qian, C., Yu, W., Lu, C., Griffith, D., and Golmie, N. (2022). Toward generative adversarial networks for the industrial internet of things. *IEEE Internet Things J.* 9, 19147–19159. doi:10.1109/jiot.2022.3163894

Ramanujan, D., and Bernstein, W. Z. (2018). Vesper: visual exploration of similarity and performance metrics for computer-aided design repositories. *Int. Manuf. Sci. Eng. Conf.* 51371, V003T02A034. doi:10.1115/MSEC2018-6527

Rardin, R. L., and Uzsoy, R. (2001). Experimental evaluation of heuristic optimization algorithms: a tutorial. *J. Heuristics* 7, 261–304. doi:10.1023/a:1011319115230

Research, G. V. (2023). Synthetic data generation market size and growth report. Technical Report. Available at: https://www.grandviewresearch.com/industry-analysis/synthetic-data-generation-market-report.

Rio-Torto, I., Campaniço, A. T., Pereira, A., Teixeira, L. F., and Filipe, V. (2021). "Automatic quality inspection in the automotive industry: a hierarchical approach using simulated data," in 2021 IEEE 8th International Conference on Industrial Engineering and Applications (ICIEA) (IEEE), April 23–26, 2021, 342–347.

Santos, M. C., Borges, A. I., Carneiro, D. R., and Ferreira, F. J. (2021). "Synthetic dataset to study breaks in the consumer's water consumption patterns," in Proceedings of the 2021 4th International Conference on Mathematics and Statistics, Paris France, June 24 - 26, 2021, 59–65.

Shetve, D., VaraPrasad, R., Trestian, R., Nguyen, H. X., and Venkataraman, H. (2021). Cats: cluster-aided two-step approach for anomaly detection in smart manufacturing. *Adv. Comput. Netw. Commun.* 2, 103–115. doi:10.1007/978-981-33-6987-0_9

Shrawanty, Y. R. (2022). Synthetic data generation market research. *Tech. Rep.*

Sibona, F., and Indri, M. (2021). "Data-driven framework to improve collaborative human-robot flexible manufacturing applications," in IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society (IEEE), Toronto, Ontario, Canada, 13-16 October 2021.

Sikora, J., Wagnerová, R., Landryová, L., Šíma, J., and Wrona, S. (2021). Influence of environmental noise on quality control of hvac devices based on convolutional neural network. *Appl. Sci.* 11, 7484. doi:10.3390/app11167484

Singh, S. K., Chakrabarti, S. K., and Jayagopi, D. B. (2020). "Automated testing of refreshable braille display," in Human-Centric Computing in a Data-Driven Society: 14th IFIP TC 9 International Conference on Human Choice and Computers, HCC14 2020, Tokyo, Japan, September 9–11, 2020, 181–192.

Sisca, F. G., Fiasché, M., and Taisch, M. (2015). "A novel hybrid modelling for aggregate production planning in a reconfigurable assembly unit for optoelectronics," in Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, 571–582.

Suhail, S., Hussain, R., Jurdak, R., Oracevic, A., Salah, K., Hong, C. S., et al. (2022). Blockchain-based digital twins: research trends, issues, and future challenges. *ACM Comput. Surv. (CSUR)* 54, 1–34. doi:10.1145/3517189

Sun, N., Kopper, A., Karkare, R., Paffenroth, R. C., and Apelian, D. (2021). Machine learning pathway for harnessing knowledge and data in material processing. *Int. J. Metalcasting* 15, 398–410. doi:10.1007/s40962-020-00506-2

Syafrudin, M., Fitriyani, N. L., Alfian, G., and Rhee, J. (2018). An affordable fast early warning system for edge computing in assembly line. *Appl. Sci.* 9, 84. doi:10.3390/app9010084

Tao, F., Qi, Q., Liu, A., and Kusiak, A. (2018). Data-driven smart manufacturing. *J. Manuf. Syst.* 48, 157–169. doi:10.1016/j.jmsy.2018.01.006

Tayi, G. K., and Ballou, D. P. (1998). Examining data quality. *Commun. ACM* 41, 54–57. doi:10.1145/269012.269021

Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B. D., et al. (2022). A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies. *Struct. Multidiscip. Optim.* 65, 354. doi:10.1007/s00158-022-03425-4

van Doorn, M., Duivestein, S., Mamtani, D., and Pepping, T. (2020). Infinite machine creativity. Available at: https://labs.sogeti.com/research-topics/infinite-machine-creativity/.

Whitley, E. (2022). *Manufacturers must make data-driven decisions: 6 reasons why.*

Xu, J., Kovatsch, M., Mattern, D., Mazza, F., Harasic, M., Paschke, A., et al. (2022). A review on ai for smart manufacturing: deep learning challenges and solutions. *Appl. Sci.* 12, 8239. doi:10.3390/app12168239

Zhang, Z., Pan, L., Du, L., Li, Q., and Lu, N. (2021). "Catnet: scene text recognition guided by concatenating augmented text features," in Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, 350–365.

Zheng, Z., Zhang, Z., and Pan, W. (2020). Virtual prototyping-and transfer learning-enabled module detection for modular integrated construction. *Automation Constr.* 120, 103387. doi:10.1016/j.autcon.2020.103387