2024 IEEE 29th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD) | 979-8-3503-7764-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CAMAD62243.2024.10942898

Optimizing Industry 5.0 Machine Learning-Based Applications via Synthetic Data Generation

Lorenzo Colombi*, Matteo Brina*, Michela Vespa*, Filippo Tabanelli*, Simon Dahdal*,
Elena Bellodi*, Riccardo Venanzi†, Mauro Tortonesi*, Massimiliano Vignoli‡, Cesare Stefanelli*
* University of Ferrara, Ferrara, Italy, Email: {firstname.lastname}@unife.it
† University of Bologna, Bologna, Italy, Email: riccardo.venanzi@unibo.it
‡ Bonfiglioli Group, Calderara di Reno, Italy, Email: massimiliano.vignoli@bonfiglioli.com

***Abstract*—Machine Learning (ML) innovations are revolutionizing industrial processes by improving productivity and competitiveness, particularly by creating predictive maintenance applications and Artificial Intelligence (AI) powered services. In the Industry 5.0 paradigm, which strives to reach Zero Defect and Zero Waste Manufacturing, the role of ML is crucial for optimizing these processes. However, developing ML solutions for production-ready industrial settings presents various challenges, such as data scarcity and dataset imbalance, which are further amplified when dealing with tabular data. To overcome such issues we investigate the use of various Deep Generative Models (DGMs) to generate synthetic data that closely mimics real-world conditions. Despite extensive studies of DGMs in the literature, there is still limited knowledge of their practical suitability in Industry 5.0 environments. Our research includes a detailed evaluation of several DGMs, such as Generative Adversarial Networks, Variational Autoencoders, and Diffusion Models, implemented in the gearbox assembly and testing line in the Bonfiglioli EVO plant. Based on our results, the evaluated DGMs demonstrate significant potential in generating high-quality synthetic data, that allows training a high-performance classifier to distinguish faulty gearboxes from well-functioning ones.**

***Index Terms*—Industry 5.0, Machine Learning, Deep Generative Model, Generative Adversarial Network, Diffusion Model, Variational Autoencoder.**

I. Introduction

Innovations in Machine Learning (ML) enable businesses to boost productivity and gain competitive advantages [1], [2] by predicting equipment failures [3] and enhancing supply chain efficiency through real-time forecasting. ML applications and associated Operations (MLOps) optimize industrial processes by leveraging data-driven insights to refine production methods, improve product quality, and increase operational efficiency [4]. Research is advancing toward the Industry 5.0 paradigm, which includes Zero Defect Manufacturing and Zero Waste Manufacturing, aiming for more sustainable processes [5]–[9]. Lastly, ML has also found significant applications in network management [10] and security, optimizing resource allocation, and detecting potential security threats in real-time [11].

Maximizing the potential of ML in industrial settings requires addressing the specific issues posed by these environments. A significant challenge for effective ML adoption in Industry 5.0 applications is imbalanced and skewed datasets. This imbalance often arises from the scarcity of data on faulty components, given the high output quality of manufacturing processes [12]. Moreover, most ML models assume access to extensive and balanced datasets, and imbalance could lead to limited or biased models [13]. This issue arises, of course, because manufacturing companies cannot be expected to tackle it by deliberately causing component failures or operational disruptions for data generation, as this approach is costly and risky.

A possible solution is rebalancing datasets through augmentation by generating synthetic faulty component data using Deep Generative Models (DGMs) [14]. These models can learn the distribution of a dataset and generate new synthetic examples following that distribution. When well-trained, the synthetic dataset will have a distribution similar to the real one, making its examples nearly indistinguishable from real data. However, improperly trained DGMs may overfit, memorizing and reproducing fixed patterns [15], especially when trained on highly unbalanced datasets. To address this risk, synthetic datasets are also evaluated using privacy metrics.

In addition, industrial settings often require the analysis of tabular data that includes continuous and categorical attributes, presenting a unique complexity in data handling and requiring innovative modelling approaches. Continuous variables in these datasets tend to show heavily skewed distributions, making it difficult to model and authentically reproduce the data, which is fundamental for predicting rare events like equipment failures. For these reasons, training a high-quality model for tabular data can be more challenging than for computer vision or NLP tasks because of the diversity of individual features and the typically smaller sizes of tabular datasets [16].

Although many DGMs have been proposed in the literature, there is limited knowledge about their suitability for Industry 5.0 applications. This paper explores the implementation of different DGMs to address data scarcity within a challenging industrial scenario. The study presents a comprehensive solution tailored for the gearbox assembly and testing line at the Bonfiglioli EVO plant. The proposed solution aims to enhance the Bonfiglioli pre-testing phase of a real assembly line at the EVO plant, where ML models are employed to classify gearboxes as faulty or well-functioning, by using synthetic data generation.

We performed an extensive evaluation of the synthetic data generated by the different DGMs by first measuring

979-8-3503-7764-4/24/$31.00 ©2024 IEEE

their quality using a variety of metrics, including those assessing data distribution similarity and privacy. Specifically, we tested Conditional Tabular and Wasserstein Generative Adversarial Networks (CTGAN and WGAN), Tabular Variational AutoEncoder (TVAE) and Denoising diffusion probabilistic models (TabDDPM). Then, we employed a logistic regression model to perform the classification of gearboxes. The latter was also used to evaluate ML utility metrics, such as precision and recall. This allowed us to assess whether the synthetic data improved the model's performance compared to the baseline training done on the unbalanced dataset.

In summary, the primary contributions of this paper are as follows: (*i*) we evaluated several DGMs in a real industrial environment to establish which ones were better suited for our purpose; (*ii*) we show how synthetic data generation can improve ML model performance, in our case a binary classifier, resulting in higher efficiency and sustainability of the manufacturing process.

The paper is organized as follows: Section II introduces background and related work on DGMs and evaluation metrics. Section III describes the industrial use case at Bonfiglioli and our proposed solution. Section IV presents the experimental evaluation of different DGMs to effectively balance unbalanced industrial datasets. Section V concludes the paper and outlines future work.

II. Background & Related Work

Thanks to recent advancements in deep learning, the extensive data collected from industrial plant sensors has become a major driving force. However, as modern data-driven and ML-powered applications are adopted across various industries, the reliance on substantial, high-quality tabular data becomes crucial. Challenges such as data incompleteness, poor quality, and inadequate quantity can pose significant obstacles [17]. To address these issues, various solutions are available in the literature [18], encompassing both data-centric and data-generation approaches.

Generative AI has proven to be a revolutionary field, particularly with the advent of Deep Generative Models. These cutting-edge models, which combine generative algorithms with deep learning, have remarkable capabilities to create new, realistic data samples that mimic the features and patterns of the training data, known as synthetic data. The primary objective of training a DGM is to learn an unknown probability distribution from a typically limited number of independent and identically distributed samples. Once successfully trained, a DGM can be used to evaluate the likelihood of a given sample and generate new samples that resemble those from the unknown distribution [19]. Various architectural designs of Deep Generative Models (DGMs) can be found in the literature, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Energy-based models, Autoregressive models, and Diffusion models.

The Synthetic Minority Over-sampling Technique (SMOTE) [20] is a pioneering algorithm designed to address class imbalance in machine learning datasets. It works by generating synthetic observations of the minority class based on the K-NN algorithm. The ADASYN (Adaptive Synthetic Sampling) [21] algorithm helps balance imbalanced datasets by focusing specifically on generating more synthetic observations for those parts of the minority class that are harder for models to learn.

Diffusion models [22], [23] are a type of DGM that transforms data from a simple distribution into a complex one by progressively adding noise and then learning to reverse this process. They function through two primary processes: during the forward process noise is incrementally added to the data over multiple steps, eventually turning it into a simple distribution like Gaussian Noise; during the reverse process, the model learns how to systematically remove this noise and gradually refining it to approximate the real data. Specifically for tabular data, the Tabular Data Diffusion Probabilistic Model (TabDDPM) [16] applies this method to create synthetic data samples. By gradually introducing and then removing noise, TabDDPM effectively learns the distribution of tabular datasets.

Variational Autoencoders (VAEs) have also been exploited to address synthetic data generation. In particular, the Tabular Variational Autoencoder (TVAE) [24] is a specific adaptation of the traditional VAE designed to handle tabular data efficiently. TVAE leverages the principles of VAEs to learn the distribution of real data and generate synthetic data that maintains statistical similarities with the real dataset.

Various types of GANs have been introduced in the literature, including Deep Convolutional GANs, Conditional GANs, Pix2Pix GANs, Cycle GANs, and others [25]. One notable variant is the Wasserstein GAN (WGAN), introduced by Arjovsky et al. in 2017 [26]. WGAN utilizes the Wasserstein distance as a loss function, which stabilizes training and reduces the mode collapse phenomenon. The WGAN approach provides a more meaningful measure of convergence and better insights into generator performance.

Another significant GAN-based model is the Conditional Tabular GAN (CTGAN) [24], designed to generate synthetic tabular data. CTGAN addresses data scarcity issues and mitigates mode collapse by introducing 'mode-specific normalization' for processing continuous features. This technique involves fitting a variational Gaussian mixture model to each feature and normalizing values using the corresponding Gaussian component's mean and variance. CTGAN also enhances the generation of categorical features by modifying the GAN's loss function to include the cross-entropy between the one-hot encoding of input and generated data. Additionally, CTGAN's Training-by-Sampling process replicates the real dataset's feature distribution through strategic data sampling and conditional vector construction. Despite their advantages, GANs face limitations, particularly in generating discrete data accurately. Misclassified data points in the training dataset, common in industrial settings, can negatively impact the learning process of CTGANs, leading to increased false positives and false negatives. This underscores the importance of clean and accurately classified training data for the effective performance of CTGANs [27].

Ensuring that the generated synthetic data is high quality and closely resembles the real data is essential. This requires reliable evaluation metrics to measure the similarity (or distance) between synthetic and real datasets [28], [29]. The evaluation of synthetic data is usually established using three different metrics: (*i*) ML utility, (*ii*) statistical similarity, and (*iii*) privacy preservation. The first two metrics assess the synthetic data's effectiveness as a substitute for the real data. ML utility in particular helps to determine how well a model works in terms of prediction accuracy when trained on synthetic or altered datasets. Examples of such metrics are Accuracy, Precision, Recall, and F1-Score. Metrics for statistical similarity assess how closely synthetic data mirrors the statistical properties of the real data. The Kolmogorov-Smirnov Distance (KS-D) is used to compare two probability distributions by quantifying the maximum difference between their Cumulative Distribution Functions, quantifying how much two distributions diverge. On the other hand, the KS-complement is calculated as $1-$(KS-D), with a higher value indicating greater similarity, thus providing a method to assess how closely the synthetic data matches the real data. Similarly, the Wasserstein Distance (WD) can be used to compare two probability distributions by measuring the minimum cost to transform one distribution into the other, providing another metric to capture how closely the generated data matches the real data. The Correlation Similarity (CS) measures the similarity between two datasets by computing the Pearson Correlation Coefficient between corresponding elements, quantifying the linear relationship between the real and the synthetic data. Lastly, the Jensen-Shannon Divergence (JS) is calculated by taking the average of two distributions and measuring how much each distribution diverges from this average using the Kullback-Leibler Divergence (KL). Additionally, the Distance to Closest Record (DCR) is employed to measure the average distance between each synthetic data point and its nearest neighbor in the real data. A low DCR indicates that the synthetic data is very close to the real data, which may be undesirable if the goal is to generate diverse, high-quality synthetic data that captures the full distribution of the real data.

III. Industrial Use Case

Bonfiglioli S.p.A. (https://bonfiglioli.com) is a prominent manufacturing company specializing in the design and production of a comprehensive range of gear motors, drive systems, planetary gear motors, reducers, and inverters, boasting over 130 years of experience. Bonfiglioli is progressively aligning with Industry 5.0 best practices, integrating efficient and environmentally sustainable processes.

Among the various manufacturing lines at Bonfiglioli, the gearbox assembly and testing line at the EVO Plant stands out. This line employs advanced machinery for automated precision assembly and rigorous testing processes. The assembly line comprises three workstations (WS), as depicted in Fig. 1. WS1—Differential Assembly is responsible for assembling the differential part of the gearbox. This workstation collects data on insertion forces and tightening torque. WS2—Gearbox Assembly focuses on assembling the entire gearbox and gathers data on insertion forces and tightening torque. Lastly, WS3 - End-of-Line Testing is the final workstation, where the assembled gearbox undergoes testing.

WS3 is further divided into two distinct machines. The first machine generates cycle data through specific stress tests on the component under examination, performing two phases of static analysis at 800 revolutions per minute (RPM) and two phases of ramp analysis at 11,000 RPM. These tests are crucial for evaluating the component's performance under varying operational conditions. Concurrently, the second machine monitors vibrational data, tracking vibration levels at different RPM thresholds. This data is essential for understanding the vibrational characteristics of the component, ensuring its reliability and structural integrity.

To enhance the efficiency of the Bonfiglioli EVO gearbox assembly and testing line, a pre-testing phase has been integrated between WS2 and WS3. This phase offers substantial benefits in terms of sustainability and cost reduction. The pre-testing phase is an early filter to identify defects, thereby conserving resources and energy that would otherwise be used in the resource-demanding final testing phase. Detecting quality issues at an earlier stage minimizes waste through rectification or recycling, aligning with sustainable manufacturing practices and improving overall product quality.

Implementing an effective pre-testing phase necessitates addressing the challenge of training an accurate classifier using a highly imbalanced dataset collected from the WS1 and WS2 machines.

The dataset collected from WS1 and WS2 focuses on two processes: tightening torques and press-fit forces. The dataset comprises approximately 700 features critical to the assembled gearbox's functionality. Through a rigorous evaluation process involving domain experts, 66 key features were identified and prioritized based on their influence on the gearbox's performance and readiness. Using these input features and the output from the WS3 testing machine, which assesses faulty and well-functioning components, we constructed the final dataset for training and testing the pre-test classifier.

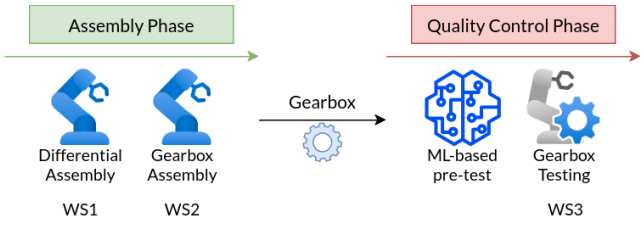


Fig. 1: Bonfiglioli gearbox assembly line at the EVO plant.

To classify faulty from well-functioning gearboxes we trained a logistic regression model as a binary classifier. However, since this preliminary experiment on the original unbalanced dataset resulted in suboptimal performance, we decided to apply several well-known DGMs to generate synthetic examples belonging to the minority class (faulty

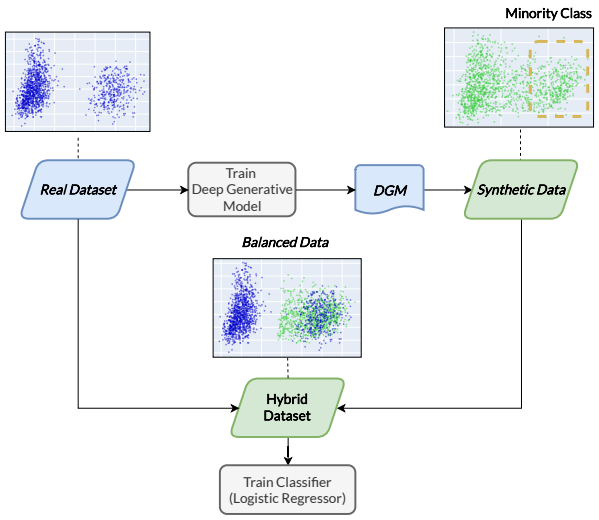


Fig. 2: Diagram of the synthetic data generation and integration process used to balance the real dataset.

gearboxes), following the workflow in Fig. 2. In detail, we tested many widely used DGMs such as GANs (CTGAN and WGAN), Diffusion Models (TabDDPM), and a Variational Autoencoder (TVAE).

Concerning the WGAN, it was built from scratch using Keras and TensorFlow. The number of layers and neurons in each layer, together with other hyperparameters, such as epochs, noise dimension and batch size, were thoroughly searched using a hyperparameters optimization tool configured to minimize the Kolmogorov-Smirnov distance between the distribution of real data and generated data. This choice was made due to the extensive array of hyperparameters and their substantial influence on the model performance, particularly when training GANs. Specifically, we used Optuna (https://optuna.readthedocs.io/), a robust and comprehensive hyperparameter optimization tool. Optuna facilitates the systematic exploration and fine-tuning of these parameters, ensuring that the model's performance is maximized. CTGAN and TVAE are part of the Synthetic Data Vault (SDV) framework (https://sdv.dev/), a comprehensive Python library tailored for generating tabular synthetic data that offers various synthesizers. SDV models are accessible using a convenient Python API that allows defining, or automatically finding, the metadata associated with the real dataset. This metadata includes, for example, the number and the value type of each column, and is used to improve the training process.

Lastly, to test TabDDPM we used the official implementation provided by the authors [16], that again uses Optuna to select the best hyperparameters.

IV. Experimental Results

A. Setup

To evaluate the quality of the generated datasets we calculated several statistical metrics which measure the similarity between the synthetic and the real dataset. We used the Kolmogorov-Smirnov Complement (KS-complement), which ranges from 0 to 1 with values closer to 1 indicating higher similarity; Wasserstein Distance, where smaller values denote greater similarity; Correlation Similarity, which also ranges from 0 to 1 with values closer to 1 indicating stronger pairwise correlations; and Jensen-Shannon Divergence, bounded by 1, with 0 representing identical distributions. These metrics, as highlighted in Section II, are used to assess synthetic data quality because they provide a comprehensive evaluation of how well the synthetic data replicates the characteristics of the real data. Moreover, we used a distance metric, specifically the Distance to Closest Record, to determine the generator's overfitting. This choice was made because DCR is intuitive and easy to understand [30]. Specifically, DCR compares the smallest real-to-synthetic distance (RSD) and the real-to-real distance (RRD). We identified and counted the number of suspiciously similar synthetic examples as shown in [15]. Precisely, we mark the example $\hat{d}$ as suspicious when $\mathrm{RSD}(\hat{d}) < \mathrm{RRD}(d^*)$, with $d^* := \arg\min_{d \in D} \mathrm{Dist}(\hat{d}, d)$, where D is the real dataset [15] and Dist(a, b) function returns the Euclidean distance between a and b. Then, we calculated the DCR ratio (DCRR) over the total number of examples in the synthetic dataset.

Subsequently, we evaluated data quality by training a Logistic Regression model on the real dataset balanced with the synthetic data belonging to the minority class, which represents faulty gearboxes. We tested the data generated by each of the DGMs (WGAN, CTGAN, TVAE, TabDDPM) in addition to SMOTE and ADASYN. To do so, we compared the logistic regressor trained on hybrid (real + synthetic) datasets with the same model trained exclusively on real data. In detail, we filtered the synthetic dataset generated by DGMs, keeping only examples belonging to the minority class (faulty gearboxes). This was accomplished in an unsupervised way due to the presence of mislabeled examples in the real dataset's training set. We applied the K-means algorithm to the real dataset to obtain the coordinates of the centroids of two clusters. Each cluster represents a dataset class: faulty or well-functioning. We then used these centroids to filter the synthetic data accordingly. Synthetic data is filtered by retaining at least N examples closest to the centroid, within a maximum distance of d_{max}. N represents the difference between the number of examples in the majority and minority classes in the original dataset, and d_{max} is the empirically chosen maximum distance a point can have from the centroid to be classified as belonging to the minority class. After obtaining this filtered dataset, we split the real dataset into training and test sets. The training set is then balanced with the previously filtered synthetic examples, resulting in a hybrid dataset composed of both real and synthetic examples.

B. Results

Table I provides a comparative analysis of all the tested DGMs across several metrics designed to measure data quality and similarity to real datasets. On the other hand

Table II shows the performance of a logistic regression model trained on the real dataset and on different artificially balanced datasets. Together these two tables aim to evaluate the quality of the generated data.

TABLE I: Synthetic data quality comparison of WGAN, CTGAN, TVAE, TabDDPM using different metrics: Kolmogorov-Smirnov Complement (KS-C), Correlation Similarity (CS), Wasserstein Distance (WD), Jensen-Shannon Divergence (JS), Distance to Closest Record Ratio (DCRR). In bold the highest scores.

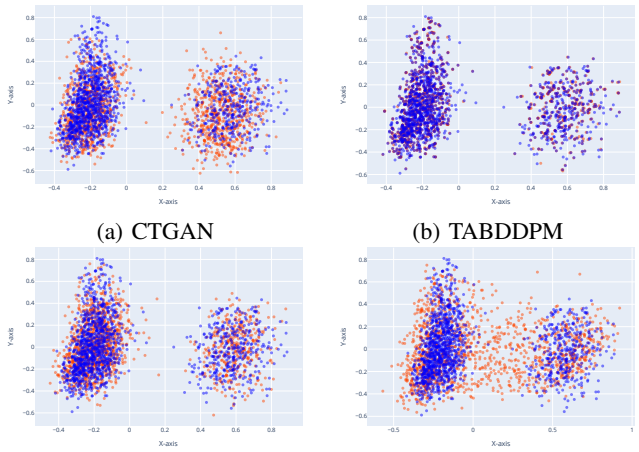
	KS-C	CS	WD	JS	DCRR
WGAN	0.795	**0.946**	0.024	0.300	21.590
CTGAN	0.828	0.940	0.021	0.246	**7.732**
TVAE	0.862	0.945	0.011	0.227	21.736
TabDDPM	**0.873**	0.937	**0.005**	**0.222**	99.198

TABLE II: Performance of a Logistic Regression model trained on the real dataset and on the real dataset balanced with synthetic examples generated by WGAN, CTGAN, TVAE, TabDDPM, SMOTE and ADASYN. In bold the highest scores.

	Precision	Recall	F1-Score	F2-Score
UNBALANCED	0.65	0.81	0.72	0.77
WGAN	0.64	1.00	0.78	0.90
CTGAN	**0.67**	**1.00**	**0.80**	**0.91**
TVAE	0.63	0.95	0.76	0.86
TabDDPM	0.64	1.00	0.78	0.90
SMOTE	0.65	0.98	0.78	0.89
ADASYN	0.65	0.98	0.78	0.89

Moreover, to visually compare the synthetic datasets, we employed dimensionality reduction techniques. Specifically, the Principal Component Analysis (PCA) algorithm was used to reduce the dataset's dimensionality from 66 features to 2 dimensions. As shown in Fig. 3 the PCA applied to the real dataset (represented in blue) successfully formed two distinct clusters: the denser left cluster represents the good gearboxes, while the right cluster indicates the broken gearboxes. The same happens, with some differences with the synthetic datasets.

Based on the results from Table I, CTGAN and TVAE both demonstrate strong performance across all similarity metrics; still, the former shows the best score for the DCRR, ensuring that individual data points are not closely replicated. Following the results in Table II, CTGAN stands out as the best overall model for our task. Instead, TabDDPM shows a high similarity (KS-C), very low Wasserstein Distance, and a high DCRR, suggesting it generates highly similar synthetic data with potential overfitting. This observation is further supported by Fig. 3b, where there is a noticeable overlap of points between the synthetic and real datasets, illustrating how closely the synthetic data mirrors the real data. The WGAN presents a reasonable balance between similarity metrics and divergence while providing average scores in ML performance metrics. However, as shown in Fig. 3d, the synthetic data distribution (in orange) does not properly follow the real dataset (in blue). This discrepancy is undesirable because it suggests that the synthetic data fails to capture the true underlying patterns and characteristics of the real data.



(a) CTGAN (b) TABDDPM

(c) TVAE (d) WGAN

Fig. 3: Visualization of synthetic and real datasets after the dimensionality reduction. Real examples are represented in blue, and synthetic in orange.

Lastly, let us point out that in previous work we used the CTGAN's conditional output to generate only examples belonging to the minority class instead of the whole dataset distribution, obtaining significantly lower performance [31]. This once again demonstrates the worth of the unsupervised filtering approach.

V. Conclusions and Future Work

Our extensive evaluation of DGMs, specifically CTGAN, TVAE, TABDDPM, and WGAN, on a real industrial use case provided by Bonfiglioli, demonstrates substantial differences in their performance when employed to generate high-fidelity synthetic data. Therefore, we can conclude that selecting the appropriate tool for the task is essential.

Building on the promising results of this study, several avenues for future work could explore the applicability of DGMs in other use cases provided by different industrial partners. DGMs could also generate enough data for timely training, considerably reducing the time needed to train an ML model. For example, generating a training set using these technologies could be possible starting from just a few examples of broken components. Another future work could be exploring different techniques, such as unsupervised anomaly detection, to reach the same goal of improving the manufacturing plant efficiency.

Acknowledgements

Research funded by the Italian Ministry of University and Research through PNRR, Mission 4, Component 2, Investment 1.3, Partenariato Esteso PE00000013 – "FAIR: Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI" – CUP J33C22002830006 (Grant Assignment Decree n. 341 adopted 15/03/2022), and Investment 1.4, Call for

tender No. 1409 published on 14/9/2022, "National Centre for HPC, Big Data and Quantum Computing (HPC)" – CUP D43C22001240001 (Grant Assignment Decree No. 1031 adopted on 17/06/2022), under NextGeneration EU programme grants.

REFERENCES

[1] D. Kiel *et al.*, "Sustainable industrial value creation: Benefits and challenges of industry 4.0," *International Journal of Innovation Management*, vol. 21, no. 08, p. 1740015, 2017. doi: 10.1142/S1363919617400151. [Online]. Available: https://doi.org/10.1142/S1363919617400151

[2] C. Hegedűs *et al.*, "Enabling Scalable Smart Vertical Farming with IoT and Machine Learning Technologies," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, 2023. doi: 10.1109/NOMS56928.2023.10154269 pp. 1–4.

[3] J. L. Herrera *et al.*, "A machine learning-based framework to estimate the lifetime of network line cards," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, 2020. doi: 10.1109/NOMS47738.2020.9110455 pp. 1–5.

[4] A. Frankó *et al.*, "Applied Machine Learning for IIoT and Smart Production Methods to Improve Production Quality, Safety and Sustainability," *Sensors*, vol. 22, no. 23, 2022. doi: https://doi.org/10.3390/s22239148

[5] L. Colombi *et al.*, "A machine learning operations platform for streamlined model serving in industry 5.0," in *NOMS 2024-2024 IEEE Network Operations and Management Symposium*, 2024. doi: 10.1109/NOMS59830.2024.10575103 pp. 1–6.

[6] R. Venanzi *et al.*, "Enabling adaptive analytics at the edge with the Bi-Rex Big Data platform," *Computers in Industry*, vol. 147, p. 103876, 2023. doi: https://doi.org/10.1016/j.compind.2023.103876

[7] G. Dimitrakopoulos *et al.*, "Industry 5.0: Research areas and challenges with artificial intelligence and human acceptance," *IEEE Industrial Electronics Magazine*, pp. 2–13, 2024. doi: 10.1109/MIE.2024.3387068

[8] K. Duran *et al.*, "Digital twin enriched green topology discovery for next generation core networks," *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 4, pp. 1946–1956, 2023. doi: 10.1109/TGCN.2023.3282326

[9] L. Colombi *et al.*, "Multivariate time series anomaly detection in industry 5.0," in *ITADATA 2024-The 3rd Italian Conference on Big Data and Data Science*, 2024.

[10] K. Duran *et al.*, "Digital twin-native ai-driven service architecture for industrial networks," in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 1297–1302.

[11] A. G. Avran *et al.*, "Securing southbound interface in sdns: Utilizing support vector machines for openflow packet classification," in *2023 IEEE 28th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2023. doi: 10.1109/CAMAD59638.2023.10478400 pp. 258–263.

[12] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, 2023. doi: 10.1186/s40537-023-00727-2

[13] T. B. Nyíri *et al.*, "What can we learn from Small Data," *Infocommunications Journal*, vol. 15, no. SI, pp. 27–34, 2023. doi: 10.36244/ICJ.2023.5.5

[14] S. Bond-Taylor *et al.*, "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, p. 7327–7347, Nov. 2022. doi: 10.1109/tpami.2021.3116668

[15] A. T. P. Boudewijn *et al.*, "Privacy measurements in tabular synthetic data: State of the art and future research directions," in *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

[16] A. Kotelnikov *et al.*, "Tabddpm: Modelling tabular data with diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17 564–17 579.

[17] Y.-T. Chen *et al.*, "On the Private Data Synthesis Through Deep Generative Models for Data Scarcity of Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 551–560, 2023. doi: 10.1109/TII.2021.3133625

[18] Z. Pödör *et al.*, "A Practical Framework to Generate and Manage Synthetic Sensor Data," *Infocommunications Journal*, vol. XIV, no. 2, pp. 64–72, June 2022. doi: 10.36244/ICJ.2022.2.7

[19] L. Ruthotto *et al.*, "An introduction to deep generative modeling," 05 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/gamm.202100008

[20] N. V. Chawla *et al.*, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002. doi: 10.1613/jair.953. [Online]. Available: http://dx.doi.org/10.1613/jair.953

[21] H. He *et al.*, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008. doi: 10.1109/IJCNN.2008.4633969 pp. 1322–1328.

[22] J. Sohl-Dickstein *et al.*, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 2256–2265.

[23] J. Ho *et al.*, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[24] L. Xu *et al.*, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[25] Y. Li *et al.*, "The theoretical research of Generative Adversarial Networks: an overview," *Neurocomputing*, vol. 435, pp. 26–41, 2021. doi: https://doi.org/10.1016/j.neucom.2020.12.114

[26] M. Arjovsky *et al.*, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup *et al.*, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[27] O. Habibi *et al.*, "Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105669, 2023. doi: https://doi.org/10.1016/j.engappai.2022.105669

[28] J. Fonseca *et al.*, "Tabular and latent space synthetic data generation: a literature review," *Journal of Big Data*, vol. 10, no. 1, p. 115, Jul 2023. doi: 10.1186/s40537-023-00792-7. [Online]. Available: https://doi.org/10.1186/s40537-023-00792-7

[29] M. Hernandez *et al.*, "Standardised metrics and methods for synthetic tabular data evaluation," *Authorea Preprints*, 2023.

[30] A. Kunar, "Effective and privacy preserving tabular data synthesizing," 2021. [Online]. Available: https://arxiv.org/abs/2108.10064

[31] S. Dahdal *et al.*, "An mlops framework for gan-based fault detection in bonfiglioli's evo plant," *Infocommunications Journal*, vol. 16, no. 2, pp. 2–10, 2024. doi: https://doi.org/10.36244/ICJ.2024.2.1