Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 134 (2025) 831–836

www.elsevier.com/locate/procedia

Proceedings of the 58th CIRP Conference on Manufacturing Systems 2025

Synthetic Data Generation Using Causal Models for Injection Molding Processes

Christoph Hennebold[a], Jonas Krauß[a], Marco F. Huber[a,b,]

[a]*Artificial Intelligence and Machine Vision, Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Nobelstr. 12, 70569 Stuttgart, Germany*
[b]*Institute for Industrial Manufacturing and Management IFF, University of Stuttgart, Allmandring 35, 70569 Stuttgart, Germany*

* Corresponding author. *E-mail address:* christoph.hennebold@ipa.fraunhofer.de

Abstract

The generation of synthetic data is particularly interesting for areas such as industrial production, where the generation of new data is associated with high effort, costs and possible production downtimes. By generating additional data, potential problems associated with the use of small sample sizes when applying machine learning algorithms can be avoided. This allows machine learning-based approaches to be better utilized and processes to be trained more efficiently. In order to address the problem of small sample sizes and the time-consuming generation of additional data, a process model based on causal mechanisms is generated in this work, which can be used to generate new synthetic data that is based on the available observational data. To make this possible, a minimal model is generated, which can be trained using existing real data and used for further data generation. The generated synthetic data are tested on a real-life application use-case of an injection molding machine and evaluated using various utility metrics.

Keywords: Causal Inference; Graphical Causal Model; Synthetic Data

1. Introduction

In manufacturing, the amount of data collected on the shop floor is constantly growing due to increasing digitalization. Collected data often requires preprocessing to ensure quality and usability. A sufficient amount of data is essential, because the recognition of patterns in the data is proportional to the size of the dataset [10]. In manufacturing this often poses a problem due to ever shorter product life cycles in which production systems have to be adapted and optimized [15], and because the generation of additional data is associated with additional costs and production downtime [10]. To mitigate this problem, extensive work has been done to generate synthetic data. Among other things, these artificially generated data can be used to enhance production processes by using the data for monitoring, scheduling, and process optimization [4]. Existing approaches can be divided into the two categories of *physical simulation* and *numerical modeling*. *Physical simulation* examples include discrete-event simulation models for the simulation and optimization of flow-shop scheduling problems [1] or the generation of audio data that describe different types of errors based on frequencies [19]. In addition, in the automotive sector, 3D CAD models [14] and 3D-simulation software [17] proofed to be very valuable for generating synthetic data. In contrast, *numerical modeling* based approaches include data-driven methods that attempt to simulate the existing conditions of the real world in the best possible way [4]. In the field of Generative Adversarial Networks (GANs), there is already a large body of literature demonstrating the applicability of GANs to generate synthetic data [6]. Bayesian networks (BNs) [7] and purely statistical approaches [9] to preserve existing correlations in the data have also proven useful. There are also promising approaches for process optimization through improved resource allocation as well [13]. However, the use of mathematical simulation models to utilize physical effects and data-driven approaches requires in-depth process knowledge and depends on the application area and the respective problem complexity. Data-driven models often neglect causal relationships, causing deviations from real data. Additionally there exist possible disadvantages due to longer model training times. The mentioned approaches have in common that the actual cause-and-effect relationships are not always taken into account and relations between variables may not be correct, because the generated data does not

10.1016/j.procir.2025.02.202

correspond to the reality. Therefore, in this work an approach based on the Graphical Causal Model (GCM) implementation of DoWhy-GCM [2] is presented, to extract the causal mechanisms of the underlying data generation process and build a causal process model in order to generate new synthetic data. The applicability of this approach is demonstrated by evaluating a use-case of a real-life injection molding machine. A minimal model of the process is generated using relevant parameters, which are then used for data generation and the subsequent evaluation by various utility measures. This is particularly interesting for future process optimization and generalizing over comparable machines by generating new data based on causal relations instead of only correlations.

This article is structured as follows: Sec. 2 introduces the theoretical background such as utilized metrics and the basics of GCM on which this work is based. Sec. 3 presents the use-case applied in the evaluation, along with the data utilized. The use-case is then evaluated in Sec. 4 with the previously introduced metrics, and finally in Sec. 5, the work is summarized and an outlook is provided.

2. Background

In this section the basics of causal mechanisms (Sec. 2.1), the kernel density estimation (Sec. 2.2), and utility measures (Sec. 2.3) are introduced, which are used for the subsequent evaluation of both machine learning models and dataset similarities.

2.1. Graphical Causal Model (GCM)

DoWhy-GCM [2] is an extension of the Python DoWhy library [18] that allows the reconstruction of the data generating process (so-called causal mechanism) of each variable in the model using directed acyclic graphs (DAGs) and available already observed data of the specific process in question. The reconstructed model allows to ask and answer causal questions, such as finding the root cause of errors, asking what-if questions or the effect of interventions on target variables. The basis of causal mechanisms is the causal Markov condition, which, given a causal graph G with variables $\{X_1, X_2, ..., X_N\} \in X$ allows a factorization of the joint distribution through

$$P_{X_1,...,X_n} = \prod_{j=1}^{n} P_{X_j|PA_j}. \tag{1}$$

A causal mechanism is defined by $P_{X_j|PA_j}$ with the variable X_j and its direct parents $PA_j \subset X$. According to Budhathoki et al. [3], each mechanism is invariant under interventions in other variables, allowing to analyze the influence of individual mechanisms on the overall system. Due to the known model structure and the connections between the mechanisms, new samples can be generated that are based on the existing data and can thus be used to generate new synthetic data. This approach uses the available dataset and model structure to either compute empirical distributions for root-nodes or additive noise models (ANM) for dependant nodes. In this case an ANM is defined as

$$Y = f(X_1, X_2, ..., X_n) + N, \tag{2}$$

with a function f, cause variables X_i, the effect variable Y, and a noise term N. The function f describes a causal mechanism and is calculated in this regard using a set of regression functions such as Random Forest Regressor, KNN Regressor or Support Vector Regressor based on the available observation data in order to calculate the unknown true function f. The model with the smallest calculated mean squared error (MSE) value that best approximates the function f is then selected to represent the causal mechanism.

2.2. Kernel Density Estimation

The Kernel Density Estimation (KDE) can be used to determine the probability density distribution of a random variable. This allows the distribution of the data points to be visualized and used for further analysis. The KDE $\hat{f}(x)$ is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

where n is the number of data points, h a smoothing parameter, K the kernel function used (e.g., Gaussian or rectangular function), x the points at which the density is estimated and x_i the individual points from the sample that are used to estimate the density.

2.3. Utility Measures

Utility measures are metrics that are used to compare and evaluate synthetic and real data, respectively. More precisely, they can be divided into similarity metrics (Sec. 2.3.1) and machine learning specific metrics (Sec. 2.3.2).

2.3.1. Similarity Metrics

The following section introduces the similarity metrics used for the evaluation.

Kullback-Leibler Divergence. The Kullback-Leibler (KL) divergence allows the calculation of the deviation of a probability distribution Q from another probability distribution P. In detail, the loss of information is calculated that occurs when, for example, distribution P is used to approximate distribution Q. In this paper, the KL divergence on discrete probability distributions is used, which is defined as

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \tag{4}$$

Table 1. Overview of utilized machine learning metrics

Metric	Description
Accuracy	$= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$
Precision	$= \frac{\text{True Positives}}{\text{True Positives + False Positives}}$
Recall	$= \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
F1-Score	$= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$
Mean Average Error (MAE)	$= \frac{1}{n} \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
Mean Squard Error (MSE)	$= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

where X is the set of possible events (values), $P(x)$ is the probability of the event x according to the distribution P and $Q(x)$ is the probability of the event x according to the distribution Q.

Hellinger Distance. The Hellinger distance is a measure for determining the similarity of two probability distributions. It allows the measurement of the difference between two distributions, whereby the metric assumes a value range between 0 and 1, where 0 means that the distributions are identical and 1 means that they are completely different from each other. The Hellinger distance is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} \left(\sqrt{P(x_i)} - \sqrt{Q(x_i)} \right)^2}, \tag{5}$$

with the discrete probability distributions P and Q and a finite set $x_1, x_2, ..., x_n$ of elements in the distributions.

2.3.2. *Machine Learning Metrics*

The metrics listed in Table 1 are used to evaluate the different machine learning tasks in Sec. 4. The used metrics both reflect classification specific metrics such as accuracy, precision, recall, and f1-score, as well as regression specific metrics such as the Mean Average Error (MAE) or the Mean Squared Error (MSE).

3. Use-Case

The use-case is from an injection molding process. The dataset is introduced in Sec. 3.1, and the associated causal model is further presented in Sec. 3.2.

3.1. *Dataset*

The used dataset holds the recorded process data of a real injection molding machine, and was generated during the research project “ProBayes” [12]. Based on expert knowledge, from over 300 features, a number of key features were selected that have a significant influence on the quality of the produced parts. The variables are divided into adjustable machine parameters, process parameters, and the target variable. Machine parameters can be changed by the user in order to influence the process, process parameters are influenced by the machine parameters and can only be observed. Table 2 provides an overview and detailed descriptions of all variables. The dataset is comprised of a total of 204 samples, each representing a produced part that was classified with a binary quality label.

3.2. *Causal Process Model*

The graphical causal model, which is the basis for the further evaluation, was created in two steps. In the first step, various causal discovery methods such as Fast Causal Inference (FCI) [20], Greedy Equivalence Search (GES) [5], and Greedy Relaxations of the Sparsest Permutation (GRaSP) [11] were used for data-driven modeling to relieve the process expert from unnecessary work, whereby the model created by the GRaSP algorithm, as depicted in Fig. 1, is aligned closest to the actual process. It is important to note, that there already exists a large body of work for causal discovery [8].

After the initial model was generated by the algorithm, in the second step, final adjustments were made by the process expert based on available process knowledge. The resulting model is shown in Fig. 2 and includes the basic components of the injection molding process under investigation that affect the quality of the produced parts.

In detail, the figure shows the adjustable machine parameters on the left, the process parameters on the right and the quality label `LBL_NOK` at the bottom right, which represents a binary distinction between good and faulty products based on the two most commonly occurring quality errors.

By utilizing the adjusted DAG from the second stage and combining it with the available process data, a GCM is generated by utilizing the DoWhy library [2]. The resulting model

Table 2. Overview of relevant machine and process parameter variables in the injection molding process

Name	Param. Type	Description
`E77_CylinderTemperature11`	Machine	Temperature of a specific zone of the cylinder
`E77_CoolingTime`	Machine	Required time for cooling the molded part
`E77_InjectionSpeed1`	Machine	Speed at which the melt is injected into the cavity
`E77_HoldingPressureMax`	Machine	Maximum pressure during holding phase
`E77_CycleTime`	Process	Duration of the entire cycle including all phases
`E77_InjectionTime`	Process	Required time to inject the melt into the cavity
`E77_InjectionPressureMax`	Process	Peak pressure during the injection phase
`E77_CavityPressure1Max`	Process	Highest pressure during the injection process inside the cavity
`LBL_NOK`	Target	Target feature which indicates good or bad parts

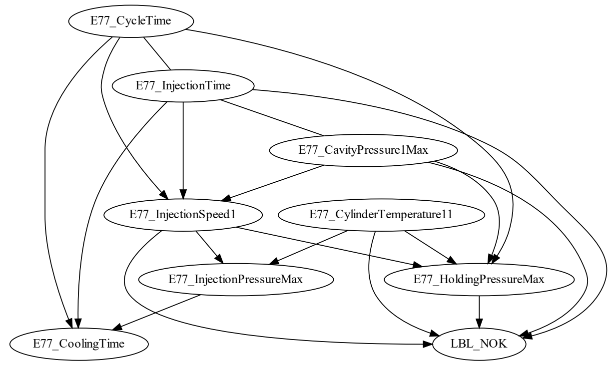

Fig. 1. The automatically generated process model by the GRaSP algorithm

is comprised of the different causal mechanisms that are fitted with the available data, as described in Sec. 2.1.

4. Evaluation

For the evaluation, a synthetic dataset with 204 samples, identical in size to the real dataset, was generated using the causal model presented in Sec. 3.2. The data generation follows the causal structure of the model by recursively sampling values from the marginal distributions of root nodes and propagating them through the defined causal mechanisms (see Sec. 2.1). This ensures that the generated data adheres to the causal dependencies of the model and is suitable for further analysis. A 10-fold cross-validation approach was applied to separate training and evaluation. The causal model was trained on training folds only, while the synthetic data was evaluated on unseen validation sets. This prevents data leakage and ensures unbiased performance assessment. The comparison of real data (R) and synthetic data (S) is presented in Sec. 4.1 using utility measures. Additionally, Sec. 4.2 evaluates machine learning tasks to assess the applicability of synthetic data for various use cases.

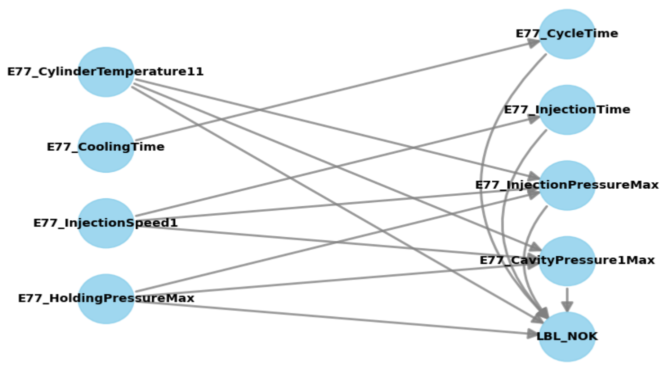

Fig. 2. Overview of the causal process model that results from the dataset used and the most important features of the injection molding process

Table 3. KL-divergence and Hellinger distance results

Feature	$D_{KL}(P\|\|Q)$	$H(P,Q)$
`E77_CylinderTemperature11`	$1.13 \cdot 10^{-2}$	$7.38 \cdot 10^{-2}$
`E77_CoolingTime`	$3.53 \cdot 10^{-2}$	$2.16 \cdot 10^{-2}$
`E77_InjectionSpeed1`	$9.49 \cdot 10^{-1}$	$5.20 \cdot 10^{-2}$
`E77_HoldingPressureMax`	$3.40 \cdot 10^{-2}$	$1.82 \cdot 10^{-1}$
`E77_CycleTime`	$8.00 \cdot 10^{-6}$	$2.20 \cdot 10^{-3}$
`E77_InjectionTime`	$6.75 \cdot 10^{-1}$	$1.89 \cdot 10^{-2}$
`E77_InjectionPressureMax`	$5.62 \cdot 10^{-2}$	$5.88 \cdot 10^{-1}$
`E77_CavityPressure1Max`	$7.16 \cdot 10^{-2}$	$5.66 \cdot 10^{-1}$

4.1. *Similarity Evaluation*

In this section, the results of the similarity evaluation are listed. This includes the comparison of the kernel desity estimation 4.1.1, Kullback-Leibler divergence 4.1.2, and the Hellinger distance 4.1.3.

4.1.1. *Kernel Density Estimation*

The comparison between the Kernel Density Estimation (KDE) of the real data and the synthetic data was carried out using a sample size of 204 in each case. The comparison shown in Fig. 3 between the machine and process parameters used in each case shows predominantly similar distributions. The results indicate that the modeled causal mechanisms are capable of generating data that exhibit similar distributions to the original real data.

It is noteworthy that the process parameters show greater deviations from the original data. One possible explanation is the structure of the causal model and the fact that only a minimal model was created, which is subject to the influence of hidden confounding, leading to deviations from the original data [21, 22]. Potential hidden variables that could influence these differences include material properties (e.g., viscosity, thermal characteristics), environmental factors (e.g., ambient temperature, humidity), or machine-specific variations (e.g., wear, calibration). While the minimal causal model used in this study captures key process-relevant variables, incorporating additional variables representing these factors could reduce discrepancies and enhance the robustness of the causal model.

4.1.2. *Kullback-Leibler Divergence*

The results of the calculation of the Kullback-Leibler divergence are listed in Table 3. It is evident that, apart from `E77_InjectionTime` and `E77_InjectionSpeed1`, there are no major deviations present. A possible explanation for this deviation is the existence of hidden confounders. However, it is important to emphasize that the deviation between real and synthetically generated data is only marginal and remains within an acceptable range.

4.1.3. *Hellinger Distance*

The calculated results of the Hellinger distance are listed in Table 3. Most values remain close to zero, which means that there is only a small deviation between the examined

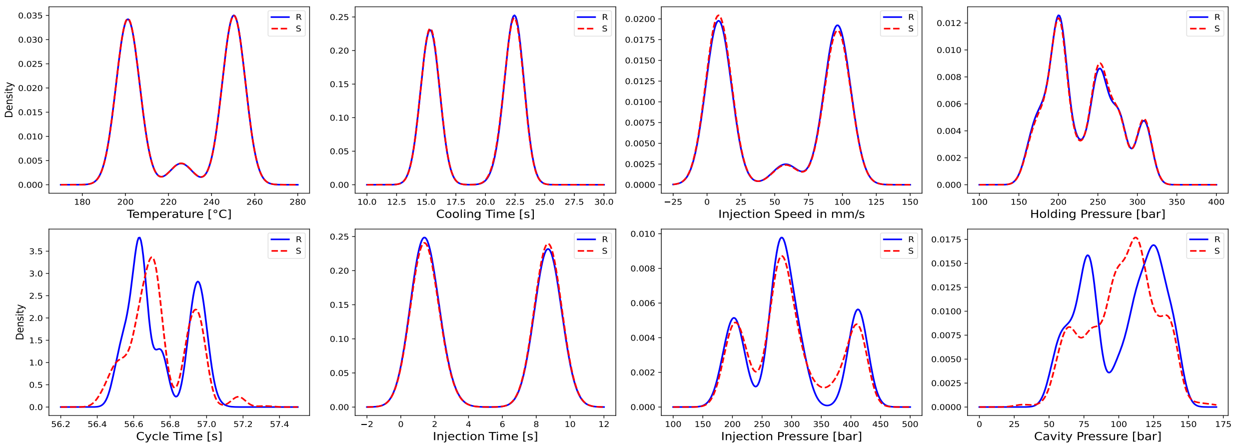

Fig. 3. KDE comparison for the different model parameters between the real dataset (R), and the dataset generated using the causal model (S)

distributions. This in turn is a sign of high similarity between the distributions. However, `E77_CavityPressure1Max` and `E77_InjectionPressureMax` show noticeable deviations, likely due to the minimal causal model used in this study, which does not incorporate all relevant process variables, highlighting potential effects of hidden confounders.

4.2. Machine Learning Task Evaluation

In this section, the results of both the machine learning classification task (Sec. 4.2.1) and regression task (Sec. 4.2.2) are evaluated in detail.

4.2.1. Classification Task

This experiment examines the prediction accuracy of the end product quality using different classification methods. The focus is on how used machine learning methods differ in terms of accurately classifying when using real and synthetic data and whether the use of synthetic data results in comparable results. Different well-known classification algorithms such as K-Nearest Neighbor (KNN), Gaussian Process Classification (GPC), Gaussian Naive Bayes (GNB), Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC), and Support Vector Machine (SVM) implemented in scikit-learn [16] are used for further evaluation.

Looking at the classification results listed in Table 4 the calculated metrics for every algorithm on the real data R and synthetic data S are close to each other and there are no major deviations, which suggest that the used data is suitable for this type of use-case.

4.2.2. Regression Task

For this task, the cylinder temperature, injection speed, and maximum holding pressure machine parameters are used to calculate the cavity pressure process parameter. A prior causal effect analysis determined that the cavity pressure process parameter influences the final result by more than 20%. While this analysis is not part of the current study, it provided important insights into the significant causal relationships between these variables. Based on this finding, a regression analysis is used to compute the cavity pressure depending on these parameters. The original data and synthetically generated data are compared to determine their respective deviations from the target value. Regression methods such as Linear Regressor (LR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR) are used for this calculation. The regression algorithms implemented in scikit-learn [16] are used in the evaluation.

The results listed in Table 5 show good comparable results between the original and synthetically generated data, whereby the linear regression shows the greatest deviation from the optimum, as the existing complexity in the data cannot be modeled in detail. In principle, however, differences are to be expected, as the synthetically generated data are relying on a minimal causal model that does not take into account all process relevant

Table 4. Product quality prediction results

Name	Data	Acc.	Precision	Recall	F1
KNN	R	0.98	0.98	0.97	0.97
	S	0.87	0.91	0.80	0.85
GPC	R	0.98	0.99	0.97	0.98
	S	0.93	0.94	0.89	0.91
GNB	R	0.84	0.87	0.80	0.82
	S	0.92	0.91	0.93	0.92
RFC	R	0.98	0.98	0.98	0.98
	S	0.97	0.98	0.95	0.96
GBC	R	0.98	0.99	0.97	0.98
	S	0.97	0.97	0.96	0.96
SVM	R	0.79	0.84	0.68	0.74
	S	0.87	0.87	0.84	0.85

Table 5. Cavity pressure regression analysis results

Name	Data	MAE [bar]	MSE
LR	R	18.1	433.3
	S	17.9	525.6
RFR	R	3.9	95.3
	S	6.7	112.9
GBR	R	4.1	97.1
	S	6.8	100.9

parameters and thus possible effects due to hidden confounding remain.

5. Outlook and Conclusion

"This paper presented a novel approach for generating synthetic data in manufacturing using a minimal causal model based on DoWhy-GCM. Despite possible effects due to hidden confounding factors, the results show that even a subset of process-relevant variables is sufficient to generate synthetic data for further analysis in a manufacturing environment. Once a suitable causal model is built, data is generated and evaluated by comparing the original and synthetic datasets using utility measures and machine learning tasks.

The evaluation confirmed the similarity between real and synthetic data and demonstrated the applicability of synthetic data for predictive modeling. These findings underscore the potential of synthetic data to supplement real data in scenarios with limited availability, while maintaining high accuracy and generalizability. Additionally, synthetic data provides opportunities for generating interventional data samples not present in the original dataset, enabling the exploration of new parameter spaces and process optimization without extensive real-world experimentation.

However, the presented approach is not without limitations. The causal model relies on balanced and diverse training data to approximate distributions effectively. Larger datasets with high-dimensional features can increase computational cost and time, while smaller datasets may challenge the model's robustness. Although the model performed well with 204 samples, future work will explore scalability for larger datasets, robustness for smaller datasets, and the establishment of formal thresholds for effective training. Adapting this approach to other manufacturing processes requires identifying process-specific causal relationships and collecting relevant data. Future work will address these challenges and aim to generalize the model. Despite these limitations, expanding the model to incorporate additional variables will further reduce discrepancies, align synthetic data more closely with real distributions, and provide a scalable solution for data generation and process optimization in manufacturing.

Acknowledgements.

The authors would like to thank the Baden-Wuerttemberg Ministry of Economic Affairs, Labor and Tourism for funding this work concerning the project KI-Fortschrittszentrum "Lernende Systeme und Kognitive Robotik" under Grant No. 036-140100.

References

[1] Apornak, A., Raissi, S., Pourhassan, M.R., 2021. Solving flexible flowshop problem using a hybrid multi criteria taguchi based computer simulation model and dea approach. Journal of industrial and systems engineering 13, 264–276.

[2] Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A.A., Janzing, D., 2024. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. Journal of Machine Learning Research 25, 1–7.

[3] Budhathoki, K., Janzing, D., Bloebaum, P., Ng, H., 2021. Why did the distribution change?, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 1666–1674.

[4] Buggineni, V., Chen, C., Camelio, J., 2024. Enhancing manufacturing operations with synthetic data: a systematic framework for data generation, accuracy, and utility. Frontiers in Manufacturing Technology 4, 1320166.

[5] Chickering, D.M., 2002. Optimal structure identification with greedy search. Journal of machine learning research 3, 507–554.

[6] Figueira, A., Vaz, B., 2022. Survey on synthetic data generation, evaluation methods and gans. Mathematics 10, 2733.

[7] Gogoshin, G., Branciamore, S., Rodin, A.S., 2021. Synthetic data generation with probabilistic bayesian networks. Mathematical biosciences and engineering: MBE 18, 8603.

[8] Hasan, U., Hossain, E., Gani, M.O., 2023. A survey on causal discovery methods for iid and time series data. arXiv preprint arXiv:2303.15027 .

[9] Jävergård, N., Lyons, R., Muntean, A., Forsman, J., 2024. Preserving correlations: A statistical method for generating synthetic data. arXiv preprint arXiv:2403.01471 .

[10] Kokol, P., Kokol, M., Zagoranski, S., 2022. Machine learning on small size samples: A synthetic knowledge synthesis. Science Progress 105, 00368504211029777.

[11] Lam, W.Y., Andrews, B., Ramsey, J., 2022. Greedy relaxations of the sparsest permutation algorithm, in: Uncertainty in Artificial Intelligence, PMLR. pp. 1052–1062.

[12] Lambers, J., Schüder, J., Krauß, J., 2022. Injection-molding production data with quality labels. URL: https://b2share.eudat.eu/records/03133fb279294389a15baefd55e4257a, doi:10.23728/B2SHARE.03133FB279294389A15BAEFD55E4257A.

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

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

[15] Overbeck, L., Graves, S.C., Lanza, G., 2024. Development and analysis of digital twins of production systems. International Journal of Production Research 62, 3544–3558.

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

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

[18] Sharma, A., Kiciman, E., 2020. Dowhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216 .

[19] Sikora, J., Wagnerová, R., Landryová, L., Šíma, J., Wrona, S., 2021. Influence of environmental noise on quality control of hvac devices based on convolutional neural network. Applied Sciences 11, 7484.

[20] Spirtes, P., Meek, C., Richardson, T.S., 2013. Causal inference in the presence of latent variables and selection bias. CoRR abs/1302.4983. URL: http://arxiv.org/abs/1302.4983, arXiv:1302.4983.

[21] Yu, J., Pychynski, T., Barsim, K.S., Huber, M.F., 2024a. Causal knowledge in data fusion: Systematic evaluation on quality prediction and root cause analysis, in: 2024 27th International Conference on Information Fusion (FUSION), pp. 1–8. doi:10.23919/FUSION59988.2024.10706429.

[22] Yu, J., Pychynski, T., Huber, M.F., 2024b. Causal knowledge in data fusion subject to latent confounding and measurement error, in: 2024 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 1–8. doi:10.1109/MFI62651.2024.10705789.