Applied Computer Science, vol. X, no. X, pp. X–X
doi: X

Submitted: YYYY-MM-DD | Revised: YYYY-MM-DD | Accepted: YYYY-MM-DD

Keywords: Industrial Manufacturing systems, generative doppelgangers, stochastic data, operational optimization

Richard NASSO TOUMBA [0009-0006-7596-1860]*, *Maxime MOAMISSOAL SAMUEL*, *Achille EBOKE*, *Boniface ONDO***, *Timothée KOMBE* *

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 quantities of stochastic data, often impeding operations optimization, particularly 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 employing 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, collectively contributing to OEE, is highlighted. The research validates this approach using real-world data from an industrial sugar plant, demonstrating its potential for providing valuable insights into system behavior under various operational scenarios for proactive optimization.

1. INTRODUCTION

Complex industrial manufacturing systems are characterized by intricate networks of interconnected processes, machinery, and human interactions (Mbohwa, 2020; Train & Salehin, 2023). Their complexity arises from factors like redundancy, scale, desynchronizations, environmental variability, and human errors, as highlighted in the context of system reliability (Zhou et al., 2021). This inherent complexity leads to the generation of vast 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,

* The university of Douala, Laboratory of technologies and Applied Sciences, Cameroon, richardnassotoumba4@gmail.com,moamissoalmaxime@gmail.com,ebokeachille@gmail.com, tkombe@yahoo.fr
** The university of Masuku , department of electrical engineering, Gabon, bonitoondo@gmail.com

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 forecasting based on past trends but often reliant 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 are contrasted with GAN capabilities (Creswell et al., 2018)), often rely on deterministic or linear assumptions or lack the capacity to fully capture the complex, non-linear, 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 settings (Wen et al., 2023). Consequently, there is a significant need for novel approaches capable of effectively learning and simulating the stochastic behavioral patterns of these systems to enable proactive optimization and informed decision-making.

To address this critical gap, this paper proposes a novel approach employing 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 like 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 the complex joint distributions of multivariate time series data characteristic of 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 forecasting future values rather than generating synthetic, yet realistic, sequences that preserve the intricate stochastic properties of the original data (AIGC for Industrial Time Series, 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, non-linear, and stochastic patterns observed in industrial operations (Yoon et al., 2019). This generative capability is crucial for conducting realistic what-if analyses and exploring potential future scenarios that go beyond simple forecasting. The doppelganger architecture, with its ability to jointly model both the temporal dynamics and the inter-dependencies between different operational parameters, allows for the creation of "behavioral doppelgangers" that can simulate the nuanced impact of various factors on OEE components like availability, performance, and quality (Lin et al., 2020).

Unlike traditional digital twins that often focus on the static physical model, the "behavioral doppelganger" aims to decipher the hidden patterns and stochastic nuances within the system's operational data. The application of generative doppelgangers for the 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 earlier 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 tackle some particular problems in many engineering domains, among which:

- Image Generation and Synthesis: GANs have been extensively employed to produce realistic images. They can be employed to generate novel artistic images, create new images that resemble a specific dataset, or even 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 create extra samples to improve training data. When the dataset is limited, this is particularly helpful as it improves the generalizing power of machine learning models (Kuntalp & Düzyel, 2024; Wang & Yan, 2024; Vdoviak & Giedra, 2024).
- Video Generation: By sequentially generating each frame, GANs have been expanded 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 evaluations, dialogue, and language translation (Bahrum et al., 2024; Ren et al., 2024; Saiz et al., 2021).
- Style Transfer: GANs have the ability to acquire the style of a specific image or artwork and adapt it to another image, thereby achieving style transfer. This method has been employed for image modification, artistic depiction, 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. The potential benefits of this include the provision of supplementary training data for medical image analysis tasks and the resolution of privacy concerns (Makhlouf et al., 2023; Fukaya et al., 2023).
- Game development: GANs have been used in game creation to generate levels, characters, and graphics among other game elements. They may also be used to create human-like behaving non-player characters (NPCs) (Fukaya et al., 2023).
- Data Anonymization: GANs have been explored for the means of privacy-preserving data distribution. They have the ability to produce synthetic data that preserves the statistical properties of the original data while safeguarding sensitive information (Hellmann et al., 2024).

- Additionally, the utilization of Generative Adversarial Networks (GANs) for the prediction of industrial process failures has the potential to provide numerous advantages:
- Enhanced precision: GANs are capable of learning intricate patterns and producing genuine samples. By training GANs on historical data from industrial processes, it is possible to comprehend the fundamental patterns and correlations that contribute to failure. This can lead to a more precise prediction of failure than conventional methods (Jiang et al., 2019; Hu et al., 2023; Zhang et al., 2023).
- Early detection: GANs have the potential to identify minute changes in data patterns that may indicate an impending error.
- The process data can be perpetually monitored and compared to the learnt patterns by GANs, which can minimize disruptions and enable proactive maintenance. This enables them to issue early cautions (Chung et al., 2024; Mumbelli et al., 2023; Kusiak, 2020).
- Anomaly detection: By contrasting the produced samples with the original data, GANs may be utilized to find abnormalities in industrial processes. To identify a possible mistake or anomalous conduct, label any departure from the assimilated patterns as an anomaly (Jiang et al., 2019; Zhao et al., 2019; Kumarage et al., 2019).
- Cost savings: GANs may be used to streamline maintenance schedules, minimize unscheduled downtime, and avoid expensive repairs by foreseeing problems. Process owners may be able to save a lot of money and time by making their tools work better and cutting down on production costs (Zhao et al., 2019). It's important to know that using GANs to predict failure in an industrial process needs collecting data, preparing data, and training the model. For implementation to go well, subject understanding and cooperation between data analysts and topic experts are also very important (Fu et al., 2023).
- Anomaly detection: GANs can be used to identify anomalies in industrial processes by comparing the generated samples to the original data. Any variation from previously acquired patterns might be marked as an anomaly to highlight a probable mistake or aberrant conduct (Chung et al., 2024; Farady et al., 2023; Noor et al., 2023).
- The development of predictive maintenance strategies is significantly improved by the implementation of failure prediction, which allows organizations to foresee and prevent equipment failures in advance. Through the utilization of sophisticated analytics and machine learning algorithms, failure prediction models can analyze historical data, sensor readings, and other pertinent factors to predict the likelihood of a specific component or system failing (Zhang et al., 2023; Rezaei et al., 2024; Qian et al., 2022).
- Enhanced Equipment Availability: Organizations can anticipate potential issues and implement preventative measures by means of failure prediction. Organizations may enhance the availability and dependability of essential assets by taking proactive measures to resolve these concerns, which will result in a reduction in unexpected equipment outage (Goodfellow et al., 2014; Song et al., 2023). Optimized Maintenance Planning: Failure prediction enables businesses to arrange maintenance activities based on real-time requirements rather than predetermined timetables, thereby giving valuable information about the well-being and state of

equipment. This approach is condition-based, which optimizes maintenance planning, minimizes superfluous maintenance duties, and maximizes the utilization of maintenance resources (Ye et al., 2020).

- Predictive Maintenance: Researchers at the University of Michigan have utilized GAN-based doppelgangers to generate synthetic sensor data representing various stages of machine degradation. By comparing real-time sensor data to these synthetic patterns, the remaining useful life of machines can be estimated, allowing for proactive maintenance and reducing unplanned downtime (Salierno et al., 2024; Ntavelis et al., 2020).
- Quality Control: Doppelgangers have been explored to augment datasets 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, 2015).
- Power Grid Management: Researchers have investigated using GANs to model the complex behavior of power grids, including variations in load 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: Personnel and assets may be at risk of injury due to unanticipated equipment malfunctions. In an effort to mitigate safety hazards, failure prediction is employed to identify prospective failure modes and implement corrective measures. Organizations may create a safer workplace for their staff via early identification of possible mistakes. Improved judgment: The expected failures offer insightful analysis of equipment performance and deterioration trends. The data presented helps companies to make informed decisions about general asset management, equipment replacement, maintenance plans, and spare parts inventories. By seeing potential faults ahead of time, companies may either stop or lessen the effects of mistakes, hence increasing the lifetime of equipment. This proactive strategy helps to extend the lifetime of equipment, therefore lowering the need for early replacements and producing significant cost savings (Zhao et al., 2019; Luo et al., 2021).

3. MATERIALS AND METHODS

3. 1. Classical Generative GAN and DOPPELGANGERS

A remarkable subfield of machine learning is the Generative Adversarial Network (GAN) (Goodfellow et al., 2014). These models comprise two neural networks in competition: a generator and a discriminator. To maximize realism, the generator is built to create unique data samples made of text and photos among other things (Sun, 2024).

On the other side, it is the discriminator's job to distinguish these artificial samples apart from real data. This back-and-forth between challenge and response allows the generator to get better at making plausible impersonations and the discriminator to get better at detecting them. The two networks are trained in a competitive capacity in an aim to improve the generator's ability to produce real-world samples and the discriminator's capacity to correctly

identify them. Because GANs can generate varied and realistic data, they have been used in many different fields.

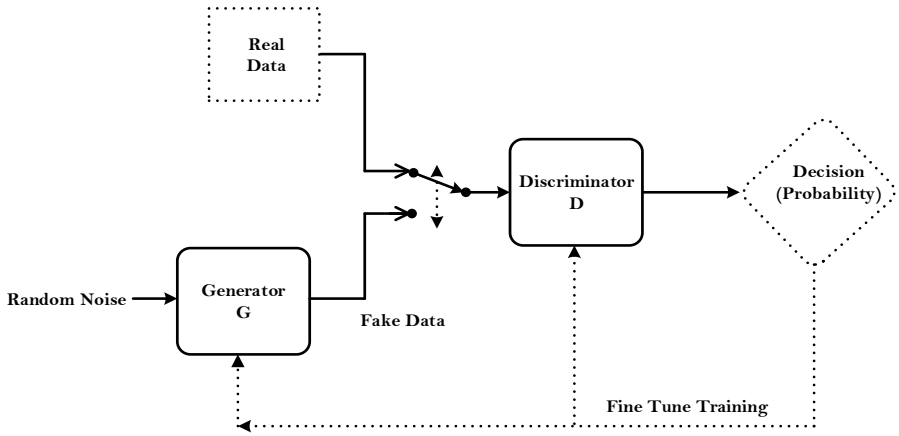

Fig.1. Classical GAN principle (source:adapted from *Air Liquid)*

In this architecture, Z represent the Noise to feed into the generator, G (Z) the fake data which is the output of the generator, x the real data. When fake data is fed into the discriminator, the outcome is $D(G(z))$ which probability that the fake sample is real is. Likewise, when real data is fed into the discriminator, the outcome D(x) represent the probability that the real sample is real as estimated by the discriminator.

In the Neural Networks adversarial principle, the expectation of the log of discriminator output on the input of real data is given by:

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

Likewise,

$$\mathbb{E}_z \sim p_z(z)\left[\log(1-D(G(z)))\right]. \qquad (2)$$

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

$$\mathbb{E}_x \sim pdata(x)\left[\log D(x)\right] + \mathbb{E}_z \sim p_z(z)\left[\log(1-D(G(z)))\right] \qquad (3)$$

Definitely, the optimisation objective function is given by:

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

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

1. Mode collapse: GANs may provide restricted output, unable to capture the entire range of training data (Salimans et al., 2016).
2. GANs need a big quantity of data for training in order to identify meaningful patterns and provide high-quality results. The first major issue with GANs using recurrent Neural Networks as generators (RNN) is their incapacity 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 solve this problem, the generator component of the Doppelganger model incorporates Long Short-Term Memory (LSTM) networks to effectively capture the temporal dependencies inherent in the industrial operational data. LSTMs were chosen for their well-established ability to model long-range dependencies in sequential data, which is crucial for understanding the evolving patterns and stochastic trends that influence OEE (Hochreiter & Schmidhuber, 1997; Gers et al., 2000). Unlike traditional Recurrent Neural Networks (RNNs) that suffer from the vanishing gradient problem (Bengio et al., 1994), LSTMs utilize 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 impact of past operational 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 intricate temporal patterns. Temporal Convolutional Networks (TCNs) present an alternative approach using convolutions, but might require very deep networks for long dependencies (Bai et al., 2018). Attention-based mechanisms, while powerful for long-range dependencies (Vaswani et al., 2017), could introduce greater 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 complexities of industrial data, they were selected for the Doppelganger generator.

The Doppelganger architecture also utilizes mini-batch training (process of feeding small data points at once). This helps to improve the efficiency of the model by reducing the 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 the query of relevant failures while studying or analyzing complex industrial systems. Another relevant feature of Doppelganger is that it does not only generate synthetic data based on the real ones but also jointly generates attributes (metadata) for that data (Lin et al., 2020). This is important to give context to the data. For example, the context of analyzing critical components' data is not the same as that of tolerant components. Besides, poorly categorized

data leads to inaccurate insight. Finally, it helps analyze trends, optimize processes, and make relevant data-driven decisions for the analysis of dependability components of complex industrial systems.

Model Training

Hyperparameter Specification: The doppelganger model, employing 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 ($\beta1=0.9$, $\beta2=0.994$).

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 might be possible through the implementation of GPU optimization techniques. Strategies such as utilizing 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.

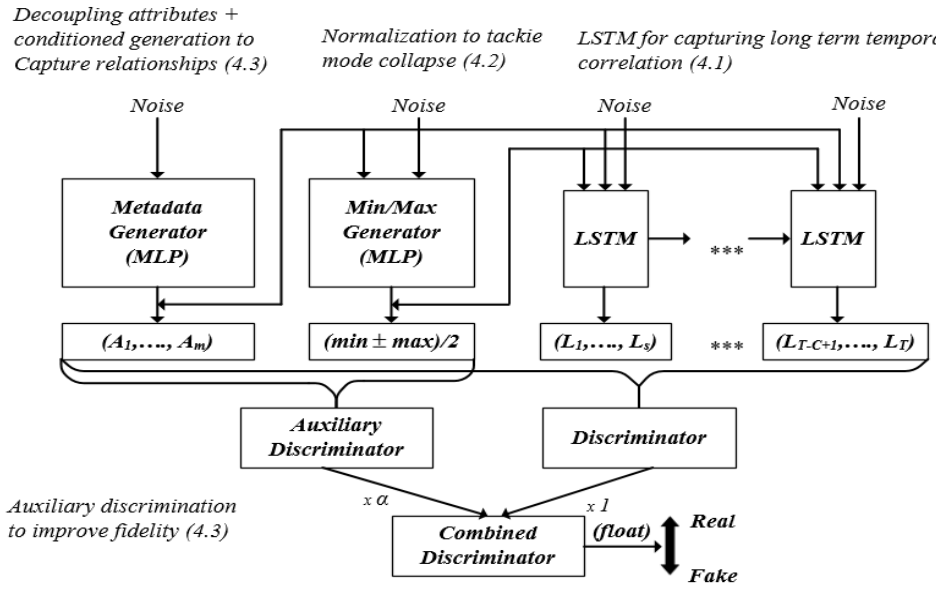

Fig.2. the proposed doppelgänger architecture

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 potentially metadata) as input and discriminates between real and fake.
  - Auxiliary Discriminator: Improves fidelity 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 their corresponding 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 its gradient (minimize loss_G)

- End for

3. Generation

- Provide desired metadata
- Generate time series using the trained Metadata Generator, Min/Max Generator, and LSTMs

Contrarily to the Classical GANs,in DOPPELGANGERS, the generator takes metadata as input, allowing for controlled generation of time series based on specific attributes or conditions; through the normalization, the Min/Max Generator provides normalization to tackle mode collapse, ensuring 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 makes 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 plant from GABON Republic, SUCAF, a medium size plant with a production figure of 24320 tons of sugar per year and employs 1152 staffs at its peak period.it is made of 60 major equipment pieces active in the process: The workflow of the process is made of seven stages:

Reception and Cleaning subsystem:
Made of several structures among which:

- The Cane carrier which is an Electromechanical structure (conveyor belts, 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)
- The stone and trash removers: Mechanical structure (screens, sieves)
- The Magnetic separators: Electromechanical (magnets, conveyor belts)

Extraction subsystem: Composed of several pieces of equipment:

- The Milling tandem : Being an Electromechanical nature (rollers, motors, gears)
- The Bagasse diffuser : That is a Mechanical strucutre (conveyor, diffusers)
- The Diffusion towers/Continuous diffusers (sugar beets):Constituted of mechanical components Mechanical (towers, conveyors, pumps)

Clarification Subsystem: Made of lime mixers and dosing system, is an electromechanical structure made of pumps and mixers

The Evaporation Ssubsystem: made of Multiple-effect evaporators: Mechanical (heat exchangers, vessels).

The Crystallisation subsystem: Constituted of:

- Vacuum pans/Crystallizers: Mechanical (vessels, heating coils)
- Seed slurry preparation tanks: Mechanical (tanks, agitators)
- Magma pumps: Electromechanical (pumps, motors)

The Centrifugation subsystem :

- The, Batch centrifuges is an Electromechanical structure made of: centrifuge bowls, motors, controls components
- The Molasses pumps: made of pumps, motors is an electromechanical system

The Drying and Packaging Subsystem: Made of rotary sugar dryers, sugar coolers Screening/Grading Equipment, Bagging/Bulk loading systems.

The following images respectively Fig.3. and Fig.4. represents respectively the sugar production process and the Real images of the Industrial sugar Plant.

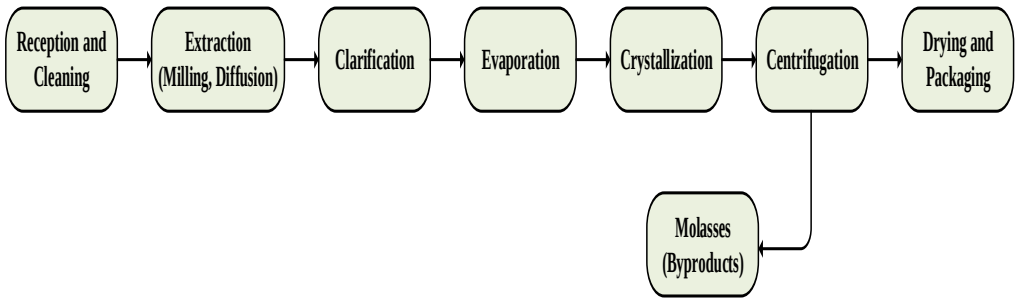

Fig.3. The sugar production process

Fig.4. Real images of the industrial sugar plant

3.3. Data Acquisition and Preprocessing

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

Key performance indicators (KPIs) related to Overall Equipment Effectiveness (OEE) were derived from the collected data. The following metrics 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 on a daily basis.

3.3.1. Equipment Representation

In this study, the primary focus was on the main equipment critical to the sugar production process. The selection of these key equipment pieces was based on a preliminary importance factor analysis, which assessed their impact on overall production output and potential for contributing to OEE losses. While the dataset encompasses various equipment within the facility, the analysis and modeling efforts were concentrated on those identified as having the most significant influence on the system's performance. It is important to note that due to the inherent structure of the production line and the varying criticality of different 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 main equipment, crucial for core operations, have a substantial representation to allow for meaningful analysis. Future research could explore incorporating 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 employed to identify and handle outliers within the key operational parameters. Extreme values that deviated significantly from the typical range, potentially due to recording errors or unusual operational events, were addressed. The primary approach for outlier management involved removal of data points identified as extreme outliers based on interquartile range (IQR) analysis.
- Missing Data Imputation: Instances of missing data were observed across the dataset. To mitigate the impact of these gaps, a nearest neighbor imputation technique was applied. This method replaced missing values with the values from temporally adjacent data points, assuming a 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 contributed equally to the model training process and to improve the convergence of the neural networks, min-max scaling was applied to normalize the data within the range of 0 to 1.

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

Fig.a. Dataset overview

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

3.4. Data description

3.4.1. MTBF, MTTR

Fig.b. Brief statistic of the dataset

	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

(1) MTBF (Mean Time Between Failures):

The mean MTBF is 0.204, which is relatively low. This suggests that the system experiences failures fairly frequently, the value of the median (0.215) is slightly higher than the mean, indicating some skewness in the data distribution, with a few longer periods between failures, and a range of 0.223 highlights variability in the system's reliability.

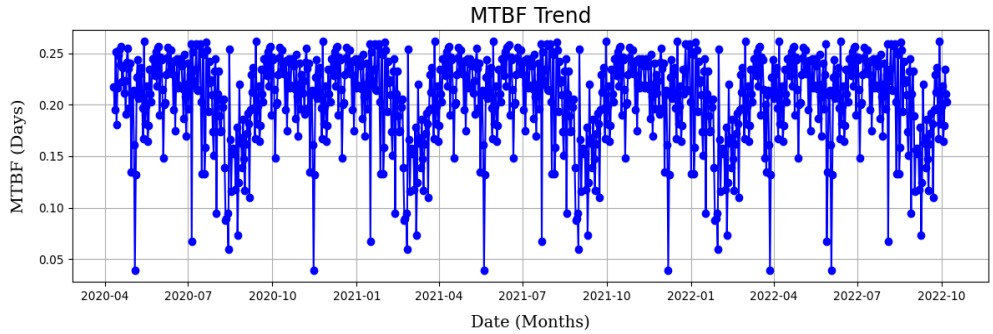

Fig.5. Trends of the MTBF over time

(2) MTTR (Mean Time To Repair):

The average MTTR is 0.044, which is quite good. This means that when failures 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.

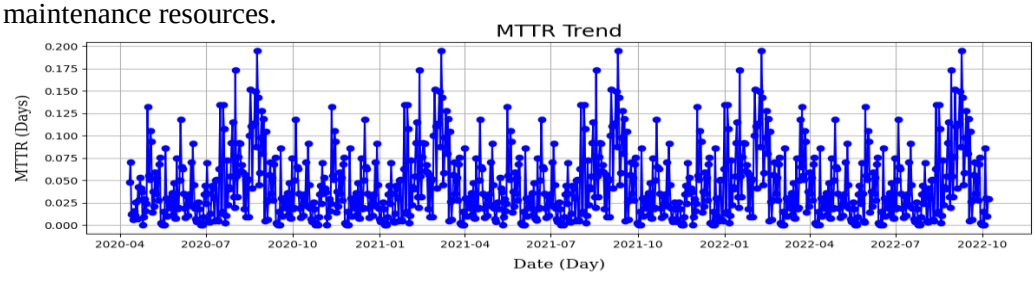

Fig.6. Trends of the MTTR over time

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, leading to 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 graph presented (Fig.7) highlights many relevant points of the operating system such as breakdowns, disruptions, and the pattern trend of the system availability for prediction.

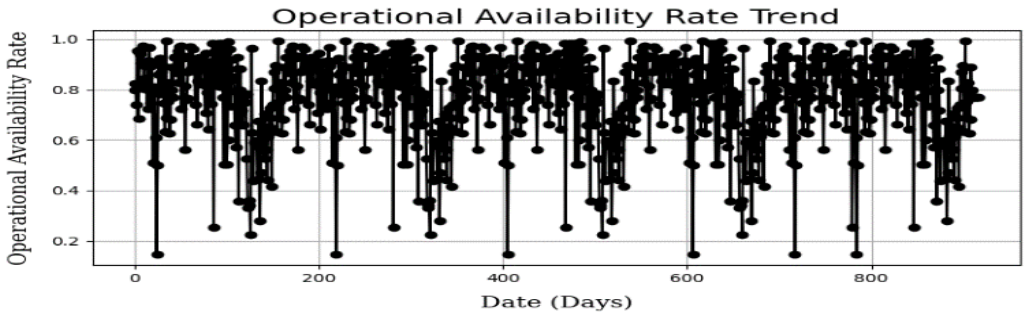

Fig.7. Operational availability trend over time

There is a High Variability of the operational availability fluctuating significantly, ranging from lows around 0.2 to highs near 1.0. It is suggesting frequent disruptions, process bottlenecks, inefficiencies in equipment reliability, aging machinery, inadequate maintenance, or improper operation issues leading to the reduction of the sugar production output. In addition, the metric frequently drops to lower levels, implying 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 indicates that there are stochastic underlying issues affecting availability that haven't been systematically addressed or that new challenges are continually arising.

3.4.2.2. Performance

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

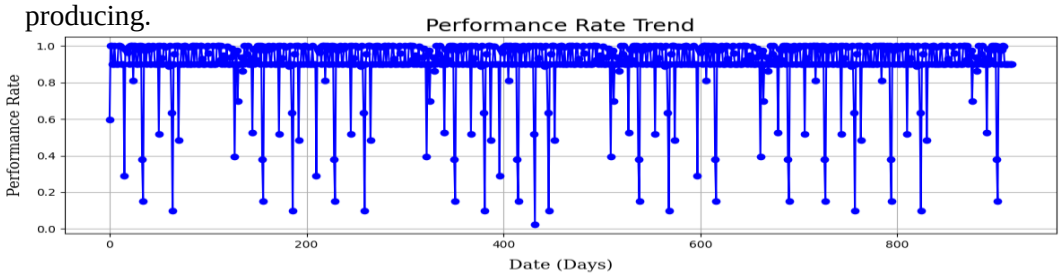

Fig.8. Performance rate trend over time

The performance rate trend graph shown in Fig.8 presents a High initial performance with Subsequent Decline**:** The performance starts very high, close to 1.0, but then experiences a sharp drop. Then stabilizes at a lower level, fluctuating between 0.2 and 0.4 for a significant duration. This implies an initial period of optimal operation followed by a significant decline in its performance. The initial drop and subsequent low performance suggest potential problems with equipment reliability or process stability. However, There is a persistent low values. The extended period of low performance indicates persistent issues impacting the system's efficiency. These could stem from equipment malfunctions, process inefficiencies, or other operational challenges. Alternatively, we can observe sporadic increases in the performance rate, potentially signifying temporary improvements or successful interventions. However, these gains are not sustained, pointing to underlying problems that haven't been fully resolved.

In this context, the prolonged period of low performance will likely translate to decreased sugar production. This could lead to financial losses and impact the ability to meet market demands. In conclusion, 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 industrial manufacturing plant refers to the percentage of sugar output that meets the desired specifications or standards.

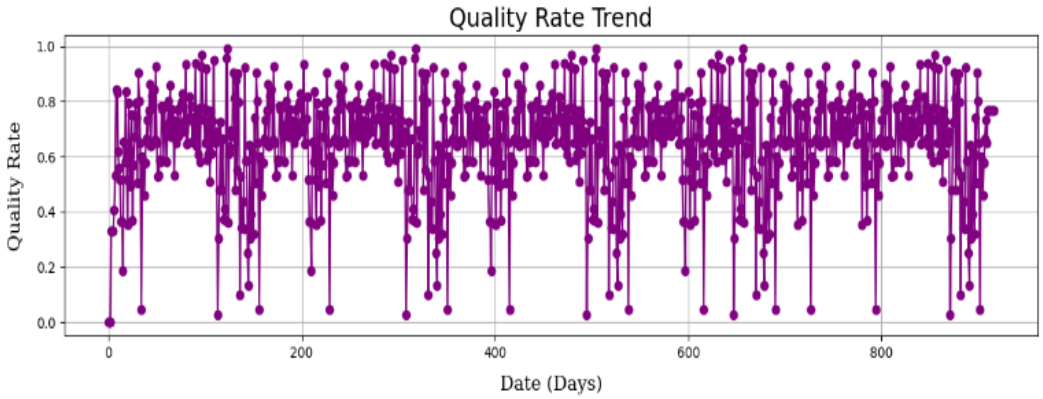

Fig.9. Quality rate trend over time

The quality rate shows substantial fluctuation, ranging from near 0.0 to 1.0. This clearly indicates inconsistency in the production process, leading to varying levels of product quality. The high variability and frequent drops indicate a lack of control over critical process parameters. This could result in product recalls, customer dissatisfaction, and financial losses. In addition, there are frequent drops, implying variations in raw materials, equipment malfunctions, process instability, or human error. Likewise with the Operational availability, there is no discernible upward or downward trend indicating stochasticity. It indicates clearly that underlying issues affecting quality have not been systematically addressed or new challenges are still emerging. Overall, the quality rate trend of the product plant exhibits a significant challenge in maintaining consistent sugar quality.

3.4.2.3. Overall Equipment Effectiveness

Composed of the Operational availabity, Performance rate and Quality rate, The OEE is a key performance indicator, a metric that measures how well a manufacturing operation is 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, enhance quality control measures to minimize the production of defective products. Besides, it helps to identify and eliminate bottlenecks or inefficiencies in the production process.

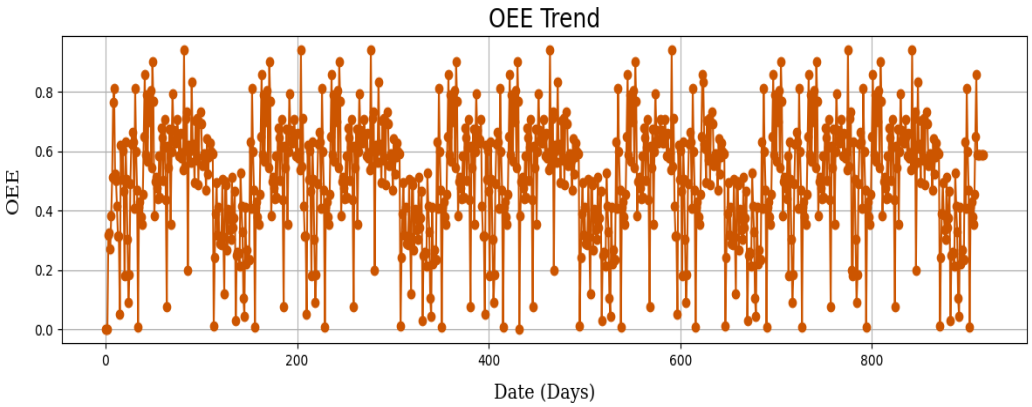

Fig.10. OEE trend over time

As in the previous metrics, a significant variability can be observed. The OEE values exhibit substantial fluctuations, ranging from near 0 to around 0.85, implying suboptimal utilization and potential losses. Also, it doesn’t have a clear upward or downward trend. Indicating that underlying factors influencing the OEE haven't been systematically identified or that new there are new challenges coming forth.

4. PROPOSED METHODOLOGY ANALYSIS USING DOPPELGANGERS

The methodology is made of a cyclical process where a doppelganger is used to generate synthetic data that mimics the behavior of a real-world industrial system. This synthetic data can then be leveraged to conduct in-depth analyses of various KPIs, ultimately leading to the identification of potential areas for improvement within the industrial pant. The insights obtained from these analyses can be subsequently used to refine and enhance the doppelganger model, thus creating a continuous feedback loop that drives ongoing optimization of the model.

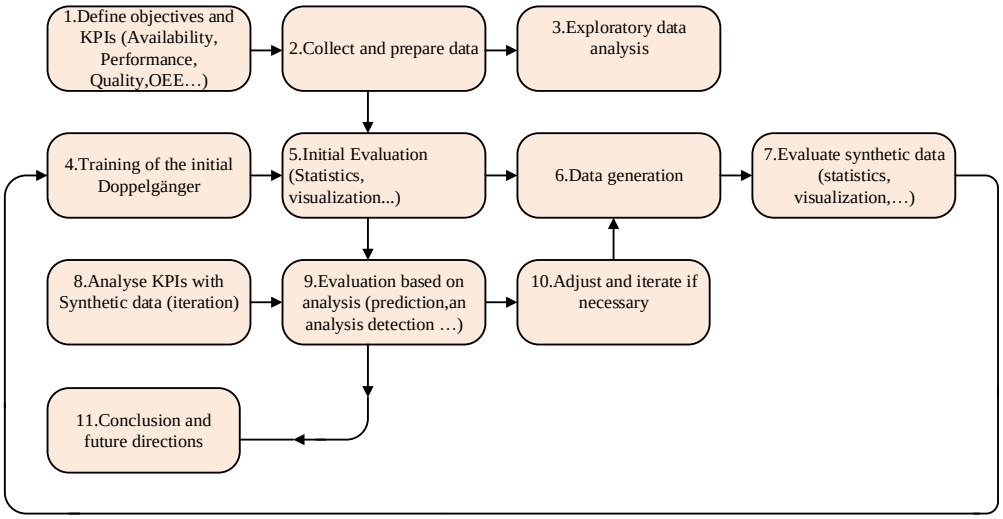

Fig.11. Doppelgänger-based KPI analysis framework

4.1. Real system and data collection

The process starts with the real industrial system, from which operational data is gathered. This data serves as the foundation for training the doppelganger model and forms the basis for comparison and validation later in the process.

4.2. Doppelganger Model and Training

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

4.3. Synthetic data generation

Once trained, the doppelganger model is employed to produce synthetic data. While being artificial, the dataset should closely resemble the real-world data in terms of statistical distributions and underlying trends.

4.4. KPI analysis

Both the real-world 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.

4.5. Comparison and validation

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

4.6. Identification of improvement areas

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

4.6. Feedback and improvement

The identified improvement areas 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, where 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 to handle stochastic data trends in complex manufacturing system.

5.1. Operational availability

Fig.12. Statistics comparison of operational availability (Real vs. Synthetic)

From the chart, we can observe the statistics of Real Operational Availability metrics as follow Mean: 0.774, Variance: 0.029. Implying that on average, the actual sugar plant is operational and ready to produce about 77.4% of the time. The variance of 0.029 suggests there's some fluctuation in this availability. Likewise the statistics of the same metric

generated by the Doppelganger indicates Synthetic Operational Availability Mean: 0.771, Variance: 0.032.It is very close to the real availability, It indicates that the model is doing capturing the real-world dynamics of the plant. The slightly higher variance in the synthetic data suggests the model might be predicting a bit more variability in availability than what's actually observed. In addition, the close alignment 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 the plant's behavior. Also, the model can potentially be used to predict how changes to the system (e.g., new equipment, different maintenance schedules) might impact the operational availability, to help identify equipment that's more prone to failure or predict when maintenance is needed, improving overall availability; by simulating different operating conditions or process changes, the model could help identify ways to further increase operational availability and improve production efficiency.

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

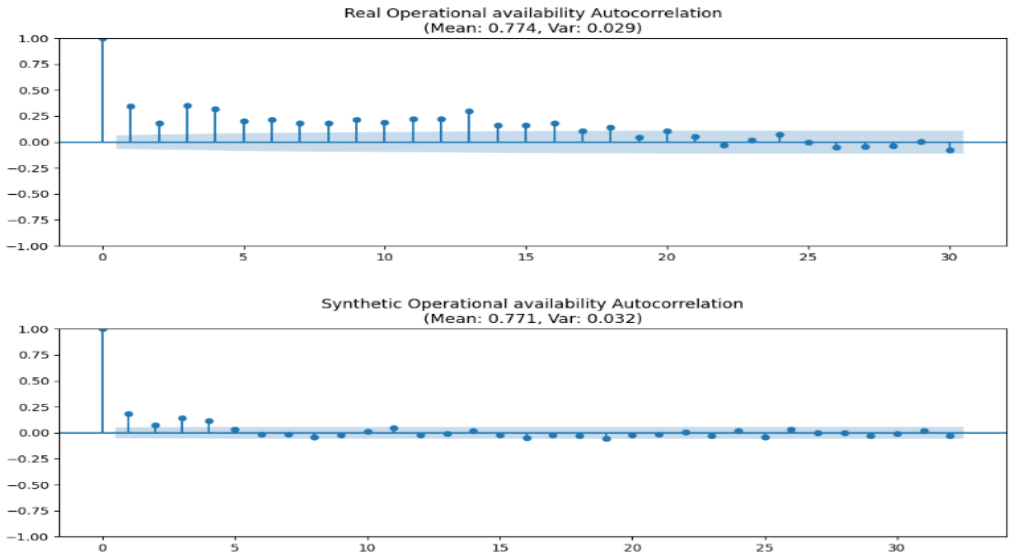

Fig.13. Operational availability autocorrelation plots

The blue sky shaded area represents the confidence interval; and if the autocorrelation values fall within this band, it suggests 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 in the autocorrelation patterns between the real and synthetic data suggests that the model capturing the temporal dynamics of the real system's operational availability.

The line charts in Fig14 discloses the resemblance between real and synthetic operational availability suggesting that the doppelganger model effectively captures the system's operational availability dynamics; both the real (blue) and synthetic (red) operational availability values fluctuate over time, largely remaining within 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.

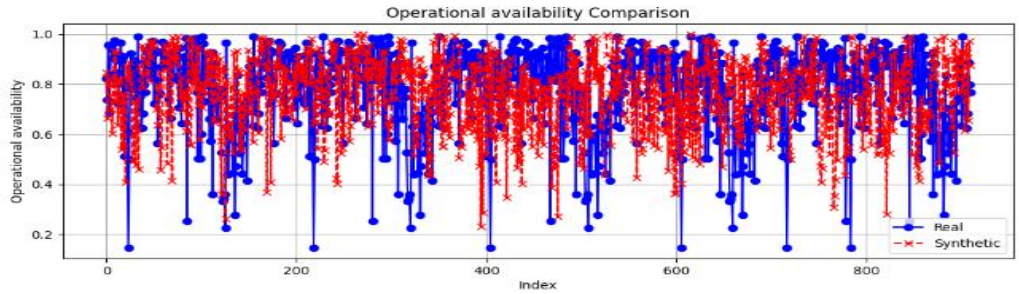

Fig.14. operational availabily line charts (Real vs. Synthetic)

In conclusion, the synthetic operational availability data s a promising tool for analyzing and optimizing the complex industrial system**.**

5.3. Performance

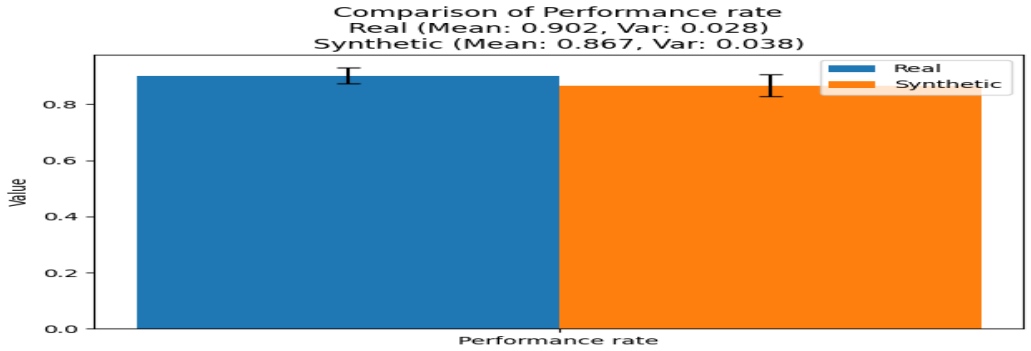

Fig.15. Statistics comparison of performance (Real vs. Synthetic)

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

In addition, the examination of autocorrelations plots (Fig.16) of both synthetic and real, performance data shows that both data are weekly correlated to their past beyong the first few lags, indicating that the model is capturing the temporal dynamics of the performance metrics.

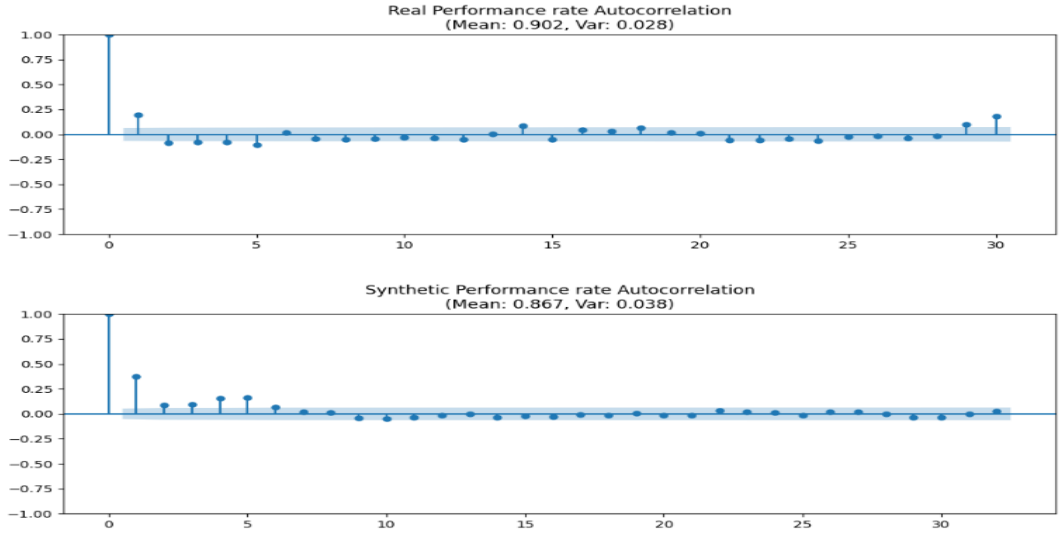

Fig.16. Performance autocorrelation plots

Furthermore, the line charts as shown in Fig.17 suggests that the model generating the synthetic data has captured the underlying dynamics influencing performance.

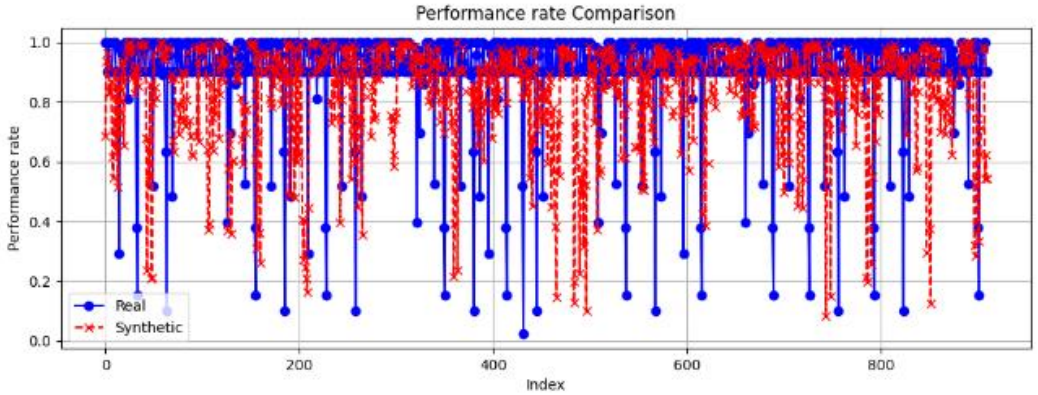

Fig.17. Performance line charts (Real vs. Synthetic)

In conclusion, 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 can be relevant to: Stress-test the system under different scenarios, simulating the effect of process improvements or changes on performance, and train machine learning models for predictive maintenance or performance optimization.

5.3. Quality

The Figure 18(Fig.18) indicates 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 there's some fluctuation 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 be slightly underestimating the actual quality performance of the system. The relatively close alignment between real and synthetic quality rates indicates that the doppelganger model is capturing the essential aspects of the quality dynamics in the sugar production process

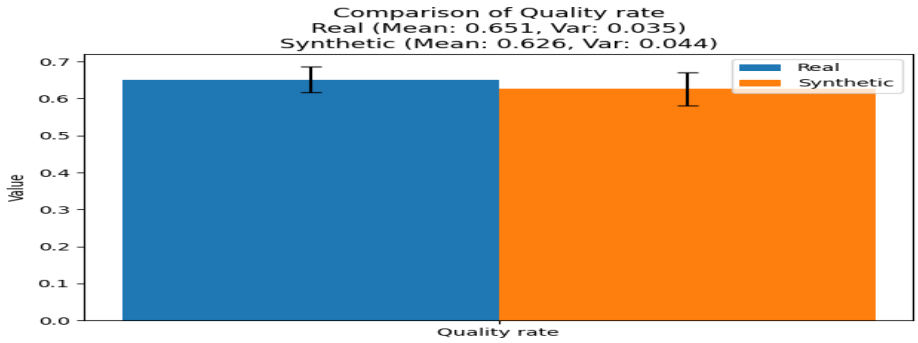

Fig.18. Statistics comparison of quality (Real vs. Synthetic)

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

Moreover, the autocorrelations plots (Fig.19) indicate that after the first few lags both synthetic and real data very weakly correlated to their past. Indicating that the temporal dynamics of the metric have been captured by the model.

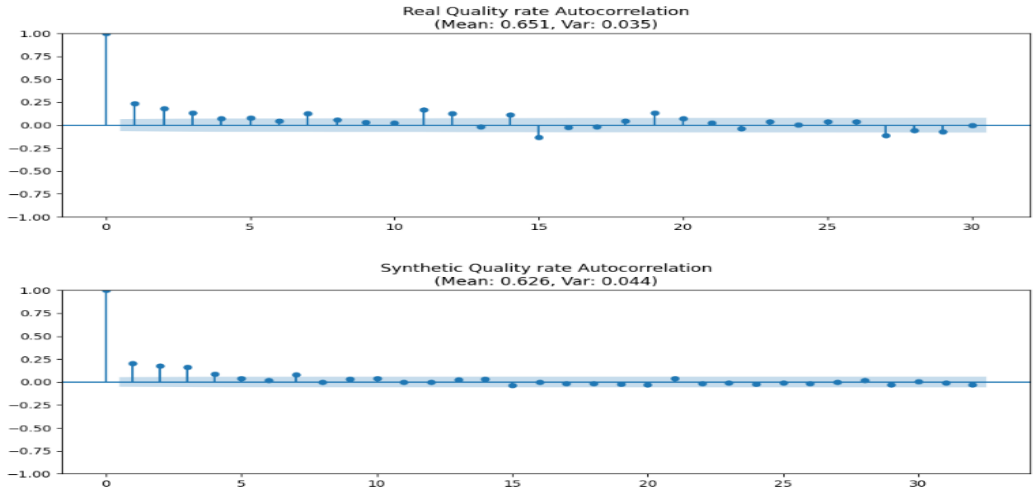

Fig.19. Quality autocorrelation plot

The line charts of the Figure 20 (Fig.20) confirms that the synthetic data seems to capture the upward and downward trends of 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, demonstrates the potential of the doppelganger model in capturing essential aspects of the 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.

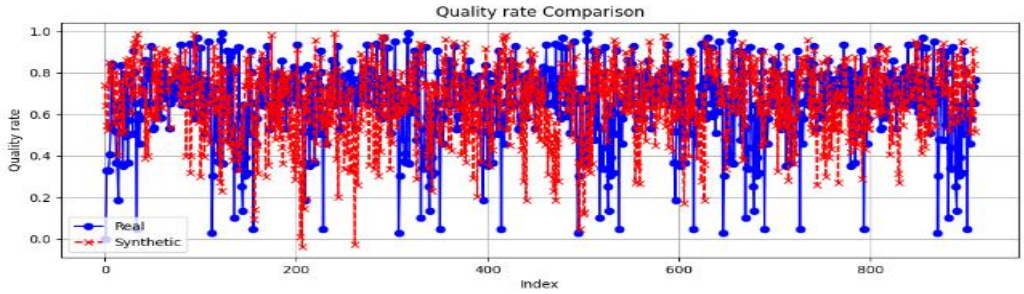

Fig.20. quality line charts (Real vs. Synthetic)

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

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

Table 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. Furthermore, 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 supplementary findings, when considered alongside the initial analysis of variance, mean, and autocorrelation, provide a more comprehensive assessment of the synthetic data's ability 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 presents 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.

Table 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

Table 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 indicate that the Doppelgänger model demonstrates the strongest ability to replicate the mean values of the real data across all three metrics, with its synthetic data closely matching the real data's average values, outperforming the LSTM models which show some discrepancies. Similar to the mean, Doppelgänger effectively replicates the variance of the real data, whereas 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 across all three metrics, indicating the highest accuracy in predicting the real data values, compared to both LSTM models which exhibit higher RMSE, suggesting less accurate predictions.. In summary, Doppelgänger outperforms the Vanilla LSTM and LSTM VAE models in generating synthetic data that closely resembles the real industrial performance data, demonstrating superior performance in terms of matching the mean and variance, and achieving the lowest RMSE across all metrics. These results highlight Doppelgänger's effectiveness in capturing both the statistical properties and temporal dependencies of the industrial performance data, making it a promising tool for generating high-fidelity synthetic data.

6. CONCLUSIONS AND PERSPECTIVES

This article makes several significant contributions in the field of complex industrial systems management among which:

Novel Modeling Approach**:** It proposes the use of generative doppelgangers, 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: The ability of doppelgangers to generate realistic synthetic data, the article 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: The article demonstrates how doppelgangers can be used to analyze various key performance indicators (KPIs) such as operational availability, 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 operations optimization.

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

In summary, this article proposes an innovative and promising approach to managing the complexity of stochastic data in industrial systems. The use of generative doppelgangers opens new perspectives for operations optimization, improved decision-making, and significant efficiency gains in the industry. However, the future direction of this work aim to explore the following potentials of generative doppelgangers in industrial systems:

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

Integration with Emerging Tech: The integration of doppelgangers with IoT, edge computing, and cloud platforms offers exciting possibilities. This could lead to more efficient data collection and processing, enabling faster and more accurate decision-making within industrial settings.

Enhanced Model Interpretability: Addressing the "black-box" nature of GANs is crucial for wider adoption in industry. Techniques to improve the interpretability of doppelganger

models will increase trust and understanding of their outputs, making them more valuable for decision-makers.

Advanced Evaluation Metrics: Developing more sophisticated metrics to evaluate the quality, diversity, and temporal accuracy of generated data will be essential. This will ensure that doppelgangers accurately capture the complexities of industrial processes and provide reliable insights.

Addressing Ethical & Societal Concerns: As with any powerful technology, the use of doppelgangers raises ethical and societal concerns. Addressing issues like data privacy, algorithmic bias, and potential job displacement will be crucial to ensure responsible and equitable deployment of this technology.

Overall, generative doppelgangers have the potential to transform how industries manage and optimize complex systems. Further research and development in this area, focusing on the perspectives mentioned above, could lead to significant advancements 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

Kuntalp, M., & Düzyel, O. (2024). A new method for GAN-based data augmentation for classes with distinct clusters. *Expert Systems with Applications*, *235*, 121199.

Wang, Y., & Yan, P. (2024). RegGAN: A Virtual Sample Generative Network for Developing Soft Sensors with Small Data. *ACS Omega*, *9*(10), 5954–5965.

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

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.

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.

Ren, L., Wang, H., Tang, Y., & Yang, C. (2024). AIGC for Industrial Time Series: From Deep Generative Models to Large Generative Models. *arXiv preprint arXiv:2407.11480*.

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.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Calix, R., Ugarte, O., Wang, H., & Okosun, T. (2024). A Dataset of CFD Simulated Industrial Furnace Images for Conditional Automatic Generation with GANs. In *TMS Annual Meeting & Exhibition* (pp. 775–783).

Zhou, R., Jiang, C., & Xu, Q. (2021). A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, *451*, 316–336.

Makhlouf, A., Maayah, M., Abughanam, N., & Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Computing and Applications*, *35*(24), 24055–24068.

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. *arXiv preprint arXiv:2311.10129*.

Hellmann, F., Mertes, S., Benouis, M., Hustinx, A., Hsieh, T.-C., Conati, C.,... & Kirchbuchner, F. (2024). Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *ACM Transactions on Multimedia Computing, Communications and Applications*.

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.

Hu, C., Sun, Z., Li, C., Zhang, Y., & Xing, C. (2023). Survey of Time Series Data Generation in IoT. *Sensors*, *23*(13), 6976.

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.

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*(10), 2387–2406.

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.

Kusiak, A. (2020). Convolutional and generative adversarial neural networks in manufacturing. *International Journal of Production Research*, *58*(6), 1594–1604.

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.

Kumarage, T., Ranathunga, S., Kuruppu, C., De Silva, N., & Ranawaka, M. (2019, April). Generative adversarial networks (GAN) based anomaly detection in industrial software systems. In *2019 Moratuwa Engineering Research Conference (MERCon)* (pp. 43–48). IEEE.

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X.,... & Liu, X. (2022). GAN-based anomaly detection: A review. *Neurocomputing*, *493*, 497–535.

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*.

Farady, I., Islam, J., Tuarob, S., Ng, H.-F., & Lin, C.-Y. (2023). GANs in Industrial Surface Defect Detection: Insights and Challenges. *Available at SSRN 4516131*.

Noor, S., Sajid, A., Khan, I., Javaid, J., & Tabasusum, I. (2023). Comparative Analysis of Anomaly Detection Techniques Using Generative Adversarial Network. *Sir Syed University Research Journal of Engineering & Technology (SSURJET)*, *13*.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017, June). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging* (pp. 146–157). Springer, Cham.

Zhang, H., Dereck, S. S., Wang, Z., Lv, X., Xu, K., Wu, L.,... & Huang, G. B. (2023). Large scale foundation models for intelligent manufacturing applications: a survey. *arXiv preprint arXiv:2312.06718*.

Rezaei, S., Cornelius, A., Karandikar, J., Schmitz, T., & Khojandi, A. (2024). Using GANs to predict milling stability from limited data. *Journal of Intelligent Manufacturing*, 1–35.

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*(22), 19147–19159.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.,... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

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*(14), 3001–3014.

Ye, Y., Yong, Z., & Han, D. (2020). Research on key technology of industrial artificial intelligence and its application in predictive maintenance. *Acta Automatica Sinica*, *46*(10), 2013–2030.

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*.

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

Radford, A. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

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.

Antonucci, D., Conselvan, F., Mascherbauer, P., Harringer, D., & Pozza, C. (2024). Synthetic data on buildings. In *Machine Learning Applications for Intelligent Energy Management: Invited Chapters from Experts on the Energy Field* (pp. 203–226). Springer.

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.

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

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733.

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*.

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.

AIGC for Industrial Time Series: From Deep Generative Models to Large Generative Models. (2024). *arXiv preprint arXiv:2407.11480*.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020). DoppelGANger: Generating High-Fidelity Synthetic Time Series Data. *arXiv preprint arXiv:2003.03453*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, *32*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.