Applied Computer Science 2025, vol. 21, no. 3, pp. 1–22
https://doi.org/10.35784/acs_7202

Submitted: 2025-02-03 | Revised: 2025-04-24 | Accepted: 2025-05-06

Keywords: industrial manufacturing systems, generative doppelgangers, stochastic data, operational optimization

Richard NASSO TOUMBA [1*], *Maxime MOAMISSOAL SAMUEL* [1], *Achille EBOKE* [1], *Boniface ONDO* [2], *Timothée KOMBE* [1]

[1] University of Douala, Cameroon, richardnassotoumba4@gmail.com, moamissoalmaxime@gmail.com, ebokeachille@gmail.com, tkombe@yahoo.fr
[2] Masuku University of Science and Technology, Gabon, bonitoondo@gmail.com
[*] Corresponding author: richardnassotoumba4@gmail.com

Taming complexity: Generative doppelgangers for stochastic data trends in complex industrial manufacturing systems

Abstract

The defining characteristics of complex industrial systems are interconnected processes that generate immense amounts of stochastic data, often hindering operational optimization, especially metrics such as Overall Equipment Effectiveness (OEE). To address the limitations of traditional methods and earlier machine learning techniques in capturing this complexity, this paper proposes a novel approach using generative doppelgangers, a Generative Adversarial Network (GAN)-based model, to simulate the operational behavior of these systems. This "behavioral doppelganger" learns intricate relationships within historical operational data from a production facility, enabling proactive what-if analyses for OEE optimization. The proposed framework's ability to replicate the impact of process parameters on availability, quality, and performance, which collectively contribute to OEE, is highlighted. The research validates this approach using real data from an industrial sugar plant, demonstrating its potential to provide valuable insights into system behavior under different operational scenarios for proactive optimization.

1. INTRODUCTION

Complex industrial manufacturing systems are characterized by intricate networks of interconnected processes, machines, and human interactions (Mbohwa, 2020; Train & Salehin, 2023). Their complexity arises from factors such as redundancy, scale, desynchronization, environmental variability, and human error, as highlighted in the context of system reliability (Zhou et al., 2021). This inherent complexity leads to the generation of large amounts of stochastic data, making the prediction and optimization of key performance indicators (KPIs) such as overall equipment effectiveness (OEE) a significant challenge for decision making, resource allocation, and efficiency improvement (Mbohwa, 2020; Lee et al., 2024). Traditional methods, including statistical process control (SPC) charts (Montgomery, 2020) (effective for monitoring process stability but less adept at modeling complex interactions), time series analysis (e.g., ARIMA, exponential smoothing (Hyndman & Athanasopoulos, 2018) - useful for predictions based on past trends but often dependent on linear assumptions and struggling with multivariate stochastic dependencies (Wen et al, 2023)), rule-based expert systems (limitations discussed in Ignatov, 2023), and even some earlier machine learning techniques such as shallow artificial neural networks (ANNs) and support vector machines (SVMs) (limitations in handling complex stochastic data contrasted with GAN capabilities (Creswell et al., 2018)) often rely on deterministic or linear assumptions or lack the ability to fully capture the complex, nonlinear, and stochastic dependencies present in these complex systems (Lee et al., 2024). These methods may not adequately learn and model the intricate probabilistic relationships and evolving stochastic trends that significantly impact OEE in modern industrial environments (Wen et al., 2023). As a result, there is a significant need for novel approaches that can effectively learn and simulate the stochastic behavior of these systems to enable proactive optimization and informed decision making.

To address this critical gap, this paper proposes a novel approach using generative doppelgangers, a specific application of Generative Adversarial Networks (GANs) (Creswell et al., 2018), to model and simulate the operational behavior of complex industrial systems. While other machine learning techniques such as XGBoost and Random Forest excel at predictive tasks by learning direct mappings from input to output features (e.g., Breiman, 2001; Chen & Guestrin, 2016), they often struggle to inherently model the underlying stochastic processes and complex joint distributions of multivariate time-series data that characterize industrial systems. Deep learning time series models such as temporal convolutional networks (TCNs) and transformers have shown promise in capturing temporal dependencies (e.g., Vaswani et al., 2017; Bai et al., 2018), but their primary focus is often on predicting future values rather than generating synthetic yet realistic sequences that preserve the complex stochastic properties of the original data (Ren et al., 2024).

Generative Adversarial Networks, particularly the doppelganger model (Lin et al., 2020), offer distinct advantages in this context. Unlike discriminative models, GANs learn the underlying probability distribution of the real data (Goodfellow et al., 2014), enabling them to generate synthetic data that closely mimics the complex, nonlinear, and stochastic patterns observed in industrial operations (Yoon et al., 2019). This generative capability is critical for conducting realistic what-if analyses and exploring potential future scenarios beyond simple predictions. The doppelganger architecture, with its ability to jointly model both the temporal dynamics and the interdependencies between different operational parameters, enables the creation of "behavioral doppelgangers" that can simulate the nuanced impact of different factors on OEE components such as availability, performance, and quality (Lin et al., 2020).

Unlike traditional digital twins, which often focus on the static physical model, the behavioral doppelganger aims to decipher the hidden patterns and stochastic nuances in the system's operational data. The application of generative doppelgangers for holistic simulation of stochastic operational behavior to specifically enable proactive OEE optimization through scenario analysis in complex industrial systems represents a novel contribution to the field (potential highlighted in Zhang et al., 2023). This framework offers the potential to overcome the limitations of deterministic and linear models, as well as the limited stochastic modeling capabilities of some previous machine learning methods, by learning and replicating the complex relationships between process parameters and their impact on OEE, thereby enabling proactive what-if analyses and targeted optimization strategies.

2. LITERATURE REVIEW

Generative Artificial Intelligence has been used extensively recently to solve some specific problems in many engineering domains, including

- Image Generation and Synthesis: GANs have been widely used to generate realistic images. They can be used to generate novel artistic images, to create new images that resemble a particular data set, or even to perform image-to-image translation tasks, such as converting images from one domain to another (e.g., transforming a horse image into a zebra image). One use of GANs is to generate additional samples to improve the training data. This is particularly helpful when the dataset is limited, as it improves the generalization power of machine learning models (Kuntalp & Düzyel, 2024; Wang & Yan, 2024; Vdoviak & Giedra, 2024).
- Video Generation: By sequentially generating each frame, GANs have been extended to produce realistic videos. This is applicable to video manipulation, video synthesis, and video prediction (Wang & Yan, 2024; Branytskyi et al., 2022).
- Text generation: GANs have the potential to produce coherent and realistic text. They have been implemented in various applications, including the production of product reviews, dialog, and language translation (Bahrum et al., 2024; Ren et al., 2024; Saiz et al., 2021).
- Style transfer: GANs have the ability to learn the style of a particular image or artwork and apply it to another image, thereby achieving style transfer. This method has been used for image modification, artistic representation, and the development of personalized filters (Ekman & Friesen, 1971; Calix et al., 2024; Zhou et al., 2021).
- Medical image analysis: GANs have been implemented in medical imaging to generate synthetic images that correspond to actual patient data. Potential benefits include providing additional training data for medical image analysis tasks and addressing privacy concerns (Makhlouf et al., 2023; Fukaya et al., 2023).

- Game development: GANs have been used in game development to generate levels, characters, and graphics, among other game elements. They can also be used to create non-player characters (NPCs) with human-like behavior (Fukaya et al., 2023).
- Data Anonymization: GANs have been explored as a means of privacy-preserving data distribution. They have the ability to generate synthetic data that preserves the statistical properties of the original data while protecting sensitive information (Hellmann et al., 2024).
- In addition, the use of Generative Adversarial Networks (GANs) to predict industrial process failures has the potential to provide numerous benefits:
- Increased accuracy: GANs are capable of learning complex patterns and producing real samples. By training GANs on historical data from industrial processes, it is possible to understand the fundamental patterns and correlations that contribute to failure. This can lead to more accurate prediction of failure than traditional methods (Jiang et al., 2019; Hu et al., 2023; Zhang et al., 2023).
- Early Detection: GANs have the potential to detect minute changes in data patterns that may indicate an impending failure.
- The process data can be continuously monitored and compared to the learned patterns by GANs, which can minimize disturbances and enable proactive maintenance. This allows for early warnings (Chung et al., 2024; Mumbelli et al., 2023; Kusiak, 2020).
- Anomaly detection: By contrasting the produced samples with the original data, GANs can be used to find anomalies in industrial processes. To identify a possible error or anomalous behavior, any deviation from the assimilated patterns is labeled as an anomaly (Jiang et al., 2019; Zhao et al., 2019; Kumarage et al., 2019).
- Cost savings: GANs can be used to streamline maintenance schedules, minimize unplanned downtime, and avoid costly repairs by predicting problems. Process owners may be able to save a lot of money and time by making their tools work better and reducing production costs (Zhao et al., 2019). It's important to note that using GANs to predict failure in an industrial process requires data collection, data preparation, and model training. Understanding the domain and collaboration between data analysts and domain experts are also very important for successful implementation (Fu et al., 2023).
- Anomaly detection: GANs can be used to identify anomalies in industrial processes by comparing the generated samples with the original data. Any deviation from previously acquired patterns can be flagged as an anomaly to highlight a likely error or anomalous behavior (Chung et al., 2024; Farady et al., 2023; Khan et al., 2023).
- The development of predictive maintenance strategies is greatly enhanced by the implementation of failure prediction, which allows organizations to anticipate and prevent equipment failures in advance. Using sophisticated analytics and machine learning algorithms, predictive models can analyze historical data, sensor readings and other relevant factors to predict the likelihood of a specific component or system failing (Zhang et al., 2023; Rezaei et al., 2024; Qian et al., 2022).
- Increased Equipment Availability: Organizations can use failure prediction to anticipate potential problems and implement preventive measures. Organizations can improve the availability and reliability of critical assets by taking proactive measures to address these issues, resulting in a reduction in unexpected equipment failures (Goodfellow et al., 2014; Song et al., 2023). Optimized maintenance planning: Failure prediction allows organizations to plan maintenance activities based on real-time needs rather than predetermined schedules, providing valuable information about the health and condition of equipment. This is a condition-based approach that optimizes maintenance planning, minimizes redundant maintenance tasks, and maximizes the use of maintenance resources (Yuan et al., 2020).
- Predictive maintenance: Researchers at the University of Michigan have used GAN-based doppelgangers to generate synthetic sensor data representing different stages of machine degradation. By comparing real-time sensor data to these synthetic patterns, the remaining useful life of machines can be estimated, enabling proactive maintenance and reducing unplanned downtime (Salierno et al., 2024; Ntavelis et al., 2020).
- Quality control: Doppelgangers have been explored to augment data sets for quality control in manufacturing processes. By generating synthetic images of defective products, they can improve the training of machine learning models for defect detection, potentially leading to increased accuracy and efficiency in quality control systems (Kuntalp & Düzyel, 2024; Radford et al., 2015).

– Power Grid Management: Researchers have explored the use of GANs to model the complex behavior of power grids, including load fluctuations and renewable energy generation. These models can help predict fluctuations, optimize power distribution, and identify potential vulnerabilities in the grid (Hobbie & Lieberwirth, 2024; Antonucci et al., 2024). Enhanced safety: Unexpected equipment failures can put personnel and assets at risk of injury. In an effort to mitigate safety hazards, failure prediction is used to identify potential failure modes and implement corrective actions. Organizations can create a safer workplace for their employees by identifying potential failures early. Improved assessment: Expected failures provide insightful analysis of equipment performance and deterioration trends. The data presented helps companies make informed decisions about overall asset management, equipment replacement, maintenance schedules and spare parts inventory. By identifying potential failures in advance, companies can either prevent or mitigate the effects of failures, thereby extending the life of equipment. This proactive strategy helps to extend the life of equipment, thereby reducing the need for early replacement and generating significant cost savings (Zhao et al., 2019; Luo et al., 2021).

3. MATERIALS AND METHODS

3.1. Classical generative GAN and doppelgangers

A notable subfield of machine learning is the generative adversarial network (GAN) (Goodfellow et al., 2014). These models consist of two competing neural networks: a generator and a discriminator. To maximize realism, the generator is designed to create unique data samples, including text and photographs (Sun, 2024).

On the other hand, it is the discriminator's job to distinguish these artificial samples from real data. This back and forth between challenge and response allows the generator to get better at producing plausible imitations and the discriminator to get better at detecting them. The two networks are trained in competition to improve the generator's ability to produce realistic samples and the discriminator's ability to correctly identify them. Because GANs can generate diverse and realistic data, they have been used in many different fields.

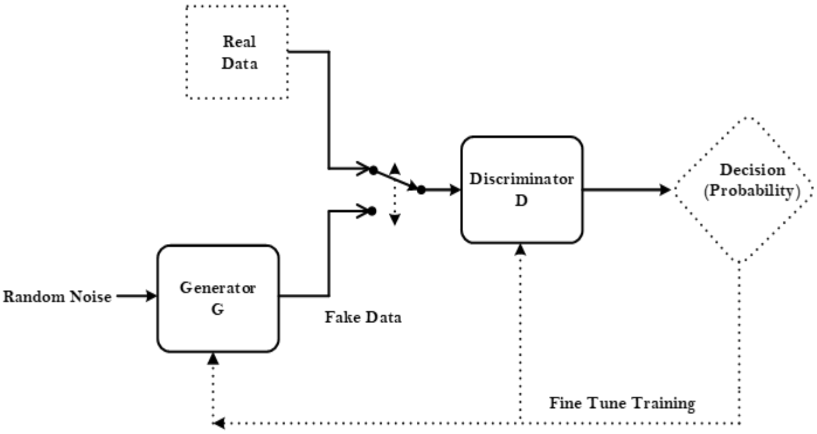


Fig. 1. Classical GAN principle (source:adapted from Air Liquid)

In this architecture, Z represents the noise to be fed into the generator, G(Z) the dummy data, which is the output of the generator, x is the real data. When fake data is fed into the discriminator, the result is $D(G(z))$ is the probability that the dummy sample is real. Similarly, when real data is fed into the discriminator, the output D(x) represents the probability that the real sample is real, as estimated by the discriminator.

In the adversarial principle of neural networks,the expectation of the logarithm of the discriminator output on the input of real data is given by:

$$\mathrm{E}_x \sim pdata(x)\left[\log D(x)\right]. \tag{1}$$

Likewise,

$$\mathrm{E}_z \sim p_z(z)\big[\log(1-D(G(z)))\big]. \tag{2}$$

This expectation is the average of the prediction of the input noise at the generator. For the discriminator, $D(G(z))$ should be minimized, while it should be maximized by the generator. Similarly, the discriminator wants to maximize $\log D(x)$. So, $1-D(G(z))$ is the set the discriminator wants to maximize. The final expectation should be the set the discriminator wants to maximize and a set the generator wants to minimize. So we have

$$\mathrm{E}_x \sim pdata(x)\big[\log D(x)\big] + \mathrm{E}_z \sim p_z(z)\big[\log(1-D(G(z)))\big] \tag{3}$$

Definitely, the optimisation objective function is given by:

$$\min_G \quad \max_D \quad V(D,G) = \mathrm{E}_x \sim pdata(x)\big[\log D(x)\big] + \mathrm{E}_z \sim p_z(z)\big[\log(1-D(G(z)))\big] \tag{4}$$

Classical Generative Adversarial Networks (GANs) present various issues. Some of the frequent issues are:

1. Mode collapse: GANs may provide limited performance, unable to capture the full range of training data (Salimans et al., 2016).
2. GANs require a large amount of data for training in order to identify meaningful patterns and produce high quality results. The first major problem with GANs using recurrent neural networks (RNNs) as generators is their inability to capture long-term correlations, often attributed to the vanishing or exploding gradient problem (Bengio et al., 1994; Figueira & Vaz, 2022; Dash et al., 2023; Alqahtani et al., 2021).

To address this issue, the generator component of the Doppelganger model incorporates Long Short-Term Memory (LSTM) networks to effectively capture the temporal dependencies inherent in industrial operational data. LSTMs were chosen because of their well-established ability to model long-range dependencies in sequential data, which is critical for understanding the evolving patterns and stochastic trends that affect OEE (Hochreiter & Schmidhuber, 1997; Gers et al., 2000). Unlike traditional recurrent neural networks (RNNs), which suffer from the vanishing gradient problem (Bengio et al., 1994), LSTMs use a gating mechanism that allows them to selectively learn and retain information over extended time steps. This capability is particularly important in our context, where the effects of past operating conditions can persist and influence future system behavior and potential failures.

While Gated Recurrent Units (GRUs) offer a more streamlined architecture with fewer parameters and can sometimes achieve comparable performance to LSTMs in capturing temporal dependencies (Cho et al., 2014), LSTMs have been shown to be effective in scenarios with very long sequences and complex temporal patterns. Temporal convolutional networks (TCNs) provide an alternative approach using convolutions, but may require very deep networks for long dependencies (Bai et al., 2018). Attention-based mechanisms, while powerful for long dependencies (Vaswani et al., 2017), could introduce more complexity. Given the proven track record of LSTMs in generative modeling of sequential data within GAN frameworks (Yoon et al., 2019; Mogren, 2016) and their suitability for capturing the temporal complexity of industrial data, they were selected for the Doppelganger generator.

The Doppelganger architecture also makes use of mini-batch training, a process in which a small number of data points are fed at once. This helps to improve the efficiency of the model by reducing expensive computation by splitting the data into smaller batches, and optimizes the model by iteratively adjusting the model's internal parameters to minimize errors, leading to faster convergence and potentially optimal solutions. This is a very critical criterion in retrieving relevant failures when studying or analyzing complex industrial systems. Another relevant feature of Doppelganger is that it not only generates synthetic data based on the real data, but also jointly generates attributes (metadata) for these data (Lin et al., 2020). This is important to provide context to the data. For example, the context of analyzing data from critical components is not the same as that of tolerant components. In addition, poorly categorized data leads to inaccurate insights. Finally, it helps analyze trends, optimize processes, and make relevant, data-driven decisions when analyzing the reliability components of complex industrial systems.

3.1.1. Model training

Hyperparameter Specification: The Doppelganger model, which uses a Generative Adversarial Network (GAN) with Long Short-Term Memory (LSTM) layers for sequential data modeling, was trained with the following key hyperparameters: Generator LSTM layers: 2, Generator hidden units per LSTM layer: 128, discriminator LSTM layers: 2, discriminator hidden units per LSTM layer: 128, learning rate (generator): 0.0002, Learning Rate (Discriminator): 0.0002, batch size: 64, training epochs: 400, and optimizer: Adam(β1=0.9, β2=0.994).

GPU Optimization: Our experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU. While this setup allowed us to achieve the reported results, we acknowledge that further performance improvements may be possible through the implementation of GPU optimization techniques. Strategies such as exploiting the power of two hidden units or exploring model parallelization could potentially lead to more efficient training, especially when working with our dataset of 911 samples, each with 7 features. These optimization avenues represent interesting directions for future research.

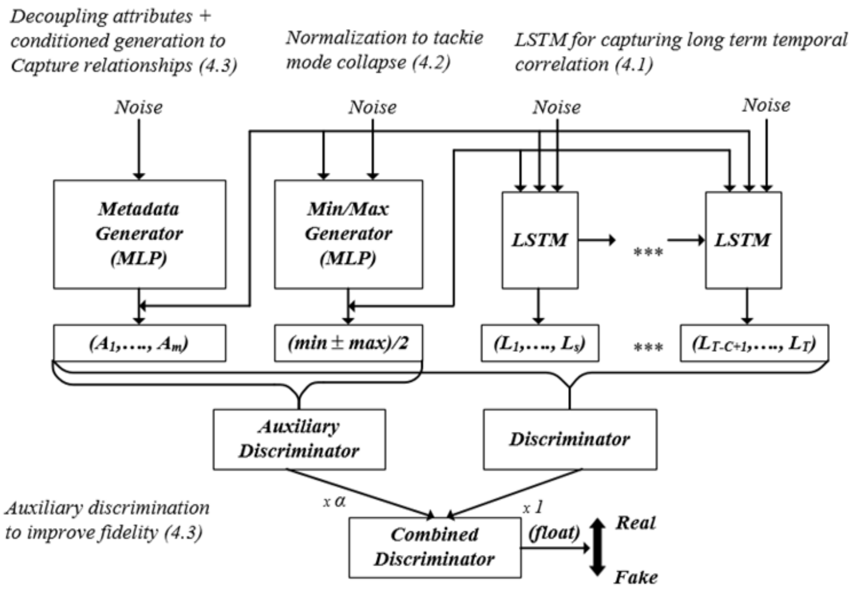

Decoupling attributes + conditioned generation to Capture relationships (4.3)
Normalization to tackie mode collapse (4.2)
LSTM for capturing long term temporal correlation (4.1)
Noise
Noise
Noise
Noise
Metadata Generator (MLP)
Min/Max Generator (MLP)
LSTM
***
LSTM
$(A_1, \ldots, A_m)$
$(\min \pm \max)/2$
$(L_1, \ldots, L_s)$
***
$(L_{T-C+1}, \ldots, L_T)$
Auxiliary Discriminator
Discriminator
Auxiliary discrimination to improve fidelity (4.3)
x α
x 1
Combined Discriminator
(float)
Real
Fake

Fig. 2. The proposed doppelgänger architecture

The implementation of the doppelganger architecture obey the following Algorithm:

1. **Initialization**

Define network architectures

- Metadata Generator (MLP): Takes random noise and metadata as input and generates attributes.
- Min/Max Generator (MLP): Takes random noise as input and generates min/max values for normalization.
- LSTM(s): LSTM network(s) to capture long-term temporal correlations.
- Discriminator: Takes generated time series (and possibly metadata) as input and discriminates between real and fake.
- Auxiliary Discriminator: Improves accuracy by focusing on specific aspects of the generated data.
- Initialize Network Weights: Randomly initialize all network weights.

Define Loss Functions:

2. **Training:**

- For each training iteration do:
  - For K steps do: (train discriminator)
    - Sample real data:
      - Sample a batch of real-time series data and its associated metadata.
    - Generate fake data:
      - Sample noise vectors.

```
                    - Generate attributes using the Metadata Generator (MLP).
                    - Generate min/max values using the Min/Max Generator (MLP).
                    - Generate fake time series using the LSTM(s), conditioned on the generated
                      attributes and normalized using the min/max values.
               ❖ Update discriminator
                    - Compute discriminator loss on real and fake data:

                      $\nabla_{\theta_\alpha} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right]$

                    - Compute auxiliary discriminator loss (If applicable).
                    - Update discriminator weights by ascending its gradient (maximize loss_D).
          • End for
          • Sample fake data:
               ❖ Sample noise vectors.
               ❖ Generate attributes using the Metadata Generator (MLP).
               ❖ Generate min/max values using the Min/Max Generator (MLP).
               ❖ Generate fake time series using the LSTM(s), conditioned on the generated
                 attributes and normalized using the min/max values.
                    • Update Generator:
                         ❖ Compute generator loss based on discriminator output (and
                           potentially auxiliary discriminator output).

                           $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[ \log(1 - D(G(z^{(i)}))) \right]$

                         ❖ Update generator weights (including Metadata Generator, Min/Max
                           Generator, and LSTMs) by descending their gradient (minimize
                           loss_G).
     ▪ End for

  3. Generation
     ▪ Provide desired metadata
     ▪ Generate time series using the trained Metadata Generator, Min/Max Generator, and LSTMs
```

Unlike classical GANs, the generator in DOPPELGANGERS takes metadata as input, allowing controlled generation of time series based on specific attributes or conditions; through normalization, the min/max generator provides normalization to combat mode collapse, ensuring that the generated data has a diverse range of values; the Auxiliary Discriminator enhances the training process by focusing on specific aspects of the generated data, potentially leading to improved fidelity; the use of multiple LSTMs allows the model to capture different levels or aspects of temporal dependencies in the data.

3.2. Complex industrial system description

The industrial system studied in this case is a complex sugar factory from the Republic of Gabon, SUCAF, a medium size factory with a production figure of 24320 tons of sugar per year and employing 1152 people at its peak.It is made up of 60 main pieces of equipment active in the process: The workflow of the process is made up of seven stages:

Reception and cleaning subsystem:

Made up of several structures, including:

- The beam, which is an electromechanical structure (conveyors, motors, sensors).
- The weighing scales: Electromechanical structure (load cells, digital display).
- The cane cutters: Mechanical (knives, rotating drums).
- Washers: Mechanical structure (rotating drums, water sprays).
- Stone and garbage removers: Mechanical structure (screens, sieves).

- Magnetic separators: Electromechanical (magnets, conveyors).

Extraction subsystem:

Consists of multiple devices:

- The milling tandem : It is an electromechanical structure (rollers, motors, gears).
- The bagasse diffuser : This is a mechanical structure (conveyor, diffusers).
- The diffusion towers/continuous diffusers (sugar beets):Consisting of mechanical components Mechanical (towers, conveyors, pumps).

Clarification subsystem: Consists of lime mixers and dosing system, is an electromechanical structure consisting of pumps and mixers.

The evaporation subsystem: consisting of multiple-effect evaporators: Mechanical (heat exchangers, vessels).

The crystallization subsystem:

Consists of:

- Vacuum pans / crystallizers: Mechanical (vessels, heating coils).
- Seed slurry preparation tanks: Mechanical (tanks, mixers).

Magma Pumps: Electromechanical (pumps, motors).

The centrifugation subsystem :

- The batch centrifuges is an electromechanical structure composed of: centrifuge bowls, motors, control components.
- The molasses pumps: consisting of pumps, motors is an electromechanical system.

The drying and packaging subsystem: Consisting of rotary sugar dryers, sugar coolers, screening/grading equipment, bagging/bulk loading systems.

The following pictures respectively Fig. 3. and Fig. 4. represent the sugar production process and the real pictures of the industrial sugar plant.

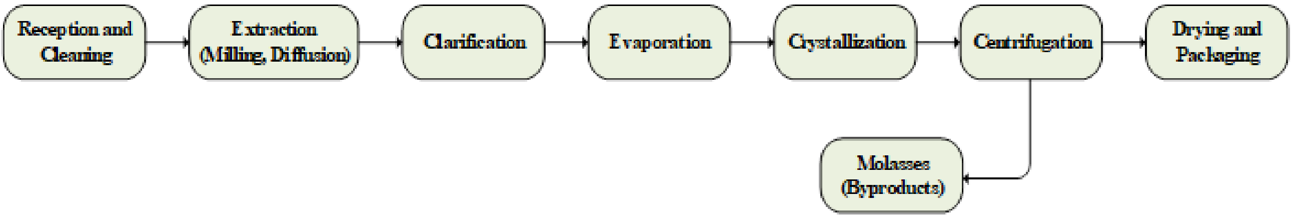


Fig. 3. The sugar production process

Fig. 4. Real images of the industrial sugar plant

3.3. Data acquisition and preprocessing

The operational data used in this study were collected from an industrial sugar plant over a period from April 2020 to October 2022. The data collection in this study is based on a production time approach. This method recognizes that different pieces of equipment may operate for different durations. However, it's important to note that this approach does not explicitly account for the interdependencies between different pieces of equipment in the initial data collection and metric calculation.

Key Performance Indicators (KPIs) related to Overall Equipment Effectiveness (OEE) were derived from the collected data. The following KPIs were calculated:

- Overall Mean Time Between Failures (MTBF): Defined as the total production time across all considered equipment divided by the total number of failures across the same equipment.
- Overall Mean Time To Repair (MTTR): Defined as the total downtime due to failures across all considered equipment divided by the total number of failures across the same equipment.
- Performance Rate, Quality Rate and Operational Availability.

Measurements for the underlying data points used to calculate these metrics (e.g., production time, failure timestamps, downtime duration, good/total counts) were logged daily.

3.3.1. Equipment representation

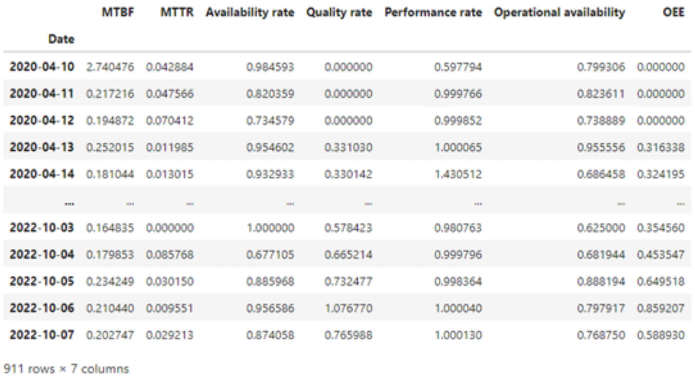
This study focused on key equipment critical to the sugar production process. The selection of these key pieces of equipment was based on a preliminary importance factor analysis that assessed their impact on overall production output and their potential to contribute to OEE losses. While the data set includes various pieces of equipment within the plant, the analysis and modeling efforts were focused on those identified as having the most significant impact on plant performance. It is important to note that due to the inherent structure of the production line and the varying criticality of different pieces of equipment, not all equipment types are equally represented in the dataset in terms of the number of data points or failure events. However, the selected key equipment, which is critical to core operations, has a substantial representation to allow for meaningful analysis. Future research could explore the inclusion of a more comprehensive range of equipment and their interdependencies.

3.3.2. Data preprocessing

Prior to model training, the raw operational data underwent a series of preprocessing steps to ensure data quality and suitability for the generative doppelganger model.

- Outlier handling: Statistical methods were used to identify and handle outliers in key operational parameters. Extreme values that deviated significantly from the typical range, possibly due to recording errors or unusual operational events, were addressed. The primary approach to outlier management was to remove data points identified as extreme outliers based on interquartile range (IQR) analysis.
- Missing data imputation: Instances of missing data were observed throughout the dataset. To mitigate the impact of these gaps, a nearest neighbor imputation technique was applied. This method replaced missing values with the values of temporally adjacent data points, assuming some degree of temporal consistency in the operational data. The extent of imputation was relatively limited to ensure the integrity of the original data patterns.
- Data Normalization: To ensure that all input features contribute equally to the model training process and to improve the convergence of the neural networks, min-max scaling was applied to normalize the data in the range of 0 to 1.

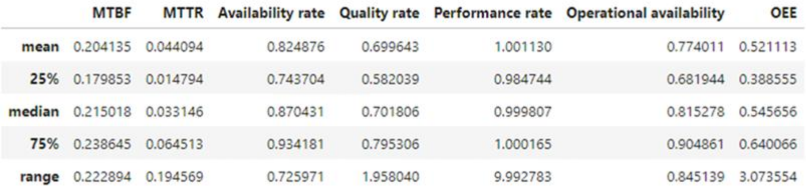
These preprocessing steps were critical in preparing the industrial operational data for effective training of the generative doppelganger model, enabling it to more accurately learn the underlying patterns and stochastic behaviors.



Date	MTBF	MTTR	Availability rate	Quality rate	Performance rate	Operational availability	OEE
2020-04-10	2.740476	0.042884	0.984593	0.000000	0.597794	0.799306	0.000000
2020-04-11	0.217216	0.047566	0.820359	0.000000	0.999766	0.823611	0.000000
2020-04-12	0.194872	0.070412	0.734579	0.000000	0.999852	0.738889	0.000000
2020-04-13	0.252015	0.011985	0.954602	0.331030	1.000065	0.955556	0.316338
2020-04-14	0.181044	0.013015	0.932933	0.330142	1.430512	0.686458	0.324195
...	...	...	...	...	...	...	...
2022-10-03	0.164835	0.000000	1.000000	0.578423	0.980763	0.625000	0.354560
2022-10-04	0.179853	0.085768	0.677105	0.665214	0.999796	0.681944	0.453547
2022-10-05	0.234249	0.030150	0.885968	0.732477	0.998364	0.888194	0.649518
2022-10-06	0.210440	0.009551	0.956586	1.076770	1.000040	0.797917	0.859207
2022-10-07	0.202747	0.029213	0.874058	0.765988	1.000130	0.768750	0.588930

911 rows × 7 columns

Fig. 5a. Dataset overview



	MTBF	MTTR	Availability rate	Quality rate	Performance rate	Operational availability	OEE
mean	0.204135	0.044094	0.824876	0.699643	1.001130	0.774011	0.521113
25%	0.179853	0.014794	0.743704	0.582039	0.984744	0.681944	0.388555
median	0.215018	0.033146	0.870431	0.701806	0.999807	0.815278	0.545656
75%	0.238645	0.064513	0.934181	0.795306	1.000165	0.904861	0.640066
range	0.222894	0.194569	0.725971	1.958040	9.992783	0.845139	3.073554

Fig. 5b. Brief statistic of the dataset

3.4. Data description

3.4.1. MTBF, MTTR

(1) MTBF (Mean Time Between Failures):

The mean MTBF is 0.204, which is relatively low. The median (0.215) is slightly higher than the mean, indicating some skewness in the data distribution, with some longer periods between failures, and a range of 0.223 indicates variability in system reliability.

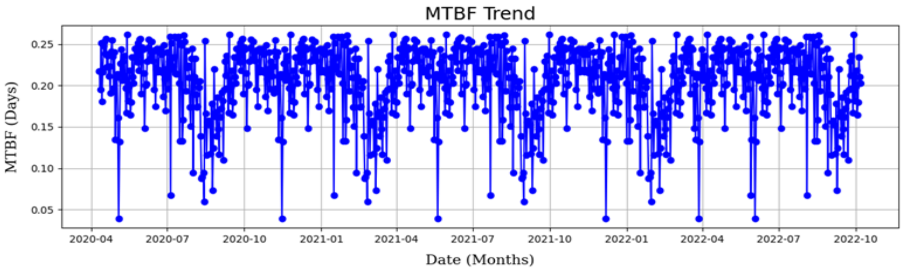


Fig. 6. Trends of the MTBF over time

(2) MTTR (Mean Time To Repair):

The average MTTR is 0.044, which is quite good. This means that when outages do occur, they are generally repaired quickly. However, the range of 0.195 indicates some variation in repair times, which could be due to the nature of the failures or the availability of maintenance resources.

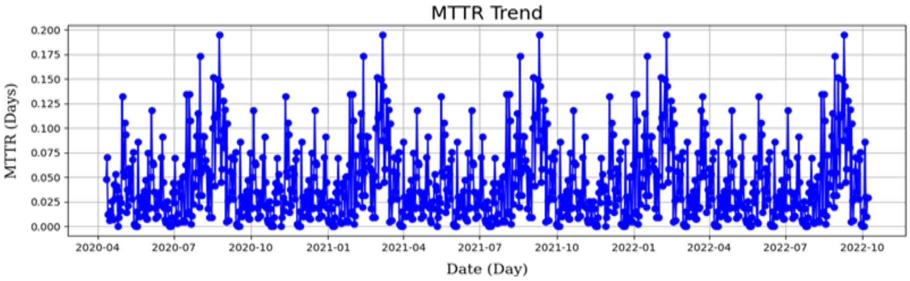


Fig. 7. Trends of the MTTR over time

Overall, there is no clear upward or downward trend in MTTR and MTBF over the period studied. This could mean that efforts to improve repair processes have had mixed results, and that the complexity of failures varies, resulting in fluctuating repair times.

3.4.2. Operational Availability, performance, quality, and OEE

3.4.2.1. The operational availability

In such a complex industrial system, producing the presented graph (Fig. 8) highlights many relevant points of the operating system, such as breakdowns, failures, and the pattern trend of system availability for prediction.

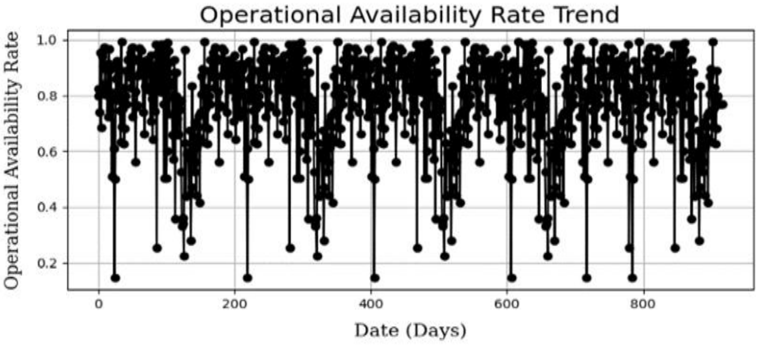


Fig. 8. Operational availability trend over time

There is a high Variability of Operational Availability, which varies significantly from lows around 0.2 to highs near 1.0. It indicates frequent disruptions, process bottlenecks, equipment reliability inefficiencies, aging equipment, inadequate maintenance, or improper operation issues that result in reduced sugar production output. In addition, the metric frequently drops to lower levels, indicating recurring problems that impact production. These could be equipment failures, critical maintenance issues, or other unforeseen events. Most importantly, there is no discernible upward or downward trend over time. This suggests that there are stochastic issues affecting availability that haven't been systematically addressed, or that new challenges are constantly emerging.

3.4.2.2. Performance

Performance is a measure of how efficiently the equipment runs when it's actually producing.

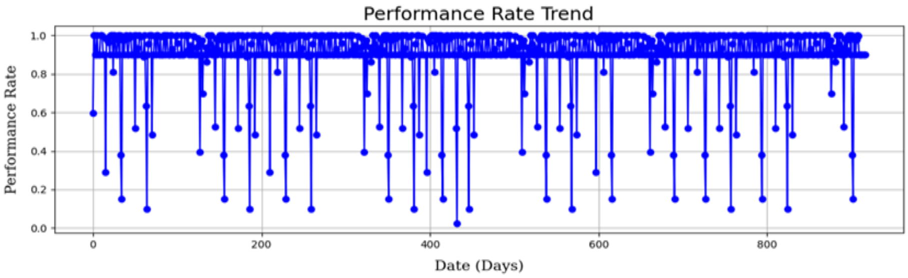


Fig. 9. Performance rate trend over time

The performance rate trend graph in Figure 9 shows a high initial performance followed by a decline**:** The performance starts very high, close to 1.0, but then experiences a sharp drop. It then stabilizes at a lower level, fluctuating between 0.2 and 0.4 for a significant period of time. This indicates an initial period of optimal operation followed by a significant drop in performance. The initial drop and subsequent low performance indicate potential problems with equipment reliability or process stability. However, there is a persistent low value.The extended period of low performance indicates persistent problems affecting system efficiency. This could be due to equipment malfunction, process inefficiencies, or other operational challenges. Alternatively, we may see sporadic increases in performance rate, which may indicate temporary improvements or successful interventions. However, these gains are not sustained, indicating that the underlying problems haven't been fully resolved.

In this context, the prolonged period of low performance is likely to result in lower sugar production. This could result in financial losses and impact the ability to meet market demands. In summary, the performance rate trend highlights a significant challenge in maintaining optimal production levels within the sugar manufacturing system.

3.4.2.3. Quality

The quality rate, in the context of an industrial manufacturing plant , refers to the percentage of sugar production that meets the desired specifications or standards.

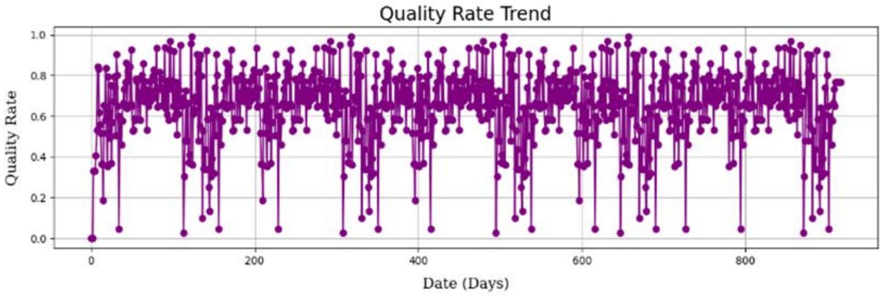


Fig. 10. Quality rate trend over time

The quality rate shows significant variation, ranging from nearly 0.0 to 1.0. This clearly indicates inconsistency in the production process, resulting in varying levels of product quality. The high variability and frequent setbacks indicate a lack of control over critical process parameters. This can lead to product recalls, customer dissatisfaction, and financial loss. In addition, there are frequent dips, indicating raw material fluctuations, equipment malfunctions, process instability, or human error. Similarly, Operational Availability shows no discernible upward or downward trend, indicating stochasticity. This clearly indicates that the underlying issues affecting quality have not been systematically addressed or that new challenges are still emerging. Overall, the trend in product plant quality rates indicates a significant challenge in maintaining consistent sugar quality.

3.4.2.4. Overall equipment effectiveness

Composed of operational availability, performance rate, and quality rate, OEE is a key performance indicator, a metric that measures how well a manufacturing operation is being utilized compared to its full potential. It can be used to track performance trends, identify areas for improvement, and make informed decisions regarding maintenance, process optimization, and resource allocation, as well as to improve quality control measures to minimize the production of defective products. It also helps identify and eliminate bottlenecks or inefficiencies in the production process.

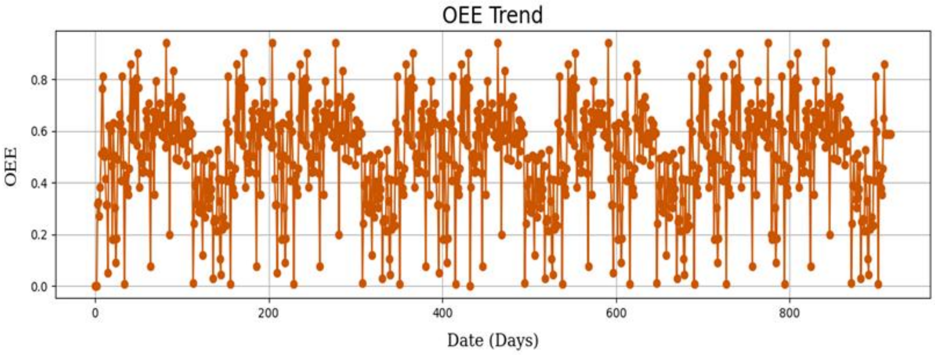


Fig. 11. OEE trend over time

As with the previous metrics, there is significant variability. The OEE values show significant fluctuations, ranging from close to 0 to around 0.85, indicating suboptimal utilization and potential losses. There's also no clear upward or downward trend. This indicates that the underlying factors influencing OEE haven't been systematically identified or that new challenges are emerging.

4. PROPOSED METHODOLOGY ANALYSIS USING DOPPELGANGERS

The methodology consists of a cyclical process in which a duplicate is used to generate synthetic data that mimics the behavior of a real industrial system. This synthetic data can then be used to perform in-depth analyses of various KPIs, ultimately leading to the identification of potential areas for improvement within the industrial system. The insights gained from these analyses can then be used to refine and enhance the doppelganger model, creating a continuous feedback loop that drives continuous optimization of the model.

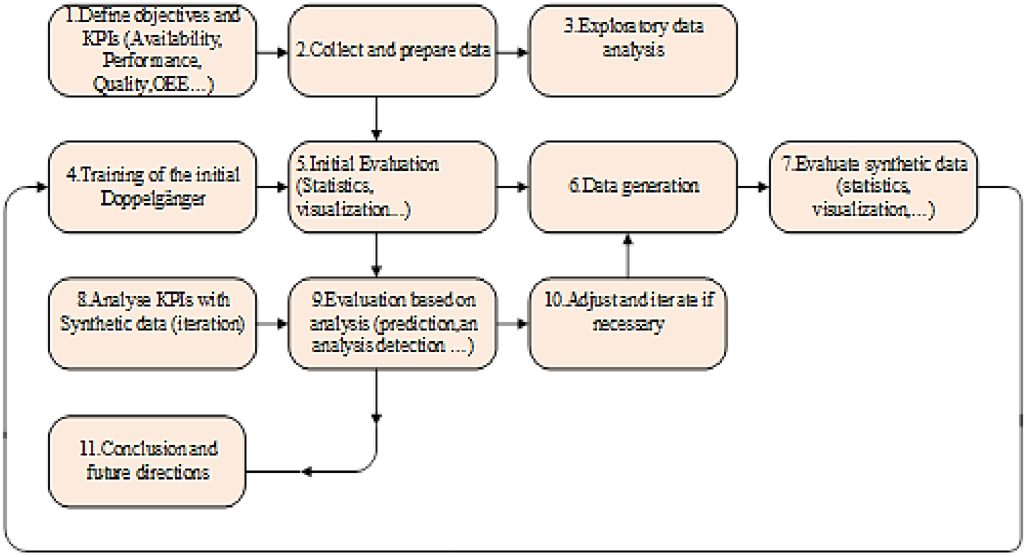

1.Define objectives and KPIs (Availability, Performance, Quality,OEE…)
2.Collect and prepare data
3.Exploratory data analysis
4.Training of the initial Doppelgänger
5.Initial Evaluation (Statistics, visualization…)
6.Data generation
7.Evaluate synthetic data (statistics, visualization,…)
8.Analyse KPIs with Synthetic data (iteration)
9.Evaluation based on analysis (prediction,an analysis detection …)
10.Adjust and iterate if necessary
11.Conclusion and future directions

Fig. 12. Doppelgänger-based KPI analysis framework

1. Real system and data collection

The process starts with the real industrial system, from which operational data is collected. This data serves as the basis for training the duplicate model and provides the basis for comparison and validation later in the process.

2. Doppelganger Model and Training

The collected real-world data is then used to train a doppelganger model. The goal is to create a digital replica capable of generating synthetic data that reflects the statistical properties and behavioral patterns observed in the real-world system.

3. Synthetic data generation

Once trained, the doppelganger model is used to generate synthetic data. Although the data set is artificial, it should closely resemble the real data in terms of statistical distributions and underlying trends.

4. KPI analysis

Both the real data and the synthetic data generated by the doppelganger are subjected to KPI analysis. This involves calculating and evaluating various performance metrics relevant to the industrial system under consideration.

5. Comparison and validation

The KPI values obtained from the real and synthetic data are then compared. This comparison helps validate the accuracy and effectiveness of the doppelganger model. If the model is performing well, the KPIs calculated from the synthetic data should closely match those calculated from the real data.

6. Identification of improvement areas

The insights gained from the KPI analysis are then used to address areas within the industrial system where performance improvements can be realized. This may include bottlenecks, inefficiencies, or opportunities for optimization.

7. Feedback and improvement

Identified areas of improvement are then fed back into the process, informing refinements to both the real system and the doppelganger model. This creates a continuous cycle of improvement in which the doppelganger becomes increasingly accurate and valuable in facilitating complex system optimization.

5. RESULTS, INTERPRETATION AND VALIDATION

This section presents the results, interpretation, and validation of the proposed methodology for handling stochastic data trends in complex manufacturing systems.

5.1. Operational availability

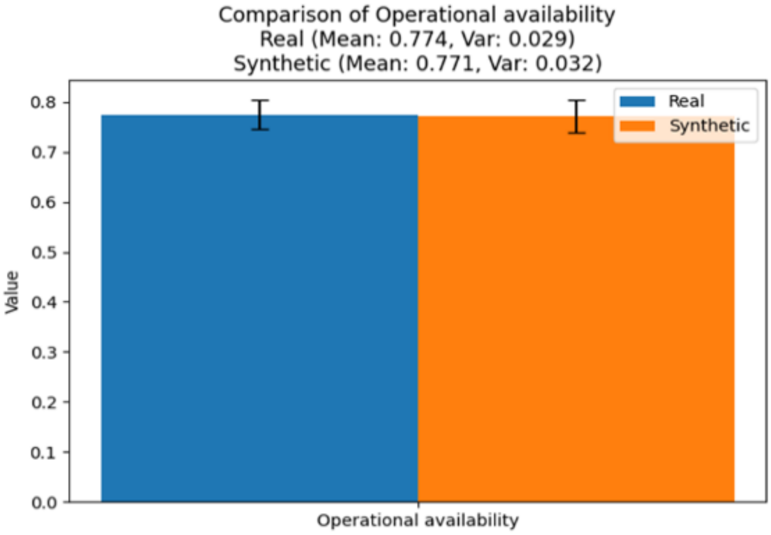


Fig. 13. Statistics comparison of operational availability (Real vs. Synthetic)

From the chart, we can observe the statistics of the Real Operational Availability metrics as follows: Mean: 0.774, Variance: 0.029. This means that, on average, 's actual sugar plant is operational and ready to produce about 77.4% of the time. The variance of 0.029 suggests that there is some variation in this availability. Similarly, the statistics of the same metric generated by the doppelganger show Synthetic Operational Availability Mean: 0.771, Variance: 0.032.It is very close to the real availability, indicating that the model does a good job of capturing the real-world dynamics of the plant. The slightly higher variance in the synthetic data suggests that the model may be predicting a bit more variability in availability than what's actually observed. In addition, the close agreement between the real and synthetic data is a strong indicator that the doppelganger model is accurately representing the real-world system. This gives us confidence in the model's ability to simulate and predict plant behavior. In addition, the model can potentially be used to predict how changes to the system (e.g., new equipment, different maintenance schedules) might affect uptime, to help identify equipment that's more prone to failure or predict when maintenance is needed, improving overall uptime; by simulating different operating conditions or process changes, the model could help identify ways to further increase uptime and improve production efficiency.

Beyond the similarities in means and variances, the autocorrelations of both the real and synthetic data are very low, indicating that the two data sets don't depend strongly on their respective historical values.

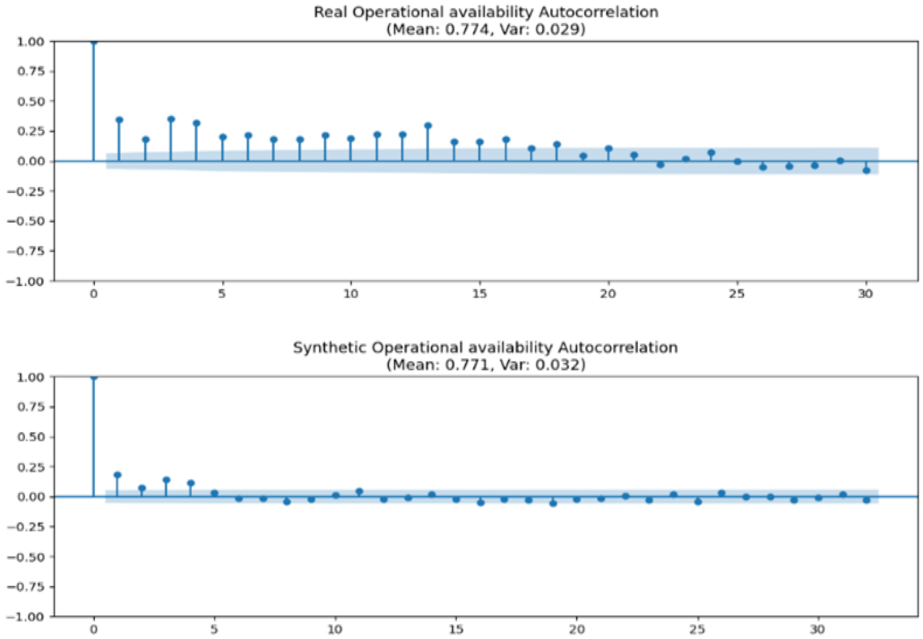


Fig. 14. Operational availability autocorrelation plots

The blue shaded area represents the confidence interval; if the autocorrelation values fall within this band, it indicates that the correlation at that lag is not statistically significant and could be due to chance. In both plots, most of the autocorrelation values after the first lag fall within this band, reinforcing the observation of weak autocorrelation.

The similarity of the autocorrelation patterns between the real and synthetic data suggests that the model captures the temporal dynamics of the operational availability of the real system.

The line graphs in Figure 15 shows the similarity between the real and synthetic operational availability, suggesting that the doppelganger model effectively captures the dynamics of the system's operational availability; both the real (blue) and synthetic (red) operational availability values fluctuate over time and remain largely in the 0 to 1 range (as expected for this metric). The synthetic data appears to follow the general trend and variability of the real data fairly well.

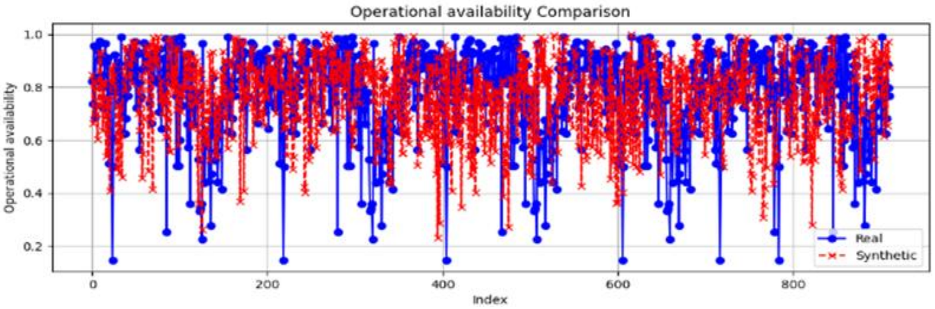


Fig. 15. operational availabily line charts (Real vs. Synthetic)

In conclusion, the synthetic operational availability data is a promising tool for the analysis and optimization of the complex industrial system.

5.2. Performance

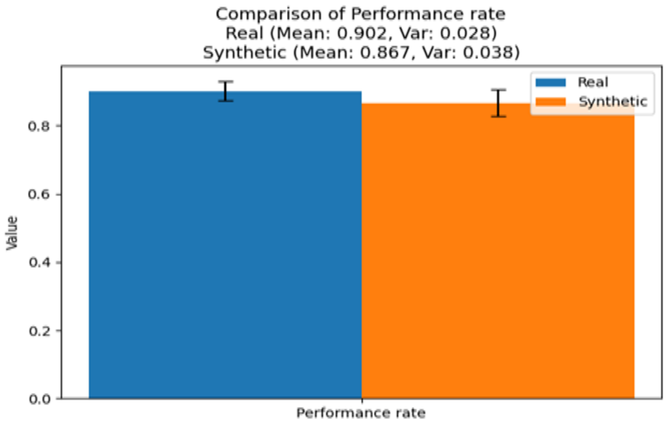


Fig. 16. Statistics comparison of performance (Real vs. Synthetic)

The relatively close agreement between real and synthetic performance rates indicates that the doppelganger model is reasonably accurate in capturing the behavior of the system. However, the slight underestimation and higher variance in the synthetic data suggest opportunities for further refinement and calibration of the model.

In addition, examination of the autocorrelation plots (Fig. 17) of both synthetic and real performance data shows that both data are correlated with the past on a weekly basis beyond the first few lags, suggesting that the model captures the temporal dynamics of the performance metrics.

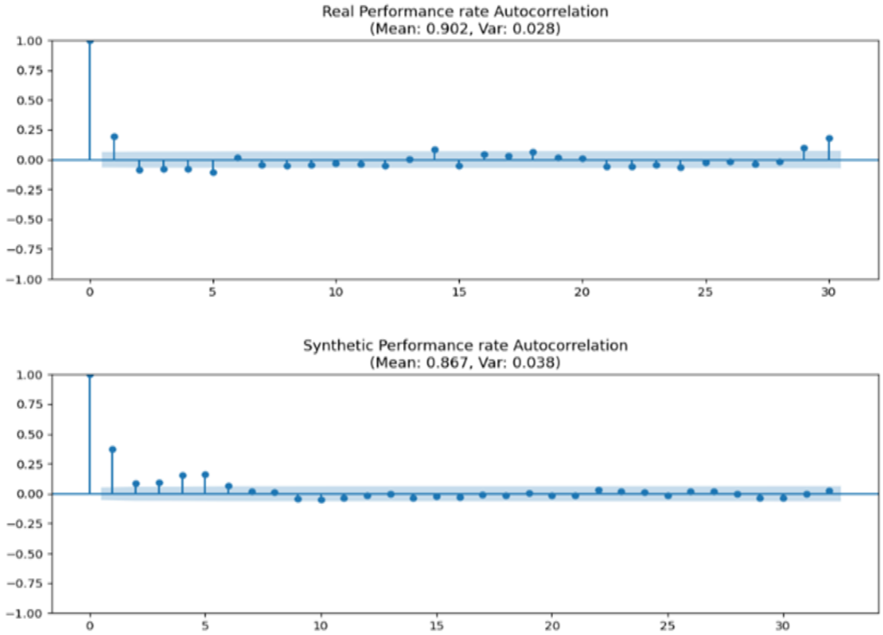


Fig. 17. Performance autocorrelation plots

In addition, the line graphs in Figure 18 suggest that the model generating the synthetic data has captured the underlying dynamics affecting performance.

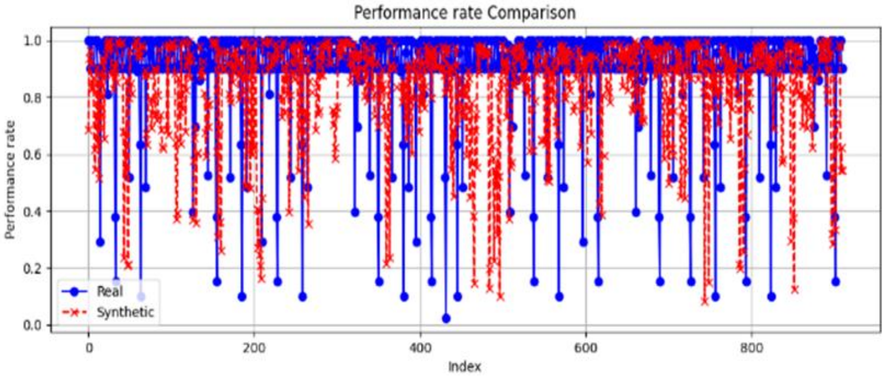


Fig. 18. Performance line charts (Real vs. Synthetic)

In summary, the general agreement between real and synthetic performance rates indicates that the model has captured the essential aspects of the system's performance dynamics that may be relevant: Stress testing the system under different scenarios, simulating the impact of process improvements or changes on performance, and training machine learning models for predictive maintenance or performance optimization.

5.3. Quality

Figure 19 shows that, on average, 65.1% of the output from the actual sugar production process meets the defined quality standards. The variance of 0.035 suggests that there's some variation in the quality rate, which could be due to factors such as variations in raw material quality, process inconsistencies, or equipment performance. However, the quality rate predicted or simulated by the doppelganger model. It's slightly lower than the real quality rate, suggesting that the model might slightly underestimate the real quality performance of the system. The relatively close match between the real and synthetic quality rates indicates that the doppelganger model captures the essential aspects of the quality dynamics in the sugar production process.

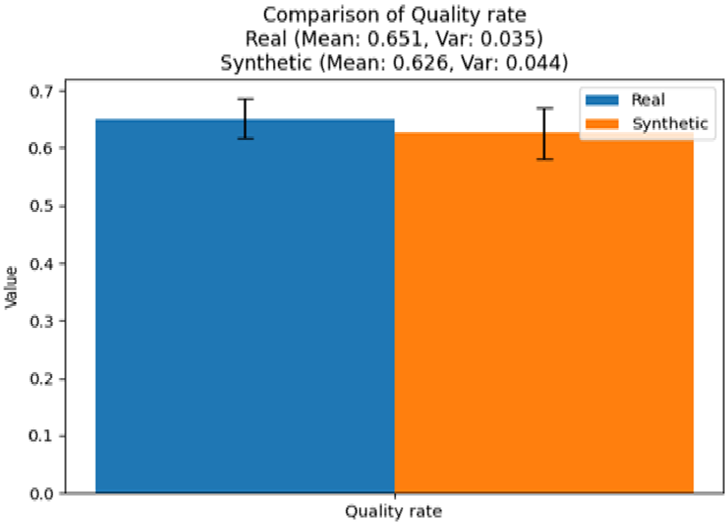


Fig. 19. Statistics comparison of quality (Real vs. Synthetic)

This means that the synthetic data provides a valuable benchmark for assessing the quality performance of the real plant. While the actual quality rate is already acceptable, the model shows that there may be room for improvement to achieve even higher quality levels. Again, by analyzing the factors that contribute to the differences between the real and synthetic quality rates, it may be possible to identify areas where the real plant can be optimized to improve product quality. This could include improving raw material quality control, fine-tuning process parameters, or addressing equipment-related issues that may be affecting quality.

In addition, the autocorrelation plots (Fig. 20) show that after the first few lags, both synthetic and real data are very weakly correlated with their past. This indicates that the temporal dynamics of the metric have been captured by the model.

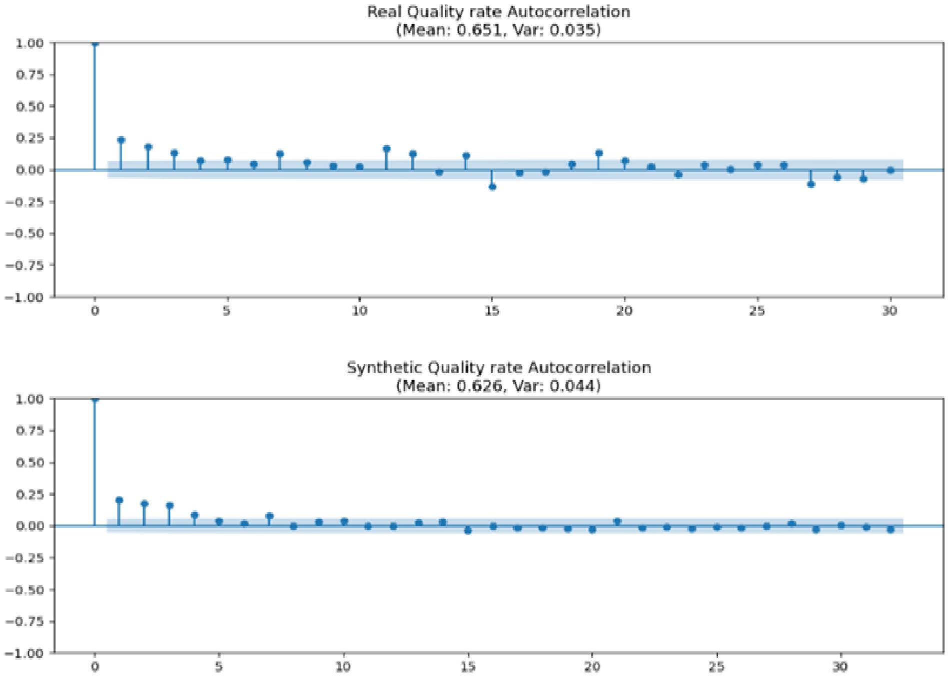


Fig. 20. Quality autocorrelation plot

The line charts in Figure 21 confirm that the synthetic data seem to capture the upward and downward trends in the real quality rate reasonably well. This indicates that the doppelganger model is reasonably effective in capturing the quality dynamics of the production system. In summary, the synthetic quality rate data, while not a perfect replica, demonstrate the potential of the doppelganger model to capture essential aspects of real-world quality dynamics. With further refinement, it could serve as a valuable tool for understanding, predicting, and ultimately improving the quality of the production process.

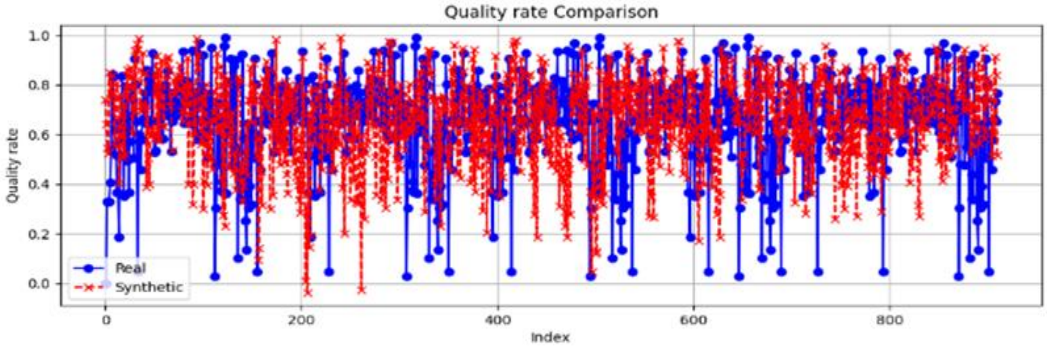


Fig. 21. Quality line charts (Real vs. Synthetic)

5.4. Evaluation via RMSE and DTW (Dynamic Time Warping)

Building on the initial evaluation of the synthetic data, which considered the variance, mean, and autocorrelation parameters of the generated time series, we conducted an additional analysis focusing on Root Mean Square Error (RMSE) and Dynamic Time Warping (DTW) to further assess the fidelity of the synthetic data to the real operational data from the SUCAF Gabon plant. The results of this additional evaluation are shown in Table 1.

Tab. 1. Doppelganger evaluation on RMSE and DTW

Metrics	RMSE	DTW
Operational availabiliy	0.07	0.2
Quality rate	0.078	0.23
Performance rate	0.09	0.3

The RMSE values when comparing the synthetic data to the real data were 0.07 for Operational Availability, 0.078 for Quality Rate, and 0.09 for Performance Rate. These values indicate a good level of similarity in the magnitude of the synthetic data compared to the real data. In addition, the DTW values between the synthetic and real time series were 0.2, 0.23, and 0.3 for the corresponding metrics, suggesting that the synthetic data also effectively captured the temporal patterns observed in the real operational data.

These additional results, when considered together with the initial analysis of variance, mean, and autocorrelation, provide a more comprehensive assessment of the ability of the synthetic data to replicate both the statistical properties and the temporal dynamics of the real operational indicators for the SUCAF Gabon plant. However, the slightly higher RMSE and DTW values observed for Performance Rate suggest that replicating the magnitude and temporal patterns of this specific metric with the same level of accuracy as Operational Availability and Quality Rate represents a potential area for further refinement of the synthetic data generation process.

5.5. Benchmarking results

The results of the benchmarking are presented in Table 2 and Table 3.

Tab. 2. Evaluation results of the Vanilla LSTM model

Metric	Quality rate	Performance rate	Operational availability
Mean (Real)	0.6941	0.7005	0.9913
Mean (Synthetic)	0.7469	0.7684	1.0499
Variance (Real)	0.0678	0.0839	0.3602
Variance (Synthetic)	0.0001	0.0024	0.0002
RMSE	0.0799	0.3004	0.6005

Tab. 3. Evaluation Results of the LSTM-Variational autoencoder model

Metric	Quality rate	Performance rate	Operational availability
Mean (Real)	0.531	0.610	0.7912
Mean (Synthetic)	0.5326	0.663	0.642
Variance (Real)	0.573	0.138	0.3001
Variance (Synthetic)	0.0254	0.0344	0.0013
RMSE	0.0813	0.4224	0.862

The results show that the Doppelgänger model has the strongest ability to replicate the mean of the real data for all three metrics, with its synthetic data closely matching the mean of the real data, outperforming the LSTM models which show some discrepancies. Similar to the mean, Doppelgänger effectively replicates the variance of the real data, while the LSTM Vanilla model significantly underestimates the variance, and the LSTM VAE model provides a better approximation, but is still less accurate than Doppelgänger. Doppelgänger achieves the lowest RMSE values for all three metrics, indicating the highest accuracy in predicting the real data values, compared to both LSTM models, which exhibit higher RMSE values, indicating less accurate predictions. In summary, Doppelgänger outperforms the Vanilla LSTM and LSTM VAE models in generating synthetic data that closely resembles real industrial performance data, demonstrates superior performance in terms of mean and variance matching, and achieves the lowest RMSE across all metrics. These results highlight the effectiveness of Doppelgänger in capturing both the statistical properties and temporal dependencies of industrial performance data, making it a promising tool for generating high-fidelity synthetic data.

6. CONCLUSIONS AND PERSPECTIVES

This article makes several significant contributions to the field of complex industrial systems management, including

Novel modeling approach**:** It proposes the use of generative doubles, a variant of Generative Adversarial Networks (GANs), to model and simulate the stochastic behavior of complex industrial systems. This approach overcomes the limitations of traditional modeling techniques, which often struggle to capture the inherent dynamics and uncertainty of these systems.

Proactive Optimization: Doppelganger's ability to generate realistic synthetic data paves the way for proactive optimization of operations. By simulating different operational scenarios, decision makers can anticipate potential problems, assess the impact of process changes, and make informed decisions to improve system efficiency.

In-Depth KPI Analysis: This article demonstrates how doppelgangers can be used to analyze various key performance indicators (KPIs) such as uptime, performance, quality, and overall equipment effectiveness (OEE). This analysis helps identify bottlenecks, inefficiencies, and opportunities for improvement within the system.

Validation on a real-world case: The proposed methodology is validated on a real industrial sugar factory, reinforcing its relevance and applicability in concrete industrial contexts. The results obtained show that doppelgangers can effectively capture the complex dynamics of the system and provide valuable insights for operational optimization.

Continuous Improvement Loop: The article highlights the possibility of using the information from the KPI analysis to improve both the real system and the doppelganger model. This continuous feedback loop allows gradual refinement of the model and continuous optimization of the industrial system.

In conclusion, this paper proposes an innovative and promising approach to manage the complexity of stochastic data in industrial systems. The use of generative doppelgangers opens new perspectives for operational optimization, improved decision making, and significant efficiency gains in industry. However, the future direction of this work aims to explore the following potentials of generative doppelgangers in industrial systems:

Real-time applications: The article suggests the possibility of using doppelgangers for real-time monitoring and control of industrial processes. This could revolutionize the way industries respond to dynamic situations, allowing for immediate adjustments and optimizations based on real-time data and predictions.

Integration with emerging technologies: There are exciting opportunities to integrate doppelgangers with IoT, edge computing and cloud platforms. This could lead to more efficient data collection and processing, enabling faster and more accurate decision making in industrial environments.

Enhanced model interpretability: Addressing the "black box" nature of GANs is critical for wider industrial adoption. Techniques to improve the interpretability of GAN models will increase confidence and understanding of their results, making them more valuable to decision makers.

Advanced evaluation metrics: The development of more sophisticated metrics to evaluate the quality, variety, and temporal accuracy of the data generated will be essential. This will ensure that doublets accurately capture the complexity of industrial processes and provide reliable insights.

Addressing ethical and societal concerns: As with any powerful technology, the use of doppelgangers raises ethical and societal concerns. Addressing issues such as privacy, algorithmic bias, and potential job displacement will be critical to ensuring the responsible and equitable use of this technology.

Overall, generative doppelgangers have the potential to transform the way industries manage and optimize complex systems. Further research and development in this area, focusing on the above perspectives, could lead to significant advances in industrial efficiency, productivity, and decision making.

Acknowledgment

The authors would like to thank SUCAF GABON for the operational data support for this study.

Funding Sources

The authors declare that no funds, grants, or other support were received during this work.

Conflict of Interest

The authors report there are no competing interests to declare.

REFERENCES

Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (GANs): An updated review. *Archives of Computational Methods in Engineering*, *28*(2), 525–552. http://dx.doi.org/10.1007/s11831-019-09388-y

Antonucci, D., Conselvan, F., Mascherbauer, P., Harringer, D., & Pozza, C. (2024). Synthetic data on buildings. *Machine Learning Applications for Intelligent Energy Management: Invited Chapters from Experts on the Energy Field*, *35*, 203–226. https://doi.org/10.1007/978-3-031-47909-0_7

Bahrum, N. N., Setumin, S., Othman, N. A., Maruzuki, M. I. F., Abdullah, M. F., & Ani, A. I. C. (2024). Performance evaluation of generative adversarial networks for generating mugshot images from text description. *Bulletin of Electrical Engineering and Informatics*, *13*(1), 300–311. https://doi.org/10.11591/eei.v13i1.5895

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv, abs/1803.01271*. https://doi.org/10.48550/arXiv.1803.01271

Branytskyi, V., Golovianko, M., Malyk, D., & Terziyan, V. (2022). Generative adversarial networks with bio-inspired primary visual cortex for Industry 4.0. *Procedia Computer Science*, *200*, 418–427. https://doi.org/10.1016/j.procs.2022.01.240

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Calix, R., Ugarte, O., Wang, H., & Okosun, T. (2024). A dataset of CFD simulated industrial furnace images for conditional automatic generation with GANs. *TMS Annual Meeting & Exhibition*, 775–783. https://doi.org/10.1007/978-3-031-50349-8_66

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *22nd acm sigkdd international conference on knowledge discovery and data mining (KDD '16)* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chung, J., Shen, B., & Kong, Z. J. (2024). Anomaly detection in additive manufacturing processes using supervised classification with imbalanced sensor data based on generative adversarial network. *Journal of Intelligent Manufacturing*, *35*, 2387–2406. https://doi.org/10.1007/s10845-023-02163-8

Dash, A., Ye, J., & Wang, G. (2023). A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: from medical to remote sensing. *IEEE Access*, *12*, 18330-18357. https://doi.org/10.1109/ACCESS.2023.3346273

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129. https://psycnet.apa.org/doi/10.1037/h0030377

Farady, I., Islam, J., Tuarob, S., Ng, H.-F., & Lin, C.-Y. (2023). GANs in industrial surface defect detection: Insights and challenges. https://dx.doi.org/10.2139/ssrn.4516131

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Fu, W., Chen, Y., Li, H., Chen, X., & Chen, B. (2023). Imbalanced fault diagnosis using conditional wasserstein generative adversarial networks with switchable normalization. *IEEE Sensors Journal*, *23*(23), 29119-29130. https://doi.org/10.1109/JSEN.2023.3322040

Fukaya, K., Daylamani-Zad, D., & Agius, H. (2023). Intelligent generation of graphical game assets: A conceptual framework and systematic review of the state of the art. *ACM Computing Surveys*, *57*(5), 118. https://doi.org/10.1145/3708499

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Communications of the ACM*, *63*(11), 139-144. http://dx.doi.org/10.1145/3422622

Hellmann, F., Mertes, S., Benouis, M., Hustinx, A., Hsieh, T.-C., Conati, C., Krawitz, P & André, E. (2024). Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *ACM Transactions on Multimedia Computing, Communications and Applications*, *21*(1), 6. https://doi.org/10.1145/3641107

Hobbie, H., & Lieberwirth, M. (2024). Compounding or Curative? Investigating the impact of electrolyzer deployment on congestion management in the German power grid. *Energy Policy*, *185*, 113900. https://doi.org/10.1016/j.enpol.2023.113900

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, C., Sun, Z., Li, C., Zhang, Y., & Xing, C. (2023). Survey of time series data generation in IoT. *Sensors*, *23*(15), 6976. https://doi.org/10.3390/s23156976

Jiang, W., Hong, Y., Zhou, B., He, X., & Cheng, C. (2019). A GAN-based anomaly detection approach for imbalanced industrial time series. *IEEE Access*, *7*, 143608–143619. https://doi.org/10.1109/ACCESS.2019.2944689

Khan, I. U., Noor, S., Sajid, A., Javaid, J., & Tabasusum, I. (2023). Comparative analysis of anomaly detection techniques using generative adversarial network. *Sir Syed University Research Journal of Engineering & Technology*, *13*(2), 8-17 http://dx.doi.org/10.33317/ssurj.615

Kumarage, T., Ranathunga, S., Kuruppu, C., De Silva, N., & Ranawaka, M. (2019). Generative adversarial networks (GAN) based anomaly detection in industrial software systems. *2019 Moratuwa Engineering Research Conference (MERCon)* (pp. 43–48). IEEE. http://dx.doi.org/10.1109/MERCon.2019.8818750

Kuntalp, M., & Düzyel, O. (2024). A new method for GAN-based data augmentation for classes with distinct clusters. *Expert Systems with Applications*, *235*, 121199. https://doi.org/10.1016/j.eswa.2023.121199

Kusiak, A. (2020). Convolutional and generative adversarial neural networks in manufacturing. *International Journal of Production Research*, *58*(6), 1594–1604. https://doi.org/10.1080/00207543.2019.1662133

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2019). Using GANs for sharing networked time series data: Challenges, initial promise, and open questions. *ArXiv, abs/2003.03453*. https://doi.org/10.48550/arXiv.1909.13403

Luo, J., Huang, J., & Li, H. (2021). A case study of conditional deep convolutional generative adversarial networks in machine fault diagnosis. *Journal of Intelligent Manufacturing*, *32*(2), 407–425. https://doi.org/10.1007/s10845-020-01579-w

Makhlouf, A., Maayah, M., Abughanam, N., & Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Computing and Applications*, *35*, 24055–24068. https://doi.org/10.1007/s00521-023-09100-z

Mumbelli, J. D., Guarneri, G. A., Lopes, Y. K., Casanova, D., & Teixeira, M. (2023). An application of generative adversarial networks to improve automatic inspection in automotive manufacturing. *Applied Soft Computing*, *136*, 110105. https://doi.org/10.1016/j.asoc.2023.110105

Ntavelis, E., Kastanis, I., Van Gool, L., & Timofte, R. (2020). Same same but different: Augmentation of tiny industrial datasets using generative adversarial networks. *2020 7th Swiss Conference on Data Science (SDS)* (pp. 17–22). IEEE. https://doi.org/10.1109/SDS49233.2020.00011

Qian, C., Yu, W., Lu, C., Griffith, D., & Golmie, N. (2022). Toward generative adversarial networks for the industrial internet of things. *IEEE Internet of Things Journal*, *9*(19), 19147–19159. https://doi.org/10.1109/JIOT.2022.3163894

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv, abs/1511.06434*. https://doi.org/10.48550/arXiv.1511.06434

Ren, L., Wang H., Li, J., Tang, Y., & Yang, C. (2024). AIGC for industrial time series: From deep generative models to large generative models. *ArXiv, abs/2407.11480*. https://doi.org/10.48550/arXiv.2407.11480

Rezaei, S., Cornelius, A., Karandikar, J., Schmitz, T., & Khojandi, A. (2024). Using GANs to predict milling stability from limited data. *Journal of Intelligent Manufacturing*, *36*, 1201–1235. https://doi.org/10.1007/s10845-023-02291-1

Saiz, F. A., Alfaro, G., Barandiaran, I., & Graña, M. (2021). Generative adversarial networks to improve the robustness of visual defect segmentation by semantic networks in manufacturing components. *Applied Sciences*, *11*(14), 6368. https://doi.org/10.3390/app11146368

Salierno, G., Leonardi, L., & Cabri, G. (2024). A big data architecture for digital twin creation of railway signals based on synthetic data. *IEEE Open Journal of Intelligent Transportation Systems*, *5*, 342-359. https://doi.org/10.1109/OJITS.2024.3412820

Song, J., Lee, Y. C., & Lee, J. (2023). Deep generative model with time series-image encoding for manufacturing fault detection in die casting process. *Journal of Intelligent Manufacturing*, *34*, 3001–3014. http://dx.doi.org/10.1007/s10845-022-01981-6

Sun, C. (2024). *Deep Generative Models for Network Data Synthesis and Monitoring*. The University of Edinburgh.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *ArXiv, abs/1706.03762*. http://dx.doi.org/10.48550/arXiv.1706.03762

Vdoviak, G., & Giedra, H. (2024). Review and experimental comparison of generative adversarial networks for synthetic image generation. *New Trends in Computer Sciences*, *2*(1), 1–18. https://doi.org/10.3846/ntcs.2024.20516

Wang, Y., & Yan, P. (2024). RegGAN: A virtual sample generative network for developing soft sensors with small data. *ACS Omega*, *9*(5), 5954–5965. https://doi.org/10.1021/acsomega.3c09762

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. *33rd International Conference on Neural Information Processing Systems* (pp. 5508-5518). Curran Associates Inc.

Yuan, Y., Zhang, Y., & Ding, H. (2020). Research on key technology of industrial artificial intelligence and its application in predictive maintenance. *Acta Automatica Sinica*, *46*(10), 2013–2030. http://dx.doi.org/10.16383/j.aas.c200333

Zhang, H., Dereck, S. S., Wang, Z., Lv, X., Xu, K., Wu, L., Jia, Y., Wu, J., Long, Z., Liang, W., M, X. G., & Huang, G. B. (2023). Large scale foundation models for intelligent manufacturing applications: a survey. *ArXiv, abs/2312.06718*. https://doi.org/10.48550/arXiv.2312.06718

Zhang, Y., Schlueter, A., & Waibel, C. (2023). SolarGAN: Synthetic annual solar irradiance time series on urban building facades via Deep Generative Networks. *Energy and AI*, *12*, 100223. https://doi.org/10.1016/j.egyai.2022.100223

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

Zhou, R., Jiang, C., & Xu, Q. (2021). A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, *451*, 316–336. https://doi.org/10.1016/j.neucom.2021.04.069