technologies

MDPI

Article

Systematic Generation and Evaluation of Synthetic Production Data for Industry 5.0 Optimization

Solomiia Liaskovska [1,2,*], Sviatoslav Tyskyi [1], Yevgen Martyn [3], Andy T. Augousti [2] and Volodymyr Kulyk [4]

[1] Department of Artificial Intelligence, Lviv Polytechnic National University, Kniazia Romana Street, 5, 79905 Lviv, Ukraine
[2] Department of Mechanical Engineering, Faculty of Engineering, Computing and the Environment, Kingston University, Room RV MB 215, Main Building (RV), Roehampton Vale, Kingston, London SW15 3DW, UK; augousti@kingston.ac.uk
[3] Department of Project Management, Information Technologies and Telecommunication, Lviv State University of Life Safety, 79007 Lviv, Ukraine; evmartyn@gmail.com
[4] Department of Materials Science and Engineering, Lviv Polytechnic National University, 12 S. Bandera Street, 79013 Lviv, Ukraine
* Correspondence: solomiam@gmail.com; Tel.: +380-676737755

Abstract: Our research focused on analyzing and advancing information technologies to identify ecological parameters in production. The primary goals were to enhance efficiency, reduce waste, and minimize the environmental impact of manufacturing processes. By incorporating the results of the study, we observed and systematized changes occurring in the transition from Industry 4.0 to Industry 5.0. Special attention was given to studying processes and technologies related to the generation of synthetic data and analyzing the implementation of cutting-edge technologies. The research object includes new parameters introduced within the framework of Industry 5.0, encompassing automation and cognitive technologies. Our scientific interests also extended to synthetic data used in modeling various production processes, including optimizing device performance in manufacturing and forecasting abnormal situations in industrial equipment operations. The subject of the research involves algorithms for generating synthetic data and methods for validating them to ensure their statistical similarity to real-world data. During the study, we also analyzed the impact of artificial intelligence implementation on improving the efficiency and adaptability of manufacturing systems.

Keywords: manufacturing processes; synthetic data; artificial intelligence; Industry 5.0; parameters; cognitive technologies

Academic Editor: Lipo Wang

Received: 12 January 2025
Revised: 11 February 2025
Accepted: 15 February 2025
Published: 18 February 2025

Citation: Liaskovska, S.; Tyskyi, S.; Martyn, Y.; Augousti, A.T.; Kulyk, V. Systematic Generation and Evaluation of Synthetic Production Data for Industry 5.0 Optimization. *Technologies* **2025**, *13*, 84. https://doi.org/10.3390/technologies13020084

1. Introduction

Modern industry is undergoing significant and forward-looking technological transformations. The transition from Industry 4.0 to Industry 5.0 is gaining increasing relevance and is impacting production automation [1]. A defining feature of these transformations is the active use of artificial intelligence and cognitive technologies. This evolution affects not only the technical characteristics of production but also introduces new challenges, particularly in the areas of social responsibility and environmental safety [2]. The relevance of this scientific research lies in the need to understand and evaluate the impact of the rapid technological changes occurring during the transition from Industry 4.0 to Industry 5.0 [3]. This transition involves the adoption of new technologies such as Artificial Intelligence (AI), cognitive systems, and automation. In our view, these advancements have the poten-

tial to significantly improve production processes, enhance manufacturing efficiency, and optimize costs.

The authors of article [2] describe the development of a tool for assessing the life cycle of production, which analyzes the environmental impact of food supply chains. The tool covers all stages, from raw material extraction to waste management, including production, distribution, and retail. The article also explores the use of Internet of Things (IoT) technologies to optimize these chains and reduce environmental impact, with a focus on the potential to minimize food waste [4].

The changes during the transition from Industry 4.0 to Industry 5.0 are minor but aim to enhance readability and precision [5]. It is important to emphasize that, on the one hand, these changes promise improvements and advancements in production capacities, reductions in energy and material costs, and higher product quality. On the other hand, they pose new challenges to society, such as the need to increase workplace safety, adapt the workforce to new working conditions, and find ways to reduce environmental impacts [6].

The relevance of our study is determined by the need to develop strategies for effectively implementing technological innovations.

The research described in reference [7] examines the evolution of manufacturing from traditional 3D technologies to innovative 4D printing, which incorporates the ability of materials to adapt or change properties in response to external influences. This development opens new possibilities for production in complex industries such as aerospace, biomedical engineering, and electronics. The article concludes that 4D printing could radically transform design and manufacturing approaches, offering significant advantages in product flexibility and adaptability, which are crucial for the transition to Industry 5.0.

These strategies must consider not only economic but also social and environmental aspects to ensure sustainable development and improved quality of life [6].

Article [8] explores the Critical Success Factors (CSFs) for implementing the concept of the circular economy (CE) in small and medium-sized enterprises (SMEs). The study employs the Fuzzy Decision-Making Trial and Evaluation Laboratory (DEMATEL) methodology to analyze and determine the structural interactions between CSFs and conducts a case study to validate the results. The authors identify that new government policies and CE regulations, consumer awareness, demand for CE products, and economic incentives for CE products are the leading causal groups of CSFs.

The authors of article [9] describe the Optimized Multi-Level Multi-Type Ensemble (OMME) forecasting model for one-dimensional time series, specifically for predicting energy consumption. The authors presented a novel approach combining various machine learning algorithms and statistical methods to create an ensemble model aimed at reducing forecasting errors. The study demonstrated that hybrid ensemble methods, such as bagging and hybrid approaches using algorithms such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Multilayer Perceptron (MLP) achieve superior prediction accuracy. The authors presented a novel approach that combines various machine learning algorithms and statistical methods to create an ensemble model. This model is designed to reduce forecasting errors. The research demonstrated that using hybrid ensemble methods (such as bagging and hybrid techniques) with algorithms like LSTM, GRU, and MLP achieves better prediction accuracy. The application of optimization algorithms, such as Bayesian optimization and Tabu search (TS) is an iterative neighborhood search algorithm, where the neighborhood changes dynamically, and helps in selecting the optimal hyperparameters for the models.

The findings of the study suggest that employing ensemble methods in machine learning can significantly enhance the efficiency of time series forecasting, making it an essential tool for managing energy consumption in a more stable and cost-effective

manner. Additionally, the authors showed that GRU played a key role in forecasting energy consumption, while Autoregressive integrated moving average (ARIMA) models were less effective due to the high noise levels in the data.

The use of optimization algorithms [10], including Bayesian Optimization and Tabu Search, enables the selection of optimal model hyperparameters. The study concludes that employing ensemble methods in machine learning significantly enhances the efficiency of time series forecasting and serves as an essential tool for managing energy consumption in a more stable and cost-effective manner [11].

The transition from Industry 4.0 to Industry 5.0 [3] signifies an important shift in the integration of advanced technologies into manufacturing processes [12]. Industry 4.0 primarily focuses on the digitalization of production [13] and leveraging big data [14], IoT, and artificial intelligence to achieve higher levels of operational efficiency and automation. One of the challenges in the development and analysis of modern technological systems is the lack of sufficiently high-quality data on equipment operation and failures. In real-world conditions, the number of equipment failure cases may be limited, complicating the training process for machine learning models. This is where synthetic data come into play. The generation of synthetic data [15] allows for the creation of artificial datasets that can replicate real scenarios and equipment behavior. This provides the opportunity to expand existing datasets, balance them, and improve failure prediction accuracy, thereby ensuring more effective and reliable operation of monitoring systems.

Synthetic data are artificially created data that mimic real data. They are generated using statistical models and machine learning algorithms and do not contain identifying information about individuals or objects. Synthetic data can serve as an alternative to real data when access to the latter is limited or when the use of real data raises concerns about privacy or legal issues. One of the advantages of synthetic data is that they allow researchers and developers to create and test algorithms and models without the risk of exposing sensitive or private information. Synthetic data can also be used to create diverse and representative datasets and to fill existing data gaps.

The generation of synthetic data involves creating a model based on existing data and then using this model to produce new data with similar statistical properties. The model can be adjusted to generate data with varying levels of complexity, variation, and noise. However, it is important to ensure that synthetic data accurately reflect real data and do not introduce any biases or inaccuracies [16].

In contrast, Industry 5.0 builds on these technological foundations to introduce more complex interactions between humans and machines, such as the use of collaborative robots (cobots) and advanced interfaces that enable more intuitive communication with machines [17].

The contributions of this work can be summarized as follows:

- Providing a systematic analysis and determination of production parameters with a focus on their synthetic generation to optimize production processes such as ambient air temperature, operating temperature of the equipment, continuous usage time of the device, total usage time of the device, load on the device, and the level of the device network voltage.
- The primary objective was to create synthetic datasets that replicate the structure and statistical relationships of real data, while also evaluating their performance in classification tasks. Various generation methods, such as Gaussian Copula and Generative Adversarial Networks (GANs), were employed to produce synthetic samples for each target variable category. The quality of the generated data was assessed using metrics like statistical similarity and distribution of variables, and it was determined that the TVAE method produced the most satisfactory results [18].

- This study presents a comprehensive methodology for identifying critical production parameters, selecting appropriate models for generating synthetic data [19], and developing reliable methods for evaluating the quality and validity of the generated data. These contributions address key challenges in the shift from Industry 4.0 to Industry 5.0, with a focus on sustainable and adaptive manufacturing practices.

2. Materials and Methods

One of the critical aspects of improving production efficiency is the timely prediction and identification of potential equipment failures. Breakdowns in manufacturing can lead to significant downtime, negatively affecting productivity, and increasing costs. Therefore, the implementation of monitoring and predictive maintenance systems is a key task for many enterprises.

Modern technologies enable not only the timely detection of issues but also the prediction of potential failures based on the analysis of equipment performance data. This process leverages various machine learning methods, such as classification and regression, to analyze large datasets and accurately identify potential failures at early stages [20].

However, one challenge in developing such systems is the availability of sufficient high-quality data on equipment operation and failures. In real-world conditions, the number of equipment failure cases may be limited, complicating the training process for machine learning models. This is where synthetic data come into play. Generating synthetic data allows the creation of artificial datasets that can replicate real scenarios and equipment behavior. This enables the expansion and balancing of existing datasets, enhancing the accuracy of failure prediction and ensuring more efficient and reliable operation of monitoring systems. For the research, we used a dataset with real data [21] that include the following metrics taken from a milling machine and consists of 10,000 data points. The metrics used include the following variables: Air Temperature, Rotational Speed, Torque, and Tool Wear [22].

2.1. The Methods of Synthetic Data Generation

2.1.1. Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) [13] is a probabilistic model used to represent data as a mixture of several normal (Gaussian) distributions. It is commonly used in clustering tasks, especially when the data consist of several subgroups, each of which follows its own normal distribution with unknown parameters.

The GMM assumes that the data are generated as a mixture of several Gaussian distributions, which are called components. Each component has its own parameters:

- The mean (μ), which defines the center of the distribution;
- The variance (σ) for single-variable cases;
- The covariance matrix (Σ), which characterizes the shape and distribution of data around the mean for multivariate cases;
- The weight coefficient (ϕ), defines the probability that a random sample belongs to a particular component.

The GMM represents the overall probability density of the data as a weighted sum of several Gaussian distributions. The formula for the probability that a given data point x belongs to the model in the univariate case is as follows (1):

$$p(x) = \sum_{k=1}^{K} \phi_i N(x|\mu_i \sigma_i) \quad (1)$$

where

K is the number of components (Gaussian distributions) in the model.

$N(x|\mu_i\sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}}e^{(-\frac{(x-\mu_i)^2}{2\sigma_i^2})}$ —the probability density of a normal distribution.

$$\sum_{k=1}^{K}\phi_i = 1$$

For the multivariate case (2):

$$p\left(\vec{x}\right) = \sum_{k=1}^{K}\phi_i N(\vec{x}|\vec{\mu}_i\Sigma_i) \quad (2)$$

GMMs are trained using the Expectation–Maximization (EM) algorithm.

Maximization (M-step): Update the parameters (means, covariance matrices, and weights) of each component to maximize the likelihood of the data, based on the calculated probabilities. This process is repeated until the parameter changes become negligible or a predefined number of iterations is reached.

The number of Gaussian components (K) is usually set before training the model. Methods like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) can be used to determine the optimal number of components by balancing likelihood and model complexity. The main advantages are the ability to model clusters of different shapes, sizes, and densities. The disadvantages include sensitivity to the choice of initial parameters [23], possibility of becoming stuck in local optimum during training, and being computationally demanding, especially for large datasets. The Gaussian Mixture Model (GMM) is applied to clustering tasks with a small number of components and well-defined groups, data with simple dependencies between variables, and where soft classification is required. However, the GMM is less effective for large multidimensional data or nonlinear dependencies between variables.

2.1.2. Gaussian Copula Model

The Gaussian Copula Model is one of the most popular methods for modeling dependencies between random variables, especially when each variable has its own unique distribution. Copulas are used to split complex multivariate distributions into two parts: the univariate marginal distributions of each variable and the dependency structure between them. This method is effective for simulating correlations and relationships between variables in a dataset.

The copula is a mathematical tool that describes the dependence between random variables independently of their marginal distributions. The use of copulas allows for the modeling of complex correlations and dependencies between variables, even when their distributions are not normal.

The Gaussian copula is a specific type of copula that operates based on normal distributions. The concept of the Gaussian copula involves transforming any random variables into a space of standard normal distributions (i.e., distributions with a mean of 0 and a variance of 1), which allows for easy modeling of dependencies between them using a correlation matrix. Since the copula operates in a normalized space, this space can be used to establish any complex relationships between variables.

Let there be a dataset of n random variables. X_1, $X_2, \ldots, X_n$, each of which has its own unique distribution. Initially, to transform the original values into a uniform space, for each variable. X_i, the empirical distribution function is used F_i (3),

$$U_i \rightarrow F_i(X_i) \quad (3)$$

where U_i—a uniformly distributed variable on the interval [0, 1]. This step allows the distribution of each variable to be "fitted" across a uniform interval.

After obtaining the uniform variables U_i, they are transformed into the standard normal space using the inverse quintile function of the standard normalized distribution [ ]. In this normalized space, the correlation between variables is established by constructing a correlation matrix Σ. This matrix defines how the variables interact with each other. The correlation matrix can be estimated based on empirical data. Using the obtained correlation matrix, new synthetic datasets can be generated in the normal space. The generation process includes generating a vector $Z = (Z_1, Z_2, \ldots, Z_n)$ from a multivariate normal distribution with the correlation matrix. Σ, transformation Z_i into a uniform space by using the quantile function of the standard normal distribution. (4):

$$U_i \rightarrow \Phi(Z_i) \tag{4}$$

The use of the inverse empirical distribution function to obtain synthetic values in the original space (5):

$$X_i^{synthetic} = F_i^{-1}(U_i) \tag{5}$$

The generated synthetic data $X_i^{synthetic}$ preserve the dependency structure and the marginal distributions of the original variables, allowing them to be used for further analysis and modeling.

For our forecasting task, where it is important to maintain the correlation between variables. For the datasets we use, which have clearly defined dependencies and a small number of variables. It should be noted that the method is less effective for complex nonlinear dependencies or large dimensions.

2.1.3. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) consist of two neural networks: the generator and the discriminator (Figure 1). The generator creates synthetic data that mimic real data, while the discriminator evaluates whether the data are real or generated. This process is akin to a competition between a counterfeiter and a fraud detector.

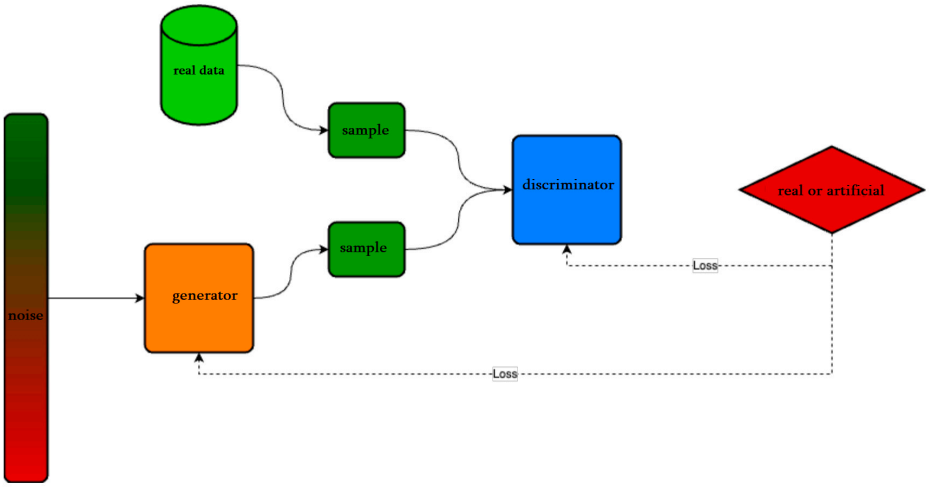


Figure 1. Models for Generating Synthetic Data: GAN Architecture.

As the generator improves its ability to produce realistic data, the discriminator enhances its skills in identifying fakes. Over time, this adversarial training leads to the generation of highly realistic synthetic data. The dynamic between these two networks is essential for advancing applications in various fields, including image synthesis and

data augmentation. GANs are used for the generation of images, text, or other data with complex structures and the creation of data with a high level of variability, for example, in biomedical or financial research.

The research consists of two parts, each covering different aspects of working with synthetic data and their impact on the quality of classification models.

The first part is dedicated to the generation of synthetic data and their evaluation using commonly accepted metrics. The main goal of this phase of the study is to create artificial data that closely resemble real data in terms of statistical properties and structure. Various models for generating synthetic data are used, such as Gaussian copula models and Generative Adversarial Networks (GANs). The generation process involves not only creating the data but also assessing their quality by comparing them with the original data using metrics such as statistical similarity, the Kolmogorov–Smirnov test, total variation distance, and others. These metrics help verify how well the synthetic data reflect the distributions and relationships between variables, which is important for their further use in real-world analytical and modeling tasks.

The second part of our research focuses on evaluating the quality of synthetic data in the context of a specific task—classification. At this stage, the primary emphasis is on how synthetic data can improve or worsen the performance of machine learning models. A comparison is made between the results of classification models trained only on real data and those augmented with synthetic data. A key point is analyzing the impact of synthetic data on various classification metrics, such as accuracy, recall, F-measure, and others. It is important to understand whether the use of synthetic data helps improve classification results, especially in cases where real data are limited or classes are imbalanced. The approach we have formulated allows for a comprehensive evaluation of both synthetic data generation and its practical usefulness in real tasks, such as classification. Comparing modeling results with and without synthetic data makes it possible to draw conclusions about their effectiveness and the justification of their use to improve machine learning models.

3. Results

This Section is dedicated to the generation of synthetic data and their evaluation using widely accepted metrics. The primary goal of this stage is to create artificial data that closely resemble real data in terms of statistical properties and structure. To achieve this, various models for generating synthetic data are employed, such as Gaussian copula models and Generative Adversarial Networks (GANs). The second part of the study focuses on assessing the quality of synthetic data in the context of a specific task—classification. At this stage, the emphasis is on how synthetic data can improve or hinder the performance of machine learning models. A comparison is made between the results of classification models trained solely on real data and those supplemented with synthetic data. A key aspect is analyzing the impact of synthetic data on various classification metrics such as accuracy, recall, F-measure, and others. It is important to determine whether the use of synthetic data enhances classification results, particularly in cases where real data are limited or class distributions are imbalanced. This approach allows for a comprehensive evaluation of both synthetic data generation and their practical utility in real-world tasks like classification. Comparing the modeling results with and without synthetic data enables conclusions about their effectiveness and justification for their application to improve machine learning models.

3.1. Generation of Synthetic Data

The objective of this stage is to generate 10,000 new data rows, with 1000 entries for each target category, and 7000 entries where the target column equals 0.

The first step involved data cleaning. Incorrect or unnecessary rows with categories "RNF" and "TWF" were removed. Columns that were not needed for generating synthetic data were also eliminated from the dataset.

Subsequently, synthetic data were generated using the Gaussian Mixture Model (GMM). For each category, such as "Machine failure", "HDF", "PWF", and "OSF", synthetic samples were generated based on the data distribution in real examples. This process was carried out separately for each category to ensure that synthetic data matched real distributions in each category. The optimal number of components for each class was selected using information criteria such as AIC and BIC.

Several different methods for data generation were utilized, including GANs, TVAE, and Copula; however, the results were less satisfactory compared to GMM. These models demonstrated lower-quality synthetic data, which were evaluated using various metrics, including statistical similarity and distribution coverage metrics.

For each of these methods, generation models were built and assessed using standard quality metrics for synthetic data. The generation of synthetic samples for each category was conducted using parametric and non-parametric distributions, such as Gaussian KDE, beta distribution, and normal distribution for various variables.

Figure 2 demonstrates a comparison of the ratio between the columns Torque [Nm] and Rotational speed [rpm] for both real and synthetic data.

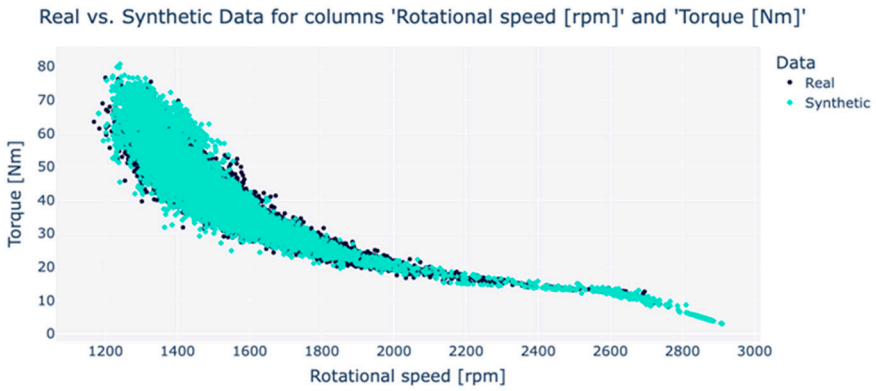


Figure 2. Comparison of the ratio of the columns Torque [Nm] to Rotational speed [rpm] for real and synthetic data.

For each of these methods, generation models were built, and their performance was assessed using standard synthetic data quality metrics. The generation of synthetic samples for each category was carried out using parametric and non-parametric distributions, such as Gaussian KDE, beta distribution, and normal distribution for different variables. The results were collected in Tables 1 and 2. Table 1 demonstrates data generation quality metrics.

Table 2 demonstrates aggregated metrics for data generation quality.

The presented results show that the quality metrics for generating synthetic data significantly differ depending on the method used. In particular, for the models, GausMixture, GaussianCopulaSynthesizer, TVAESynthesizer, CTGANSynthesizer, and CopulaGANSynthesizer, the values of the Column Pair Trends and Column Shapes metrics indicate some shortcomings in preserving dependencies between columns and their initial distributions. This happens because the target functions were generated separately, and the contingency with the target column was not preserved. However, all other metrics are within normal

limits, so it was decided to retain all generated datasets for testing their suitability for classification [23].

Table 1. Data generation quality metrics.

	Column Pair Trends	Column Shapes	Data Structure	Data Validity
GausMixture	0.26	0.87	1	0.99
GaussianCopula Synthesizer	0.26	0.86	1	1
TVAESynthesizer	0.26	0.84	1	1
CTGANSynthesizer	0.24	0.86	1	1
CopulaGANSynthesizer	0.24	0.85	1	1
GaussianCopulaSynthesizer-Categorical	0.86	0.91	1	1
CopulaGANSynthesizerCategorical	0.82	0.86	1	1
TVAESynthesizerCategorical	0.81	0.87	1	1
CTGANSynthesizerCategorical	0.83	0.88	1	1

Table 2. Aggregated metrics for data generation quality.

	Data Quality Score	Diagnostic Score
GausMixture	0.57	0.99
GaussianCopulaSynthesizer	0.56	1
TVAESynthesizer	0.55	1
CTGANSynthesizer	0.55	1
CopulaGANSynthesizer	0.55	1
GaussianCopulaSynthesizerCategorical	0.88	1
CopulaGANSynthesizeCategorical	0.84	1
TVAESynthesizerCategorical	0.84	1
CTGANSynthesizerCategorical	0.86	1

3.2. Data Quality Control in Practice (the Quality of the Generated Synthetic Data, for Which We Conducted a Series of Experiments)

In order to check the quality of the generated synthetic data, we conducted a series of experiments using different classification models. In particular, the models used were KNeighborsClassifier, for which the number of neighbors was selected using the "elbow method", DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, and MLPClassifier with a neural network architecture including hidden layers (25, 50, 25), activation of 'relu' and optimization algorithm 'adam'. The purpose of the experiment was to evaluate how well synthetic data reproduce real statistical dependencies and whether they can be used to improve classification in real problems [24].

The experimental process was divided into several stages. First, a test dataset was created from the original dataset, which was used as a basis for comparison. The first step was to train the models only on real data to obtain basic performance metrics. This made it possible to assess how well the models classify real data without the intervention of synthetic ones.

In the second step, the models were trained exclusively on synthetic data to check whether these data contained significant statistical dependencies. This was an important step for assessing the quality of synthetic data: if models can be successfully trained on synthetic data, this indicates that they reproduce the structures and dependencies present in real data.

The last step consisted of combining real and synthetic data to train the models. This allowed us to assess whether synthetic data could help improve the overall performance of

the models. The models were then tested on a test set of real data, and performance metrics were collected for each model.

Standard metrics, such as accuracy, precision, recall, F1-measure, and MCC score, as well as specialized metrics for our task, such as "Broken accuracy", "Broken precision", and "Broken recall" were used to assess the quality of the classification. The last metrics evaluated the performance of the model in the classification of (broken) objects. The "Broken accuracy" metric was defined as the ratio of the number of correctly classified broken objects to all objects that were classified as broken. "Broken precision" and "Broken recall" were also adapted to evaluate the performance of the model in detecting and correctly classifying broken objects. In addition to these metrics, training and prediction time information was collected for each model, allowing us to assess the overall efficiency and performance of the models in different scenarios.

The results of all experiments were recorded in Table 3, highlighting the top 5 options.

Table 3. Metrics for evaluating the quality of the most effective models.

Dataset	Broken Accuracy	Broken Precision	Broken Recall	Precision	1-Score	MCC Score
TVAESynthesizer	48.27	81.90	54.03	98.15	7.70	65.18
real	47.69	54.39	79.49	98.18	8.11	65.04
TVAESynthesizer	46.92	82.43	52.14	98.10	7.59	64.27
real	46.18	57.47	70.17	97.91	7.95	62.25
TVAESynthesizer	44.89	90.18	47.20	98.20	7.31	64.03

By evaluating the results using metrics important for this case, such as MCC, Broken accuracy, Broken precision, and Broken recall, we can see that synthetic data indeed helped the models better distinguish failures, specifically heat dissipation violations (ПВТ), power failures (ЗЖ), and overvoltage (ПНП). The MCC metric indicates that the models predict membership in a specific class better and are not biased toward any one particular class.

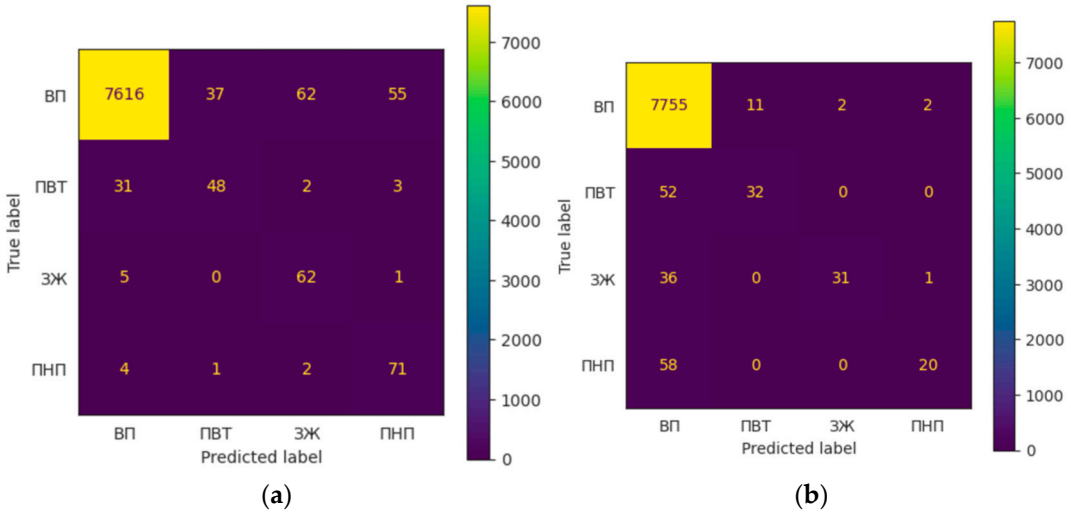
Additionally, to explore avenues for further optimization, the metrics for methods that showed poorer results were analyzed and recorded in Table 4.

Table 4. Metrics for evaluating model quality.

Dataset	Broken Accuracy	Broken Precision	Broken Recall	Precision	F1-Score	MCC Score
GaussianCopulaSynthesizer	42.09	60.18	58.33	97.88	97.52	58.13
TVAESynthesizer_Manual	41.65	78.54	46.99	46.99	97.22	59.35
TVAESynthesizer	41.31	75.58	47.67	97.79	97.27	58.55

Models such as Gaussian Copula, TVAE with a manual approach to calculating the target column, and GAN also showed decent results; however, due to the specificity of the data, they were deemed less effective compared to TVAE with an automatic approach to calculating the target column according to the metrics.

The evaluation was also conducted using the confusion matrix. A comparison of the results of models trained using synthetic data (Figure 3a) and models that used only the original dataset (Figure 3b) shows that synthetic data increased the ability of the model to more accurately and frequently predict device failures in cases of their actual occurrence (True Positive). At the same time, the number of false predictions of failure in cases of its absence (False Positive) has increased. However, for this particular task, these results are satisfactory.



(**a**) (**b**)

Figure 3. Discrepancy of a matrix: (**a**) Mismatch matrices for synthetic data; (**b**) Discrepancy matrices for the original data.

Figure 3a shows the prediction of failures using synthetic data, while Figure 3b shows the prediction of failures using original data. The parameters are as follows: ВП—No error, ПВТ—Heat dissipation failure, ЗЖ—Power failure, and ПНП—Overvoltage.

To assess the quality of the generation of synthetic data, we used data visualization, which compares the distributions of real and synthetic data according to various characteristics. Figure 4 shows a comparison of synthetic data distributions.

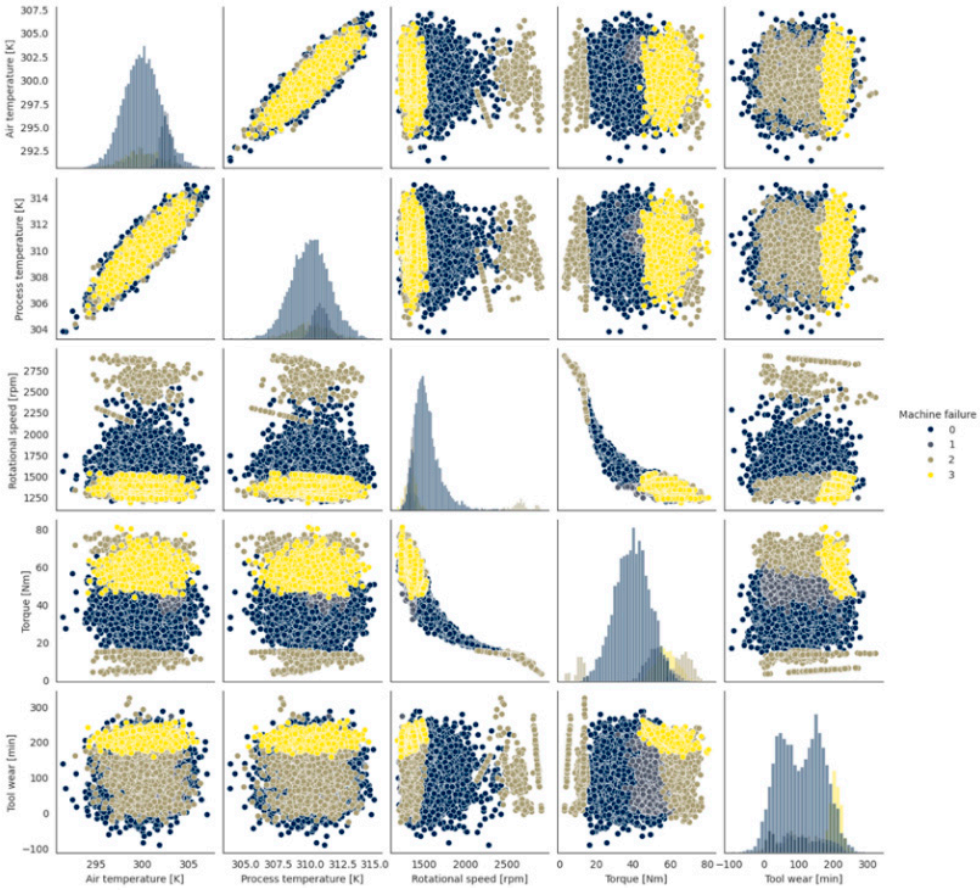


Figure 4. Comparison of synthetic data distributions.

In Figure 4 we can see the original data. The different categories of the "Machine failure" target function are color-coded:

- Blue color (0) corresponds to the absence of breakdowns,
- Gray (1), brown (2), and yellow (3) are different types of breakdowns.

In Figure 4, it can be observed that the distributions of the synthetic data generally preserve similar relationships between variables, as seen in the original data presented in Figure 5. However, certain areas in the synthetic data appear more 'dense' or 'uniformly filled'. This suggests that the generation model may have attempted to reproduce the full range of values for each feature. For example, the relationship between "Process temperature [K]" and "Air temperature [K]" in both images has a similar linear nature, indicating the successful reproduction of this correlation by the synthetic data. However, the synthetic data demonstrate greater variability in some areas, which may explain the better results in predicting failures (True Positives), but also the increased number of false alarms (False Positives), as mentioned earlier. Overall, the synthetic data in Figure 4 effectively reproduce the general trends and correlations, but some details, such as natural variations in distributions or the presence of clusters, may differ from the original data in Figure 5.

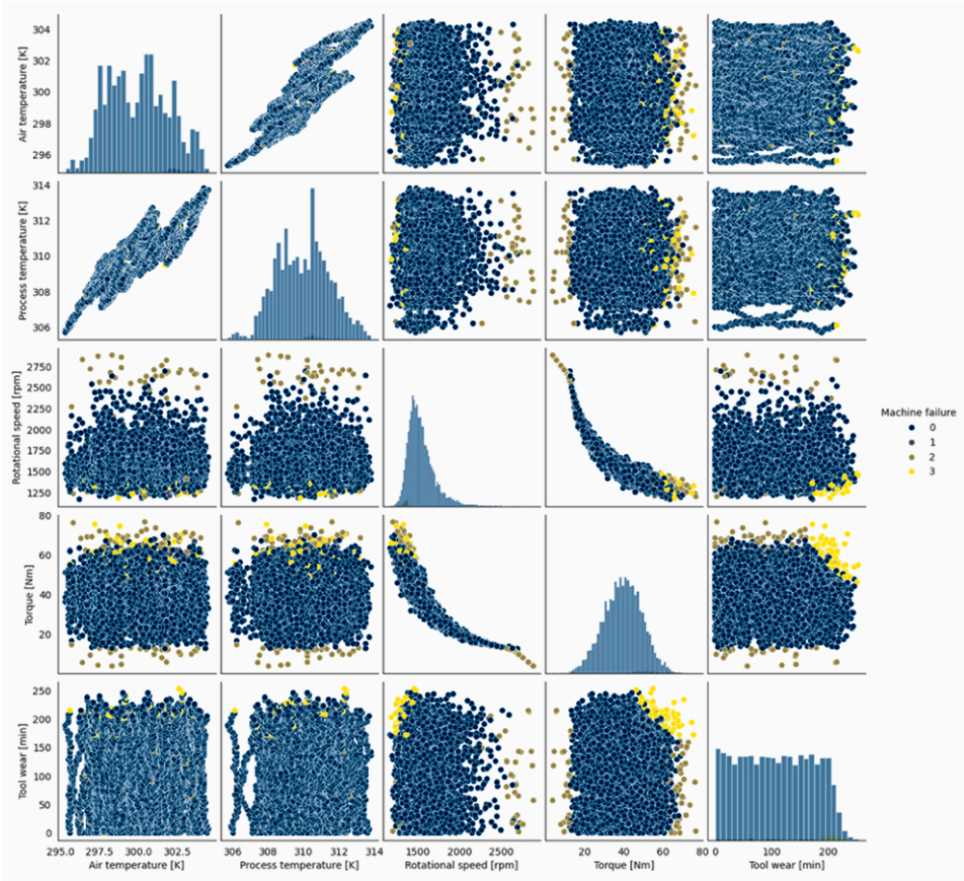


Figure 5. Comparison of original data distributions.

4. Discussion

Using various generation methods, such as Gaussian Copula and Generative Adversarial Networks (GANs), synthetic samples were generated for each category of the target variable. The quality of the generated data was assessed using metrics such as statistical similarity and variable distributions, and it was found that the most satisfactory results were achieved using the TVAE method. We conducted experiments on the generation

of synthetic data and investigated their impact on the quality of classification models. Synthetic datasets were created to replicate the structure and statistical dependencies of real data, as well as to evaluate their effectiveness in classification tasks. Using various generation methods, such as Gaussian Copula and Generative Adversarial Networks (GANs), synthetic samples were generated for each category of the target variable. The quality of the generated data was assessed using metrics such as statistical similarity and variable distributions, revealing that the most satisfactory results were achieved using the Gaussian Mixture method. The original data exhibit natural heterogeneity and some gaps in the distributions. For instance, the relationship between 'Process temperature [K]' and 'Air temperature [K]' shows a similar linear pattern in both figures. This indicates a successful reproduction of this correlation by the synthetic data. However, the synthetic data sometimes demonstrate greater variability, which may account for better results in predicting failures (True Positives) but also a higher number of false alarms (False Positives), as noted earlier.

Overall, the synthetic data in Figure 4 effectively replicate general trends and correlations. However, certain details, such as natural variations in distributions or the presence of clusters, may differ from the original data in Figure 5.

5. Conclusions

This study carried out a sequential process, starting from the analysis of scientific sources to practical experiments, to achieve the primary objective: investigating environmental parameters in production. After systematizing and selecting the materials, it was established that emerging trends, particularly in Industry 5.0, focus on sustainable development, reducing the environmental impact of production, automation, and improving human–machine interaction. Key technologies that enhance the efficiency of production processes and minimize waste were also identified.

This phase focused on creating a synthetic dataset of 10,000 records, ensuring balance across the target variable categories with 1000 records for each of the three categories and 7000 records for the "normal state". The dataset structure aimed to balance rare categories and the dominant class, maintaining a realistic distribution and enhancing modeling effectiveness. Data cleaning was performed to remove irrelevant or incorrect entries, including categories "RNF" and "TWF", reducing noise and optimizing the dataset for further analysis. Various data generation methods were tested, with GANs, TVAN, and Copula models yielding less satisfactory results compared to Gaussian Mixture Models (GMMs), which provided better synthetic data quality, assessed through standard metrics. Synthetic samples were generated using parametric and non-parametric distributions such as Gaussian KDE, beta, and normal distributions.

A detailed implementation of experiments on synthetic data generation and their impact on the quality of classification models was carried out. The main goal was to create synthetic datasets that would replicate the structure and statistical dependencies of real data, as well as evaluate their effectiveness in classification tasks. Using various generation methods, such as Gaussian Copula and Generative Adversarial Networks (GANs), synthetic samples were generated for each target variable category. The quality of the generated data was assessed using metrics like statistical similarity and variable distributions, and it was found that the most satisfactory results were achieved using the TVAE method. A series of experiments were conducted to train classification models on real, synthetic, and combined datasets. The results showed that synthetic data can improve model performance, especially in cases of insufficient real data or class imbalance. Model performance was evaluated using metrics such as accuracy, F1-score, and MCC, as well as specialized metrics for faulty objects ("Broken accuracy", "Broken precision", "Broken

recall"). The best results were shown by models trained on combined datasets, confirming the feasibility of using synthetic data to improve classification quality. Thus, the conducted research demonstrates that synthetic data, generated using the correct methods, can be effectively used to enhance the accuracy of machine learning models, especially in cases where real data are limited or imbalanced.

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

Funding: This work is funded by the European Union's Horizon Europe research and innovation program under grant agreement No. 101138678, project ZEBAI (Innovative Methodologies for the Design of Zero-Emission and Cost-Effective Buildings Enhanced by Artificial Intelligence). Some of the authors (S.L. and A.T.A.) should also like to thank the British Academy for the award of a Researcher at Risk Fellowship Aware Reference: RaR\100791.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this study. These data can be found here: https://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020 (accessed on 20 October 2024).

Acknowledgments: The authors thank the anonymous reviewers for their insightful and constructive recommendations, which helped improve the paper.

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

References

1. Aguiar, M.C.D.; Fernandes, M.C.S.; Sant'Ana, M.A.K.; Sagrillo, V.P.D.; Anastácio, A.D.S.; Gadioli, M.C.B. Eco-Efficient Artificial Stones Produced Using Quartzite Quarry Waste and Vegetable Resin. *Sustainability* **2023**, *16*, 247. [CrossRef]
2. Costa, T.P.D.; Gillespie, J.; Pelc, K.; Adefisan, A.; Adefisan, M.; Ramanathan, R.; Murphy, F. Life Cycle Assessment Tool for Food Supply Chain Environmental Evaluation. *Sustainability* **2022**, *15*, 718. [CrossRef]
3. Tzampazaki, M.; Zografos, C.; Vrochidou, E.; Papakostas, G.A. Machine Vision—Moving from Industry 4.0 to Industry 5.0. *Appl. Sci.* **2024**, *14*, 1471. [CrossRef]
4. Durkin, A.; Otte, L.; Guo, M. Surrogate-Based Optimisation of Process Systems to Recover Resources from Wastewater. *Comput. Chem. Eng.* **2024**, *182*, 108584. [CrossRef]
5. Jain, V.; Mitra, A. Development and Application of Machine Learning Algorithms for Sentiment Analysis in Digital Manufacturing: A Pathway for Enhanced Customer Feedback. In *Advances in Logistics, Operations, and Management Science*; Hassan, A., Dutta, P.K., Gupta, S., Mattar, E., Singh, S., Eds.; IGI Global: Hershey, PA, USA, 2024; pp. 26–38. ISBN 979-8-3693-0920-9.
6. Nykoniuk, M.; Basystiuk, O.; Shakhovska, N.; Melnykova, N. Multimodal Data Fusion for Depression Detection Approach. *Computation* **2025**, *13*, 9. [CrossRef]
7. Khaira, A. From 3D to 4D: The Evolution of Additive Manufacturing and Its Implications for Industry 5.0. In *Advances in Logistics, Operations, and Management Science*; Hassan, A., Dutta, P.K., Gupta, S., Mattar, E., Singh, S., Eds.; IGI Global: Hershey, PA, USA, 2024; pp. 39–53. ISBN 979-8-3693-0920-9.
8. Kumar, R.; Gupta, S.; Ur Rehman, U. Circular Economy a Footstep toward Net Zero Manufacturing: Critical Success Factors Analysis with Case Illustration. *Sustainability* **2023**, *15*, 15071. [CrossRef]
9. Usmani, M.; Memon, Z.A.; Danyaro, K.U.; Qureshi, R. Optimized Multi-Level Multi-Type Ensemble (OMME) Forecasting Model for Univariate Time Series. *IEEE Access* **2024**, *12*, 35700–35715. [CrossRef]
10. Maio, R.; Araújo, T.; Marques, B.; Santos, A.; Ramalho, P.; Almeida, D.; Dias, P.; Santos, B.S. Pervasive Augmented Reality to Support Real-Time Data Monitoring in Industrial Scenarios: Shop Floor Visualization Evaluation and User Study. *Comput. Graph.* **2024**, *118*, 11–22. [CrossRef]

11. Melnykova, N.; Kulievych, R.; Vycluk, Y.; Melnykova, K.; Melnykov, V. Anomalies Detecting in Medical Metrics Using Machine Learning Tools. *Procedia Comput. Sci.* **2022**, *198*, 718–723. [CrossRef]
12. Penchel, R.A.; Aldaya, I.; Marim, L.; Dos Santos, M.P.; Cardozo-Filho, L.; Jegatheesan, V.; De Oliveira, J.A. Analysis of Cleaner Production Performance in Manufacturing Companies Employing Artificial Neural Networks. *Appl. Sci.* **2023**, *13*, 4029. [CrossRef]
13. Shakhovska, N.; Mochurad, L.; Caro, R.; Argyroudis, S. Innovative Machine Learning Approaches for Indoor Air Temperature Forecasting in Smart Infrastructure. *Sci. Rep.* **2025**, *15*, 47. [CrossRef] [PubMed]
14. Izonin, I.; Muzyka, R.; Tkachenko, R.; Dronyuk, I.; Yemets, K.; Mitoulis, S.-A. A Method for Reducing Training Time of ML-Based Cascade Scheme for Large-Volume Data Analysis. *Sensors* **2024**, *24*, 4762. [CrossRef] [PubMed]
15. Takyar, A. Synthetic Data: Types, Generation, Evaluation, Use Cases and Applications. Available online: https://www.leewayhertz.com/what-is-synthetic-data/ (accessed on 7 February 2025).
16. Liang, O. How to Measure Statistical Similarity on Tabular Data?—Demonstrated Using Synthetic Data. *Medium* 2020. Available online: https://medium.com/@olivia.liang032/how-to-measure-statistical-similarity-on-tabular-data-demonstrated-using-synthetic-data-66a1aa60084d (accessed on 20 October 2024).
17. Velusamy, S.; Raguvaran, S.; Vinoth Kumar, S.; Suresh Kumar, B.; Padmapriya, T. From Industry 4.0 to 5.0: Digital Management Model of Personnel Archives Based on Transition From Digital Manufacturing. In *Advances in Logistics, Operations, and Management Science*; Hassan, A., Dutta, P.K., Gupta, S., Mattar, E., Singh, S., Eds.; IGI Global: Hershey, PA, USA, 2024; pp. 1–25. ISBN 979-8-3693-0920-9.
18. Olamendy, J.C. Understanding Gaussian Mixture Models: A Comprehensive Guide. *Medium* 2024. Available online: https://medium.com/@juanc.olamendy/understanding-gaussian-mixture-models-a-comprehensive-guide-df30af59ced7 (accessed on 20 October 2024).
19. Shang, C.; You, F. Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era. *Engineering* **2019**, *5*, 1010–1016. [CrossRef]
20. Izonin, I.; Tkachenko, R.; Yemets, K.; Havryliuk, M. An Interpretable Ensemble Structure with a Non-Iterative Training Algorithm to Improve the Predictive Accuracy of Healthcare Data Analysis. *Sci. Rep.* **2024**, *14*, 12947. [CrossRef] [PubMed]
21. Matzka, S. Explainable Artificial Intelligence for Predictive Maintenance Applications. In Proceedings of the 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 21–23 September 2020; pp. 69–74.
22. Predictive Maintenance Dataset (AI4I 2020). Available online: https://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020 (accessed on 7 February 2025).
23. Khan, A.A.; Laghari, A.A.; Alroobaea, R.; Baqasah, A.M.; Alsafyani, M.; Bacarra, R.; Alsayaydeh, J.A.J. Secure Remote Sensing Data With Blockchain Distributed Ledger Technology: A Solution for Smart Cities. *IEEE Access* **2024**, *12*, 69383–69396. [CrossRef]
24. Liaskovska, S.; Gumen, O.; Martyn, Y.; Zhelykh, V. Investigation of Microclimate Parameters in the Industrial Environments. In *Advances in Artificial Systems for Logistics Engineering III.*; Hu, Z., Zhang, Q., He, M., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer Nature: Cham, Switzerland, 2023; Volume 180, pp. 448–457. ISBN 978-3-031-36114-2.