UTILIZING SYNTHETIC DATA GENERATION TECHNIQUES TO IMPROVE THE AVAILABILITY OF DATA IN DISCRETE MANUFACTURING FOR AI APPLICATIONS: A REVIEW AND FRAMEWORK

by

VISHNUPRIYA BUGGINENI

(Under the Direction of Jaime Andres Camelio)

Abstract

Acquiring high-quality data for AI applications in the manufacturing sector is challenging due to the complex and proprietary nature of manufacturing processes, as well as the cost of gathering real-world data, resulting in data scarcity. Synthetic data generation offers a promising solution to this issue by creating artificial datasets for AI model training and testing. This study investigates various synthetic data generation methods for manufacturing assembly lines and introduces a comprehensive framework for producing and validating synthetic datasets. The proposed framework consists of four stages: data collection, preprocessing, synthetic data generation, and validation. Through the case study, it was found that synthetic data can significantly improve model performance on imbalanced datasets for assembly line processes. The study concludes that the proposed framework for synthetic data generation can be a valuable resource for researchers seeking to generate synthetic data and conduct studies on assembly line processes.

Index words: Synthetic Data, Simulations, Data-Driven, Framework, Evaluation

UTILIZING SYNTHETIC DATA GENERATION TECHNIQUES TO IMPROVE THE AVAILABILITY OF DATA IN DISCRETE MANUFACTURING FOR AI APPLICATIONS: A REVIEW AND FRAMEWORK

by

VISHNUPRIYA BUGGINENI

B.Tech., Amity University, India, 2021

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

Athens, Georgia

2023

©2023

Vishnupriya Buggineni

All Rights Reserved

UTILIZING SYNTHETIC DATA GENERATION TECHNIQUES TO IMPROVE THE AVAILABILITY OF DATA IN DISCRETE MANUFACTURING FOR AI APPLICATIONS: A REVIEW AND FRAMEWORK

by

VISHNUPRIYA BUGGINENI

Approved:

Major Professor:	Jaime Andres Camelio
Committee:	Khaled Rasheed Beshoy Morkos

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2023

Acknowledgments

My heart is overflowing with thankfulness for everyone who have made this trip of mine memorable and instructive.

The foundation of my research path has been my advisor's consistent support and direction. His helpful criticism and rigorous attention to detail assisted me in shaping my thoughts and achieving my research objectives. I attribute my current level of success to his great guidance.

I am deeply grateful to the members of my committee for their insightful comments, valuable feedback, and constructive criticism. Their different viewpoints and experience were important in enhancing the quality of my study and refining my thoughts.

For my study, the Institute of Artificial Intelligence at the University of Georgia offered a vibrant and supportive atmosphere. The institute's resources and opportunities have been vital to my academic development and success.

I would also want to express my gratitude to my classmates and lab mates for their support and companionship. Their enthusiasm for education and varied backgrounds have widened my viewpoint and encouraged me to investigate new study possibilities.

My family and friends have provided me with unfailing love, encouragement, and faith in my skills. Their confidence in my abilities has propelled me to this success.

Thank you to each and everyone who are part of my journey from the deep down of my heart.

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Challenges 4

1.3 Knowledge Gap 6

1.4 State-of-art solutions 7

1.5 Organization of this thesis 9

2 Background **10**

2.1 Systematic Review Search Process 10

2.2 Synthetic data generation methods 12

2.3 Application of synthetic data 30

3 Comparative Study **40**

3.1 Manufacturing Challenges 41

3.2 Data Challenges 60

4 Research Design **69**

4.1 Framework for Synthetic data generation 69

4.2 Synthetic Data Generation for Different Types of Data on Assembly Lines and Opportunities 101

5 Case study **110**

5.1 Data Collection 110

5.2 Pre-processing 111

5.3 Synthetic data generation 112

5.4 Evaluation 113

6 Future Scope and Conclusion **116**

6.1 Future Scope 116

6.2 Conclusion 117

Bibliography **129**

List of Figures

2.1 Systematic Review Search Process 11

4.1 Data generation framework 70

5.1 Data and its Features 111

5.2 Outlier detection 112

5.3 Data distribution 112

List of Tables

1.1 Research questions 8

2.1 Literature on Synthetic data generation 13

4.1 Synthetic data generation techniques, examples, and opportunities for various data types on assembly line 102

5.1 Fidelity comparison 113

5.2 Utility Comparison 114

Chapter 1

Introduction

Modern manufacturing has undergone a remarkable transformation, thanks to the advent of assembly lines[1]. From automobiles to smartphones, these systems have revolutionized production by seamlessly fitting together every component with clockwork precision to create a flawless end product. However, optimizing assembly line procedures is a delicate balancing act that requires engineers and researchers to navigate complex tradeoffs between speed, accuracy, and quality control[2]. Significant advancements in increasing production, decreasing waste, and improving product quality have been accomplished through years of devoted research and development. Despite these advances, there are still significant challenges associated with creating and testing new assembly line models and algorithms. One of the major obstacles is the need to obtain and utilize actual production data, which is essential for effective optimization[3]. As such, ongoing research in this area remains critical to ensure the continued success of modern manufacturing.

1.1 Motivation

Efficient and rapid production of intricate products heavily relies on assembly line systems, but the optimization of these systems requires precise and dependable data[3]. Unfortunately, practical scenarios present numerous challenges in acquiring such data, including high costs, data privacy concerns, and proprietary information constraints[4]. As a result, manufacturers must seek innovative solutions to address these challenges and ensure the effective optimization of assembly line procedures.

However, the challenges of data scarcity and its proprietary nature are significant obstacles for developers and academic researchers in the manufacturing industry when it comes to running machine learning models and conducting data analysis. In addition, assembly line manufacturers are also confronted with a significant challenge of inadequate data, particularly for small and medium-sized businesses with limited capacity to gather extensive volumes of data. Obtaining enough authentic real-world data can be an arduous and costly process, which restricts the ability to optimize assembly line procedures effectively[4]. Furthermore, even when data is available, it may be restricted in breadth or not fully representative of the range of operating circumstances, which can limit the performance of assembly line systems[4]. Incomplete or inaccurate data can be a potential issue due to factors like sample selection or measurement errors, leading to data bias. Hence, it becomes crucial to consider alternative solutions that can overcome these challenges and gather the required data for efficient and effective assembly line optimization.

Additionally, it is important to consider that data obtained from real-world scenarios may be limited by proprietary restrictions, preventing sharing with external entities or utilization for research and development purposes[5]. This can hinder collaboration and impede innovation within the manufacturing industry. Finding other approaches is therefore important to get beyond these limitations and promote increased sharing and cooperation between

industry players. The industrial sector may gain from increased productivity, creativity, and competitiveness by encouraging more data exchange and cooperation.

Over the past few years, there has been an increasing focus on leveraging synthetic data as a means of addressing these challenges and enhancing the efficacy of assembly line systems[5]. Synthetic data refers to data that is artificially generated to replicate the statistical characteristics of actual data. In the realm of manufacturing assembly lines, synthetic data can be utilized to complement or substitute real-world data, particularly in instances where genuine data is limited or deficient[6].

With the advancements in machine learning and artificial intelligence, generating synthetic data has become an alternative solution that can replace the costly processes of data collection and storage. The ability to create synthetic data at a high scale and accuracy level has enabled the creation of realistic digital copies of manufacturing assembly lines. This flexibility facilitates the development of data sets that represent diverse circumstances, including those that may be arduous or expensive to obtain through real-world data. The digital replicas can be used to simulate various scenarios and identify optimization areas to enhance productivity and efficiency within the manufacturing industry[7].

The use of synthetic data in industrial assembly line systems has the potential to completely transform the sector by offering a quick and affordable way to analyze data, improve workflows, and improve collaboration. Synthetic data is not subject to the same proprietary restrictions as real-world data, enabling easy sharing between industry players, which in turn fosters innovation and enhances the competitiveness of the manufacturing industry. Therefore, the incorporation of synthetic data in manufacturing assembly line systems can have far-reaching implications for the industry, driving productivity and efficiency improvements while also promoting collaboration and knowledge sharing.

This thesis aims to examine the existing literature on synthetic data generation in manufacturing assembly lines and present a framework for producing synthetic data that can

replicate various production scenarios. This approach allows researchers and engineers to assess and optimize assembly line procedures in a secure, regulated, and effective manner.

The objective of this study is to demonstrate the effectiveness of synthetic data generation as a robust tool for designing and testing machine learning algorithms and models for manufacturing assembly lines, while also overcoming issues arising from data scarcity and sensitivity. By offering a method for producing high-quality data, we strive to make a valuable contribution towards developing more productive, secure, and dependable assembly line processes.

1.2 Challenges

Synthetic data has gained prominence as a valuable tool across several industries, including manufacturing. This kind of data is produced by algorithms that create artificial data that resembles actual data. Synthetic data may replace costly and time-consuming physical tests in the manufacturing industry when it comes to testing and verifying processes. However, generating synthetic data for manufacturing assembly lines is not without its challenges. These challenges need to be addressed to ensure the accuracy and usefulness of the data. This discussion will explore the various hurdles encountered when generating synthetic data for manufacturing assembly lines.

1. Complexity: Manufacturing assembly lines are intricate and multi-layered systems that comprise numerous components and connections between production stages. The interactions among these components and stages are highly dynamic and non-linear, posing a challenge in capturing all variations and interactions in a synthetic dataset. Considering the complexity of these assembly lines, it is essential to use advanced algorithms that can precisely model how each step of the procedure behaves[5]. It is still difficult to accurately

represent all the nuances of the manufacturing process, even with sophisticated modeling approaches.

2. Realism: To ensure that synthetic data is an accurate representation of real-world data, it must mimic the characteristics of the actual manufacturing processes. This requires a profound comprehension of the underlying physics and mechanics of the process to generate realistic synthetic data[4]. In generating synthetic data, the simulation models used must rely on precise and dependable data sources such as sensor readings, production logs, and other operational data. Failure to access these data sources can result in synthetic data that does not accurately depict the behavior of the manufacturing process.

3. Data volume: Manufacturing assembly lines produce an enormous volume of data comprising sensor readings, production logs, and other operational data. Generating synthetic data that effectively captures the diversity and variability of the process can be quite challenging[3]. Synthetic data must be produced at scale to facilitate the effective testing and validation of manufacturing processes. However, generating large volumes of synthetic data can be quite resource-intensive, necessitating substantial computational resources.

4. Cost: Generating synthetic data demands a considerable amount of computational resources and expertise, which can be expensive for several manufacturing organizations[5]. The generation of synthetic data necessitates advanced simulation algorithms, high-performance computing resources, and skilled data scientists to develop and validate the models. These resources can be costly, rendering them prohibitive for many organizations, particularly smaller manufacturers with restricted budgets.

5. Legal and ethical considerations: The generation of synthetic data can potentially infringe on intellectual property rights or personal privacy, highlighting the need to guarantee that synthetic data produced on manufacturing assembly lines complies with legal and ethical regulations[3]. Synthetic data should exclude any proprietary or confidential information that could jeopardize a company's competitive edge. Moreover, synthetic data generated

from manufacturing processes should not expose any personal information about employees or customers. It is incumbent upon manufacturers to recognize these legal and ethical considerations and guarantee that synthetic data is generated and utilized appropriately.

1.3 Knowledge Gap

The field of synthetic data generation for assembly lines is constantly developing and has the potential to revolutionize manufacturing. Significant gains in sustainability, productivity, and efficiency may be made by producing a sizable amount of realistic synthetic data. Then, these revelations may be utilized to solve actual-world issues. Notwithstanding the benefits, there are still a lot of knowledge gaps that need to be addressed in order to achieve these goals. In this regard, some of the knowledge gaps in synthetic data generation for assembly lines are discussed below.

1. Generating synthetic data that can effectively capture the complexity and variability of assembly line production processes requires accurate and realistic models. However, the current models have limitations in accurately representing real-world production processes, which poses a significant gap in knowledge. Addressing this gap is crucial to developing more effective synthetic data generation methods.

2. Human operators play a critical role in assembly line production processes, and incorporating their behavior and physiology into synthetic data generation is important. However, there is limited research on accurately modeling human factors in synthetic data generation, which is a significant gap in knowledge. Understanding the behavior and physiology of human operators and incorporating this knowledge into synthetic data generation models can help generate more realistic data that better represents actual assembly line production processes.

3. The validation and testing of synthetic data is essential to ensure that it accurately represents the underlying data distribution. Comparing and contrasting various synthetic data production techniques can be challenging since there are presently no standardized validation and testing procedures for synthetic data. For the purpose of creating more dependable and precise synthetic data creation methods, this offers a huge knowledge gap that has to be filled.

4. Synthetic data must be impartial and accurately represent the distribution of the underlying data to be effective in assembly line production processes. However, there is limited research on identifying and mitigating biases in synthetic data generation, which is a significant gap in knowledge. Developing techniques to reduce potential biases in synthetic data generation methods can help generate more accurate and reliable synthetic data.

5. Physiological data, such as heart rate, breathing, or electrodermal activity, can provide valuable insights into human operator behavior and improve decision-making and prediction models in assembly line production processes. However, the accurate modeling of physiological responses in synthetic data generation is still a significant gap in knowledge. Understanding the complex interactions between physiological responses and assembly line production processes can help develop more sophisticated synthetic data generation models that better represent real-world scenarios. Table 1.1 summarizes the research questions that need to be addressed to fill the knowledge gaps in synthetic data generation for assembly lines.

1.4 State-of-art solutions

Manufacturing researchers are continually exploring ways to generate realistic synthetic data for machine learning model training. Physical simulations are one state-of-the-art method for generating synthetic data by creating a virtual environment that simulates real-

Research Questions
What are potential approaches for generating synthetic data in manufacturing?
What is an effective framework design for generating, validating, and testing synthetic data sets for Industry 4.0 practices?
How can the effectiveness of generated synthetic data for downstream tasks be evaluated through a case study?

Table 1.1: Research questions

world scenarios, but accurately simulating the complexity and diversity of assembly lines remains challenging.

Data augmentation, which modifies existing data to create new variations, is another approach for generating synthetic data. However, it may not be sufficient for generating highly diverse and complex data.

Generative Adversarial Networks (GANs) are widely used to generate synthetic data by learning statistical patterns of the source data to create synthetic data that closely resemble the original data. GANs have been successfully applied to synthesize images and videos in assembly line settings.

Variational Autoencoders (VAEs) are a deep learning model that can generate synthetic data by learning the underlying latent space of the data. VAEs can produce data that is similar but not identical to the original data, making them useful for data augmentation in assembly line quality control.

Transfer learning is another method for generating synthetic data by fine-tuning pre-trained models on similar datasets. When there are few data sources for a particular assembly line, this method might be helpful.

The generation of synthetic data using these cutting-edge approaches is promising, but each methodology has advantages and disadvantages of its own. Because of this, it is essential that researchers and assembly line makers thoroughly assess each method and select the one that would best suit their unique requirements. In an assembly line scenario, doing so can aid in producing high-quality synthetic data that will enhance the performance of machine learning models.

1.5 Organization of this thesis

The organization of this thesis is as follows. Chapter 1 provides an introduction to the challenges of data scarcity and bias in assembly line optimization, and highlights the need for alternative solutions such as synthetic data generation. Chapter 2 discusses the systematic review process and the current state-of-the-art in synthetic data generation literature. Chapter 3 presents an application of manufacturing where synthetic data is used, and discusses data challenges related to synthetic data. Chapter 4 describes a detailed framework for synthetic data generation on assembly lines, including various types of data that can be collected and synthetic data techniques that can be used for them. This chapter also explores opportunities for analyzing synthetic data. Chapter 5 demonstrates a case study using the proposed framework, and provides a detailed analysis of the results. Finally, Chapter 6 concludes the thesis and suggests directions for future research in the field of synthetic data generation in manufacturing.

Chapter 2

Background

2.1 Systematic Review Search Process

The main focus of this research paper is to investigate the application of synthetic data generation in the context of assembly line manufacturing. Assembly line manufacturing involves the production of products through a series of repetitive tasks performed by workers or machines and is a critical component of many manufacturing industries. By modeling and optimizing various production situations, it may be able to increase the efficiency and accuracy of assembly line operations by generating synthetic data. In applications involving machine learning and artificial intelligence, synthetic data is widely used to train algorithms and improve performance.

To achieve my objective, I utilized Google Scholar as a repository to explore appropriate papers. Initially, I employed the keywords "Synthetic Data" AND "Generation" AND ("Manufacturing" OR "Production") which generated 35,400 articles. Nonetheless, to customize the search specifically to the topic of manufacturing, I eliminated all articles unrelated to it, resulting in 10,600 articles as shown in Figure 2.1.

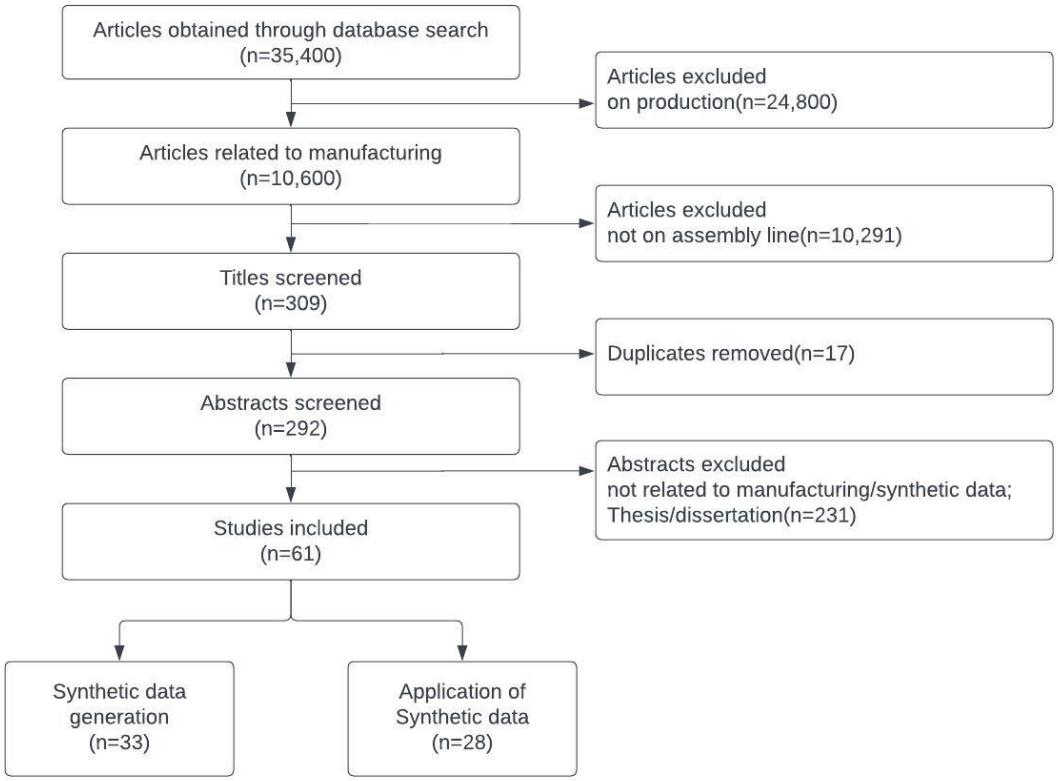

Figure 2.1: Systematic Review Search Process

When manufacturing goods, machinery and technology are utilized to transform raw materials into completed products. Before the final product is created, the manufacturing process typically involves several stages such as design, engineering, prototyping, and testing. Production, on the other hand, encompasses all actions related to creating goods or services and is a more comprehensive term than manufacturing[8]. By filtering out articles related to production, the search results were refined, and only articles most relevant to the research topic and offering valuable insights and information concerning the use of synthetic data in assembly line manufacturing were obtained.

To further narrow down the results, the keyword "Assembly Line" was used, resulting in 309 articles that were relevant to assembly line manufacturing. The titles of these articles were screened, and any duplicates were removed, leaving 292 articles. The abstracts of the

remaining articles were then screened, and any articles that were not related to manufacturing/synthetic data or were thesis/dissertation were excluded. The study identified 61 papers pertinent to the subject matter, which were classified into two groups: those that produced synthetic data (33 papers), and those that employed synthetic data in their investigations (28 papers). The former category of papers was centered on generating simulated datasets that imitated real-world data attributes, while the latter category of papers utilized synthetic data for experimentation and analysis. These two classifications serve as a valuable framework for comprehending the diverse applications of synthetic data in research.

2.2 Synthetic data generation methods

There are two primary methods used in scientific research to create synthetic data: physical simulations and data-driven algorithms.

2.2.1 Physical Simulations

In manufacturing and assembly line industries, the use of physical simulations has become increasingly popular for generating synthetic data. Physical simulations employ mathematical models to replicate real-world phenomena. This enables researchers to generate synthetic data that perfectly mimics the behavior of the system under investigation.

Assembly lines are intricate systems that require the interaction of various components and processes. Physical simulations can model these interactions and create synthetic data that accurately reflects the system's behavior. Through simulations with various parameters and configurations, researchers can produce a synthetic dataset that can be utilized to test and enhance assembly line processes.

Moreover, physical simulations can aid in enhancing the performance and efficiency of assembly line processes. By simulating various assembly line configurations and parameters,

	Type of data	Synthetic data generation method	Application of manufacturing
Discrete	Binary	ADASYN[9],Random Sampling[10],Digital Twin Simulation[11], GMM[12], SMOTE[13]	Quality Control(Fault Detection)[9, 10, 11, 13],Process monitoring[12]
	Point-cloud	IPS cable simulation and Blender[14]	Quality Control(Automated optical inspection)[14]
	2D coordinates	GANs[15]	Human Robot Collaboration[15]
	Multi-class	Simufact additive software[16]	Quality Control[16]
Continuous	Image	Unity 3D and Revit software[17], Unity 3D and CAD models[18, 19], Domain Roandomization[20], Geometric transformations[21], By varying levels of environmental noise[22], Through rotating and modifying the colors of the images[23], GANs[24], and Blender software and Domain randomization[25],Blender software [26]	Process Optimization[17],Increasing Productivity[18], Quality Control[21, 22, 25, 26], Production Monitoring[19], Operator Guidance[20], Braille Display[23], Industrial Internet of Things[24]
	Time-series	By varying parameters[27, 28], GANs[29], Promodel-PC simulation[30], Simul8 simulation[31], SIMIO simulation[32, 33], Hidden Markov Models[5], Simpy Library[34], Virtual Factory Prototype[35], Taguchi simulation[36], and Wiener Process, Gaussian Noise and by varying standard deviation[37]	Quality Control(Fault detection)[27],Production Scheduling[28],Activity Recognition[29], Component Delivery[30], Preventive Maintenance[31], Production Planning[32, 33], Pipe-Spooling[5], Quality Control[34], Cycle-Time estimation[35], Process Optimization[36], Stream Processing[37]
	Video	GANs[38]	Defect detection[38]
	3D image	Ksim9[39], Unreal Engine4[40]	Quality Control[39], Autonomous industrial mobile manipulator[40]

Table 2.1: Literature on Synthetic data generation

researchers can determine the most efficient and effective processes for manufacturing a specific product, resulting in increased productivity and decreased waste.

There is a good amount of research on synthetic data generation using physical simulations. Such as Ademujimi & Prabhu (2022) discussed the use of physical simulation to train Bayesian networks for fault diagnostics in manufacturing systems[11]. The authors propose a digital twin model that simulates the behavior of a manufacturing system, including its sensors, actuators, and control systems. The digital twin is then used to generate synthetic data that captures the behavior of the system under various operating conditions and fault scenarios. The authors prove the efficacy of their method by utilizing synthetic data generated by the digital twin to train a Bayesian network. The network can accurately detect faults in the manufacturing system, even when working under conditions that were not previously present in the training data. Additionally, the authors demonstrate that their method is applicable to various types of manufacturing systems, such as robotic assembly lines and CNC machining centers.

Bikes, Williams, and O'Connor (1994) used a simulation tool called ProModel-PC to generate synthetic data for their study[30]. The authors use synthetic data to analyze the sensitivity of assembly systems to variations in component delivery times. ProModel-PC is a simulation software tool commonly used in manufacturing and logistics studies to simulate complex systems, such as assembly lines or supply chains. By utilizing simulation-based synthetic data generation, the authors were able to conduct their study in a cost-effective and controlled manner without the requirement of expensive real-world experiments. They were able to generate synthetic data that closely mimicked real-world data and use it to gain insights into assembly system behavior under different delivery scenarios.

Jain, Narayanan, and Lee (2018) presented a comparison of different data analytics approaches using simulation[35]. The authors used a virtual factory prototype to generate synthetic data for testing the various data analytics approaches. The virtual factory was

designed to closely mimic a real-world manufacturing system, and different scenarios were simulated to capture the variability in the system. The authors compared four different data analytics approaches, including decision trees, random forests, support vector machines, and neural networks, to determine their effectiveness in predicting the quality of manufactured products. The synthetic data generated in the virtual factory was used to train and test each of the four models. The results showed that neural networks had the highest accuracy in predicting product quality, followed closely by support vector machines and random forests. Decision trees, on the other hand, had lower accuracy compared to the other three models. They demonstrate the usefulness of using a virtual factory prototype to generate synthetic data for testing and comparing different data analytics approaches in a manufacturing setting.

Biczó, Felde, and Szénási (2021) focused on predicting distortion in the additive manufacturing process using machine learning methods[16]. Synthetic data is generated using the Simufact Additive software, which simulates the printing process with varying parameters such as laser power and scanning speed to capture the variability in the process. The synthetic data is then used to train a convolutional neural network (CNN) to predict the distortion of the printed parts. The trained model showed promising results in predicting the distortion of new parts with high accuracy. The authors suggest that this approach can reduce the cost and time required for physical experimentation in the additive manufacturing process.

Maliks and Kadikis (2021) explored the use of synthetic data for the classification of multispectral data in the context of plastic bottle sorting[26]. Specifically, they focus on using deep convolutional neural networks (CNNs) to classify plastic bottles based on their material composition using multispectral images. In order to train and evaluate the CNN models, the authors generated a synthetic dataset using Blender, a popular 3D graphics software. The synthetic dataset consists of images of plastic bottles rendered under a variety

of lighting and camera conditions, as well as images of plastic bottle labels to aid in the classification task. The authors found that training the CNN models on the synthetic dataset improved their accuracy on real-world multispectral data, indicating that synthetic data can be a useful tool for training machine learning models for plastic bottle sorting applications.

Outón et al. (2021) described a paradigmatic industrial application that combines accurate autonomous navigation and 3D perception for pose estimation in an unstructured industrial environment[40]. They explain that it is still possible to automate or semi-automate many low-value processes in modern industry by working together safely between robots and humans. The authors conducted real-world tests on their proposed method, which achieved an 83.33% success rate using a combination of several technologies fused into an AIMM (autonomous industrial mobile manipulator). The authors generated a synthetic dataset using Unreal Engine 4 (UE4), which is a popular game engine that has been increasingly used for generating synthetic data in various domains, including robotics and industrial automation. The synthetic data allowed the authors to train and validate their system in a controlled and safe environment, without the risk of damaging actual equipment or causing harm to humans.

Grappiolo, Pruim, Faeth, and de Heer (2021) proposed a novel approach for in vitro assembly search based on an artificial intelligence framework[20]. The proposed framework, named ViTroVo, generates synthetic data using a virtual environment to train machine learning models for in vivo adaptive operator guidance. To generate synthetic data, ViTroVo uses a virtual environment that simulates the assembly table, assembly components, and distractor objects in 3D. The synthetic data is generated by randomizing the position, orientation, and appearance of the assembly components and distractor objects. The background plane simulates the assembly table and is kept constant in each synthetic scene. The authors demonstrate the effectiveness of ViTroVo using two case studies: (1) assembly of an electronic module and (2) assembly of a fuel pump. The results show that the proposed

framework can effectively learn the assembly process and can be used to generate adaptive operator guidance for in vivo assembly tasks. The use of synthetic data generated by ViTroVo reduces the need for physical testing, which can be costly and time-consuming, and allows for highly customized manufacturing.

Zheng, Zhang, and Pan (2020) proposed a method for detecting modules in modular integrated construction using virtual prototyping and transfer learning[17]. The authors generate synthetic data by modeling the construction process in a virtual environment using a 3D CAD software tool called Revit. The synthetic data was used to train a convolutional neural network (CNN) for module detection. Transfer learning was used to improve the performance of the CNN by leveraging pre-trained models on large datasets. The authors conducted a comparative analysis of their method with conventional computer vision techniques and established its superiority in terms of both accuracy and efficiency. Additionally, they performed experiments on real-world datasets to confirm the effectiveness of their approach. The results showed that their method achieved high accuracy in module detection, and could be applied to modular integrated construction to improve efficiency and quality.

Sisca, Fiasché, and Taisch (2015) proposed a novel hybrid model for aggregate production planning in a reconfigurable assembly unit for optoelectronics[32]. The model integrates a data-driven neural network model with a simulation-based optimization model. In order to train the neural network model, the authors generated synthetic data using a simulation model. Specifically, they developed a discrete-event simulation model of the production system using the SIMIO simulation software. The simulation model took into account various factors such as machine availability, production capacity, and product mix. The synthetic data generated from the simulation model was then used to train the neural network model. The combination of the simulation-based optimization model and the data-driven neural network model allowed for more accurate and flexible aggregate production planning. The results of the study demonstrate the effectiveness of using synthetic data generated from

simulation models to train data-driven models for aggregate production planning in reconfigurable assembly units.

Apornak, Raissi, and Pourhassan (2021) proposed a hybrid multi-criteria Taguchi-based computer simulation model and DEA approach to solve the flexible flow-shop scheduling problem[36]. To generate the required synthetic data, the authors utilized a simulation model based on a continuous time-series dataset. The simulation model incorporated the hybrid Taguchi-DEA approach and produced a set of data that could be used to optimize the flow-shop scheduling problem. The synthetic data were then used to evaluate the performance of the proposed approach in terms of makespan, flow-time, and tardiness criteria. Overall, the use of synthetic data allowed the authors to test their approach under a variety of scenarios and assess its effectiveness in solving the flexible flow-shop problem.

Fiasché, Ripamounti, Sisca, Taisch, and Tavola (2016) used synthetic data generated from a simulation to evaluate their proposed approach[33]. They used SIMIO to construct the white'R environment, which served as the basis for the synthetic dataset. The synthetic dataset was then used to evaluate the effectiveness of the proposed hybrid fuzzy multi-objective linear programming (FMOLP) method for aggregate production planning. By utilizing synthetic data, the authors were able to test their model in a controlled setting with various uncertain parameters, such as market demands, production capacities, workforce levels, unit costs, and product prices. By assessing the model's performance against the synthetic data, the authors were able to pinpoint areas of improvement and fine-tune their methodology.

Guner, Chinnam, and Murat (2016) used simulation-based synthetic data generation to develop a decision support system for plant-level maintenance[31]. The authors use a simulation software tool called Simul8 to generate synthetic data for their study. Simul8 is a commonly used simulation software tool in manufacturing and service industries for simulating complex systems, such as production lines, supply chains, and service operations.

The authors generate synthetic data by simulating a manufacturing system for a specific case study. They simulate the system under different maintenance policies and record the system performance metrics, such as machine downtimes, maintenance costs, and production throughput. The authors then use the synthetic data to develop a decision support system that helps plant managers select the best maintenance policy for their manufacturing system.

The authors leveraged simulation-based synthetic data generation to conduct their study in a controlled and cost-effective manner, without the need for expensive real-world experiments. They created synthetic data that accurately resembled real-world data and utilized it to develop a customizable decision support system for diverse manufacturing systems. This study emphasizes the advantages of simulation-based synthetic data generation for building decision support systems in both manufacturing and service industries.

Sikora et al. (2021) focused on the quality control of HVAC devices based on environmental noise using a convolutional neural network (CNN)[22]. To train the CNN, the authors generated synthetic data that simulates different noise levels and patterns that can be present in real-world HVAC systems. The synthetic data was generated using a combination of white noise and different frequency bands to mimic the environmental noise. The authors used the synthetic data to train and test the CNN for quality control of HVAC devices. The results showed that the trained CNN was able to accurately classify the quality of the HVAC devices based on the environmental noise, demonstrating the effectiveness of using synthetic data for training deep learning models in quality control applications.

Nguyen, Habiboglu, and Franke (2022) presented a case study on using synthetic data to enable deep learning in the context of automotive wiring harness manufacturing[14]. The authors use a combination of computer-aided design (CAD) software and a physics-based simulation tool called IPS Cable Simulation to generate synthetic data, where they modeled the manufacturing process and used the simulation to generate a large number of synthetic images of wiring harnesses. The synthetic images were then labeled using a semi-automatic

labeling approach, where an operator reviewed and corrected the labels generated by an automatic labeling algorithm.

The authors used synthetic data to train a deep-learning model that can identify faulty wiring harness images. They discovered that the model trained on synthetic data performed better than the one trained on a limited amount of actual data. Therefore, they propose that using physics-based simulations to generate synthetic data could be an advantageous method for integrating deep learning into manufacturing, particularly when acquiring real-world data is difficult or costly.

Rio-Torto et al. (2021) proposed a hierarchical approach for automatic quality inspection in the automotive industry using simulated data[39]. The authors utilized a physical simulation environment to generate synthetic data. A 3D computer-aided design (CAD) software was employed to create a virtual environment that mimics the physical production line, and this served as the foundation for the simulation model. The simulated data comprised of diverse attributes, such as object surface area, color, texture, and intensity. The synthetic data was then used to train and test the proposed hierarchical approach for quality inspection, which consisted of two levels: the first level for defect detection and the second level for defect classification. The experimental results showed that the proposed approach achieved high accuracy in detecting and classifying defects, which validates the effectiveness of the synthetic data generated through physical simulation.

Lai, Tao, Leu, and Yin (2020) presented a smart augmented reality instructional system for mechanical assembly that uses synthetic data to train deep learning models[18]. To generate the synthetic data, the authors used CAD models and physics engines to simulate different assembly scenarios in a virtual environment created using Unity3D. The virtual environment was created to closely mirror the real assembly line, and several situations were simulated to represent the diversity of the assembly process.

The synthetic data generated was used to train a convolutional neural network (CNN) for recognizing different assembly tasks and a long short-term memory (LSTM) network for tracking the progress of the assembly task. The trained models were tested on real-world data and showed promising results in accurately recognizing and tracking the assembly tasks. The use of synthetic data in this study allowed for the development of a worker-centered intelligent manufacturing system that can effectively train and track assembly tasks in real-time. The results demonstrate the potential of using synthetic data for training deep learning models in assembly line applications and highlight the benefits of using virtual environments to simulate complex manufacturing processes.

Overall, physical simulations offer a powerful tool for generating synthetic data and optimizing assembly line processes, but they require expertise in mathematics, physics, and computer science, as well as significant computational resources. When used effectively, physical simulations can lead to significant improvements in manufacturing and assembly line industries.

2.2.2 Data-driven algorithms

The process of generating synthetic data for assembly lines through data-driven techniques involves using machine learning algorithms and other similar methods to simulate real-world scenarios in assembly lines. Before deploying machine learning models in real assembly line settings, the resultant synthetic data may be used to train and evaluate machine learning models.

Data-driven synthetic data generation provides a considerable benefit by supplying vast volumes of diverse and complicated data that might be difficult or costly to acquire in real-world settings. Hence, this enables more comprehensive training of machine learning models, resulting in improved performance in real-world applications.

Multiple data sources, including sensor data, CAD models, and production line layout data, can be utilized to generate synthetic data for assembly lines. These data sources can be used to create realistic simulations of assembly line processes and generate synthetic data that captures the variability and complexity of real-world scenarios. For instance, Han, Choi, Choi, and Oh (2019) focused on developing a data-driven approach for diagnosing faults in planetary gear carrier packs[9]. The authors used vibration signals collected from accelerometers attached to the planetary gear carrier packs to train a machine learning classifier to diagnose faults.

As the dataset was imbalanced, the authors used synthetic data generation techniques to balance the dataset. They used a combination of SMOTE and ADASYN algorithms to generate synthetic samples of the minority classes. After balancing the dataset, they trained a ResNet-50 deep learning model using the augmented dataset. The authors reported that their approach achieved a high classification accuracy for diagnosing faults in planetary gear carrier packs. They concluded that their data-driven approach using synthetic data generation can be an effective tool for fault diagnosis in manufacturing systems.

Kim, Lee, Tama, and Lee (2020) proposed a method for classifying camera lens modules using a semi-supervised regression method[10]. The authors utilized synthetic data generation techniques to increase the amount of available training data for the regression model. The synthetic data was generated using 3D modeling software and then processed to simulate various types of lens modules with different shapes and configurations. The authors then combined the synthetic data with real-world data to train the regression model. The results showed that the proposed method outperformed other classification methods and demonstrated the effectiveness of using synthetic data generation techniques in enhancing the reliability of the classification model.

Fecker, Märgner, and Fingscheidt (2013) addressed the problem of imbalanced datasets in machine learning, where the number of examples in one class significantly outweighs

the number in another class[12]. The authors propose a novel oversampling method called Density-Induced Oversampling (DIO), which uses Gaussian Mixture Models (GMMs) to generate synthetic data samples for the minority class. The DIO method involves first fitting a GMM to the feature space of the minority class. The GMM is then used to generate new samples for the minority class by sampling from the GMM's probability density function. The number of new samples generated is determined by the desired level of oversampling. The authors compare the performance of DIO with several other oversampling methods, including Random Oversampling, Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN), using several imbalanced datasets. The results show that DIO outperforms the other methods in terms of classification performance and robustness across different datasets.

Kohtala and Steinert (2021) discussed the use of synthetic data generated from CAD models to train object detection models for virtual reality (VR) applications[19]. The authors propose a method for generating synthetic data by creating CAD models of objects, augmenting the models to produce multiple variations, and rendering the models with varying lighting conditions and camera angles to produce a diverse set of images. These images are then used to train object detection models, specifically YOLOv3, for VR applications.

The authors assessed the effectiveness of their method by comparing the performance of models trained on both synthetic and real data. They observed that the models trained on synthetic data attained comparable performance to those trained on real data. Moreover, they discovered that employing synthetic data enabled them to generate a more extensive and diverse dataset than what could be achieved using only real data.

de la Rosa, Gómez-Sirvent, Sánchez-Reolid, Morales, and Fernández-Caballero (2022) proposed the use of geometric transformation-based data augmentation techniques to improve the classification of defects in segmented images of semiconductor materials using a ResNet50 convolutional neural network[21]. Due to the high cost of obtaining labeled data,

the study utilizes a synthetic dataset generated by combining different transformations, such as rotation, translation, and scaling, applied to the original images. The authors demonstrate that the use of synthetic data combined with data augmentation techniques improved the accuracy of the model in detecting and classifying defects in semiconductor materials. The proposed approach also allowed for a better understanding of the factors affecting the performance of the classification model. The study suggests that synthetic data generation combined with data augmentation techniques can be a useful tool for improving the performance of machine learning models, particularly in cases where obtaining large amounts of labeled data is not feasible.

Laxman, Sastry, and Unnikrishnan (2007) presented a method for discovering frequent generalized episodes in event sequences, where events can persist for different durations[27]. The authors generate synthetic data to evaluate the performance of their proposed algorithm. The synthetic data is generated by randomly generating event sequences with varying episode lengths and frequencies, and then adding noise to the sequences to simulate real-world scenarios. The authors used the generated synthetic data to test the performance of their proposed algorithm and compared it with existing algorithms for discovering frequent episodes. The results showed that their proposed algorithm outperformed the existing algorithms in terms of accuracy and efficiency. Overall, the paper demonstrates the usefulness of synthetic data in evaluating the performance of algorithms for discovering frequent episodes in event sequences.

Andres, Guzman, and Poler (2021) proposed a mixed integer linear program (MILP) model to minimize costs associated with the production of automotive plastic components[28]. To generate synthetic datasets for the model, the authors defined a set of parameters and their values. The goodness of the model was evaluated based on computational time and deviation from optimal results. The model demonstrated remarkable efficiency in addressing small and medium datasets. Nonetheless, its efficacy was found to be restricted when handling

large datasets. Overall, the use of synthetic datasets allowed for efficient evaluation of the MILP model's performance and optimization of the production process for automotive plastic components.

Qian, Yu, Lu, Griffith, and Golmie (2022) proposed the use of Generative Adversarial Networks (GANs) for generating synthetic data in the context of the Industrial Internet of Things (IIoT)[24]. Specifically, the authors focus on the task of anomaly detection in IIoT systems, which is an important aspect of ensuring the reliability and safety of industrial processes. The authors propose the utilization of GANs as a potential solution to tackle the issue of scarce data availability in the IIoT field by generating synthetic data that can supplement the existing dataset. The suggested technique involves training a GAN on the existing data to produce synthetic samples that closely resemble the real data, which can then be utilized to train an anomaly detection model. The paper presents a proof-of-concept implementation of the proposed approach and evaluates its performance on a dataset of sensor readings from a wind turbine. The results show that the use of synthetic data generated by the GAN improves the performance of the anomaly detection model compared to using only real data.

Overall, the authors suggest that GANs have the potential to be a useful tool for generating synthetic data in the IIoT domain, where data availability is often limited but where reliable anomaly detection is critical for ensuring the safety and efficiency of industrial processes.

Ameperosa & Bhounsule(2020) presented a method for estimating the positions of bolts using deep neural networks and domain randomization[25]. The authors propose a method that uses synthetic data generated from a simulator as input to train the deep neural network. They use domain randomization to simulate variations in lighting conditions, surface textures, and other environmental factors to improve the robustness of the network. The trained network is then used to estimate the positions of bolts in real-world images. The authors

establish the efficacy of their approach by contrasting its performance with other cutting-edge techniques on a collection of real-world images. Their approach surpasses the other methods, attaining superior accuracy and improved generalization to novel data. The authors conclude that their approach could be useful for a range of industrial applications where accurate localization of objects is required.

da Silva et al. (2021) focused on the application of spatio-temporal deep learning-based methods for defect detection in an industrial setting[38]. One challenge in this application is the limited availability of labeled data, which is essential for training machine learning models. To address this challenge, the authors proposed the use of synthetic data generated by a GAN to augment the limited labeled dataset. Specifically, they used the GAN to generate anomalous sequences similar to those produced by defective devices, which were then used to train the model. The results of their study demonstrated that the use of synthetic data generated by a GAN can improve the performance of machine learning models in detecting defects, particularly when labeled data is limited. This approach can potentially reduce the cost and time required to obtain sufficient labeled data, making it a promising approach for defect detection in industrial settings.

Singh, Chakrabarti, and Jayagopi (2020) used a synthetic data generation technique to train CNNs for detecting malfunctions in a refreshable Braille display[23]. They generated synthetic data by rotating and modifying the colors of the images. The synthetic data was used in combination with real data to train the model. The results showed that the model achieved an accuracy of 97.3% in detecting malfunctions during the experiment. This demonstrates the effectiveness of using synthetic data in combination with real data for training machine learning models for automated testing of devices.

Sibona and Indri (2021) presented a new data-driven framework that uses both real-world and synthetic data to optimize the performance of human-robot collaborative systems in flexible manufacturing applications[15]. The authors utilize Generative Adversarial Networks

(GANs) to create synthetic data that can supplement the limited quantity of real-world data available for training machine learning models. The GANs are trained on actual data to generate new data that has a similar distribution but contains differences that can enhance the variety and extent of the training data. This synthetic data is used to train machine learning models for various purposes, including anticipating the probability of a collision and optimizing the motion and placement of the robot arm. The authors showcase that incorporating both genuine and synthetic data in their data-driven framework can boost the performance of the human-robot collaborative system and enable more economical and efficient training of machine learning models.

Cai, Bernstein, Wu, and Chandramouli (2021) investigated the optimization of threshold functions over streams[37]. They propose a new method for threshold selection in a stream join operation that reduces the memory required for the operation. The authors generated synthetic data using different methods, including the Wiener process, Gaussian noise, and varying the standard deviation. By generating synthetic data, they were able to experiment with different scenarios and evaluate their proposed method's effectiveness. The results showed that the proposed method achieved a significant reduction in memory usage for stream join operations.

Martin, Depaire, Caris, and Schepers (2020) presented a method for automatically retrieving resource availability calendars from event logs containing information about process execution[34]. The retrieved calendars capture two dimensions of availability: the time of day when resources are available and intermediate interruptions, such as breaks. The authors evaluated their method using synthetic data generated based on the Business Process Model and Notation (BPMN), which closely matched the key outputs of their method to their realistic equivalents. The synthetic data was generated using the SimPy simulation library in Python, which allowed for the creation of realistic process models and the simulation of process execution with varying levels of noise and randomness. The authors found that

their method performed well on both real and synthetic data and could be used to support process improvement and optimization efforts in various industries. The use of synthetic data allowed for rigorous evaluation and testing of the method, even when real-world data was scarce or difficult to obtain.

Malekzadeh, Clegg, and Haddadi (2017) presented a privacy-preserving algorithm for analyzing sensory data using autoencoder neural networks[29]. The authors propose the replacement autoencoder (RAE) which is trained on the original sensory data to learn a compressed representation of it. The RAE is utilized to create artificial data examples that replace the original data while retaining the statistical characteristics of the original data. This enables analysis to be carried out on the synthetic data rather than the original data, safeguarding the privacy of the individuals from whom the data was gathered. The paper also presents experiments demonstrating the effectiveness of the RAE in preserving privacy while still allowing for useful analysis of the data.

Syafrudin, Fitriyani, Alfian, and Rhee (2018) presented a study on an affordable fast early warning system for edge computing in assembly line operations[13]. The study aims to identify abnormal events during the assembly line operations using an affordable fast early warning system based on edge computing. To achieve this, the authors generated a balanced dataset using the Synthetic Minority Over-sampling Technique (SMOTE) to overcome the class imbalance problem commonly encountered in industrial datasets. The SMOTE algorithm generates synthetic samples by interpolating between the feature vectors of the minority class, creating new synthetic samples that are similar to the original minority samples. The generated synthetic samples are added to the original dataset, resulting in a balanced dataset that can be used to train machine learning models.

The authors demonstrated that the use of the SMOTE algorithm to generate synthetic data leads to improved performance in detecting abnormal events in assembly line operations. The results showed that the proposed early warning system achieved a high detection

accuracy of 97.8% on the balanced dataset generated using the SMOTE algorithm. Overall, the study shows that the use of synthetic data generation techniques, such as SMOTE, can be a valuable tool in improving the performance of machine learning models in detecting abnormal events in assembly line operations.

Mubarak, Mohamed, and Bouferguene (2020) proposed a data generator for industrial pipelines, which can be used to create synthetic data for the optimization of the pipe spooling process[5]. The authors explain the functionality of the generator and how it can produce pipelines of different types, characterized by features like length, diameter, curvature, and roughness. Subsequently, the generated data is employed to evaluate a pipe spooling optimization algorithm, which seeks to minimize the number of pipe fittings necessary to construct a pipeline. the paper presents an extensive experimental setup and analysis of the optimization algorithm's outcomes, indicating the effectiveness of the proposed approach.

In general, the utilization of data-driven synthetic data generation presents a favorable solution for enhancing the effectiveness and precision of assembly line operations, in addition to lowering expenses and advancing safety measures in manufacturing settings.

Although physical simulations are very precise and give complete models of complicated systems, they may be computationally intensive and need substantial system knowledge. In contrast, data-driven algorithms can generate synthetic data quickly and efficiently. But, they may not capture all the intricate details that exist in the actual system.

Both physical simulations and data-driven algorithms possess their own set of advantages and disadvantages. However, both methods are essential tools for generating synthetic data across various research domains, including engineering, physics, biology, and social sciences. Complicated systems that would be difficult or impossible to grasp with just real-world data may be understood utilizing artificial data. Furthermore, it can be utilized to train and assess machine learning algorithms, ultimately resulting in the creation of novel computational techniques for analyzing and predicting data.

2.3 Application of synthetic data

This section provides a brief overview of how synthetic data has been utilized in experimentation in the literature. While some papers generate synthetic data to create artificial datasets, the papers in this section utilize pre-existing synthetic data to perform experiments and analysis. Synthetic data has numerous benefits in experimentation, such as cost reduction, enhanced accessibility, and improved privacy protection. This section delves into the various domains and research fields where synthetic data has been employed, emphasizing its potential impact on different areas of research.

Marazopoulou, Ghosh, Lade, and Jensen (2016) proposed a framework to discover causal relationships among variables in manufacturing processes using observational data[6]. The authors used a Bayesian network technique to explain the causal links between variables in the manufacturing process and then tested the efficacy of their proposed framework using both the actual dataset and a collection of synthetically produced datasets. Nevertheless, the authors did not describe a specific approach for producing synthetic data. Overall, the paper highlights the potential of using causal discovery techniques and synthetic data generation in manufacturing domains to improve process optimization and quality control.

Mihai et al. (2021) proposed a digital twin framework that uses synthetic data to train machine learning models for predictive maintenance[7]. Although the research does not disclose particular details on the techniques used to produce synthetic data, the authors state that a data production engine is used to mimic a variety of situations. The authors highlight the advantages of synthetic data, such as the capacity to create situations that may be impossible or dangerous to recreate in the actual world and the ability to generate big datasets more quickly and easily. By using synthetic data, the authors are able to predict the remaining useful life of industrial assets more accurately, leading to cost savings and increased efficiency. The digital twin framework integrates various components, such

as sensor data acquisition, data processing and analysis, and machine learning models, to provide actionable insights to maintenance personnel.

Luckow et al. (2018) discussed the potential of using synthetic data in training machine learning models for automotive manufacturing applications [41]. The authors describe the challenge of obtaining and labeling large datasets of real-world data, and the potential benefits of using synthetic data instead. They provide several examples of the use of synthetic data in the automotive manufacturing industry, including the generation of synthetic images of car body panels with defects to train a CNN for defect detection. They do not specifically mention the synthetic data generation technique used for the car body panel defect detection CNN. The authors also emphasize that synthetic data may be utilized to build more diversified and intricate datasets than may be achievable with actual data. Nonetheless, the report notes that developing synthetic data may be challenging and that the quality of the data might restrict its value. Overall, the article highlights the potential of synthetic data for enhancing the precision and lowering the cost of training machine learning models for automotive production applications.

Shetve et al. (2021) used synthetic data to generate data points for performance analysis of their proposed anomaly detection method called CATS[42]. The authors reported that CATS was able to detect outliers with an accuracy of 98.58%. However, the paper did not provide any information regarding the specific method used for generating the synthetic data. Nonetheless, the use of synthetic data in evaluating the performance of CATS highlights its potential for efficient and effective anomaly detection in smart manufacturing systems.

Georgiadis, Nizamis, Vafeiadis, Ioannidis, and Tzovaras (2022) proposed a digital cognitive platform for production scheduling optimization[43]. The platform employs cutting-edge optimization algorithms and machine learning approaches to create scheduling models capable of handling big and complicated industrial settings. The authors emphasise the significance of data quality in the construction of these models and the use of both actual and

synthetic data. They indicate that synthetic data was utilized to complement the actual data and improve the training and testing of their machine learning models, however they do not provide a particular strategy for producing synthetic data. Overall, the article demonstrates that synthetic data might enhance the robustness and generalizability of machine learning models in the context of production scheduling optimization.

Gao, Wang, Helu, and Teti (2020) discussed the potential of big data analytics for smart factories of the future[4]. They highlight the importance of collecting and analyzing large volumes of data generated by sensors, machines, and other sources in order to optimize production processes and increase efficiency. Synthetic data generation is mentioned as a technique for augmenting real data sets and creating training data for machine learning algorithms. The authors suggest that synthetic data can help overcome limitations related to data availability, privacy, and security. Additionally, they discuss the importance of ensuring the quality and reliability of both real and synthetic data to maximize the value of big data analytics in the context of smart factories.

Sun et al.(2021) proposed a machine learning pathway for improving the quality and efficiency of material processing in the casting industry[44]. The authors remark that material processing is a hard endeavor involving several factors such as temperature, pressure, and chemical composition. To solve these difficulties, the authors suggest using machine learning models that can learn from data and generate predictions about material processing procedures. The authors also explore the obstacles posed by the restricted availability of data in the material processing area and suggest the use of physics-based simulations to produce synthetic data. The authors demonstrate the effectiveness of their machine learning pathway in improving the quality and efficiency of the casting process through case studies and experiments. Overall, the paper presents a novel approach to improving material processing operations by leveraging machine learning and synthetic data generation techniques.

Zhang et al. (2021) proposed a method for scene text recognition using synthetic data[45]. The proposed method consists of an auxiliary network and a text recognizer. The auxiliary network is designed to mimic traditional computer vision functions, and it extracts rich augmented features from the input image. Synthetic data is commonly used in scene text recognition to label images since labeling scene text images can be time-consuming, labor-intensive, and costly. The proposed method uses two synthetic datasets for training. The experimental results have shown that the proposed method improves the performance of the recognizers, especially for degraded text images in challenging settings. The use of synthetic data has allowed for the creation of a larger and more diverse dataset, leading to better performance in recognizing text from real-world images.

Bécue, Maia, Feeken, Borchers, and Praça (2020) proposed a new concept of digital twin that supports the optimization and resilience of factories of the future[46]. The authors employed a simulation-based approach that involves generating synthetic data to create digital twins capable of predicting and enhancing the performance of industrial systems. This method can overcome the limitations of traditional data-driven techniques, which rely on historical data and may not account for changes in the industrial environment. By utilizing synthetic data to construct digital twins, the authors successfully optimized the performance of a robotic arm system, demonstrating the potential of this approach to improve factory efficiency and resilience in the future.

Godil, Eastman, and Hong (2013) discussed the benefits of using synthetic data in object recognition and tracking[47]. Synthetic data offers the advantage of producing accurate, reliable, and reproducible ground truth data, which is particularly valuable in situations where acquiring real-world data is challenging, expensive, or incomplete. In addition, synthetic data can produce a diverse range of data, which is advantageous for training machine learning algorithms that demand extensive data to achieve high accuracy. Moreover, utilizing synthetic data may help reduce bias in training data by creating data that represents

a wide range of situations and contexts. Ultimately, synthetic data can be a useful tool in object identification and tracking, allowing researchers to construct comprehensive and precise datasets that can be employed to train and evaluate machine learning systems.

Cimino, Feretti, and Leva (2021) proposed a toolset and paradigm for harmonizing and integrating digital twins across multiple domains[48]. One of the suggested solutions is the use of synthetic data creation methods to generate digital twins that are representative of actual situations. The authors stress that using actual data to construct digital twins may be problematic owing to factors such as data privacy, data quality, and data availability. Synthetic data production may address some of these limitations by generating data that is comparable to actual data, but without privacy concerns or data quality problems. The authors propose employing machine learning methods to produce synthetic data that may be used to train digital twins, therefore increasing their accuracy and efficacy. In addition, they remark that synthetic data may be used to recreate circumstances that are difficult or impossible to duplicate in the actual world, allowing for improved decision-making and analysis.

Ramanujan and Bernstein (2018) presented a novel tool for exploring large repositories of computer-aided design (CAD) models based on similarity and performance metrics[49]. The authors emphasize the difficulty of dealing with huge datasets within the context of design repositories and the necessity for efficient exploration and analysis techniques. They present a method that integrates machine learning, visualization, and human-computer interaction approaches to enable designers and engineers to get insight into the performance and resemblance of CAD models. The tool uses synthetic data from the Drexel repository in its case study, demonstrating its effectiveness in analyzing and exploring large CAD repositories. The authors highlight the potential for their approach to enable new discoveries and innovation in the field of design, manufacturing, and engineering.

Mihai et al. (2022) discussed the importance of synthetic data in the development and training of digital twins[50]. Synthetic data may be used to enhance or replace real-world data in the development and testing of digital twins. This technique may aid in overcoming issues related to the price, accessibility, and quality of real-world data. The authors highlight that although synthetic data may be very valuable, it must be rigorously vetted and calibrated to appropriately represent the behavior of the actual system being modeled. Furthermore, the authors emphasize the necessity for continuing study and development in the area of synthetic data creation and validation, especially as digital twins gain importance across a variety of sectors and applications.

Rardin and Uzsoy (2001) emphasized the importance of empirical evaluation in selecting the most appropriate optimization algorithm for a given problem[51]. They discuss the basics of heuristic optimization methods, including genetic algorithms, simulated annealing, and tabu search. This study presents a comprehensive review of the experimental design, data analysis, and interpretation of findings in investigations of heuristic optimization. Although synthetic data was not employed in this study, the authors emphasized the issues associated with randomly produced data and the necessity for rigorous data selection to assure the reliability and use of the outcomes of empirical assessments. Overall, the paper provides a comprehensive guide for researchers and practitioners to conduct empirical evaluations of heuristic optimization methods.

Bertolini, Mezzogori, Neroni, and Zammori (2021) provided a comprehensive literature review of machine learning applications in the industrial domain[52]. Synthetic data was used in some studies to generate datasets that are representative of real-world industrial scenarios. The authors discussed several ways for generating synthetic data, including physics-based simulation, generative adversarial networks (GANs), and picture augmentation techniques. In situations when obtaining real-world data is difficult or costly, the use of synthetic data is vital, and it may increase the performance of machine learning algorithms. In addition,

synthetic data may be utilized to expand the dataset's variety and balance the distribution of classes, resulting in more accurate models. The authors also explored the difficulties and limits of synthetic data, including the difficulty of replicating the complexity of the actual world and the danger of injecting biases into the data. Overall, the work emphasizes the significance of synthetic data in industrial machine learning applications and the necessity for more research in this field.

Achar, Laxman, Viswanathan, and Sastry (2012) focused on discovering injective episodes with general partial orders [53]. To evaluate the proposed algorithm, used synthetic datasets to test the algorithm's ability to discover injective episodes with varying sizes and noise levels. The authors highlighted that synthetic datasets were used to illustrate the efficacy of the suggested method, but real-world data would be more complicated and varied. In addition, they emphasized the constraints of utilizing synthetic data and the need of validating the algorithm's effectiveness using real-world data. The use of synthetic data in this paper helped in testing and validating the proposed algorithm's ability to discover injective episodes in different scenarios.

Thelen et al. (2022) provided a comprehensive review of digital twin technology and its enabling technologies[54]. Digital twins are virtual replicas of physical systems, and they can be used for various purposes such as predicting system behavior, optimizing system performance, and reducing maintenance costs. Synthetic data plays a crucial role in the creation of digital twins, as it enables the modeling and simulation of complex physical systems. The authors discuss various synthetic data generation techniques such as finite element analysis, computational fluid dynamics, and agent-based modeling. They also emphasize the importance of using real-world data to validate the accuracy of digital twin models.

Xu et al. (2022) discussed various challenges and solutions for implementing deep learning techniques in smart manufacturing[55]. To overcome the problem of inadequate training data, they recommend the use of data augmentation, generative adversarial networks (GANs), and

simulation models. The authors recommend producing new data using data augmentation by randomly transforming the current data. GANs may create synthetic data that resembles the distribution of actual data, while simulation models can generate vast quantities of synthetic data by mimicking industrial processes. The authors underline the usefulness of synthetic data in deep learning model training and propose that it may be used to complement or replace real-world data in certain circumstances.

Suhail et al. (2022) explored the use of blockchain-based digital twins and their research trends, issues, and future challenges[56]. The authors highlight the revolutionary possibilities of digital twins in many areas, including manufacturing, healthcare, and transportation. The authors propose employing a mix of real-world data and synthetic data to produce data for simulations of digital twins. They suggest using machine learning methods and data generators to generate synthetic data that can be used to train and validate digital twin models. By merging actual and synthetic data, cost-effective and scalable digital twin models may be constructed and validated. The authors also note the problems and limits of utilizing synthetic data, such as the requirement for correct data production algorithms and the possibility of bias in the created data. The study underlines the significance of synthetic data in the creation and implementation of blockchain-based digital twins.

Botero, Wilson, Lu, Rahman, Mallaiyan, Ganji, Asadizanjani, Tehranipoor, Woodard, and Forte (2021) focused on the use of reverse engineering, image analysis, and machine learning techniques to improve hardware trust and assurance[57]. The authors present an overview of the current state of the art in these disciplines and describe many case studies in which these approaches were used to enhance the security and dependability of hardware systems. In the absence of actual data, synthetic data creation is briefly addressed as a possible method for training machine learning models. Overall, the paper highlights the importance of these techniques for ensuring the trustworthiness of hardware systems and presents future research directions.

Qian et al. (2022) proposed the use of generative adversarial networks (GANs) for the Industrial Internet of Things (IIoT) to generate synthetic data for various industrial applications such as predictive maintenance, quality control, and anomaly detection[24]. They suggest that GANs may overcome the constraints of conventional data creation techniques by producing synthetic data that closely mimics real-world data, hence eliminating the need for costly and time-consuming data collecting and labeling. The authors present a comprehensive analysis of the advantages and disadvantages of current GAN models and recommend future research topics for enhancing the performance and efficiency of GANs in IIoT applications.

Mahmoodian, Shahrivar, Setunge, and Mazaheri (2022) proposed the development of a digital twin for intelligent maintenance of civil infrastructure[58]. Synthetic data was used to create a virtual model of the physical infrastructure, which can be used to predict potential failures and improve maintenance scheduling. The digital twin was created using structural analysis software, finite element analysis, and data from sensors embedded in the infrastructure. The authors used synthetic data to simulate different scenarios and predict the behavior of the infrastructure under different conditions. The results of the simulations showed that the digital twin could accurately predict the behavior of the physical infrastructure and provide valuable insights for maintenance planning. The authors conclude that the use of digital twins with synthetic data can lead to significant improvements in infrastructure maintenance and management.

Flores, Fernández-Casal, Naya, and Tarrío-Saavedra (2021) presented a package called "qcr" for the R statistical computing environment [59]. This package is designed to help with statistical quality control and includes functions for generating synthetic data. Specifically, the package includes a function for generating normal and uniform random numbers, which can be used as synthetic data for testing statistical quality control procedures. The authors note that using synthetic data can be useful for testing quality control procedures in

situations where collecting real data is difficult or expensive. By generating synthetic data, users can test the effectiveness of their statistical quality control procedures in a controlled environment before applying them to real-world data.

Doorn, Duivestein, Mamtani, and Pepping (2020) presented an overview of machine creativity and its potential for generating synthetic data[60]. According to the authors, employing machine learning algorithms can enable computers to generate new and creative ideas that surpass what has been previously observed in the data. They emphasize the potential of generative models, such as GANs, to produce synthetic data that closely mimics actual data, enabling researchers to study and examine data while avoiding privacy breaches or data loss. In addition, the authors propose that as machine creativity improves, machines may be able to develop totally new notions and facts, resulting in new scientific discoveries and applications.

Asturias and Rossbach (2023) examined how the misallocation of resources across firms can lead to a decrease in aggregate productivity[61]. They develop a novel method to measure factor misallocation by grouping firms based on their factor share patterns and assessing the variation in those patterns across groups. They find that a significant portion of the variation in aggregate factor shares can be explained by misallocation across firms, which ultimately leads to a reduction in aggregate productivity. To test their technique and illustrate its relevance for misallocation analysis, the authors employ both actual and synthetic data. These results have significant ramifications for policymakers addressing productivity and resource allocation concerns in the economy.

Chapter 3

Comparative Study

We will examine how synthetic data creation is used in manufacturing in more detail in this section. As mentioned earlier, collecting high-quality data for manufacturing applications can be a difficult task. Synthetic data generation provides a promising solution to this challenge. Manufacturing may benefit from synthetic data production in a number of ways, including decreased data collecting costs and time, increased model accuracy, and the ability to simulate circumstances that may be challenging or impossible to recreate in the real world.

Our examination will focus on several manufacturing applications, including quality control, predictive maintenance, and process optimization, and explore how synthetic data generation has been applied in these contexts. Specifically, we will examine the benefits that synthetic data has brought to manufacturing, as well as the challenges and limitations that must be addressed. By reviewing the current literature and identifying gaps, we aim to contribute to the advancement of synthetic data generation in manufacturing.

3.1 Manufacturing Challenges

3.1.1 Quality Control

In assembly line manufacturing, quality control is essential to ensure that products meet predetermined standards and specifications. Artificial intelligence (AI) has become increasingly popular in quality control, particularly in generating synthetic data for assembly line applications.[62]. Quality control encompasses diverse tasks, including product inspection and testing, fault diagnosis with root cause analysis, utilization of statistical process control to regulate the manufacturing process, experimentation for optimizing process parameters, assurance of correct implementation of quality control procedures through quality assurance and auditing, and perpetual enhancement of product or service quality. The ultimate objective of quality control is to produce products or services that go beyond customer expectations while meeting their demands.

The use of synthetic data has been explored in several research papers as a potential solution to address challenges in quality control for assembly line applications. For instance, Han et al. (2019) investigated the application of synthetic data for improving quality control in planetary gear carrier packs[9]. The authors identified a class imbalance and multiclass classification problem in fault diagnosis, which can lead to inaccurate results. To overcome this issue, they proposed a method for generating synthetic data to balance the dataset and improve the accuracy of the fault diagnosis system. The proposed approach showed promising results and highlights the potential for using synthetic data to address class imbalance and multiclass classification problems in quality control applications. The study by Han et al. (2019) demonstrates the potential of synthetic data to enhance quality control in assembly line applications and suggests that further research in this area could yield valuable insights and solutions[9].

Kim et al. (2020) presented a novel approach to improve the reliability of camera lens module classification through the use of synthetic data in quality control[10]. The authors recognized the challenges of limited data availability and potential overfitting in machine learning models for quality control and proposed a semi-supervised regression method that incorporates both labeled and unlabeled data to improve model accuracy. To further enhance the reliability of the model, the authors generated synthetic data to augment the dataset. The results of their study showed significant improvements in classification accuracy, highlighting the potential of synthetic data generation as a valuable tool in quality control for assembly line applications.

Fecker et al. (2013) presented a method for generating synthetic data to improve the performance of machine learning algorithms for quality control in manufacturing[12]. The authors focus on highly imbalanced datasets, where the number of instances in the minority class is much smaller than that in the majority class. They propose a density-induced oversampling method that uses a Gaussian mixture model (GMM) to generate synthetic data for the minority class. The GMM is first trained on the majority class data and then adapted using Bayesian adaptation with the sparse data of the minority class. The adapted GMM is used to generate new synthetic data for the minority class, and a threshold is used to determine whether the synthetic data should be accepted or discarded. The proposed method is evaluated on a real-world dataset from a quality control application, and the results show that it outperforms other oversampling techniques and improves the classification performance of the machine learning algorithm.

Ademujimi et al. (2022) proposed a method for training Bayesian networks for fault diagnosis of manufacturing systems using a digital twin and synthetic data generation[11]. The authors highlight the importance of accurate fault diagnosis in manufacturing systems to minimize downtime and improve productivity. However, the lack of available data can hinder the development of effective fault diagnosis systems. To circumvent this difficulty, the authors

suggest generating synthetic data using a digital twin, a virtual clone of the manufacturing system. This strategy permits the production of a big and diversified dataset, which may enhance the correctness of the Bayesian network model. The authors also introduce a method of evaluating the model's performance using a confusion matrix and ROC curve analysis. The results demonstrate the effectiveness of the proposed approach and highlight the potential for using synthetic data and digital twins in quality control for manufacturing systems.

Bavelos, Kousi, Gkournelos, Lotsaris, Aivaliotis, Michalos, and Makris (2021) have proposed a system that employs synthetic data to train mobile robots for a flexible manufacturing environment [63]. This system merges shopfloor and process perception, enabling mobile robots to adjust to changes in the manufacturing environment and work in tandem with human workers. By utilizing a blend of physics-based simulation and machine learning techniques, the authors generated synthetic data that facilitates the swift development and testing of new robot behaviors, and the training of robots in difficult or perilous scenarios that would be challenging to replicate in reality. Although the specific approach or algorithm utilized to generate synthetic data is not explicitly mentioned, the authors underline the potential of synthetic data to transform the manufacturing industry by improving the adaptability and responsiveness of production processes.

Syafrudin et al. (2018) proposed an early warning system for edge computing in assembly lines to improve quality control[13]. They address the issue of slow response time and limited connectivity in traditional monitoring systems by introducing an affordable and fast system that can monitor machines and processes in real-time. The authors utilize synthetic data to train a machine learning algorithm that detects anomalies in the assembly line system and triggers an alarm to notify operators. The use of synthetic data allows for a more extensive and diverse dataset to train the machine learning model, leading to more accurate and efficient anomaly detection. The proposed system demonstrates the potential of synthetic

data generation in quality control applications and highlights the importance of early warning systems in assembly line production.

Maliks et al. (2021) proposed a deep Convolutional Neural Network (CNN) for classifying plastic bottles based on multispectral data obtained from a hyperspectral imaging system[26]. The authors highlight the importance of quality control in the plastic recycling industry, where automated bottle sorting is essential for achieving high-quality recycled products. However, acquiring real-world data for training and testing the proposed CNN model is challenging due to the cost and effort involved in collecting labeled samples from a wide variety of plastic bottles. To overcome this challenge, the authors have used synthetic data generated by a physics-based simulation tool that models the behavior of a hyperspectral imaging system. The synthetic data allows the CNN model to be trained and tested effectively, improving its accuracy and reducing the cost and effort involved in data collection. The proposed method achieves an average classification accuracy of 99.28%, demonstrating the effectiveness of the proposed approach for quality control in the plastic recycling industry.

Martin et al. (2020) discussed the use of synthetic data in the context of quality control[34]. The authors use synthetic data to evaluate the accuracy and effectiveness of their method for retrieving resource availability calendars from event logs. Comparing the findings acquired with actual data to those produced with synthetic data, they show that their technique works well in both instances. This highlights the potential benefits of adopting synthetic data in quality control since it permits complete testing and validation of procedures without the need for costly and time-consuming investigations on real-world data. The authors indicate that their method might be used in a variety of industrial and service sectors where resource availability calendars play a significant role in process optimization and quality control.

de la Rosa et al. (2022) proposed the use of geometric transformation-based data augmentation techniques to improve the classification of defects in segmented images of semiconductor materials using a ResNet50 convolutional neural network for quality control purposes[21]. Due to the high cost of obtaining labeled data, the study utilizes a synthetic dataset generated by combining different transformations applied to the original images. The authors demonstrate that the use of synthetic data combined with data augmentation techniques improved the accuracy of the model in detecting and classifying defects in semiconductor materials. The suggested method also provided for a greater comprehension of the aspects influencing the classification model's performance. The work implies that synthetic data creation paired with data augmentation approaches may be a helpful tool for boosting the performance of machine learning models, especially in situations when getting huge quantities of labeled data is impractical.

Nguyen et al. (2022) presented a case study that explores the use of synthetic data to enable deep learning for quality control in automotive wiring harness manufacturing[14]. The authors address the issue of limited data availability and the challenges of collecting sufficient labeled data for training deep learning models. To overcome these challenges, they propose a method of generating synthetic data that is similar to real-world data but provides the necessary diversity to train deep learning models effectively. By training a deep learning model on both actual and synthetic data and comparing the outcomes, the authors show the usefulness of their technique. The findings demonstrate that the introduction of synthetic data enhances the accuracy of the deep learning model in identifying manufacturing process flaws. This study highlights the potential of using synthetic data to improve quality control in manufacturing and demonstrates the importance of data diversity in training deep learning models.

Sikora et al. (2021) investigated the influence of environmental noise on the quality control of HVAC (heating, ventilation, and air conditioning) devices using a convolutional

neural network (CNN)[22]. The authors address the issue of reduced accuracy in quality control due to environmental noise in the manufacturing facility. To mitigate this issue, they propose using synthetic data to augment the dataset and improve the accuracy of the CNN-based quality control system. The authors generate synthetic data by adding noise to the existing data and using it to train the CNN model. The results show that the proposed approach improves the accuracy of the quality control system and demonstrates the potential of using synthetic data in quality control applications for manufacturing systems affected by environmental noise.

Rio-Torto et al. (2021) presented a hierarchical approach for automatic quality inspection in the automotive industry[39]. The suggested approach employs machine learning models for detecting and categorizing defects, trained on synthetic data generated through physical simulation. The synthetic data was created by taking into account the physics of the inspection process and the defects, allowing for a wide range of scenarios and defect types to be incorporated. The authors reveal that utilizing synthetic data led to a considerable enhancement in the defect detection and classification models' performance, indicating the efficacy of synthetic data in quality control applications.

The study conducted by da Silva et al. (2021) emphasized the importance of quality control and proposes the use of synthetic data to improve defect detection[38]. The authors note the limitations of traditional methods and highlight the potential of deep learning-based approaches that capture spatial and temporal features. To augment the limited amount of real-world data for training, the authors suggest using synthetic data generated through computer simulations. The proposed approach can lead to more accurate and efficient defect detection, ultimately improving product quality and reducing the risk of defects reaching customers.

Jain et al. (2018) presented a comparison of data analytics approaches using simulation for quality control in a virtual factory prototype[35]. The authors used a simulation model to

generate synthetic data for testing different data analytics methods such as logistic regression, decision trees, and neural networks. The synthetic data closely resembled the real-world data, and the results showed that the neural network approach outperformed the other methods in predicting defective products. This study highlights the potential of using synthetic data in quality control applications and the importance of choosing the appropriate data analytics method for achieving the desired accuracy.

Biczó et al. (2021) discussed the use of machine learning methods to predict distortion in additive manufacturing processes[16]. To train their model, the authors generated synthetic data using Simufact Additive software, which simulates the additive manufacturing process and predicts the resulting distortions. Subsequently, the synthetic data was employed to train a machine learning model that forecasts distortions in the actual manufacturing process. Predicting distortions in the additive manufacturing process is vital for ensuring the quality of the end product in terms of quality control. Through the use of synthetic data for training their machine learning model, the authors could simulate numerous distortion scenarios and generate a considerable amount of training data that might be challenging to obtain through physical experimentation. This allowed them to build a more robust model for predicting distortions and ultimately improve the quality of the additive manufacturing process.

In the context of quality control, the accurate estimation of bolt positions is critical in ensuring that assembled products meet specific tolerances and safety requirements. Ameperosa et al.(2020) proposed the use of domain randomization techniques combined with deep neural networks to generate synthetic data to train a model for estimating the positions of bolts in an assembly[25]. Synthetic data allows for the generation of a large and diverse dataset, which can improve the performance and generalization of the model. The utilization of synthetic data in this study facilitated the model to predict bolt positions with 99.7% accuracy for unseen samples. This method can greatly decrease the dependence on expensive

and time-consuming manual data collection and annotation while simultaneously enhancing the precision and efficiency of quality control in manufacturing processes.

Despite notable advancements in utilizing synthetic data for quality control in assembly line applications, certain gaps remain. For example, there is a necessity for more research to create more efficient techniques to handle imbalanced datasets. Moreover, robust synthetic data generation methods that can create high-quality data for machine learning models are needed. These gaps offer potential prospects for future research in this field.

In addition to the gaps mentioned earlier, there are several other areas where further research is needed to improve the use of synthetic data in quality control for assembly line applications.

One area of concern is the reliability and interpretability of machine learning models used for quality control. Even though machine learning models can attain high accuracy in identifying defects or deviations from quality standards, comprehending how the model arrived at its conclusion can be challenging. The absence of interpretability can hinder adoption in industries that require transparency in decision-making processes. Additional research is needed to create machine learning models that are not just precise but also transparent and interpretable.

The possibility of bias in machine learning models used for quality control is an additional concern. The accuracy of machine learning models relies on the data they are trained on, and if the data is biased, so will the model. This can result in incorrect decisions or perpetuate existing biases. As a result, further research is necessary to create bias detection and mitigation techniques to guarantee that machine learning models used for quality control are unbiased and impartial.

Integrating synthetic data into current quality control methods might be challenging. This is particularly true in sectors where conventional quality control systems have been in place for an extended period of time and where the introduction of new technologies might

be received with opposition. Hence, additional study is required to identify the barriers to the adoption of synthetic data-based quality control and to design successful integration techniques.

Finally, the use of synthetic data in quality control presents ethical considerations, especially in regard to employment. The automation of quality control jobs might result in the loss of employment for those now doing these duties. Consequently, it is essential to evaluate the ethical implications of synthetic data-based quality control and to devise measures to limit possible negative employment effects.

In conclusion, although the use of synthetic data in quality control for assembly line applications has shown significant potential, more study is still required in a number of areas. To unleash the full potential of synthetic data in quality control and to ensure that it is utilized in a responsible, ethical, and effective manner, it will be essential to address these gaps.

3.1.2 Predictive Maintenance

Predictive maintenance is a technique used to identify potential failures in machinery before they occur. This method involves collecting and analyzing data from sensors and other sources to predict when equipment might fail. Predictive maintenance can help reduce downtime, maintenance costs, and improve the overall efficiency of an assembly line.

One challenge in implementing predictive maintenance is the availability of data. Gathering data on the performance of equipment involves substantial work and money, and it may not always be feasible to collect sufficient data to create reliable prediction models. Synthetic data can play an important part in resolving this issue.

In predictive maintenance, the purpose of synthetic data is to generate bigger, more diversified datasets that may be used to train and evaluate prediction models. Next, these models may be used to discover patterns and trends in the data that may suggest equipment

breakdowns. Moreover, synthetic data may be utilized to recreate various situations in order to verify the robustness of prediction models and uncover any flaws.

Using synthetic data in predictive maintenance can help companies reduce the time and cost of collecting and labeling real-world data. Synthetic data can be generated quickly and efficiently, allowing companies to create large datasets that capture a broad range of scenarios and conditions. By using synthetic data, companies can also minimize the risk of damaging expensive machinery during testing.

There is a growing body of literature on the usage of synthetic data for predictive maintenance. Many researchers have recognized the potential benefits of using synthetic data to supplement real-world data in predictive maintenance applications. For example, a study conducted by Guner et al. (2016) presented a simulation platform for anticipative plant-level maintenance decision support systems[31]. The platform incorporates physics-based modeling, statistical modeling, and machine learning approaches. The authors investigated the use of synthetic data production for predictive maintenance machine learning models, including the usage of GANs for data synthesis.

The authors demonstrate the effectiveness of the proposed approach using a case study involving a manufacturing system. According to the findings, incorporating synthetic data in the training of the predictive maintenance model resulted in better accuracy compared to models that only utilized real-world data. Furthermore, the authors described the benefits of utilizing synthetic data in predictive maintenance. These advantages include lowering expenses and time spent on data gathering and enhancing the effectiveness of predictive maintenance models in circumstances when acquiring real data may be problematic or restricted.

Although significant progress has been achieved in the use of synthetic data methodologies for predictive maintenance and data creation in the industrial industry, there are still a number of gaps to fill.

The absence of standards in data gathering and processing is a significant obstacle. Effectively training predictive maintenance models requires huge volumes of data, but if the data are not gathered regularly or evaluated in a uniform manner, it becomes difficult to compare outcomes across studies. This lack of standardization makes it difficult to identify best practices for predictive maintenance and retards the field's development.

For the development of reliable predictive maintenance models, huge quantities of high-quality data are also required. In reality, it might be challenging to get sufficient data that is both relevant to the situation at hand and of sufficient quality to train a model successfully. This difficulty may be especially severe in industrial contexts, where data may be created by a variety of equipment and systems, and where distinct data sources may use different formats or protocols.

There is also a need for more research on the integration of predictive maintenance and other manufacturing processes. Guner et al. (2016) proposed a simulation platform for anticipative plant-level maintenance decision support systems, but more work is needed to develop methods for integrating these systems into larger manufacturing processes[31]. Integrating predictive maintenance into overall production planning and control systems could lead to significant improvements in efficiency and uptime, but it requires a deep understanding of the underlying processes and systems.

In addition, there is a need for more studies on the ethical and social ramifications of using synthetic data methodologies in manufacturing. Concerns like data privacy, algorithmic bias, and the effects on employees and society must be addressed as synthetic data become more prevalent in this field. Thus, it is crucial that the collection and use of this data comply with ethical and equitable norms since predictive maintenance models depend on data from a number of sources.

Overall, the application of synthetic data techniques in predictive maintenance and data generation presents numerous opportunities in the manufacturing industry. To reach their

full potential, however, it is required to overcome the gaps and difficulties in this sector. This involves standardizing data collecting and processing, discovering methods for obtaining high-quality data, incorporating predictive maintenance into broader production processes, and evaluating the ethical and societal consequences of these technologies. By addressing these challenges, the field of predictive maintenance can continue to grow and thrive, bringing significant benefits to manufacturers and society as a whole.

3.1.3 Human-machine Collaboration

Human-machine cooperation refers to the collaboration between people and machines towards a shared objective, with each contributing their unique abilities and expertise. In flexible manufacturing, this collaboration is crucial for improving productivity, flexibility, and safety. Establishing effective human-machine cooperation poses a significant challenge due to the requirement of a massive amount of data needed to train AI models to anticipate and respond to human behavior.

Recent studies have investigated the utilization of machine learning algorithms to generate synthetic data as a solution to this challenge. Synthetic data can replicate human behavior on assembly lines and provide data to train AI models, reducing the need for expensive and time-consuming data collection from actual scenarios. One study that has investigated the use of synthetic data for improving collaborative human-robot flexible manufacturing applications is the paper by Sibona et al (2021). They proposed a data-driven framework that utilizes synthetic data to train AI models for predicting human behavior and improving collaboration between humans and robots[15].

While the paper by Sibona et al. (2021) provides valuable insights into the use of AI and machine learning for improving human-machine collaboration in flexible manufacturing[15], there may be gaps in the proposed framework that need to be addressed. For instance, the efficacy of the framework may rely on the quality and amount of the synthetic data

created, which in turn may be affected by a number of variables, such as the complexity of assembly line activities and the variety of human behavior. The ideal techniques for producing synthetic data and the framework's applicability to a variety of contexts need more investigation.

Synthetic data is a potential strategy for human-robot cooperation in assembly lines, but there are research gaps that need to be filled. One possible constraint is the quality and representativeness of the synthetic data, which may be affected by variables like the precision and dependability of machine learning techniques. More study is required to build trustworthy machine learning algorithms capable of creating synthetic data for human-robot interaction on assembly lines, taking into account issues such as feature selection, hyperparameters, and quality of training data. In addition, replicating complicated human behaviors and interactions in real-world circumstances may be difficult and lead to inadequate or biased data that may not reflect the whole spectrum of probable human behavior. To guarantee that the created synthetic data are accurate and representative, more study is required.

3.1.4 Supply Chain Management

Supply chain management entails managing and organizing activities associated with the manufacturing and delivery of products and services to clients. To ensure timely and cost-effective delivery of products, various tasks such as managing suppliers, overseeing production processes, logistics, and customer relations are carried out. With supply networks becoming increasingly intricate, there is a rising trend of utilizing artificial intelligence (AI) in supply chain management to enhance efficacy and reduce expenses.

One area where AI can be used in supply chain management is in generating synthetic data on assembly lines. For example, a recent paper by Kohtala et al. (2021) explored the use of synthetic data from CAD models for training object detection models in a virtual

reality (VR) industry application case[19]. The authors demonstrate that synthetic data can be used to train object detection models more efficiently than traditional methods.

Andres et al. (2021) proposed a mathematical model to optimize the production process for automotive plastic components[28]. The model considers machine capacity, setup time, lot size, and scheduling, among others, to minimize the total production cost while meeting the demand requirements. The paper's contribution lies in the use of synthetic data to improve the accuracy of the model while maintaining privacy and confidentiality. The proposed model is applied to a case study in the automotive industry, and the results show significant cost savings. The paper demonstrates the potential of using parallel flexible injection machines with setup common operators in supply chain management and provides insights for further research.

Simulation is another commonly used tool in supply chain management to evaluate different production scenarios and optimize system performance. For instance, Bikes et al. (1994) used simulation to study the sensitivity of assembly systems to component delivery in a logistics study[30]. The authors show that delays in component delivery can have significant impacts on production efficiency and lead to increased costs.

Sisca et al. (2015) proposed a hybrid model for aggregate production planning that uses both real and synthetic data[32]. The study highlights the challenges of production planning in a reconfigurable assembly unit for optoelectronics and proposes a model that combines an analytical model, a simulation model, and a genetic algorithm. The hybrid model uses synthetic data to model customer demand and component availability while maintaining privacy and confidentiality. The study demonstrates the potential of the hybrid model for improving the accuracy and efficiency of production planning in a flexible manufacturing environment.

The paper emphasizes the significance of precise data in supply chain management and the potential of synthetic data in enhancing production planning processes. The hybrid

model suggested in the paper presents a novel way of integrating production planning that can be implemented in different sectors, such as optoelectronics, electronics, and automotive. The findings of the research have implications for supply chain management and emphasize the importance of utilizing both real and synthetic data to enhance the precision and efficiency of production planning procedures.

Fiasché et al. (2016) proposed a hybrid method for aggregate production planning that uses both real and synthetic data[33]. The study focuses on the challenges of dealing with uncertainty and imprecision in a dynamic environment and proposes a method that incorporates fuzzy logic and multi-objective optimization. The authors highlight the importance of accurate data in production planning and the challenges of dealing with uncertainty and imprecision. The proposed method uses both real and synthetic data to model customer demand, production costs, and other input parameters and incorporates fuzzy logic to represent imprecision and uncertainty. It allows for multiple objectives and constraints to be optimized simultaneously and demonstrates the potential of synthetic data and fuzzy logic for addressing uncertainty and imprecision in production planning. This research presents a novel strategy for consolidating production planning, which can be employed across different sectors such as electronics, automotive, and manufacturing. It highlights the significance of advanced optimization methods in managing supply chains.

The utilization of synthetic data for training assembly line models has gained popularity because it can replicate various scenarios without costly and time-consuming physical testing. Nonetheless, the literature has not thoroughly investigated its potential usage in supply chain management. The current studies primarily focus on optimizing production, lot sizing, and scheduling using mathematical models and simulations. While these studies are beneficial in boosting production efficiency and decreasing expenses, they may not directly tackle the issues of inventory management and quality control.

Using synthetic data generated from CAD models could help enhance accuracy and efficiency in quality control and inventory management in assembly lines. By utilizing synthetic data, models can be trained to detect defects and determine the underlying causes of quality problems, leading to increased precision in quality control procedures. Furthermore, synthetic data can aid in streamlining inventory management by forecasting demand, pinpointing supply chain constraints, and emulating diverse scenarios.

The creation of synthetic data necessitates the use of precise AI and machine learning algorithms capable of replicating real-life situations in assembly lines. Further research into the utilization of AI and machine learning algorithms for generating synthetic data could unleash synthetic data's full potential in enhancing supply chain management. By leveraging synthetic data, supply chain managers can make better-informed decisions and optimize their operations, leading to enhanced performance, lowered costs, and increased customer satisfaction.

3.1.5 Process Control

Manufacturing relies on process control, which involves monitoring, analyzing, and optimizing manufacturing processes using various techniques and technologies. One promising area where process control can be utilized is the application of synthetic data generation in assembly lines. Synthetic data generation entails using machine learning algorithms to create artificial datasets that simulate real-world data. Specifically, in assembly lines, synthetic data generation can be used to generate datasets that reflect diverse scenarios, which can be used to train machine learning models for process control purposes. By utilizing synthetic data generation for process control, manufacturers can improve their production processes, reduce errors, and minimize waste.

Recent research has explored the use of synthetic data generation to improve operator guidance in manufacturing. This method entails employing machine learning algorithms to

produce synthetic datasets that replicate the assembly process of intricate products. By utilizing these datasets, machine learning models can be trained to anticipate the ideal order of assembly tasks for a specific product and offer live instructions to human operators based on those predictions.

In a recent study, researchers proposed a novel approach for using synthetic data generation to improve operator guidance in manufacturing. The study presented a system called ViTroVo[20], which is an artificial intelligence framework designed to provide adaptive operator guidance in complex manufacturing environments. The system generates synthetic datasets that simulate the assembly process for complex products, allowing for the optimization of assembly tasks and process control. The synthetic datasets are generated through a process of domain randomization, where 3D scenes are generated with randomized assembly components and distractor objects. Machine learning models for assembly prediction are trained and assessed using these scenarios, which can subsequently guide human operators during the assembly procedure. This system can be utilized in various industries, including aerospace, automotive, and electronics.

Experimental results showed that ViTroVo was effective in improving the efficiency and quality of assembly tasks. This method effectively decreased the frequency of errors made by human operators during the assembly process while simultaneously enhancing overall productivity. The study offers a hopeful path for employing machine learning and synthetic data generation in controlling manufacturing processes, which could have notable implications for augmenting the quality and efficiency of manufacturing procedures.

Although utilizing synthetic data generation for process control in manufacturing appears encouraging, there are still various obstacles and inadequacies that require attention. One of the primary challenges is generating synthetic datasets that precisely mirror the intricacies of actual manufacturing settings. For instance, capturing all the pertinent factors that influence

assembly tasks, such as environmental circumstances, equipment inconsistencies, and human elements, may pose difficulties.

Another hurdle is the requirement for efficient machine learning algorithms that can acquire knowledge from synthetic data and offer valuable instructions to human operators. To develop such algorithms, extensive amounts of training data and precise tuning of model parameters may be necessary, which can be both expensive and time-consuming.

Lastly, it is crucial to contemplate the conceivable ethical and societal consequences of utilizing artificial intelligence and machine learning in controlling manufacturing processes. There could be apprehensions regarding the influence of automation on employment and job stability, in addition to the possibility of partiality in machine learning models that could adversely affect employees or consumers.

Addressing these gaps and challenges will be crucial for advancing the use of synthetic data generation and machine learning in manufacturing process control, and ensuring that these technologies are used in ways that are both effective and socially responsible.

3.1.6 Process Optimization

Process optimization involves improving a manufacturing process to achieve greater efficiency, higher quality, and lower costs. Attaining process optimization may entail recognizing and resolving bottlenecks, minimizing waste, refining resource usage, and enhancing production throughput. One method of achieving process optimization is by utilizing synthetic data, which can simulate various scenarios and reveal areas for enhancement.

One example of the use of synthetic data for process optimization is in assembly line manufacturing. Zheng et al. (2020) used a virtual prototyping approach to generate synthetic data for the detection of modules in modular integrated construction[17]. They used the 3D CAD software tool Revit to model the construction process in a virtual environment and generated synthetic data to train a transfer learning model for module detection. The authors

found that their approach improved the accuracy of module detection and could be applied to improve process optimization in other manufacturing contexts.

Apornak et al. (2021) presented a hybrid approach for solving the flexible flow-shop problem (FFSP) by combining Taguchi-based computer simulation and data envelopment analysis (DEA) techniques[36]. The authors utilized synthetic data generated by the computer simulation model to optimize the FFSP with multiple objectives, including minimizing makespan, total completion time, and work-in-process. The use of synthetic data allowed for efficient and effective exploration of the design space, enabling the identification of optimal FFSP solutions that would have been difficult or impossible to achieve using only real-world data. The results showed that the proposed hybrid approach could effectively solve the FFSP, providing significant improvements over traditional methods.

Mubarak et al. (2020) described a new approach for generating synthetic data to optimize the pipe spooling process[5]. To generate synthetic data, the authors developed an industrial pipelines data generator (IPDG) that uses Hidden Markov Models (HMMs) to simulate the pipe spooling process. The generated data was then used to optimize the spooling process by testing different scenarios and identifying the optimal parameters. The results showed that the IPDG was able to generate realistic data that accurately represented the pipe spooling process. The use of synthetic data allowed for more efficient and cost-effective testing of different scenarios, leading to improved process optimization. The authors suggest that this approach could be extended to other manufacturing processes to support optimization and decision-making.

While the generation of synthetic data holds promise for enhancing process optimization, significant gaps and obstacles remain to be addressed. One of the main gaps is the difficulty in accurately modeling the variability and complexity of real-world assembly line processes using synthetic data. This requires a deep understanding of the underlying process dynamics, which may be difficult to capture using traditional statistical techniques. Another

gap is the requirement to verify that the synthetic data appropriately represents the variety of inputs and circumstances that the process may experience in real-world settings. This requires careful consideration of the data input types, their distributions, and the models and algorithms used to generate synthetic data. Ultimately, the quality and quantity of available training data, as well as the complexity of the underlying process dynamics, may constrain the utility of synthetic data production for process optimization. A mix of process optimization, data analytics, and sophisticated machine learning and artificial intelligence approaches is required to address these gaps. Future research should concentrate on establishing more effective and economical ways for synthetic data creation that can correctly reflect the diversity and complexity of real-world processes, and that can be used for a broad range of industrial applications.

3.2 Data Challenges

Data challenges are ubiquitous in synthetic data generation on assembly lines, regardless of whether it is for quality control or process control applications. The shortage of accessible data that precisely depicts different scenarios and the steep expenses involved in gathering and labeling real-world data are significant challenges. The precision of the data utilized for training synthetic models is also crucial because low-quality data can produce partial or inaccurate machine learning models. Variability of the data and data privacy and security concerns further compound the challenges in creating synthetic datasets that adequately represent the wide range of possible scenarios encountered on assembly lines. In this section, we delve into these data issues in detail.

3.2.1 Data Scarcity

Data scarcity is another significant challenge that must be addressed when generating synthetic data on assembly lines. This challenge arises when the available data is not sufficient to generate accurate models and predictions, which can result in ineffective synthetic data that does not reflect the reality of the assembly line process.

Data scarcity can occur in several ways. One common cause is the lack of data from certain production scenarios or worker actions. For example, if certain types of assembly line products or components are not frequently produced, there may be a scarcity of data on how workers handle those products or components, which can lead to inaccurate synthetic data. Similarly, if certain workers or teams have unique skills or work processes that are not well-represented in the available data, this can also lead to data scarcity.

Another cause of data scarcity is the need for high-quality, labeled data for supervised learning models. This can be particularly challenging in the assembly line context, where labeled data may be scarce or difficult to obtain. For example, labeling data on worker actions or errors can be time-consuming and labor-intensive, and may require additional sensors or cameras on the assembly line.

To address the data scarcity challenge, it is important to identify and prioritize the most critical data needs for generating accurate synthetic data. This may involve collecting additional data through new sensors or cameras or collaborating with other organizations or industries to obtain relevant data. Another approach is to use unsupervised learning techniques that do not require labeled data or to use transfer learning techniques that leverage existing labeled data from other industries or applications.

In summary, data scarcity is a significant challenge when generating synthetic data on assembly lines. To tackle this challenge, it is crucial to recognize and give importance to vital data requirements and explore alternative approaches like unsupervised learning or transfer learning. By addressing the scarcity of data, it is plausible to create more precise synthetic

data that mirrors the actuality of the assembly line process and results in more efficient manufacturing processes.

3.2.2 Data Proprietary

Utilizing synthetic data to enhance manufacturing processes might be difficult if the needed data originates from many departments or firms. Each source may own data that they are hesitant to disclose or that is subject to use limitations. A manufacturing corporation, for instance, may have proprietary information about their production methods that they do not like to share with rivals or academics developing synthetic data models. Additionally, legal or ethical considerations regarding worker or customer data may limit its use for synthetic data generation.

To overcome this challenge, it's crucial to establish clear guidelines and agreements regarding data ownership, access, and usage. This could involve negotiating data-sharing agreements with different organizations or departments or developing ethical guidelines for worker or customer data usage.

In addition, it may be required to use methods such as data anonymization or differential privacy in order to secure the privacy and confidentiality of the data while still enabling the development of accurate synthetic data. Overall, the data proprietary challenge is an important consideration when generating synthetic data on assembly lines. By establishing clear guidelines and agreements, and using appropriate privacy and confidentiality techniques, it is possible to address this challenge and generate high-quality synthetic data that reflects the reality of the assembly line process.

3.2.3 Data Quality

Data quality is one of the most significant challenges in generating synthetic data for assembly lines. The accuracy and reliability of synthetic data generated for assembly lines depend largely on the quality of the source data used to train machine learning models. Poor-quality source data can result in synthetic data that is inaccurate or unreliable, potentially leading to errors in process control and reduced manufacturing efficiency.

There are several factors that can affect the quality of source data for generating synthetic data on assembly lines. These include:

1. Incomplete or missing data: If the source data used to train machine learning models is incomplete or contains missing values, the resulting synthetic data may also be incomplete or inaccurate. This may be especially troublesome if the missing data reflect crucial process parameters or other variables that might affect the correctness of the synthetic data.

2. Data errors: Mistakes in the original data may also contribute to inaccuracies in the synthetic data that is generated. For instance, if sensor data from assembly line equipment is noisy or includes measurement errors, the synthetic data derived from this data may likewise be imprecise.

3. Data biases: Biases in the source data can lead to biases in the synthetic data generated from this data. For instance, if the source data has a disproportionate number of instances of specific kinds of items or assembly processes, the synthesized data may be skewed toward these aspects.

4. Data normalization: The act of normalizing data for use in machine learning models might alter the quality of synthetic data created from these sources. The normalization procedure might add mistakes or distortions to the final synthetic data if it is not properly developed and applied.

To overcome these obstacles, it is essential to carefully choose and preprocess source data to ensure that it is precise, exhaustive, and representative of the whole spectrum of conceiv-

able assembly line events. This may need data cleansing, standardization, and augmentation approaches, as well as careful assessment of any biases in the source data. In addition, it may be required to construct machine learning models that are resistant to mistakes and biases in the original data, so that the resultant synthetic data is as accurate and trustworthy as feasible.

3.2.4 Data Volume

Data volume is another important challenge in generating high-quality synthetic data on assembly lines. Often, vast amounts of data must be collected and processed in order to develop synthetic data that correctly depicts the complete range of conceivable assembly line situations. Collecting and analyzing voluminous amounts of data could be resource- and time-intensive, leading to storage and processing difficulties.

To generate high-quality synthetic data, it is necessary to acquire and analyze an adequate quantity of data, which might be challenging. If insufficient data is obtained and analyzed, the synthetic data produced may be erroneous or biased. On the other side, excessive data collection and processing may result in noise or complexity that is unneeded. Many aspects, such as the assembly process's complexity, the variety of product standards, and the machine learning methods used, influence the determination of the right quantity of data.

One of the challenges in data generation is coping with massive volumes of data storage and processing. This method may be highly costly, requiring significant computational resources and posing scalability and performance issues. To tackle these obstacles, it's crucial to choose and deploy data storage and processing solutions that match the requirements of the data being generated.

In addition, data amount may influence the speed and effectiveness of machine learning training. The longer it may take to train a machine learning model, the bigger the dataset. This may provide difficulties in terms of computing resources, training time, and model

performance. To solve these obstacles, it may be important to choose and deploy machine learning methods suited for large-scale datasets.

The generation of high-quality synthetic data on assembly lines is hampered by the sheer quantity of data. By carefully managing the amount of data collected and processed and implementing the appropriate data storage and processing solutions, it is possible to generate synthetic data that accurately reflects the full range of possible assembly line scenarios, and that can be utilized for effective process control and quality assurance.

3.2.5 Data Diversity

The generation of high-quality synthetic data on assembly lines is hampered by the variety of data. Diversity of data refers to the variety and complexity of the data that requires synthesis. To generate a synthetic dataset that is reflective of the whole range of possible situations on an assembly line, the synthetic data must be varied enough to capture the diversity and complexity of the real-world data.

Due to the multitude of variables that might influence assembly-line events, such as changes in product configurations, manufacturing process modifications, and environmental variances, it can be challenging to collect all conceivable forms of assembly-line events. To overcome this difficulty, it may be essential to develop assembly line-specific data collecting and processing techniques. By doing so, it may be able to verify that all pertinent data is collected and used to produce correct synthetic data.

Ensuring that synthetic data is indicative of real-world data is a further difficulty associated with data variety. Artificial data that is skewed or insufficient might result in machine learning models that are ineffective at regulating the assembly process or spotting quality concerns. To solve this difficulty, it may be important to rigorously verify the synthetic data by comparing it to real-world data and ensuring that it covers the whole range of variability and complexity found in real-world data.

In addition, the variety of data may influence the precision and efficacy of machine learning models. Models of machine learning trained on inadequately diverse datasets may be prone to overfitting or lack the capacity to properly transfer to fresh or uncharted situations. To address this difficulty, it may be important to develop and implement machine learning algorithms that are optimized for several datasets and can successfully capture the data's diversity and complexity.

Overall, data variety is a significant obstacle to producing high-quality synthetic data on assembly lines. By carefully designing and executing data collection and processing procedures, validating the synthetic data, and selecting appropriate machine learning algorithms, it is possible to generate synthetic data that accurately reflects the full range of possible assembly line scenarios and can be used for effective process control and quality assurance.

3.2.6 Ethical Considerations

Data Privacy

Protecting the privacy of employees and other stakeholders while producing synthetic data on assembly lines is a serious concern that must be addressed. Assembly line data is sensitive and may include personally identifiable information, such as names and addresses, that must be safeguarded from exposure to unauthorized parties.

Data anonymization is crucial for mitigating the danger of personal information being revealed. This entails deleting any personally identifying information from the data to guarantee that it cannot be traced back to specific persons. It is essential to guarantee that the anonymization procedure is comprehensive, and that no personally identifiable information is revealed mistakenly.

The potential for data breaches is another privacy worry. Assembly line information is important and might be targeted by cyberattacks. Encryption, firewalls, and more security

measures must be used to prevent unwanted access. In addition, it is essential to keep the data in secure areas and limit access to only authorized individuals.

Also, it is crucial to evaluate the repercussions of data sharing on privacy. Sharing information with third-party contractors or service providers raises the likelihood of privacy breaches. To avoid data abuse, it is crucial to verify that the parties receiving the data have proper data security techniques and that data-sharing agreements are in place.

Overall, generating synthetic data on assembly lines requires careful consideration of data privacy. Anonymizing data, securing it, and establishing data-sharing agreements are necessary measures to protect the privacy of workers and other stakeholders.

Data Bias

Data bias is a significant challenge that must be addressed when generating synthetic data on assembly lines. When the data utilized to produce synthetic data is unrepresentative or slanted toward specific conclusions, bias may develop. This may result in inaccurate models and projections, which may have a negative impact on assembly line operations.

Data bias may occur if the data used to generate synthetic data is restricted in its range. This may occur when data is biased toward certain demographic groups or production circumstances, resulting in skewed synthesized data, faulty models, and erroneous forecasts. To tackle this challenge, it is critical to collect various data from various sources and ensure that it appropriately represents a wide range of production scenarios and worker demographics.

The use of old data that may not represent current circumstances on the assembly line is another form of data bias. For example, if a manufacturing process has recently been modified, previous data may no longer be reflective of the current process, resulting in skewed synthetic data. To overcome this issue, it is critical to use current data that reflects the actual status of production. It is also important to consider the potential impact of biased synthetic data on workers. If the synthetic data is used to make decisions about worker performance

or productivity, biased data can unfairly penalize certain groups or individuals, leading to further inequality and injustice.

To address the difficulty of data bias, it is crucial to adopt procedures to detect and rectify biased data. This may be accomplished by using algorithms that discover and rectify biases in the data, or by employing human supervision and evaluation of the data used to produce synthetic data.

In conclusion, when producing synthetic data on assembly lines, data bias is a serious obstacle that must be overcome. To alleviate this issue, it is crucial to acquire varied, current data and apply techniques to detect and rectify data biases. This will ensure that the resulting synthetic data is representative and accurate, leading to more effective assembly line processes and fair treatment of workers.

Chapter 4

Research Design

This chapter of my thesis will cover the research design framework for synthetic data generation, along with an exploration of the various types of data that can be collected from assembly stations. The chapter aims to equip readers with the necessary knowledge and tools to conduct data-driven research projects related to assembly stations, by comprehensively explaining the principles of generating synthetic data for this type of data and discussing various techniques involved. As a result, readers should have a thorough understanding of synthetic data generation by the end of the chapter.

4.1 Framework for Synthetic data generation

This section provides a comprehensive step-by-step guide to generate synthetic data for assembly lines. The generated data can be utilized for making predictions or conducting analysis. The process begins with collecting data from assembly lines, followed by a thorough cleaning process to eliminate errors and inconsistencies as shown in the Figure 4.1. The subsequent step entails selecting the most suitable synthetic data generation technique based on the nature of the data. Finally, the generated synthetic data is evaluated to ensure that

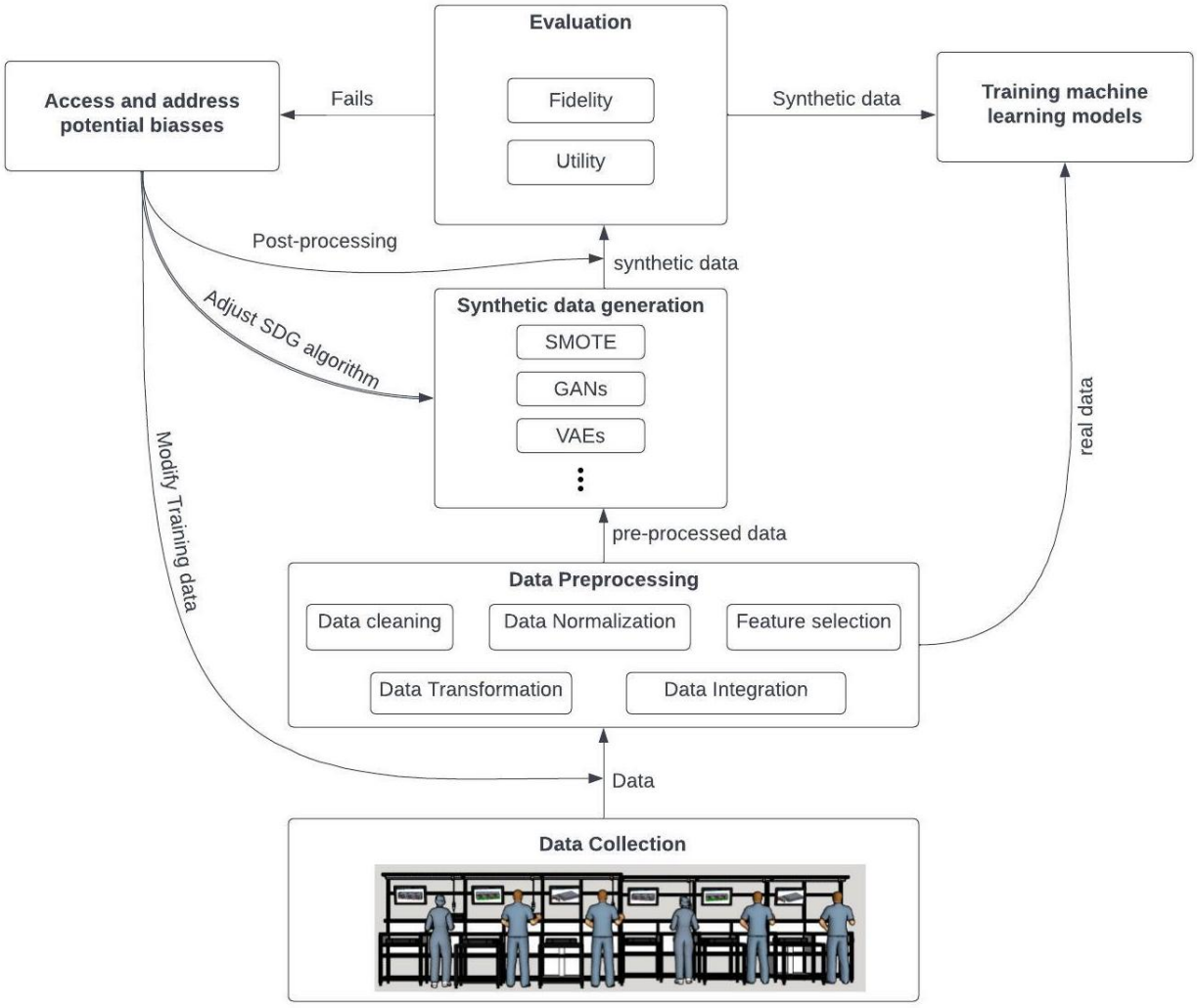

Figure 4.1: Data generation framework

it maintains realism and similarity to the original data. Further, the framework is described in detail, including a thorough explanation of each step and the various methods that can be used. This guide offers researchers the requisite knowledge to create high-quality synthetic data for assembly lines to use in their research or analysis.

4.1.1 Data Collection

To develop synthetic data that precisely resembles actual data for assembly lines, the first step is to collect data from diverse sources, such as sensors and cameras. For this synthetic data to be high quality, the obtained data must be accurate and dependable. Thus, it is vital to adhere to a rigorous and extensive data collection strategy. This method assures that we collect reliable and accurate data, which is important for synthesizing data that closely matches the original.

It is essential to examine the frequency and length of data collection, as well as the location of sensors and equipment throughout the data-gathering process. These factors may significantly affect the precision and dependability of the obtained data. Thus, we must assess them thoroughly to guarantee that we can provide high-quality synthetic data for assembly lines that precisely represents the original data. Failure to account for these variables may result in errors that have negative effects on the produced synthetic data.

Finally, gathering high-quality data is required if we want to develop accurate and trustworthy synthetic data for assembly lines[41].

4.1.2 Data Preprocessing

Once the data has been collected, the subsequent step in the framework for synthetic data generation on assembly lines is pre-processing. This entails refining and formatting the gathered data to make it suitable for utilization in the synthetic data generation procedure.

This stage is crucial for guaranteeing the precision and authenticity of the synthesized data. The pre-processing method to be employed is determined by the nature of the data.

Data Cleaning

The elimination of missing or flawed data points from the data set is a critical aspect of pre-processing, commonly known as data cleaning.

Data cleaning is usually essential because of various factors, such as inaccuracies in measurement, errors in data entry, or problems with equipment[64]. Such discrepancies may cause missing or erroneous data points that could distort the outcomes of subsequent data analysis or modeling. Therefore, identifying and eliminating such errors is crucial to guarantee the precision and credibility of the data. There are multiple methods available for cleaning data, including:

1. Imputation: Imputing missing data is a prevalent technique for handling missing values in a dataset, which can occur due to various reasons[65]. Imputation algorithms can estimate missing values by exploiting the information offered in the dataset.

Several imputation approaches exist, including mean imputation, median imputation, regression imputation, k-nearest neighbors imputation, and multiple imputation[65]. Mean imputation replaces missing values with the feature's mean, while median imputation uses the feature's median. Using a regression model based on other characteristics of the dataset, regression imputation guesses missing values. K-nearest neighbors imputation imputes the missing value using the values of the k-nearest observations to the missing observation.

Evaluating the constraints and assumptions of each imputation approach is essential since they can introduce bias and impact the variability of the dataset. Before utilizing the imputed data, it is important to assess the effects of imputation on analytical findings.

2. Deletion: Data deletion is a pre-processing method used to delete missing, irrelevant, or redundant data points from a dataset is data deletion[65]. It is useful for cleaning datasets,

but it should be used with caution to prevent unforeseen effects such as the loss of vital information, bias in outcomes, and a reduced accuracy in machine learning models. To eradicate bias, the removed data should be scattered randomly across the dataset.

3. Interpolation: Data interpolation is a pre-processing method used to fill in missing values in a dataset. It entails predicting missing data points based on current data using techniques such as linear, spline, or k-nearest neighbor interpolation[65].

Interpolation is most useful when missing data is scattered randomly throughout a big dataset and the predicted values are sufficiently credible. Nevertheless, it should be utilized with care since if not used correctly, it can produce mistakes or biases in the data. Interpolation should not be used to fill big gaps in missing data since it can provide erroneous or unreliable results.

It is critical to distinguish data interpolation from data imputation, which includes replacing missing values with predicted values based on statistical analysis or machine learning methods. Interpolation is a simpler approach that guesses missing values based on the dataset's known data points, while data imputation needs more advanced estimating algorithms.

4. Data augmentation: Data augmentation techniques involve creating new data points based on the existing data. This can be done by adding noise to the existing data or by generating synthetic data using statistical models[65].

5. Use of expert knowledge: Expert knowledge can be employed in certain instances to estimate missing values. When the missing value pertains to a measurement that is improbable to have undergone significant changes over time, an expert can make an estimate based on their knowledge of the system[65].

Regardless of the method used, it is crucial to evaluate the effect of missing or corrupted data on analysis outcomes and document any pre-processing measures taken to tackle these problems. By addressing missing values or corrupted data points, researchers can ensure that

their analysis is based on accurate and complete data, resulting in more reliable findings and conclusions.

Data Transformation

Transformation of data is a crucial stage in the preprocessing of real-world data, since it converts the data into a more useable and comprehensible format[65]. Transformation of data entails replacing the original data representation with one that is more suited for analysis.

Data transformation is used for several purposes. Secondly, it may lessen the impact of outliers or extreme values in the data, which can distort the conclusions of an investigation. Two, data transformation may assist in linearizing the connections between variables, making them more suitable to linear regression analysis.

Employing diverse data transformation methods, contingent on the data type and research objectives can help enhance the normality of the data distribution, which is commonly a precondition for conducting statistical tests. Some common data transformation techniques include:

1. Scaling: Scaling is used to alter data so that it fits inside a certain range. This may be accomplished via the use of min-max scaling or standardization.

2. Log transformation: Log transformation is used to transform data that is highly skewed or has a non-normal distribution. This method can be used to make the data distribution more normal.

3. Box-Cox transformation: Box-Cox transformation is a sort of power transformation used to alter skewed data distributions.

4. Fourier transformation: Time-series data are transformed into the frequency domain using the Fourier transform. This technique can assist in identifying any existing patterns or trends within the data.

5. Principal Component Analysis (PCA): PCA is a technique utilized to reduce the dimensionality of data, which can aid in identifying patterns and correlations present within the data.

6. Wavelet transformation: The wavelet transformation approach may help the spotting of patterns or trends by evaluating the time-frequency connections within the data.

Data Normalization

Data normalization is an essential pre-processing method used to put data into a uniform format, facilitating comparisons between various samples and minimizing the impact of outliers or discrepancies in units and scales. This is a vital stage in several data analysis and machine learning jobs. The objective of normalization is to rescale the data to a similar range or unit, often between 0 and 1 or -1 and 1, depending on the approach used[65].

Normalization methods are applicable to a variety of data types, including continuous and categorical data. When working with characteristics that have multiple sizes or units, data normalization is sometimes necessary. In such situations, normalizing the data to a similar range or unit might enhance the precision of the analysis or forecast. Normalization is also beneficial when dealing with data that has outliers or extreme values. Normalization methods may assist in reducing the impact of these outliers and making the data more typical of the underlying distribution.

Min-Max Scaling, Z-score normalization, and Decimal scaling are prominent data normalizing approaches. The choice of normalizing approach relies on the data kind and distribution, as well as the analysis's unique application and objectives. It is also crucial to evaluate the influence of normalization on the interpretability of the data, since some normalization approaches may alter the connections between variables or distort the data in a manner that makes them difficult to interpret.

Feature Selection

Feature selection is an important pre-processing technique that involves selecting a subset of relevant features from the original set of features in a dataset. The purpose of feature selection is to enhance the performance of a machine learning model by lowering the dataset's dimensionality and deleting unnecessary or redundant features that may generate noise or bias[65].

Many applications use feature selection, including image recognition, natural language processing, and signal processing. In these applications, the dataset may include several attributes, some of which may not be helpful or relevant to the current situation. By picking a subset of important characteristics, we may lower the model's computational complexity and increase its precision and generalization performance. There are several approaches available for selecting features, including:

1. Filter methods: These methods select features based on some statistical measure, such as correlation, mutual information, or chi-square. They are typically fast and computationally efficient, but they may not always select the most relevant features. This methods involve ranking features based on their relevance to the target variable, and selecting the top-ranked features.

2. Wrapper methods: These methods evaluate the performance of a machine learning model using different subsets of features. They are computationally more expensive than filter methods, but they may lead to better performance by selecting more relevant features.

3. Embedded methods: These methods incorporate feature selection as part of the model building process. For example, decision tree algorithms such as Random Forest can select important features during the training process.

4. Principal Component Analysis (PCA): This approach can be used for both feature selection and data compression. It entails translating the data into a lower-dimensional space while preserving as much data variation as feasible.

The implementation of a feature selection approach is contingent on the nature of the data and the addressed issue. In general, it is vital to choose characteristics that are pertinent to the situation at hand, while eliminating redundant or unneeded characteristics. This can assist increase the accuracy and generalizability of machine learning models.

Data Integration

The process of merging data from many sources into a cohesive format for analysis is known as data integration. Data integration is often used in pre-processing when data obtained from diverse sources or in different forms must be merged and turned into a common format for analysis[65].

Data merging, which includes joining datasets with a shared variable or key, is a typical strategy for integrating data.

Data stacking is another strategy that combines datasets with comparable variables but distinct observations. This is often used when data is obtained over various time periods or in different places, but the variables are the same or comparable.

Data fusion is a process that combines various sorts of data from several sources. This may include merging sensor data, picture data, and other data sources to generate a more complete dataset.

The ultimate objective of data integration is to establish a uniform and consistent dataset that can be readily evaluated and utilized to draw conclusions and make choices. In situations where data is acquired from various sources or in different formats, data integration may assist in decreasing redundancy, increasing data accuracy, and promoting more efficient analysis.

4.1.3 Synthetic Data Generation

The generation of synthetic data comes after the assembly line data is processed in the synthetic data production framework. To achieve the objective of generating synthetic data, various techniques are employed, such as the use of generative adversarial networks (GANs), variational autoencoders (VAEs), and other deep learning approaches. The primary aim of producing synthetic data is to create new data that can replicate the patterns and attributes of the source data while also introducing some degree of variation to avoid producing an identical copy of the original.

An advantage of synthetic data generation is that it can expand existing datasets, enhancing the quantity and diversity of the data available for training machine learning models. Furthermore, synthetic data can simulate challenging or hazardous situations that may be impractical or dangerous to replicate in the real world, leading to more comprehensive testing and validation of assembly line processes.

Some techniques that can be used for synthetic data generation include:

1. Random Sampling
2. Synthetic Minority Over-sampling
3. Adaptive Synthetic Sampling
4. Random Over-sampling
5. Safe-level Synthetic Minority Over-sampling
6. Borderline Synthetic Minority Over-sampling
7. Gaussian Mixture Model
8. Autoencoders
9. Variational Autoencoders
10. Generative Adversarial Networks
11. PointNetGAN
12. Deep Convolutional GAN

13. VAE-GAN
14. Conditional GAN
15. Recurrent Neural Network
16. StyleGAN
17. Deep Belief Network
18. Transformer-Based Time-Series GANs
19. Convolutional Autoendcoder
20. PixelCNN
21. Generative Flow Networks
22. Hidden Markov Model
23. Autoregressive Integrated Moving Average
24. Long Short-Term Memory
25. Convolutional Neural Networks
26. Recurrent Variational Autoencoder
27. Conditional Variational Autoencoder

Random Sampling

Random sampling is a fundamental method used in synthetic data creation for assembly line manufacturing applications to generate a sample of data points that are representative of the original dataset. It is particularly useful when working with large datasets when training with the entire dataset may not be feasible. The method involves selecting data points at random, with no particular criterion or sequence, from the original assembly line dataset[66].

In assembly line manufacturing, stratified random sampling is also commonly used, whereby data points are divided into categories and then selected at random within each category to create a more representative synthetic dataset.

Random sampling is critical for the development of synthetic data for assembly line applications because it helps to provide a varied dataset that mimics the underlying distribution of the original dataset. By randomly selecting a diverse range of data points, the synthetic dataset can capture the variances and patterns contained in the original dataset, thereby enhancing the accuracy and reliability of the trained machine learning models.

Synthetic Minority Over-sampling

SMOTE(Synthetic Minority Over-sampling) is a technique that can be employed in manufacturing or assembly lines to generate synthetic data for machine learning models. When dealing with imbalanced datasets, where one class has significantly more or fewer instances than the other, SMOTE can be used to address the issue[66]. By generating synthetic data for the minority class, SMOTE can create a more balanced dataset that accurately represents the distribution of the original dataset. This can be beneficial for improving the accuracy and dependability of machine learning models used in quality control, predictive maintenance, and other applications in manufacturing or assembly lines.

Adaptive Synthetic Sampling

ADASYN, or Adaptive Synthetic Sampling, is a data augmentation technique that can be applied to manufacturing or assembly lines to balance class distribution in imbalanced datasets. For example, in quality control tasks, there may be an imbalanced dataset with significantly fewer defective products than non-defective products. ADASYN generates synthetic data points for the minority class samples that are more difficult to learn than others. It uses the density distribution of minority class data points to determine the number of synthetic samples to be generated for each point. More synthetic samples are generated in areas where the minority class density is low, and fewer samples where it is high. The ADASYN algorithm has been found to be effective in addressing imbalanced datasets, particularly

when there is a significant class distribution disparity. It can improve the performance of machine learning models by increasing the volume of data for the minority class, resulting in better generalization of the model[66].

Random Over-sampling

ROS(Random Over-sampling) can be used in manufacturing and assembly lines to address imbalanced datasets, where certain components or parts might be less frequent than others. For example, in a production line for an automotive company, some components may be produced in smaller quantities than others, resulting in an imbalanced dataset. Using ROS, synthetic data can be generated for these minority components, helping to balance the dataset and improve the performance of classification models.

By randomly selecting examples from the minority class and generating synthetic examples similar to them, ROS can help create a more diverse dataset that accurately represents the manufacturing or assembly line. This can lead to better decision-making regarding maintenance schedules, equipment replacement, and other critical operations. The simplicity and computational efficiency of the ROS algorithm make it a practical and useful tool for manufacturing and assembly lines with imbalanced datasets[66].

Safe-level Synthetic Minority Over-sampling

SLSMOTE, or Synthetic minority Over-sampling Technique using SMOTE, can be applied to assembly line data where there are imbalanced classes, such as identifying faulty products in a production line. In this scenario, there may be fewer instances of faulty products than non-faulty ones, making it challenging to train a classifier that can accurately detect the faulty products[66]. SLSMOTE can help to address this issue by generating synthetic instances that are similar to the minority class (i.e., faulty products) and improving the overall balance of the dataset.

To apply SLSMOTE to an assembly line dataset, one would first identify the minority class, which in this case would be the faulty products. The next step would be to train a classifier on the original dataset, which would identify the challenging minority instances based on their distance to the decision boundary. Once the challenging minority instances have been identified, SLSMOTE can be applied to generate synthetic instances around those instances using SMOTE. The resulting balanced dataset can then be used to train a classifier that can accurately detect faulty products in the assembly line. Overall, SLSMOTE can help improve the performance of classification models in assembly line scenarios with imbalanced datasets.

Borderline-SMOTE

In manufacturing, BLSMOTE(Boderline-SMOTE) can be used to address imbalanced datasets in various contexts. For instance, in quality control, where the majority of the products produced are considered of high quality, while only a few are of low quality, BLSMOTE can be used to generate synthetic data points to balance the class distribution. By generating synthetic samples for minority class data points located near the borderline between high and low-quality products, BLSMOTE can help to identify areas where production processes need improvement to ensure consistent quality.

In addition, BLSMOTE can be used in predictive maintenance applications, where data on machine failures is often imbalanced, with a small number of machines failing compared to the number of machines that function correctly. By generating synthetic data points near the borderline between healthy and faulty machines, BLSMOTE can improve the accuracy of machine failure prediction models, allowing maintenance teams to take preventive measures before costly equipment breakdowns occur.

Overall, BLSMOTE can be a valuable tool in manufacturing to balance class distribution in imbalanced datasets and improve the performance of predictive models in various

applications; ultimately helping to enhance efficiency, reduce costs, and improve product quality.

Gaussian Mixture Model

In manufacturing, Gaussian Mixture Models (GMMs) can be used to generate artificial data that resembles the original data to simulate and test different scenarios. For example, a GMM can be trained on data from a manufacturing process to identify distinct subgroups within the data, such as products with specific features or defects. The GMM can then generate synthetic data points by sampling from the learned mixture model, creating artificial scenarios for testing the manufacturing process[67].

In addition, GMMs may be used to discover abnormalities or outliers in the data, such as uncommon product attributes or unanticipated production modifications. By spotting these irregularities, industrial processes may be modified to decrease waste and increase productivity.

It is important to note that the number of clusters selected for the GMM should be chosen carefully to avoid overfitting and generate data that accurately represents the underlying data distribution.

Autoencoders

Autoencoders are a type of neural network that learn the underlying patterns and structures in a dataset to generate synthetic data. They consist of two main parts: an encoder network that reduces the dimensionality of the input data, and a decoder network that reconstructs the original input from its encoded form[68].

During training, the autoencoder learns to extract the most significant data characteristics and builds a compressed representation that can be used to generate new synthetic data points. The network can then be used to produce fresh samples by randomly selecting

points in the lower-dimensional space and sending them through the decoder to reconstruct the relevant data points.

Autoencoders can be particularly useful in generating synthetic data for assembly lines where there is a need to model the intricate relationships between different components and optimize the manufacturing process. By creating new synthetic samples that resemble the original data, autoencoders can aid in designing more efficient assembly lines, reducing downtime and costs.

Variational autoencoders

Variational autoencoders (VAEs) are a type of generative model used for generating synthetic data during pre-processing. They work by learning an encoder and a decoder network to map input data to a latent space representation and then back to the input data. By sampling from the learned distribution in the latent space, VAEs can generate new data points similar to the training data[69].

VAEs can be used to generate synthetic data that mimics real-world sensor readings from the production line. By training the VAE on a dataset of sensor readings, it can learn the underlying patterns and structures in the data and create new synthetic sensor readings. These synthetic data points can then be utilized to augment the original dataset, allowing for a more robust modeling and analysis of the assembly line's performance.

Additionally, VAEs can also be used to detect anomalies in sensor data by comparing the reconstructed sensor readings to the original readings. Any significant discrepancies could indicate a malfunction or abnormality in the assembly line, allowing for prompt maintenance and repair. Overall, VAEs have the potential to improve the performance and reliability of assembly lines by generating more comprehensive datasets and detecting anomalies in real-time.

Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a type of deep learning model that generates synthetic data. GANs consist of two neural networks, a generator and a discriminator[70]. The generator takes in random noise and produces synthetic data samples that resemble real data. The discriminator is trained to distinguish between synthetic and real data samples.

During training, the generator aims to create realistic synthetic samples that can trick the discriminator into thinking they are real. The discriminator tries to correctly classify synthetic and real samples. As the networks improve, the generator generates more realistic synthetic samples that become harder for the discriminator to differentiate from real samples.

GANs have diverse applications such as generating images and videos of assembly lines. However, their usage comes with a downside as they require significant computational resources and large amounts of data. Furthermore, the quality of synthetic data created by GANs might vary dependent on the design and hyperparameters of the network.

PointNetGAN

PointNetGAN can be applied to the manufacturing industry in various ways, particularly in assembly line production processes that require the generation of 3D point cloud data. For example, it can be used to improve the accuracy of quality control inspections by generating realistic point cloud data that mimics real-world manufacturing defects. This can help identify and correct defects earlier in the production process, leading to a reduction in waste and improved overall efficiency.

Additionally, PointNetGAN can be applied to automated assembly processes by generating 3D point cloud data that can be used to improve the accuracy of robot or machine movements. By training the generator network on point cloud data of the desired final product, it can generate point clouds that represent the correct orientation and position of each

component in the assembly process. This can help reduce errors and increase the speed of assembly line production.

Overall, PointNetGAN's ability to generate realistic 3D point cloud data can improve the efficiency and accuracy of various manufacturing processes, particularly in the context of assembly line production.

Deep Convolutional GAN

DCGAN can be applied to manufacturing or assembly lines in various ways, particularly in situations where generating synthetic data is necessary for testing or training purposes. For instance, DCGAN may be used to produce synthetic pictures of items or components for quality assurance testing or machine learning training. This may decrease the cost and effort required with gathering and categorizing vast quantities of actual data.

In addition, DCGAN may be used to produce synthetic sensor data that can be utilized to train predictive maintenance algorithms. By generating synthetic data that closely resembles real-world sensor data, it is possible to train algorithms to detect anomalies and predict equipment failures before they occur. This can help reduce downtime and maintenance costs in manufacturing and assembly line environments.

However, as mentioned earlier, DCGAN can suffer from mode collapse, which can limit the variety of synthetic data generated. This can be addressed by using techniques such as regularizing the training process or modifying the architecture of the generator and discriminator networks. Nevertheless, DCGAN has significant potential for improving the efficiency and accuracy of various manufacturing processes, particularly in the context of data generation and predictive maintenance.

VAE-GAN

The combination of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), known as VAE-GAN, is a powerful hybrid deep learning model that can be leveraged in assembly lines or discrete manufacturing to generate synthetic data. This synthetic data can be used for training machine learning models or testing different scenarios, such as quality control testing or predictive maintenance. For instance, VAE-GAN can generate synthetic images of products or components, effectively reducing the need for large amounts of real-world data.

Moreover, VAE-GAN can also augment existing datasets to enhance the accuracy of machine learning models by producing synthetic data that closely mimics real-world data. This process increases the diversity of the training data and minimizes the risk of overfitting, leading to more efficient manufacturing processes and improved predictions.

VAE-GAN is computationally expensive and needs a large quantity of data to provide excellent results. It may also demand extra preprocessing processes to guarantee that the produced data is consistent with actual data.

Conditional GAN

Conditional Generative Adversarial Networks (CGANs) can be a useful tool in discrete manufacturing, allowing for the generation of synthetic data that is conditioned on specific information. CGANs consist of two neural networks - a generator and a discriminator - with the generator taking in both a noise vector and a conditional vector[70]. This conditional vector can represent any input information, such as an image label or a sequence of text, allowing the generator to produce synthetic data that is specific to that information.

The discriminator in a CGAN is also modified to take in the same conditional vector as the generator. This enables the discriminator to become better at distinguishing between real and synthetic data in a way that is specific to the conditioning information.

CGANs have several benefits in discrete manufacturing, such as generating high-quality synthetic data that can be used for training machine learning models or testing different scenarios. It is feasible to decrease the quantity of real-world data required for testing and training by producing synthetic data conditioned on specified facts. In addition, CGANs may be used to enrich current datasets, resulting in more accurate machine learning models and more efficient industrial processes.

Recurrent Neural Network

In assembly lines or discrete manufacturing, RNNs can be utilized to generate synthetic data by modeling the underlying probability distribution of sequential data. By training an RNN on a large dataset of real-world data, the model can learn to generate new synthetic data samples that resemble the original data[71]. This approach can be used in various applications, such as predicting machine failures or optimizing production schedules.

Generative RNNs are a particular sort of RNN that may also be used to produce synthetic data (GRNN). "Teacher forcing" is used by GRNNs to educate the network to anticipate the next value in a sequence based on the previous values. After the GRNN has been trained, it may be used to produce new data sequences by predicting the next value and utilizing it as input for the subsequent prediction.

Furthermore, RNNs can be combined with other approaches such as variational autoencoders (VAEs) to generate synthetic data that displays complex temporal relationships and patterns. This involves encoding the input data into a latent space using an RNN and then decoding this space using a decoder network to generate new synthetic data. This method can be used to generate synthetic data for training machine learning models or testing different scenarios, leading to more accurate predictions and more efficient manufacturing processes.

StyleGAN

StyleGAN can be applied to assembly lines to generate synthetic images of products or components, which can be used for quality control testing or machine learning training. By generating synthetic data, it is possible to reduce the amount of real-world data needed for testing and training purposes. StyleGAN's ability to control the appearance of generated images by separating the image synthesis process into two stages and introducing progressive growing, can enable the generation of high-quality, diverse, and lifelike images of products or components, which can aid in visual inspection and analysis.

In addition, StyleGAN can be used to augment existing datasets by generating synthetic data that closely resembles real-world data. This can increase the diversity of the training data and improve the accuracy of machine learning models for quality control and predictive maintenance on assembly lines. Overall, StyleGAN's capabilities in synthesizing high-quality, diverse, and lifelike images can be a valuable tool in assembly lines for quality control and machine learning training.

Deep Belief Network

DBNs, or Deep Belief Networks, can also be applied to discrete manufacturing by using them to generate synthetic data for various applications such as predicting machine failures and optimizing production lines. By learning the underlying distribution of sensor data collected from machines in a production line, DBNs can create synthetic data that can be used to test different scenarios and optimize production processes. This can save time and reduce costs by allowing manufacturers to identify potential issues before they occur and improve the efficiency of their operations. Additionally, DBNs can be used to generate synthetic data for training machine learning models used in predictive maintenance, anomaly detection, and quality control in discrete manufacturing.

Transormer-Based Time-Series GANs

TTSGAN, or Transformer-Based Time-series Generative Adversarial Network, is a deep learning model used for generating synthetic time-series data. It is a combination of Transformer and GAN models, which are both powerful in generating time-series data and capturing temporal dependencies[72].

The Transformer model is used as the generator, which takes in a random noise vector and outputs a synthetic time-series. The GAN model is used as the discriminator, which distinguishes between the synthetic and real time-series data.

TTSGAN can be applied to manufacturing or assembly lines to generate synthetic time-series data, which can be useful for predicting future trends and identifying potential issues. One potential use of TTSGAN in manufacturing is to train it on historical data from an assembly line so that it can create synthetic time-series data that forecasts the assembly line's future performance. By doing so, manufacturers can anticipate potential bottlenecks or problems in the assembly line and take preventative measures to mitigate them, thereby avoiding expensive delays or defects in the manufactured products. Additionally, TTSGAN can be used to generate synthetic data for testing and validating control systems or algorithms used in the assembly line, helping to ensure their effectiveness and reliability.

Convolutional Auto encoders

The Convolutional Autoencoder (CAE) is an unsupervised neural network that can encode data effectively and is well-suitable for picture data. Its design is comprised of an encoder and a decoder, each containing convolutional and deconvolutional layers to extract spatial characteristics from pictures and reconstruct them[73].

In manufacturing or assembly lines, CAE can be used to generate synthetic images of products or components for quality control and inspection purposes. By training the neural network on a set of images of high-quality products or components, the CAE can learn the

spatial features that distinguish them from defective ones. Then, it can generate new images with similar spatial characteristics, which can be used to augment the original dataset and improve the performance of machine learning models used for product inspection. This can help manufacturers reduce the number of defective products and improve overall quality control processes.

PixelCNN

PixelCNN is a generative model that may be used to generate artificial data. This neural network design is a per-pixel-operating convolutional neural network (CNN)[74]. PixelCNN creates a new picture of the same size as the output by taking an image as input and conditioning each pixel on previously created pixels. PixelCNN has potential applications in manufacturing and assembly lines, where it can be used to generate synthetic images of products, components, or equipment. This can be useful for testing and training machine vision systems or for simulating different scenarios in the production process.

For example, a manufacturer can train a PixelCNN model on images of their products and use it to generate synthetic images of the same products with varying defects or anomalies. These synthetic images can be used to train machine vision systems to identify and classify defects in real-time during production.

Furthermore, PixelCNN can also be used for predictive maintenance by generating synthetic images of machinery or equipment under different conditions. This can help manufacturers anticipate potential equipment failures and perform maintenance before any major issues occur, reducing downtime and improving overall efficiency.

Generative Flow Networks

Generative Flow Networks (GFNs) are a type of deep learning architecture used for synthetic data generation. GFNs are based on the idea of normalizing flow models, which

aim to transform a simple probability distribution into a more complex one[75]. The goal of a GFN is to learn the true data distribution by transforming a known, simple distribution into the target distribution. This is done by composing a series of invertible transformations, where each transformation maps one probability distribution to another.

GFNs can be applied to assembly lines for generating synthetic data that can be used for various purposes such as simulating the behavior of machines, optimizing production processes, and predicting equipment failures. For instance, GFNs may be trained using historical data from an assembly line in order to discover the underlying distribution of the data and produce synthetic data that mimics the original data. This synthetic data may then be utilized to simulate alternative scenarios and assess the influence of production process modifications on the performance of the assembly line. In addition, GFNs may be used to anticipate equipment failures by creating synthetic data that depicts multiple equipment states, such as normal functioning, partial failure, and total failure. This synthetic data can then be used to train machine learning models for predicting equipment failures and taking preventive maintenance actions. Overall, GFNs provide a powerful framework for generating synthetic data that can be used to improve the efficiency and reliability of assembly lines.

Hidden Markov Model

HMMs, or Hidden Markov Models, are a probabilistic model commonly used to model time series data in discrete manufacturing. They can infer the unobservable state of a system from observed data, making them useful for identifying patterns and predicting future behavior of the system.

One key application of HMMs in manufacturing is the generation of synthetic data with similar statistical properties to the original data. This involves training the HMM on the original data and then using the model to produce new data sequences with similar statistical features.

HMMs are particularly useful for generating sequential data in discrete manufacturing, as they are capable of capturing complex connections between the various parts of a data sequence. However, training the model can be computationally demanding and the output sequences may not necessarily be reflective of the original data. Despite the emergence of more advanced deep learning-based algorithms, HMMs remain a valuable tool for applications that require the production of sequential data.

Autoregressive Integrated Moving Average

ARIMA, which stands for Autoregressive Integrated Moving Average, is a widely used statistical model for time series analysis and prediction in assembly lines[76]. ARIMA models are composed of three components, including autoregression, integration, and moving average, and are used to describe stationary time series data in the manufacturing domain. Autoregression is the relationship between an observation and past observations, while moving average is the relationship between an observation and a residual error from a moving average model. Integration refers to the level of differencing used to transform the time series into a stationary form.

ARIMA is a powerful tool for generating synthetic data that follows the same statistical properties as the original data in assembly lines. In order to produce artificial data using ARIMA, the initial step involves fitting the model to the original time series data to obtain the model parameters. These parameters can then be employed to generate fresh data points that match the original data, such as the number of products manufactured per hour or the time required for a particular process. The newly generated data can be appended to the initial dataset or used to develop extra datasets for the purpose of testing and validation.

ARIMA has applications in many domains, including predicting machine breakdowns, optimizing assembly line processes, and forecasting manufacturing output. Importantly, ARIMA models are only applicable to stationary time series data, meaning that the statistical

features of the data do not vary over time. Before fitting an ARIMA model to non-stationary data, it may be required to perform extra preprocessing procedures.

Long Short-Term Memory

Long Short-Term Memory (LSTM) is an RNN architecture that is well-suited for creating synthetic data of manufacturing lines. LSTM networks excel in modeling time series data, making them a popular option for simulating the behavior of real-world production systems with synthetic data[77].

LSTM networks may create synthetic data that captures the temporal connections between different manufacturing process components in the context of assembly lines. For example, an LSTM network may be trained on historical data from an assembly line in order to comprehend the links between the many equipment, tools, and components involved in production. The trained network may then create synthetic data that simulates the behavior of the assembly line, including the timing of the various production processes.

The capacity of LSTM networks to capture long-term dependencies in the data is a key benefit for synthetic data creation[77]. This allows them to simulate complicated assembly line behavior patterns. In addition, LSTM networks can be trained on big datasets, enabling them to discover a wide variety of patterns and behaviors from previous data.

In general, LSTM networks are a useful tool for synthesizing data that may be used to test and assess assembly line performance and enhance production processes.

Convolutional Neural Networks

Synthetic data generation of assembly lines can be achieved using Convolutional Neural Networks (CNNs)[78]. These deep neural networks are primarily used for image recognition and processing tasks. In this context, CNNs learn the patterns and features of existing images of assembly lines to generate new ones.

The first step to using CNNs for synthetic data generation of assembly lines is to collect a dataset of labeled images. These images should be labeled based on their characteristics, such as the presence of components or defects.

Supervised learning techniques are then used to train the CNN model on the dataset. The objective is to teach the model to accurately classify the images based on their labels. During the training process, the CNN model identifies important patterns and features for classifying the images.

After the model is trained, it can generate new synthetic images of assembly lines. Random noise or other images are inputted into the model, and the learned patterns and features are used to generate new images.

Using CNNs for synthetic data generation offers the advantage of producing highly realistic images that closely resemble those in the dataset. However, controlling the specific characteristics of the generated images can be challenging, as they are determined by the learned patterns and features in the dataset.

Overall, CNNs are a powerful tool for synthetic data generation of assembly lines, but this approach requires a large dataset of labeled images and significant computational resources for training and generation.

Recurrent Variational Autoencoder

The Recurrent Variational Autoencoder (RVAE) is a neural network that can generate synthetic data, including assembly line data. The RVAE is an extension of the standard Variational Autoencoder (VAE), which adds recurrent layers to the encoder and decoder networks[79].

The RVAE is particularly useful for generating sequential data where each observation is dependent on previous time steps. For instance, in the case of assembly line data, observations at a particular time step depend on observations from previous time steps. The

RVAE can model this dependency and generate new sequences of data that are similar to the original data.

To generate new sequences, the RVAE's encoder network takes the input sequence and produces a latent representation for each time step. The decoder network then takes the latent representation and generates a new sequence of data that is similar to the original sequence. The latent representation serves as a compressed version of the input sequence and can generate new sequences.

Training an RVAE involves maximizing the lower bound on the log-likelihood of the data. This includes minimizing the reconstruction loss, which measures the difference between the original and generated sequences, and the KL-divergence loss, which measures the difference between the distribution of the latent representation and a prior distribution.

The RVAE is a potent technique for generating synthetic assembly line data that captures the sequential dependencies between time steps. Nonetheless, generating synthetic data using RVAE demands a considerable volume of training data and can incur high computational costs.

Conditional Variational Autoencoder

The Conditional Variational Autoencoder (CVAE) is a generative neural network that facilitates the synthesis of assembly line data. By learning the distribution of input data, CVAE can generate new samples that adhere to the same patterns and features as the original data[80]. To accomplish this, CVAE utilizes an encoder network that transforms input data into a compressed latent representation, and a decoder network that decodes the latent representation back into the original data space. Additionally, CVAE uses an extra input variable to condition the latent representation, enabling the generation of output data that is specific to the given condition. This approach allows CVAE to model complex dependencies and correlations between different features of the data, resulting in high-quality

synthetic data. However, effective utilization of CVAE requires a significant amount of training data and careful tuning of hyperparameters such as the number of layers and latent space dimensionality.

The generation of synthetic data is an essential aspect of machine learning, and the appropriate method for generating synthetic data depends on the type of data and its unique features and properties.

For example, Random Sampling is suitable for tabular data, while SMOTE and ADASYN are better for imbalanced tabular data. GANs, VAEs, and Autoencoders are better suited for image and text data. RNNs and LSTMs are better suited for time-series data, while HMMs and ARIMA are suitable for sequential data. StyleGAN and GMMs are more appropriate for generating realistic images, whereas GANs and CGANs are good for generating images with specific attributes.

Similarly, PointNetGAN and PointFlowGAN are specifically designed to generate synthetic point clouds. Furthermore, RVAE and CVAE are better suited for generating assembly line data.

Therefore, the choice of a specific method depends on the nature of the data and the specific problem to be solved.

4.1.4 Evaluation

After the generation of synthetic data for assembly lines, it is imperative to evaluate its quality and usefulness before utilizing it for analysis or model development. The assessment of synthetic data is critical to ensure that it can effectively replace the original data. The evaluation process primarily involves examining two aspects: fidelity and utility.

1. Fidelity: Fidelity is the degree to which synthetic data reflects the actual data properly. Synthetic data is often used to enhance training data or to produce extra data that is

comparable to the original data in machine learning. A model built on synthetic data that does not precisely replicate the actual data may perform poorly on the original data.

To assess the quality of synthetic data, various techniques can be utilized. By employing data visualization methods like histograms and scatter plots, it is possible to compare the synthetic data with the genuine data visually. If the synthetic data's visual traits closely resemble those of the original data, there is a high likelihood that the synthetic data's distribution corresponds to that of the genuine data[70].

Another method is to compare the statistical properties of the synthetic data with those of the actual data using statistical tests. The Kolmogorov-Smirnov and Anderson-Darling tests are examples of such tests used to compare the distributions of synthetic and actual data[81]. The correspondence between the distribution of synthetic data and genuine data reflects the synthetic data's dependability.

A third technique involves training a model using the synthetic data before evaluating it on the original data. This approach is frequently employed to assess the efficacy of synthetic data for specific applications such as speech recognition or image classification. Domain-specific metrics are used to determine the synthetic data's fidelity in such cases.

In addition to these methodologies, the F1 score and ROC-AUC score may also be used to assess the accuracy of synthetic data. The F1 score is a statistic for evaluating the effectiveness of a binary classification model that combines accuracy and recall. It ranges between 0 and 1, with higher numbers signifying superior performance. The ROC-AUC score ranges from 0.5 to 1, with larger values indicating greater performance. It quantifies the capacity of a binary classification model to discriminate between positive and negative classes. If the F1 score and ROC-AUC score of a model trained on synthetic data are equivalent to those of a model trained on the original data, then indicates that the synthetic data is trustworthy.

Ensuring the reliability of synthetic data is a critical stage in ensuring the effectiveness of a model trained on it. Multiple techniques, including visual examination, statistical tests, and domain-specific metrics, can be employed to assess the authenticity of synthetic data and verify its accurate replication of the actual data.

2. Utility: The usefulness of synthetic data in a specific application is known as utility. Assessing the utility of synthetic data is essential to guarantee its suitability for the intended application. For example, if synthetic data is used to train a machine learning model, its utility must be evaluated based on the model's performance on real-world data[82]. Several methods can be employed to assess the utility of synthetic data, such as:

Similarity metrics: Various similarity metrics can be employed to compare the synthetic and original data distributions. Mean squared error or correlation coefficient can be used to compare these distributions.

Classification accuracy: A classifier can be trained on the original data and tested on the synthetic data. If the classification accuracy is similar to that of the original data, it indicates good utility of the synthetic data.

Regression error: Similarly, a regression model can be trained on the original data and tested on the synthetic data. If the regression error is similar to that of the original data, it indicates good utility of the synthetic data.

Clustering: One can use clustering techniques to ascertain whether the synthetic data can be grouped in similar ways to the original data.

Visual inspection: Finally, visual inspection of the synthetic data can be conducted to determine whether it retains the crucial features and characteristics of the original data.

4.1.5 Access and address potential biases

If any aspect of evaluation, such as fidelity, or utility fails then it is important to assess and address potential biases. Biases can arise from various sources, such as the synthetic data generation process or the evaluation metrics used. Here are some ways to avoid biases.

1. Adjust the synthetic data generation algorithm and regenerate data: This approach requires adjusting the algorithm responsible for generating synthetic data to improve alignment with the intended outcome and then generating the data again. For instance, if the synthetic data's faithfulness is inadequate, the generation algorithm can be fine-tuned to create data that more closely mirrors the original data. Similarly, if the synthetic data's utility is subpar, the generation algorithm can be enhanced to incorporate more crucial features or variables that are essential for the model's precision.

2. Modify training data and regenerate data: In this approach, one modifies the training data utilized to develop the algorithm for generating synthetic data and generates the data again. This strategy can assist in mitigating any biases that were present in the original training data and were consequently inherited by the synthetic data. For instance, if the synthetic data exhibits bias towards a specific group or population, one can adjust the training data by incorporating a more comprehensive range of examples to mitigate the bias.

3. Apply post-processing to synthetic data: This approach necessitates adjusting the synthetic data after its generation to correct any detected shortcomings. Post-processing methods can be utilized to enhance the quality, usefulness, and confidentiality of the synthetic data. For instance, if the synthetic data's fidelity is inadequate, data smoothing or imputation techniques can be applied via post-processing to improve its quality. Correspondingly, if safeguarding the synthetic data's privacy is a priority, post-processing techniques such as data masking or anonymization can be employed to safeguard sensitive information.

4.1.6 Machine Learning

After accessing and addressing biases in synthetic data generation, the regenerated data needs to be reevaluated to ensure that the biases have been successfully eliminated. This process may involve repeating the synthetic data generation and evaluation steps until the data passes all necessary tests.

Once the synthetic data has successfully passed evaluation, it can be combined with real data to train machine learning models. The integration of synthetic and authentic data can furnish a more extensive and varied training dataset, resulting in more precise and resilient models. Employing synthetic data can also diminish the dependency on costly and time-consuming manual data collection methods.

Overall, the process of generating and evaluating synthetic data for machine learning applications requires careful consideration and attention to detail. By following a robust framework and addressing potential biases, synthetic data can be a valuable resource for training machine learning models and gaining useful insights from data.

4.2 Synthetic Data Generation for Different Types of Data on Assembly Lines and Opportunities

The framework mentioned above for synthetic data generation can be used for any kind of data on the assembly line. This section will delve into various data types that can be gathered on assembly stations, along with the corresponding techniques for generating synthetic data for each type as seen in the Table 4.1. Furthermore, we will assess the potential advantages that can be attained by analyzing and modeling each data type.

	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	Determine the uptimes and downtimes of the assembly line, Quantify process reliability
	Point cloud	Human positions, Robotic arm equipped with machine vision camera	PointNetGAN, DCGANs, VAE-GAN, VAEs, Autoencoders	Object detection, human-robot collaboration- obstacle detection, To classify the phase of assembly
Continuous	Biomedical	EEG data, Wristband data(EMPATICA-E4 data)	CGAN, DCGAN, VAE, GMM, RNNs	Assessment of correlation with other data types
	Image	Inventory images, Operator's image, Final product image	GANs, VAEs, StyleGAN, DBNs, CAE, PixelCNN, Generative Flow Networks	Identify the components within the inventory, Verify the identity of operator, Quality Control
	Time-series	Assembly time, Production count	TTSGANs, Hidden Markov Model, ARIMA, LSTM, RNN, VAE	Acquisition of the cycle time, Track production efficiency
	3D Image	Final product images	GANs, VAEs, CNNs	Quantify the number of assembled parts by the operator, Evaluate the quality deficiencies
	Video	Operators interactions	GANs, RVAE, CVAE	Verify the operator's identity, Determine the operator's count in the assembly process, Emotional state analysis of the operator, Determine the current assembled part

Table 4.1: Synthetic data generation techniques, examples, and opportunities for various data types on assembly line

4.2.1 Binary Data

One of the frequently occurring data types on assembly lines is binary data. It is characterized by being discrete in nature, with only two possible values, usually symbolized as 0 or 1. Examples of binary data collected on assembly lines include whether a certain machine is operating or not, and detecting the presence of the operator.

To generate synthetic binary data, various techniques can be used such as SMOTE, Random Sampling, ADASYN, GMM, ROSE, SLSMOTE, BLSMOTE. These techniques can help to generate synthetic binary data that accurately reflects the patterns and distributions of the real binary data.

Selecting an appropriate data generation method for binary data relies on the distinct attributes of the data and the analysis objectives. Here are some general recommendations:

1. SMOTE, ADASYN, and random sampling are suitable for imbalanced datasets where the minority class is underrepresented. These methods can create synthetic examples of the minority class to balance the dataset and prevent the model from being biased towards the majority class.

2. GMM (Gaussian Mixture Model) can be used to generate synthetic data that follows a specific distribution. This can be useful if the binary data is generated by a complex process and the distribution is not known prior.

3. ROSE (Random Over Sampling Examples) and SLSMOTE (Synthetic Least-Squares-based SMOTE) can be used to generate synthetic data that is similar to the existing data. These methods can be useful if the goal is to augment the dataset and increase the diversity of the examples without changing the overall characteristics of the data.

4. BLSMOTE (Borderline-SMOTE) can be used if the data has a clear separation between the two classes. This method generates synthetic examples along the boundary between the two classes, which can improve the model's performance.

Typically, it is advisable to experiment with various data generation methods and assess their effectiveness on a validation set before settling on the optimal approach for a particular problem.

The analysis of binary data can offer valuable insights into machine performance on assembly lines. By identifying patterns in machine behavior, system faults can be detected promptly, and maintenance schedules can be optimized to ensure the efficient operation of the assembly line. Additionally, analyzing binary data can also help to identify areas for improvement in the assembly line process, such as reducing machine downtime or improving the efficiency of quality control checks.

4.2.2 Point-Cloud Data

Point-cloud data is a type of data that can be collected on assembly lines, representing the position of objects or humans using a series of distinct points. Although point-cloud data is often considered as discrete data, each point may have associated continuous attributes, such as position coordinates or color values. Techniques such as PointNetGAN, DCGANs, VAE-GAN, PointFlowGAN, VAEs, and autoencoders can be used to generate synthetic point-cloud data.

The analysis of point-cloud data permits us to identify the real-time positions of humans and robots, recognize obstacles in the environment, and promote secure and efficient collaboration between them. Furthermore, point-cloud data can facilitate the classification of assembly phases and the identification of probable process bottlenecks.

For instance, in a manufacturing environment, point-cloud data can be utilized to track the movements of workers and machinery. The data can be analyzed to recognize the most commonly used pathways and streamline the utilization of resources and materials. Additionally, point-cloud data can be utilized to identify safety hazards and prevent accidents from occurring.

In terms of synthetic data generation techniques, PointNetGAN can be used to generate point-cloud data for object detection tasks. DCGANs and VAE-GAN can be used to generate realistic point-cloud data for training machine learning models. PointFlowGAN can be used to generate dynamic point-cloud data for simulating human-robot collaboration scenarios. Overall, point-cloud data provides a valuable source of information for optimizing assembly line processes and enabling safe and efficient collaboration between humans and robots.

4.2.3 Biomedical Data

Biomedical data, such as EEG data and wristband data (e.g., EMPATICA-E4 data), can be collected on assembly lines to monitor the health and well-being of workers. This is considered continuous data. To generate synthetic biomedical data, various techniques can be used, including CGAN, DCGAN, VAE, GMM, and RNNs.

CGAN (conditional generative adversarial network) can be used to generate realistic and diverse biomedical data by conditioning the generator on certain features or labels, such as age or gender[83]. DCGAN (deep convolutional generative adversarial network) is a variant of GANs that uses convolutional layers to generate more complex biomedical data, such as EEG signals.

VAE (variational autoencoder) is a popular technique for generating synthetic biomedical data that captures the underlying structure of the data. GMM (Gaussian mixture model) can be used to model complex distributions in biomedical data, such as the distribution of heart rate variability in wristband data.

RNNs (recurrent neural networks) can be used to generate time-series biomedical data, such as EEG signals, by learning the temporal dependencies between the data points[83].

The opportunities of synthetic biomedical data include the assessment of correlation with other data types, such as environmental sensors and assembly line productivity data,

to identify potential relationships between the health of workers and the efficiency of the assembly line.

4.2.4 Image Data

Image data is a valuable type of data that can be collected on assembly lines, including images of inventory, operators, and final products. This data is continuous, with each image consisting of pixels with continuous values representing color and intensity. Synthetic image data can be generated using various techniques such as GANs, VAEs, StyleGAN, DBNs, CAE, PixelCNN, and Generative Flow Networks.

Analyzing image data provides numerous opportunities to improve the manufacturing process. For example, image recognition techniques can be used to identify defects and anomalies in the final product images, allowing for prompt correction and optimization of the process. Additionally, image data can be used to track inventory and operators' movement, improving logistics and efficiency. Machine learning models can be trained using GANs, VAEs, and StyleGAN to generate realistic images of inventory and products.

DBNs and CAE can be used to learn feature representations of images that can be used for tasks such as object detection and classification. Moreover, PixelCNN and Generative Flow Networks are suitable for generating high-resolution images with fine details.

It is essential to consider the specific task requirements and the available data's characteristics when choosing a data generation technique. The appropriate technique can improve the accuracy and reliability of machine learning models and ultimately enhance the manufacturing process's quality and efficiency.

4.2.5 Time-series Data

In assembly lines, time-series data, such as production count and assembly time, can also be captured. This data type comprises of consecutive data points captured at regular intervals, such as seconds, minutes, or hours. Several algorithms can be used to produce synthetic time-series data, including TTGANs, Hidden Markov Model, ARIMA, LSTM, GRU, and RNN.

We can calculate the length of the assembly cycle time, monitor production efficiency, and improve the manufacturing process by analyzing time-series data. Through time-series analysis, we may discover problem areas in production, and by resolving these issues, we can decrease downtime and boost productivity and efficiency. Time-series data can also assist in estimating what future production patterns may look like, resulting in more educated decisions.

TTGANs and VAEs can be used to generate synthetic time-series data for training machine learning models. Hidden Markov Model and ARIMA are statistical models that can be used for time-series forecasting. LSTM, GRU, and RNN are deep learning models that can be used for time-series prediction and classification.

It is vital to keep in mind that the choice of data production strategy may rely on the unique needs of the current activity and the features of the available data. For instance, if the time-series data shows extended dependencies, LSTM and GRU models may be more appropriate than ARIMA.

4.2.6 3D Image Data

3D image data is a type of data that can be used to evaluate the final product images on the assembly line. This data can be used to quantify the number of assembled parts by the

operator and evaluate any quality deficiencies. Synthetic 3D image data can be generated using techniques such as GANs, VAEs, and CNNs.

By analyzing 3D image data, we can discover and quantify the number of assembled pieces, identify errors or anomalies in the final product, and improve quality control operations. Using 3D image recognition algorithms, we can properly count and identify the number of components in a final product in order to assure appropriate assembly.

GANs, VAEs, and CNNs can generate synthetic 3D picture data for training machine learning models. With these approaches, we may also enhance the quality and resolution of 3D photographs, resulting in a more precise and comprehensive examinations of the final product.

It is worth noting that 3D image data and point-cloud data are distinct. 3D image data represents the complete surface geometry of an object, while point-cloud data is a group of distinct points in space. 3D image data can provide more comprehensive and detailed information about the final product, but it may require more processing power and storage capacity than point-cloud data.

4.2.7 Video Data

Video data is a valuable resource for assembly lines as it provides a sequence of images captured over time, allowing us to evaluate the efficiency and quality of the manufacturing process. Synthetic video data can be generated using techniques such as GANs, RVAE, and CVAE. Analyzing video data can help us verify the operator's identity, count their participation in the assembly process, and assess their performance.

For example, facial recognition technology can verify the correct operator performing a task, while emotion recognition techniques can detect fatigue or distraction affecting their performance. Video data can also help identify any defects or quality issues in the product by tracking the components' movement throughout the assembly line.

It is important to consider the specific requirements of the task and the available data when selecting a data generation technique. For instance, synthetic video data generated through GANs, RVAE, and CVAE can train machine learning models for object detection and action recognition tasks.

Chapter 5

Case study

Quality control is a crucial aspect of the parts manufacturing industry, but imbalanced datasets can pose challenges for collecting and analyzing data. Synthetic data generation has emerged as a powerful tool for balancing imbalanced datasets and training machine learning models for quality control. In this case study, we demonstrate a synthetic data generation framework for balancing an imbalanced dataset in the parts manufacturing industry.

5.1 Data Collection

We obtained the Parts Manufacturing Industry Dataset from Kaggle for this case study. The dataset contains information on 500 parts produced by each of the 20 operators in one period of time. The features include length, width, height, and operator ID as shown in Figure 5.1. We used this dataset to demonstrate the effectiveness of our synthetic data generation framework for balancing imbalanced datasets

	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 5.1: Data and its Features

5.2 Pre-processing

To prepare the data for synthetic data generation, we first removed the Item_No feature as it was not relevant to the classification task. We then classified the parts as perfect or defective based on the presence of outliers in their dimensions as shown in Figure 5.2. In this study, we employed outlier detection to pinpoint faulty parts. To be more specific, we computed the standard deviation of the length, width, and height of each operator's parts and labeled any part with dimensions that deviated more than two standard deviations from the mean as defective. This approach allowed us to capture the presence of potential defects or anomalies in the part's dimensions.

We labeled the majority class as perfect and the minority class as defective. That means data is imbalanced as shown in Figure 5.3. So, We used synthetic data to balance the imbalanced dataset. This approach ensured that the synthetic data generation framework

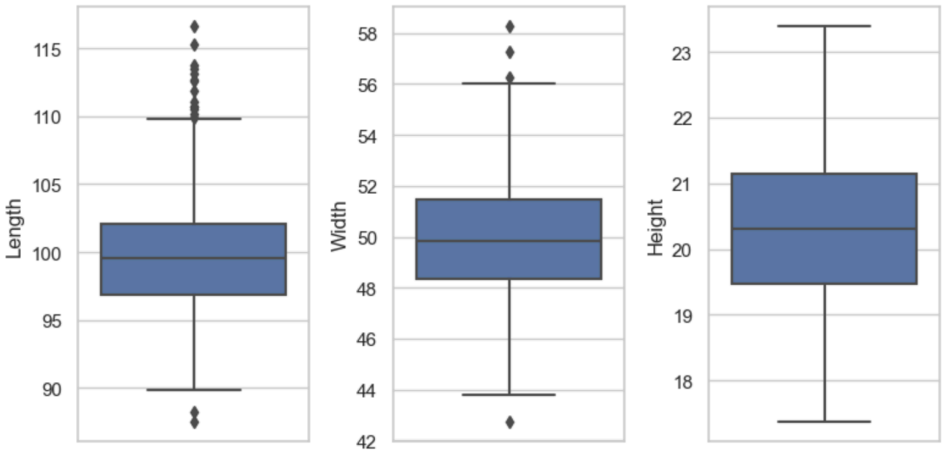

Figure 5.2: Outlier detection

focused on generating new samples of the minority class, i.e., defective parts, while preserving the distribution of the original data.

```
data['Status'].value_counts()
```

```
Perfect      480
Defective     20
Name: Status, dtype: int64
```

Figure 5.3: Data distribution

5.3 Synthetic data generation

As the Parts Manufacturing Industry Dataset is a binary dataset, we focused on generating synthetic data for the minority class (i.e., defective parts) to balance the dataset. We applied SMOTE, ADASYN, ROS, BLSMOTE, SLSMOTE, and GMM to generate synthetic data and balance the class distribution.

SMOTE and ADASYN both oversampled the minority class by creating synthetic samples through interpolation, with ADASYN generating more synthetic samples for harder-to-learn examples. ROSE applied various resampling techniques to the minority class to generate synthetic samples.

BLSMOTE and SLSMOTE focused on generating synthetic samples for the borderline instances between the minority and majority classes. BLSMOTE only considered borderline instances closer to the majority class, while SLSMOTE generated synthetic samples for borderline instances closer to the minority class.

Finally, GMM generated synthetic data by fitting a mixture of Gaussian distributions to the original data and sampling from these distributions.

By applying these techniques, we were able to generate synthetic data that balanced the dataset.

5.4 Evaluation

5.4.1 Fidelity

	Real Data			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 5.1: Fidelity comparison

To evaluate the fidelity of the various synthetic data generation techniques, we used accuracy, F-1 score, and ROC-AUC score as performance metrics as shown in the Figure 5.1. To be precise, we trained a logistic regression model on real data and evaluated it using real

data. Additionally, we trained the model using hybrid data and evaluated it using hybrid data, which is a mix of actual and synthetic data.

We found that GMM passed the fidelity test and outperformed the other synthetic data generation techniques in terms of accuracy, F-1 score, and ROC-AUC score. This indicates that the synthetic data generated by GMM was of high quality and closely resembled the real data.

5.4.2 Utility

Once the fidelity test was passed, we proceeded to assess the usefulness of the GMM-generated synthetic data in enhancing the precision of our models. We achieved this by training a random forest model with stratified 10-fold cross-validation on the real dataset and testing it on real data. We also trained the model using hybrid data and tested it on hybrid data as well as real data as shown in the Figure 5.2.

	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

Table 5.2: Utility Comparison

Our results showed that the random forest model trained and tested on the real dataset achieved high accuracy and precision but recall is still lower than the other models, including that the model is missing some of the actual positive cases.

The second model trained on hybrid data and tested on real data shows high accuracy, precision, recall, F1-score, and ROC AUC score. This indicates that the model was able to generalize well to real data and correctly identified instances of the minority class.

The third model trained on hybrid data and tested on hybrid data also shows a good performance with high accuracy, precision, recall, F1-score, and ROC AUC score. This indicates that the model was able to generalize well to the balanced dataset and correctly identify instances of the minority class.

Overall, the second and third models seem to perform well on real and hybrid data, respectively. This suggests that the synthetic data generated using GMM was effective in improving the accuracy of our models for detecting defective parts in the Parts Manufacturing Industry Dataset. The high accuracy achieved by the model trained on synthetic data and tested on hybrid data further demonstrates the usefulness of synthetic data generation techniques in addressing imbalanced datasets in the manufacturing industry.

Chapter 6

Future Scope and Conclusion

6.1 Future Scope

The synthetic data generation framework proposed in this research has the potential to bring some exciting changes to the field of machine learning in manufacturing. Moving forward, it would be beneficial to evaluate the framework's effectiveness in more complex manufacturing scenarios and explore the development of new synthetic data techniques for other types of manufacturing data beyond discrete manufacturing. It is important to validate the framework by collecting real data from manufacturing assembly lines to ensure its accuracy and provide a benchmark for comparison with synthetic data generated by the framework. Sharing synthetic datasets generated using this framework publicly can help researchers, manufacturers, and academic professionals develop and test machine learning models more efficiently. Lastly, we should investigate the potential of combining synthetic and real data to improve the machine learning model's performance in assembly line applications. Overall, these efforts could result in improved productivity and product quality in the manufacturing industry.

6.2 Conclusion

In our study, we investigated how synthetic data generation techniques can address challenges related to limited data availability, expensive data collection, and proprietary data in AI applications for discrete manufacturing. To achieve this, we delved into existing research and created a practical approach for generating synthetic data. Our framework involves an evaluation process to guarantee that the synthetic data generated resembles the real data and also its usefulness for machine learning models. We used a case study to demonstrate this framework, and our results showed that the model trained on the hybrid data outperformed the model trained on real data itself.

In addition, our study adds to the current understanding of synthetic data generation in the manufacturing industry by offering a practical methodology for generating synthetic data and identifying diverse data types that can be produced on assembly lines. We have also outlined various techniques that can be utilized to generate synthetic data for each data type and discussed the potential insights that can be gleaned from analyzing the data.

Overall, our research highlights the potential of synthetic data generation techniques to improve data availability and generate accurate and reliable results in AI applications for discrete manufacturing. The proposed framework can be used to generate high-quality synthetic data for various data types on assembly lines, enabling researchers to develop more precise machine learning models without incurring significant costs.

Bibliography

[1] J. P. Womack, D. T. Jones, and D. Roos, *The machine that changed the world: The story of lean production–Toyota's secret weapon in the global car wars that is now revolutionizing world industry*. Simon and Schuster, 2007.

[2] W. Zhang, L. Hou, and R. J. Jiao, "Dynamic takt time decisions for paced assembly lines balancing and sequencing considering highly mixed-model production: An improved artificial bee colony optimization approach," *Computers & Industrial Engineering*, vol. 161, p. 107616, 2021.

[3] F. Tao, Q. Qi, A. Liu, and A. Kusiak, "Data-driven smart manufacturing," *Journal of Manufacturing Systems*, vol. 48, pp. 157–169, 2018.

[4] R. X. Gao, L. Wang, M. Helu, and R. Teti, "Big data analytics for smart factories of the future," *CIRP annals*, vol. 69, no. 2, pp. 668–692, 2020.

[5] A.-A. Mubarak, Y. Mohamed, and A. Bouferguene, "Application of industrial pipelines data generator in the experimental analysis: Pipe spooling optimization problem definition, formulation, and testing," *Advanced Engineering Informatics*, vol. 43, p. 101007, 2020.

[6] K. Marazopoulou, R. Ghosh, P. Lade, and D. Jensen, "Causal discovery for manufacturing domains," *arXiv preprint arXiv:1605.04056*, 2016.

[7] S. Mihai, W. Davis, D. V. Hung, R. Trestian, M. Karamanoglu, B. Barn, R. V. Prasad, H. Venkataraman, and H. X. Nguyen, "A digital twin framework for predictive maintenance in industry 4.0," 2021.

[8] P. M. Blau, C. M. Falbe, W. McKinley, and P. K. Tracy, "Technology and organization in manufacturing," *Administrative science quarterly*, pp. 20–40, 1976.

[9] S. Han, H.-J. Choi, S.-K. Choi, and J.-S. Oh, "Fault diagnosis of planetary gear carrier packs: A class imbalance and multiclass classification problem," *International Journal of Precision Engineering and Manufacturing*, vol. 20, pp. 167–179, 2019.

[10] S. W. Kim, Y. G. Lee, B. A. Tama, and S. Lee, "Reliability-enhanced camera lens module classification using semi-supervised regression method," *Applied Sciences*, vol. 10, no. 11, p. 3832, 2020.

[11] T. Ademujimi and V. Prabhu, "Digital twin for training bayesian networks for fault diagnostics of manufacturing systems," *Sensors*, vol. 22, no. 4, p. 1430, 2022.

[12] D. Fecker, V. Märgner, and T. Fingscheidt, "Density-induced oversampling for highly imbalanced datasets," in *Image Processing: Machine Vision Applications VI*, vol. 8661, pp. 211–221, SPIE, 2013.

[13] M. Syafrudin, N. L. Fitriyani, G. Alfian, and J. Rhee, "An affordable fast early warning system for edge computing in assembly line," *Applied Sciences*, vol. 9, no. 1, p. 84, 2018.

[14] H. G. Nguyen, R. Habiboglu, and J. Franke, "Enabling deep learning using synthetic data: A case study for the automotive wiring harness manufacturing," *Procedia CIRP*, vol. 107, pp. 1263–1268, 2022.

[15] F. Sibona and M. Indri, "Data-driven framework to improve collaborative human-robot flexible manufacturing applications," in *IECON 2021–47th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–6, IEEE, 2021.

[16] Z. Biczó, I. Felde, and S. Szénási, "Distorsion prediction of additive manufacturing process using machine learning methods," in *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 000249–000252, IEEE, 2021.

[17] Z. Zheng, Z. Zhang, and W. Pan, "Virtual prototyping-and transfer learning-enabled module detection for modular integrated construction," *Automation in Construction*, vol. 120, p. 103387, 2020.

[18] Z.-H. Lai, W. Tao, M. C. Leu, and Z. Yin, "Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing," *Journal of Manufacturing Systems*, vol. 55, pp. 69–81, 2020.

[19] S. Kohtala and M. Steinert, "Leveraging synthetic data from cad models for training object detection models–a vr industry application case," *Procedia CIRP*, vol. 100, pp. 714–719, 2021.

[20] C. Grappiolo, R. Pruim, M. Faeth, and P. de Heer, "Vitrovo: in vitro assembly search for in vivo adaptive operator guidance: An artificial intelligence framework for highly customised manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 117, no. 11-12, pp. 3873–3893, 2021.

[21] F. L. de la Rosa, J. L. Gómez-Sirvent, R. Sánchez-Reolid, R. Morales, and A. Fernández-Caballero, "Geometric transformation-based data augmentation on defect classification of segmented images of semiconductor materials using a resnet50 convolutional neural network," *Expert Systems with Applications*, vol. 206, p. 117731, 2022.

[22] J. Sikora, R. Wagnerová, L. Landryová, J. Šíma, and S. Wrona, "Influence of environmental noise on quality control of hvac devices based on convolutional neural network," *Applied Sciences*, vol. 11, no. 16, p. 7484, 2021.

[23] S. K. Singh, S. K. Chakrabarti, and D. B. Jayagopi, "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, Proceedings 14*, pp. 181–192, Springer, 2020.

[24] C. Qian, W. Yu, C. Lu, D. Griffith, and N. Golmie, "Toward generative adversarial networks for the industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19147–19159, 2022.

[25] E. Ameperosa and P. A. Bhounsule, "Domain randomization using deep neural networks for estimating positions of bolts," *Journal of Computing and Information Science in Engineering*, vol. 20, no. 5, p. 051006, 2020.

[26] R. Maliks and R. Kadikis, "Multispectral data classification with deep cnn for plastic bottle sorting," in *2021 6th International Conference on Mechanical Engineering and Robotics Research (ICMERR)*, pp. 58–65, IEEE, 2021.

[27] S. Laxman, P. Sastry, and K. Unnikrishnan, "Discovering frequent generalized episodes when events persist for different durations," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 9, pp. 1188–1201, 2007.

[28] B. Andres, E. Guzman, and R. Poler, "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*, vol. 2021, pp. 1–17, 2021.

[29] M. Malekzadeh, R. G. Clegg, and H. Haddadi, "Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis," *arXiv preprint arXiv:1710.06564*, 2017.

[30] A. Bikes, G. Williams, and R. O'connor, "Assembly systems sensitivity to component delivery: A logistics study using simulation," in *Fourth International Conference on Factory 2000-Advanced Factory Automation*, pp. 638–644, IET, 1994.

[31] H. U. Guner, R. B. Chinnam, and A. Murat, "Simulation platform for anticipative plant-level maintenance decision support system," *International Journal of Production Research*, vol. 54, no. 6, pp. 1785–1803, 2016.

[32] F. G. Sisca, M. Fiasché, and M. Taisch, "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, Proceedings, Part II 22*, pp. 571–582, Springer, 2015.

[33] M. Fiasché, G. Ripamonti, F. G. Sisca, M. Taisch, and G. Tavola, "A novel hybrid fuzzy multi-objective linear programming method of aggregate production planning," in *Advances in Neural Networks: Computational Intelligence for ICT*, pp. 489–501, Springer, 2016.

[34] N. Martin, B. Depaire, A. Caris, and D. Schepers, "Retrieving the resource availability calendars of a process from an event log," *Information Systems*, vol. 88, p. 101463, 2020.

[35] S. Jain, A. Narayanan, and Y.-T. T. Lee, "Comparison of data analytics approaches using simulation," in *2018 Winter Simulation Conference (WSC)*, pp. 1084–1095, IEEE, 2018.

[36] A. Apornak, S. Raissi, and M. R. Pourhassan, "Solving flexible flow-shop problem using a hybrid multi criteria taguchi based computer simulation model and dea approach," *Journal of Industrial and Systems Engineering*, vol. 13, no. 2, pp. 264–276, 2021.

[37] W. Cai, P. A. Bernstein, W. Wu, and B. Chandramouli, "Optimization of threshold functions over streams," *Proceedings of the VLDB Endowment*, vol. 14, no. 6, pp. 878–889, 2021.

[38] L. A. da Silva, E. M. dos Santos, L. Araújo, N. S. Freire, M. Vasconcelos, R. Giusti, D. Ferreira, A. S. Jesus, A. Pimentel, C. F. Cruz, *et al.*, "Spatio-temporal deep learning-based methods for defect detection: An industrial application study case," *Applied Sciences*, vol. 11, no. 22, p. 10861, 2021.

[39] I. Rio-Torto, A. T. Campaniço, A. Pereira, L. F. Teixeira, and V. Filipe, "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)*, pp. 342–347, IEEE, 2021.

[40] J. L. Outón, I. Merino, I. Villaverde, A. Ibarguren, H. Herrero, P. Daelman, and B. Sierra, "A real application of an autonomous industrial mobile manipulator within industrial context," *Electronics*, vol. 10, no. 11, p. 1276, 2021.

[41] A. Luckow, K. Kennedy, M. Ziolkowski, E. Djerekarov, M. Cook, E. Duffy, M. Schleiss, B. Vorster, E. Weill, A. Kulshrestha, *et al.*, "Artificial intelligence and deep learning applications for automotive manufacturing," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 3144–3152, IEEE, 2018.

[42] D. Shetve, R. VaraPrasad, R. Trestian, H. X. Nguyen, and H. Venkataraman, "Cats: Cluster-aided two-step approach for anomaly detection in smart manufacturing," in *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2*, pp. 103–115, Springer, 2021.

[43] K. Georgiadis, A. Nizamis, T. Vafeiadis, D. Ioannidis, and D. Tzovaras, "Production scheduling optimization enabled by digital cognitive platform," *Procedia Computer Science*, vol. 204, pp. 424–431, 2022.

[44] N. Sun, A. Kopper, R. Karkare, R. C. Paffenroth, and D. Apelian, "Machine learning pathway for harnessing knowledge and data in material processing," *International Journal of Metalcasting*, vol. 15, pp. 398–410, 2021.

[45] Z. Zhang, L. Pan, L. Du, Q. Li, and N. Lu, "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, Proceedings, Part I 16*, pp. 350–365, Springer, 2021.

[46] A. Bécue, E. Maia, L. Feeken, P. Borchers, and I. Praça, "A new concept of digital twin supporting optimization and resilience of factories of the future," *Applied Sciences*, vol. 10, no. 13, p. 4482, 2020.

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

[48] C. Cimino, G. Ferretti, and A. Leva, "Harmonising and integrating the digital twins multiverse: A paradigm and a toolset proposal," *Computers in Industry*, vol. 132, p. 103501, 2021.

[49] D. Ramanujan and W. Z. Bernstein, "Vesper: Visual exploration of similarity and performance metrics for computer-aided design repositories," in *International Manufacturing Science and Engineering Conference*, vol. 51371, p. V003T02A034, American Society of Mechanical Engineers, 2018.

[50] S. Mihai, M. Yaqoob, D. V. Hung, W. Davis, P. Towakel, M. Raza, M. Karamanoglu, B. Barn, D. Shetve, R. V. Prasad, *et al.*, "Digital twins: a survey on enabling technologies, challenges, trends and future prospects," *IEEE Communications Surveys & Tutorials*, 2022.

[51] R. L. Rardin and R. Uzsoy, "Experimental evaluation of heuristic optimization algorithms: A tutorial," *Journal of Heuristics*, vol. 7, pp. 261–304, 2001.

[52] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine learning for industrial applications: A comprehensive literature review," *Expert Systems with Applications*, vol. 175, p. 114820, 2021.

[53] A. Achar, S. Laxman, R. Viswanathan, and P. Sastry, "Discovering injective episodes with general partial orders," *Data Mining and Knowledge Discovery*, vol. 25, pp. 67–108, 2012.

[54] A. Thelen, X. Zhang, O. Fink, Y. Lu, S. Ghosh, B. D. Youn, M. D. Todd, S. Mahadevan, C. Hu, and Z. Hu, "A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies," *Structural and Multidisciplinary Optimization*, vol. 65, no. 12, p. 354, 2022.

[55] J. Xu, M. Kovatsch, D. Mattern, F. Mazza, M. Harasic, A. Paschke, and S. Lucia, "A review on ai for smart manufacturing: Deep learning challenges and solutions," *Applied Sciences*, vol. 12, no. 16, p. 8239, 2022.

[56] S. Suhail, R. Hussain, R. Jurdak, A. Oracevic, K. Salah, C. S. Hong, and R. Matulevičius, "Blockchain-based digital twins: research trends, issues, and future challenges," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–34, 2022.

[57] U. J. Botero, R. Wilson, H. Lu, M. T. Rahman, M. A. Mallaiyan, F. Ganji, N. Asadizanjani, M. M. Tehranipoor, D. L. Woodard, and D. Forte, "Hardware trust and assurance

through reverse engineering: A tutorial and outlook from image analysis and machine learning perspectives," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 17, no. 4, pp. 1–53, 2021.

[58] M. Mahmoodian, F. Shahrivar, S. Setunge, and S. Mazaheri, "Development of digital twin for intelligent maintenance of civil infrastructure," *Sustainability*, vol. 14, no. 14, p. 8664, 2022.

[59] M. Flores, R. Fernández-Casal, S. Naya, and J. Tarrío-Saavedra, "Statistical quality control with the qcr package," *R Journal*, vol. 13, no. 1, pp. 194–217, 2021.

[60] M. van Doorn, S. Duivestein, D. Mamtani, and T. Pepping, "Infinite machine creativity." `https://labs.sogeti.com/research-topics/infinite-machine-creativity/`, 2020.

[61] J. Asturias and J. Rossbach, "Grouped variation in factor shares: An application to misallocation," *International Economic Review*, vol. 64, no. 1, pp. 325–360, 2023.

[62] S. Fahle, C. Prinz, and B. Kuhlenkötter, "Systematic review on machine learning (ml) methods for manufacturing processes–identifying artificial intelligence (ai) methods for field application," *Procedia CIRP*, vol. 93, pp. 413–418, 2020.

[63] A. C. Bavelos, N. Kousi, C. Gkournelos, K. Lotsaris, S. Aivaliotis, G. Michalos, and S. Makris, "Enabling flexibility in manufacturing by integrating shopfloor and process perception for mobile robot workers," *Applied Sciences*, vol. 11, no. 9, p. 3985, 2021.

[64] A. Fatima, N. Nazir, and M. G. Khan, "Data cleaning in data warehouse: A survey of data pre-processing techniques and tools," *Int. J. Inf. Technol. Comput. Sci*, vol. 9, no. 3, pp. 50–61, 2017.

[65] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, 2022.

[66] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and knowledge discovery handbook*, pp. 875–886, 2010.

[67] C. Chokwitthaya, Y. Zhu, S. Mukhopadhyay, and A. Jafari, "Applying the gaussian mixture model to generate large synthetic data from a small data set," in *Construction Research Congress 2020: Computer Applications*, pp. 1251–1260, American Society of Civil Engineers Reston, VA, 2020.

[68] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pp. 510–526, Springer, 2019.

[69] A. Desai, C. Freeman, Z. Wang, and I. Beaver, "Timevae: A variational auto-encoder for multivariate time series generation," *arXiv preprint arXiv:2111.08095*, 2021.

[70] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, vol. 10, no. 15, p. 2733, 2022.

[71] Y.-C. Tam, Y. Shi, H. Chen, and M.-Y. Hwang, "Rnn-based labeled data generation for spoken language understanding," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[72] X. Li, V. Metsis, H. Wang, and A. H. H. Ngu, "Tts-gan: A transformer-based time-series generative adversarial network," in *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*, pp. 133–143, Springer, 2022.

[73] K. Bascol, R. Emonet, E. Fromont, and J.-M. Odobez, "Unsupervised interpretable pattern discovery in time series using autoencoders," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2016, Mérida, Mexico, November 29-December 2, 2016, Proceedings*, pp. 427–438, Springer, 2016.

[74] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Information Sciences*, vol. 433, pp. 125–141, 2018.

[75] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio, "Flow network based generative models for non-iterative diverse candidate generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27381–27394, 2021.

[76] H. Singh and M. R. Ray, "Synthetic stream flow generation of river gomti using arima model," in *Advances in Civil Engineering and Infrastructural Development: Select Proceedings of ICRACEID 2019*, pp. 255–263, Springer, 2021.

[77] C. A. Libbi, J. Trienes, D. Trieschnigg, and C. Seifert, "Generating synthetic training data for supervised de-identification of electronic health records," *Future Internet*, vol. 13, no. 5, p. 136, 2021.

[78] C. Manettas, N. Nikolakis, and K. Alexopoulos, "Synthetic datasets for deep learning in computer-vision assisted tasks in manufacturing," *Procedia CIRP*, vol. 103, pp. 237–242, 2021.

[79] M. Ziatdinov, M. Y. Yaman, Y. Liu, D. Ginger, and S. V. Kalinin, "Semi-supervised learning of images with strong rotational disorder: assembling nanoparticle libraries," *arXiv preprint arXiv:2105.11475*, 2021.

[80] S. Dixit and N. K. Verma, "Intelligent condition-based monitoring of rotary machines with few samples," *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14337–14346, 2020.

[81] M. C. Santos, A. I. Borges, D. R. Carneiro, and F. J. Ferreira, "Synthetic dataset to study breaks in the consumer's water consumption patterns," in *Proceedings of the 2021 4th International Conference on Mathematics and Statistics*, pp. 59–65, 2021.

[82] M. S. N. Khan, N. Reje, and S. Buchegger, "Utility assessment of synthetic data generation methods," *arXiv preprint arXiv:2211.14428*, 2022.

[83] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, 2022.