Artificial Intelligence Review (2025) 58:32
https://doi.org/10.1007/s10462-024-11021-9

Data augmentation in predictive maintenance applicable to hydrogen combustion engines: a review

Alexander Schwarz[1,2] · **Jhonny Rodriguez Rahal**[1,2] · **Benjamín Sahelices**[1] · **Verónica Barroso-García**[3,4] · **Ronny Weis**[2] · **Simon Duque Antón**[2]

Accepted: 6 November 2024 / Published online: 3 December 2024

Abstract
Machine-learning-based predictive maintenance models, i.e. models that predict breakdowns of machines based on condition information, have a high potential to minimize maintenance costs in industrial applications by determining the best possible time to perform maintenance. Modern machines have sensors that can collect all relevant data of the operating condition and for legacy machines which are still widely used in the industry, retrofit sensors are readily, easily and inexpensively available. With the help of this data it is possible to train such a predictive maintenance model. The main problem is that most data is obtained from normal operating conditions, whereas only limited data are from failures. This leads to highly unbalanced data sets, which makes it very difficult, if not impossible, to train a predictive maintenance model that can detect faults reliably and timely. Another issue is the lack of available real data due to privacy concerns. To address these problems, a suitable data generation strategy is needed. In this work, a literature review is conducted to identify a solution approach for a suitable data augmentation strategy that can be applied to our specific use case of hydrogen combustion engines in the automotive field. This literature review shows that, among the different state-of-the-art proposals, the most promising for the generation of reliable synthetic data are the ones based on generative models. The analysis of the different metrics used in the state of the art allows to identify the most suitable ones to evaluate the quality of generated signals. Finally, an open problem in research in this area is identified and it is the need to validate the plausibility of the data generated. The generation of results in this area will contribute decisively to the development of predictive maintenance models.

Keywords Data augmentation · Predictive maintenance · Anomaly detection · Generative models

Extended author information available on the last page of the article

1 Introduction

In industrial environments it is critical that the machines in the production lines work continuously. An unplanned stop in production even for a short time can lead to significant losses. A study from Thomas and Weiss (2020) collected data from several U.S. manufacturers to determine the costs that could be avoided through a good maintenance strategy. These costs arise not only from direct maintenance costs, but also from the production losses. The reputation of a company can suffer greatly as well if delivery deadlines cannot be met due to a production stoppage which can furthermore lead to the loss of follow-up orders. The total cost that could be avoided was $119.1 billion. $18.1 billion of this can be attributed to failures and downtimes, $0.8 billion to defects and the remaining $100.2 billion is caused by contracts and deliveries that could not be fulfilled.

Companies perform maintenance to prevent or, once they occur, repair defects that would negatively impact production. According to Thomas and Weiss (2020) and Wen et al. (2022), different existing maintenance approaches can be categorized into the following three classes. *Reactive maintenance*, also known as corrective or failure-driven maintenance, is typically performed in response to equipment malfunctions or breakdowns. This approach is also employed when machinery fails to meet expected quality or production targets. *Preventive maintenance* is conducted on a regular basis, according to predefined intervals that are easily monitored. These intervals may be based on a fixed amount of time, the number of produced parts, machine cycles, or other parameters. The maintenance schedule is typically developed by experts with experience and an understanding of the historical breakdowns or failures of the machinery in question. *Predictive maintenance* entails measuring the reliability and condition of a given piece of machinery, a workcell, an assembly line, etc., or a manufacturing process itself. These measurements are frequently obtained through the use of sensors that capture data, which can then be combined with historical data in order to assess the current condition and inform maintenance decisions.

In his study, Thomas (2018) conducted a comparison between these three types of maintenance. The comparison demonstrates that reactive maintenance can be a cost-effective approach when the initial cost of equipment is low, it is easily replaceable, has high availability, has minimal impact on collateral failures, or has high redundancy. In contrast, preventive maintenance is cost-effective for equipment in process chains where the different parts rely on each other. However, there is a potential risk of over-maintenance, which can lead to excessive production downtimes. Predictive maintenance mitigates the risk of over-maintenance and downtimes by identifying the optimal time for maintenance, making it a more cost-effective alternative to preventive maintenance. However, it requires a higher upfront investment due to the hardware and software needed to capture and monitor the necessary data, as well as the training of personnel on monitoring techniques and data analysis.

Predictive maintenance can be divided into rule-based approaches which need expert knowledge and data-driven approaches. The data-driven approaches can be based on statistical or machine learning models. The trend in research of data-driven methods for predictive maintenance is towards machine learning. A literature research from Wen et al. (2022) shows that 60% of the publications about data-driven predictive maintenance approaches in the years 2015–2020 use some sort of machine learning model.

Murphy (2012) identifies three principal categories of machine learning: supervised, unsupervised and reinforcement learning. Supervised learning uses a set of labeled data

samples to train a model that can predict a value for unseen samples according to the labels. In the case that the labels represent a categorical variable the task is called classification and if the labels are continuous values, the task is known as regression. Unsupervised learning trains a model only using inputs, as there are no labels. The task is to find useful patterns in the data. Reinforcement learning works in the way that an agent learns an optimal strategy, called policy through interactions with its environment for which he receives positive or negative feedback signals. Reinforcement learning is not considered in this study, since supervised and unspervised approaches are more established in predictive maintenance.

In predictive maintenance tasks, unsupervised learning can be used to detect anomalies. In this case a model can be trained, e.g. an autoencoder, using only data of the normal behavior. When the model is used for inference, it can detect data samples that differ from the training data as anomalies. In contrast, supervised learning models can be utilized not only for the detection of anomalous conditions, but also for the identification of the specific fault that occurred, provided that the model is trained with labeled data that reflects the various conditions (Chandola et al. 2009).

These supervised models require large amounts of training data that needs to be captured by different sensors directly at the machines. One of the biggest problems with predictive maintenance models is that there is often not enough data available for quality training. This applies in particular to data on faulty behavior or defects resulting in imbalanced datasets. The reason is that faulty behavior and defects occur rather rarely during the lifetime of machines. To counteract this imbalanced data problem a good strategy to create additional data samples of fault cases is needed.

A sound strategy for the creation of synthetic data offers the possibility to create high quality training and evaluation datasets. These datasets would be well balanced and have sufficient data for all the possible error cases. With such datasets it is possible to train machine-learning-based predictive maintenance models that can not only detect faults in real world applications before they occur but also identify the specific fault.

In order to be able to develop such a strategy, a literature review is carried out in this article. The focus of this review is to develop a predictive maintenance model for an hydrogen combustion engine, as part of the WaVe research project, which is described in the next section. The main contributions of this paper are:

1. A survey of the current state of the art of data augmentation methods for predictive maintenance tasks.
2. A summary of the research gap found on the selected publications according to the WaVe use case and a summary of suitable approaches.

The remaining sections in this article are organized as follows. Section 2 provides a brief overview of the WaVe research project and the problem statement. Section 3 describes the methodology used for the literature review. In Sect. 4 the selected publications of the literature review are discussed according to several research questions. Section 5 analyzes the research gap found in the selected publications and discusses suitable data augmentation approaches to solve the imbalanced data problem in the WaVe project. Finally, Sect. 6 presents the conclusions of this work.

2 The WaVe research project

The joint research project WaVe is promoted by the German Federal Ministry for Economic Affairs and Climate Action. WaVe stands for *Wasserstoff-Verbrennungsmotor* which translates to hydrogen combustion engine. The goal of this research project is to develop a hydrogen-based drive system for commercial vehicles in the medium-duty range. The project partners are pooling their technological expertise and developing innovative individual solutions for a hydrogen-based drive system in multiple technological subprojects. The individual solutions are tested, harmonized and combined to form a functioning overall drive system (Commercial Vehicle Cluster-Nutzfahrzeug GmbH 2021). This will then be installed and tested in two different demonstrators—a Mercedes Benz Unimog U400l and a crawler vehicle.

To support the overall development of the hydrogen-based drive system and the two demonstrators, digital twins are under development by *comlet Verteilte Systeme GmbH*. The development of these digital twins is interconnected to the hardware development. A digital twin is an accurate representation of a physical system. Its application in the design, testing, and manufacturing phases of the physical system allows for the reduction of time and costs, as well as improvements in user safety (Grieves and Vickers 2017). In their work, Fuller et al. (2020) differentiate between three definitions based on their data flow characteristics. A digital model exhibits no inherent automatic data flow, while a digital shadow is characterized by an automatic exchange of data from the physical asset to the digital object. A digital twin, in contrast, represents a fully integrated data flow in both directions, encompassing both the physical and digital domains. Consequently, a change made to the physical object is reflected in the digital object, and vice versa. The data of a digital twin can be grouped into static and dynamic data. Static data is created during development and does not change significantly during the lifecycle. This includes manuals, technical specifications, product information, CAD models, circuit diagrams, simulations etc. Dynamic data, on the other hand, is collected by sensors in the physical asset in real time. Dynamic data can include not only data from the finished product, but also engine data recorded in various configurations on an engine test bench, data from field trials, and data captured during tests of the hydrogen tank system.

All these data provide a solid basis for the development of a predictive maintenance system for a hydrogen-based drive system and complete vehicles. The development of such a predictive maintenance system is a planned goal of *comlet Verteilte Systeme GmbH*. Digital twins offer several advantages for the creation of a predictive maintenance systems. The digital twin serves as a data provider for the predictive maintenance system. Thus, feature extraction, as well as feature engineering can directly access the data of a digital twin, or they could also be directly integrated into the digital twin. This makes it very easy to create training and test data sets from the existing data. Another advantage is the access to simulations which provide additional data.

This leads to the biggest problem for the intended predictive maintenance system. Even if there is plenty of data available, most of it contains only the normal condition of the hydrogen-based drive system and the vehicles. There will be little data from failure cases, which leads to strongly imbalanced training data. Generating fault data using deliberate defects is limited due to their destructive nature. Only defects that do not cause damage to the hydrogen drive, or vehicles, can be used. In addition, it is very time consuming and

costly to manually cause such defects in the physical systems to generate a sufficient amount of data. Simulations alone also cannot be used to generate a sufficient amount of defect data because they are usually computationally intensive. As defects that can result in personal injury deserve the most accurate consideration, numerous safety precautions for propulsion systems in general exist. These are the most dangerous and difficult defects to simulate.

To solve the lack of available data, this work discusses a strategy based on data augmentation. This strategy starts with simulations to generate a limited amount of data related to different failure cases. Then this data, as well as the real data recorded during the engine test bench and field trials, will be used to generate more data with specific characteristics using data augmentation methods for machine learning.

3 Data augmentation in predictive maintenance

Not all data augmentation methods are suitable for the WaVe use case. Since the data are predominantly, if not exclusively, time series, image-based methods are unsuitable. It is possible to use these methods partially for sensor data, but this would change the underlying structure of the time series. Sampling-based methods are of limited suitability because they are usually based on the weighted average of the available data and do not consider the underlying distribution, thus limiting the diversity of the generated data. The generative models appear most promising for the WaVe use case because they are trained to learn the underlying distribution of the available data and do not affect the structure of time series.

A state-of-the-art review is conducted to find suitable generative algorithms for the creation of high quality data for predictive maintenance models. This review follows the guidelines of PRISMA (Page et al. 2021a), which consists of four main steps: identification, screening, inclusion and discussion of papers. The whole process of the review from identification of the data sources until the inclusion of the papers is shown in Fig. 1. The review focuses on generative algorithms for data augmentation and answers the following research questions:

- RQ1: Which data augmentation techniques have proven to be suitable for predictive maintenance?
- RQ2: What role do generative algorithms play in predictive maintenance?
- RQ3: In which application domains are those algorithms typically used?
- RQ4: Which type of data and datasets are most commonly used for predictive maintenance?
- RQ5: Which validation methods and metrics are used to evaluate the quality of generated data?

The literature review was conducted in January 2024 and took publications into account starting from 2018. This means that only publications between 2018 and 2023 are included to reflect the current state-of-the-art and trend in data augmentation for predictive maintenance tasks. This can lead to the exclusion of techniques developed before that time frame that are still relevant today. The following scientific databases were used in the research:

- IEEExplore

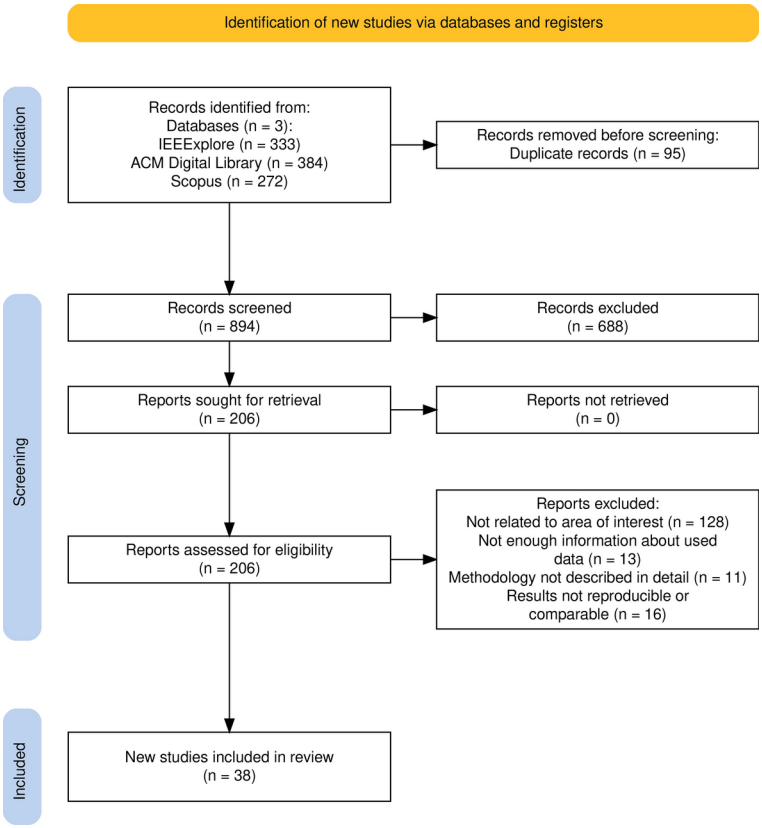

Fig. 1 Flow diagram for inclusion and exclusion of studies in literature reviews following PRISMA framework (Page et al. 2021b)

Table 1 Search query used for the literature review and the number of resulting publications

Search query	Source	Num papers
"data augmentation" AND ("predictive maintenance" OR "anomaly detection")	IEEExplore	333
	ACM Digital Library	384
	Scopus	272

- ACM Digital Library
- Scopus

In order to avoid obtaining a very high or very low number of results the search query was refined in an iterative process until a manageable number of publications was found. Table 1 shows the final search query used and the number of results from the different sources.

The search query resulted in a total number of 989 papers found in the databases. Before an in depth screening of these papers was conducted duplicate papers were removed. In

total, 95 duplicate papers were found and removed. The screening phase was split into multiple steps to discard papers that didn't fulfill the following eligibility criteria:

- Only publications in English.
- Must be published after 2018.
- Must be at least a short paper. Posters, extended abstracts, workshops, etc. are excluded.
- Must be related to the area of interest.
- Must have a focus on industrial applications or time series data.
- Must provide detailed information about the datasets used or must use public (benchmark) datasets.
- Must describe the used methodology detailed enough so that it can be reproduced.
- Must provide reproducible or comparable results, or must be a proof of concept.

The first five criteria were checked by reading the titles and abstracts of the papers. All papers found were written in English, so no papers were excluded due to this criterion. Even when the period 2018–2023 was used as filter for the search queries, three papers from 2017 and before were still in the search results and therefore discarded. The resulting papers included three extended abstracts, two posters and one workshop which were excluded. Using the titles and abstracts 679 papers were excluded because they were not related to the area of interest or had focus on other types of data. Since generative algorithms and Generative Adversarial Networks (GAN)s in particular are often used in image generation, many of the resulting papers excluded had a strong focus in image-generation-based data augmentation for non industrial applications.

The remaining 206 papers were screened in depth for eligibility according to the remaining criteria. From these papers 128 were excluded because they were not related to the area of interest. This was not evident from the abstracts of these publications alone. Another 13 papers used confidential data from companies and didn't provide enough information about the used datasets to reproduce the methods in any way and were therefore excluded. 11 papers were excluded due to not describing the used methodology in detail. This makes it impossible to reproduce their solutions and results. The last 16 papers that were discarded didn't provide reproducible or comparable results. This leaves 38 papers which are included in this state-of-the-art review.

4 Paper discussion

In this section the remaining 38 papers that are included in the review are grouped and discussed by multiple different criteria to answer the above stated research questions.

4.1 RQ1: which data augmentation techniques are used for predictive maintenance?

In Table 2 the selected papers are grouped by their general data augmentation category: image-based, sampling-based and generative and then by their specific method. Figure 2 shows the number of publications for each of these groups. The total number of the grouped papers is larger than the remaining 38 papers, since some publications use data augmenta-

Table 2 Relevant papers grouped by data augmentation method

Group	Augmentation method	Publication
Generative	GAN	Molitor et al. (2022), Fathy et al. (2021), Lu et al. (2021a, b, c), Cannizzaro et al. (2022), Lin et al. (2020), Huang et al. (2021), Li et al. (2022), Bui et al. (2021), Jiang et al. (2021), Zhang et al. (2021), Kim et al. (2023)
	WGAN	Molitor et al. (2022), Fathy et al. (2021), Lu et al. (2021b), Xu et al. (2019), Li et al. (2021a)
	CGAN	Molitor et al. (2022), Fathy et al. (2021), Ranasinghe et al. (2019), Quintana et al. (2020), Faltings et al. (2022), Zhu et al. (2022), Zheng et al. (2020), Shao et al. (2019), Behera and Misra (2021), Yan et al. (2022)
	BiGAN	Smolyak et al. (2020), Cui et al. (2021)
	AAE	Lim et al. (2018), Wu et al. (2020)
Sampling-based	DTW	Liu et al. (2022), Sadoughi et al. (2019)
	Audio data transformations	Hong and Suh (2021)
	SMOTE	Fathy et al. (2021), Martins et al. (2023), Liu et al. (2023)
	Gaussian noise	Ding et al. (2022), Martins et al. (2023)
	Amplitude scaling	Ding et al. (2022)
	Time stretching	Ding et al. (2022)
	Translations	Ding et al. (2022)
	Noise mask	Ding et al. (2022)
	Uniform noise	Mo et al. (2022)
Image-based	Cutting and pasting	Li et al. (2021b)
	Matrix transformations	Pasqualotto et al. (2021), Mahenge et al. (2021), Molitor et al. (2022)
	Brightness change	Pasqualotto et al. (2021), Molitor et al. (2022)
	Gaussian noise	Pasqualotto et al. (2021), Molitor et al. (2022)
Simulation		Dong et al. (2022)

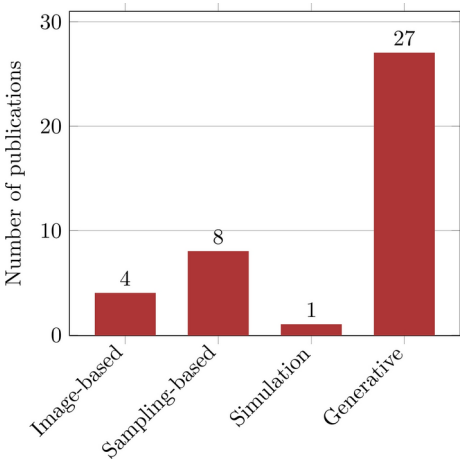

Fig. 2 Number of publications for each group of data augmentation method

tion methods from multiple groups. With a focus on industrial applications and time series data, it can be seen that in recent years most of the publications found use generative algorithms for data augmentation.

Only four publications use image-based data augmentation techniques. Image-based data augmentation usually uses matrix transformations such as rotation, translation, scaling, etc. as well as changes in color or brightness to create new images out of the already existing ones. Pasqualotto et al. (2021) use stray flux analysis images of induction motors to compare the performance of five image-based data augmentation methods: random cropping, change in brightness, time translation, frequency translation and adding Gaussian noise. Mahenge et al. (2021) use cropping, blurring and matrix transformations to augment images for their proposed road crack detection. Li et al. (2021b) propose a new data augmentation technique for defect detection in images called CutPaste. There are two variants of this method. One that cuts out a relatively large rectangular part of an image and pastes it at a random location in another image. The other pastes a long-thin rectangle with a random color. In their publication, Molitor et al. (2022) compare multiple image-based data augmentation methods with generative algorithms, more specifically different GAN architectures, and combinations of both.

Sampling-based approaches are used to balance datasets through undersampling of majority classes or oversampling of minority classes. Usually only oversampling methods are used since undersampling often leads to a loss of valuable information. These methods typically use statistical properties like the standard deviation or mean values of features in the existing data to create new data samples. Two of the most well known and successful methods are Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling approach. SMOTE developed by Chawla et al. (2002) selects a random neighbor from the K-Nearest Neighbors (KNN) for a random sample of the minority class and then generates a synthetic sample by selecting a random point in the feature space between these two samples. ADASYN was first introduced in the work of He et al. (2008) and is a variant of SMOTE that generates more samples in regions of the feature space where the density of minority samples is low and fewer samples in regions where the density is high. These two oversampling methods are used in many of the remaining publications as comparison benchmarks for their proposal. In these cases they are not listed

in the sampling-based category which leads to eight publications left that use some form of sampling-based data augmentation techniques.

In their study, Martins et al. (2023) propose a combination of SMOTE and additive Gaussian noise for data augmentation, differentiating between two approaches. In the first approach, only the minority class is augmented by creating a subset of additional samples through the addition of Gaussian noise and creating another subset using SMOTE. In the second approach, the majority class is also augmented with new samples created using additive Gaussian noise. Liu et al. (2023) present a new data augmentation technique that combines SMOTE with deep attention networks and encoder-decoder networks to generate additional abnormal time series samples. The encoder-decoder is employed to transform the raw time series into a separable feature space, thereby reducing inter-class overlap. The attention network is utilized to identify interpolation factors for SMOTE, ensuring that the generated samples are distant from the aggregation area of normal samples. Subsequently, the newly generated samples are transformed back into the original space and combined with an undersampled set of normal samples, thus forming a balanced dataset. Fathy et al. (2021) conduct a comparative analysis of SMOTE and multiple generative data augmentation methods using different classifiers based on a real world case study. In their works Mo et al. (2022), Ding et al. (2022) and Hong and Suh (2021) use time domain specific methods like time stretching, amplitude scaling, translations etc. and add Gaussian noise to augment time series data. Liu et al. (2022) create new time series data by adding and removing of a small random number of time ticks where the added time ticks are the average of the two adjacent time ticks. Sadoughi et al. (2019) propose a data augmentation approach that uses a randomized shrinkage factor to quantify the ratio of the length of the generated sample and the training sample. Then the training sample is interpolated and mapped into the generated sample.

One publication used simulations to solve the problem with lack of training data. Dong et al. (2022) implemented a mathematical simulation model for ball bearings which is used to create additional data of failure cases.

The largest group of data augmentation methods are the generative algorithms with 24 publications found. In generative data augmentation a machine learning model is trained to learn the underlying distribution of the data. This model can then be fed with random noise to generate new samples from this distribution. The group of generative algorithms mainly consists of different GAN architectures. GANs were first developed by Goodfellow et al. (2014). GANs consist of two sub models, a generator model that is trained to create new data samples and a discriminator model that tries to classify samples as either real or generated. The two models are trained adversarially until the discriminator model fails to classify the real and generated samples correctly. This means the probability that a sample belongs to the generated samples or the real samples is 50%. When this happens the generator model creates samples that are not distinguishable from real samples by the discriminator model. Since nearly every publication about GANs adds a new name to the proposed method there are more than 500 different architectures reported by The GAN ZOO repository[1] which collected all the different names of GANs and was last updated in September 2018. Therefore GANs in this review are divided by their main architecture into the following superordinate groups:

[1] https://github.com/hindupuravinash/the-gan-zoo.

- GAN: GAN, DCGAN, MSGAN, PGGAN, LSTM-GAN, CR-GAN, StyleGAN
- WGAN: WGAN, WGAN-GP
- CGAN: CGAN, WCGAN (CWGAN), C-DCGAN, ACGAN
- BiGAN: BiGAN, BiWGAN

The group of general GANs represents the architecture described by Goodfellow et al. (2014), where two models are trained adversarially. The discriminator and the generator model can use any neural network architecture like convolutional neural networks, recurrent neural networks, etc. The group of Wasserstein Generative Adversarial Networks (WGAN) differs from the general GANs in that the Wasserstein distance is used instead of the Jensen–Shannon divergence to improve convergence. Conditional Generative Adversarial Networks (CGAN) add a conditional variable as input and output to GANs, so that the generator learns a conditional distribution. This conditional variable is usually used to control the generating process by adding class labels to the input samples, but can also be used for other types of additional information. The difference of Bidirectional Generative Adversarial Networks (BiGAN) to the other GAN architectures is that they include an encoder network which learns the inverse of the generator.

In the group of general GANs are multiple publications. In their work, Cannizzaro et al. (2022) use a GAN to generate additional images for powder bed fusion, an additive manufacturing process. In a case study these images are not used to augment the training data but instead they are evaluated for quality and validity. Huang et al. (2021) use a GAN to generate additional time series samples of the minority class of rolling bearing data. The data generation is guided by variable association graphs of the majority class that are learned by an additional model. Lin et al. (2020) and Bui et al. (2021) use simple GANs to generate new time series data of fault cases in ball bearings and gearboxes. (Lu et al. 2021a, GAN-LSTM Predictor...) and (Lu et al. 2021c, A Deep Adversarial...) use a combination of GAN and Long Short-Term Memory (LSTM) to predict the Remaining Useful Life (RUL) of ball bearings. They do not use the generators of the GANs for data augmentation. Instead the generators are used to predict the degradation. The authors Zhang et al. (2021) propose a combination of LSTM layers and convolutional layers to generate multivariate time series data for noncyclic and cyclic RUL prediction. The time series data is preprocessed into a 2D matrix for noncyclic and into a 3D matrix for cyclic problems that it can be fed to the convolutional layers. Jiang et al. (2021) use a 1D Convolutional Neural Network (1D-CNN) to detect failures in rotating machinery parts and combine it with a GAN to create additional time series data samples. In their study, Kim et al. (2023) put forth a data augmentation method that initially transforms multivariate time series data into images through the use of Gramian Angular Field (GAF). Subsequently, they train a StyleGAN to learn the latent space of the time series data. New samples are then generated through interpolation between samples in the latent space. In their work, Li et al. (2022) propose Dual Multiple Generative Adversarial Networks (Dual-MGAN) for outlier detection. This approach combines Multiple Generative Adversarial Active Learning (MGAAL) and Multiple Generative Adversarial Over-Sampling (MGAOS). MGAAL is used to detect discrete anomalies. The unlabeled data is clustered into multiple classes and for each cluster a sub-GAN learns to construct a reference distribution. MGAOS detects partially labeled group anomalies and works similar to MGAAL. Instead of clustering unlabeled data only labeled anomalies are clustered. A sub-GAN for each anomaly cluster is then trained to generate additional similar

samples of these minority classes. Dual-MGAN combines both parts and adds an additional detector neural network. MGAOS increases the size of the minority classes and MGAAL and the detector are alternately optimized to separate anomalies from normal samples.

The next main group of generative models are WGANs. Lu et al. (2021b) and Xu et al. (2019) use WGANs to create additional anomalous time series data for sensor anomaly detection in industrial robots and pipeline leakage detection in petrochemical systems. In case of the pipeline leakage detection, time series data is transformed into graphs instead of using the raw sensor signals (Xu et al. 2019). Li et al. (2021a) also use WGANs for data augmentation and propose a new distance metric called Time-Regularized Hausdorff Distance (TRH) to quantify the similarity between generated and real samples. This metric is used to filter only samples with high similarity before they are passed to the discriminator to improve the overall quality of generated samples.

The third group of generative models are BiGANs. Cui et al. (2021) propose a combination of BiGAN and Wasserstein distance for fault detection in gearboxes and ball bearings. An unsupervised pre-training followed by fine-tuning with a small sample of labeled time series data is used. The pre-trained model is also used to generate additional data. Smolyak et al. (2020) combine BiGAN with Infinite Gaussian Mixture Model (IGGM) for anomaly detection and data augmentation in GPS data.

The last group are CGANs. Faltings et al. (2022), Shao et al. (2019), Behera and Misra (2021), Zheng et al. (2020), and Quintana et al. (2020) use class label information to synthesize additional images of stamps on casted steel billets, to generate additional sensor data for induction motor, aircraft engine and bearing fault detection, as well as thermal comfort data of buildings. Ranasinghe et al. (2019) propose a CGAN that uses additional auxiliary information containing expert knowledge, physics of failure and maintenance records to control the data generation process of failure samples. In their proposal, Yan et al. (2022) introduce a combination of Wasserstein Conditional Generative Adversarial Networks (WCGAN) with Variational Autoencoder (VAE) to generate additional data samples of chiller faults, which are then employed to augment the real-world data utilized for the training of an automated fault diagnosis model. The data is generated by the WCGAN, and the VAE is used to identify high-quality synthetic samples that are subsequently utilized for training the aforementioned fault diagnosis model. In their work, Zhu et al. (2022) use a combination of WGAN and CGAN to generate additional data for a polymerization reaction process of high density polyethylene. They first calculate sparse regions in the data using outliers detected by KNN algorithm. Then a Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP) is trained to generate new samples that fill these regions and after that a Cycle Structure CGAN (CS-CGAN) is used to generate and filter new data samples.

Adversarial Autoencoder (AAE) use the adversarial concept of GANs. Instead of directly generating new data samples, the generator of an AAE creates vectors in a latent space. The discriminator then predicts if this vector was generated by the autoencoder or is a random vector from the real distribution of the data. Wu et al. (2020) use an AAE to detect anomalies in ball bearing time series data. Lim et al. (2018) propose a method to augment unlabeled data for anomaly detection in tabular data using AAEs. Instead of creating additional samples of the minority class or anomalous samples, data is augmented by creating synthetic samples of infrequent nominal samples.

The last two papers conduct a comparison of different data augmentation methods. Fathy et al. (2021) compare GAN, WGAN, CGAN and WCGAN with SMOTE to test their capa-

bility of generating additional samples. In the work of Molitor et al. (2022) Conditional Deep Convolutional Generative Adversarial Networks (C-DCGAN), WGAN-GP and Progressively Growing Generative Adversarial Networks (PGGAN) are compared with image manipulation methods to create synthetical images for tool wear classification.

4.2 RQ2: what role do generative algorithms play in predictive maintenance?

In the 38 selected publications generative models are not only used for data augmentation by generating new data samples, but also directly to detect anomalies. Figure 3 shows that 20 papers (Molitor et al. 2022; Fathy et al. 2021; Cannizzaro et al. 2022; Lin et al. 2020; Bui et al. 2021; Jiang et al. 2021; Zhang et al. 2021; Lu et al. 2021b; Xu et al. 2019; Li et al. 2021a; Ranasinghe et al. 2019; Quintana et al. 2020; Faltings et al. 2022; Zhu et al. 2022; Zheng et al. 2020; Shao et al. 2019; Lim et al. 2018; Behera and Misra 2021; Yan et al. 2022; Kim et al. 2023), use generative models for data augmentation, four papers (Lu et al. 2021a, GAN-LSTM Predictor...; Lu et al. 2021c, A Deep Adversarial...; Wu et al. 2020; Liu et al. 2022) for anomaly detection and four papers (Huang et al. 2021; Li et al. 2022; Smolyak et al. 2020; Cui et al. 2021) for both combined. The remaining 10 publications did not use generative algorithms.

One of the most significant obstacles to the development of effective predictive maintenance models is the scarcity of available failure data. The scarcity of failure data can be attributed to a number of factors. One such factor is the infrequency with which failures or defects occur during operations. Another is the potential absence of suitable systems for data collection in legacy equipment. The resulting imbalance between normal data and failure data makes the latter even more valuable. Fault data can provide crucial insights into issues that are not common but can have significant consequences when they do arise. The identification of scarce or previously unknown faults can be achieved through the utilisation of anomaly detection techniques. Consequently, a generative model can be trained on nominal data, and the reconstruction error can be employed to ascertain whether a sample belongs to the nominal class or represents an anomaly.

4.3 RQ3: in which application domains are those algorithms typically used?

In Table 3 the publications are grouped by their application area and type of data. Since only publications that have a focus on predictive maintenance tasks or use time series data are included in this review, most of the papers are from the industrial area. Five publications in this area conduct predictive maintenance in vehicles, Trucks from Scania (Fathy et al. 2021; Ranasinghe et al. 2019) and planes (Huang et al. 2021; Zhang et al. 2021; Mo et al. 2022; Behera and Misra 2021; Liu et al. 2023). The other publications in the industrial area use

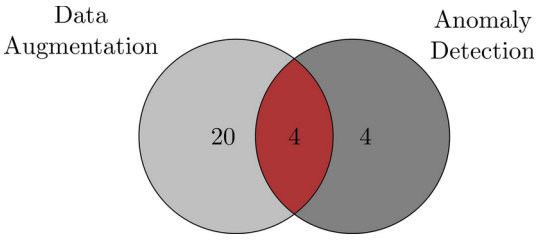

Fig. 3 Use case of generative algorithms in predictive maintenance

Table 3 Relevant papers grouped by application area and type of data

Application area	Type of data	Publication
Industrial	Time series	Lu et al. (2021a, b, c), Lin et al. (2020), Huang et al. (2021), Bui et al. (2021), Jiang et al. (2021), Zhang et al. (2021), Xu et al. (2019), Li et al. (2021a), Zhu et al. (2022), Zheng et al. (2020), Shao et al. (2019), Cui et al. (2021), Wu et al. (2020), Sadoughi et al. (2019), Hong and Suh (2021), Ding et al. (2022), Mo et al. (2022), Dong et al. (2022), Yan et al. (2022), Kim et al. (2023), Behera and Misra (2021), Liu et al. (2023), Martins et al. (2023)
	Images	Pasqualotto et al. (2021), Li et al. (2021b), Molitor et al. (2022), Cannizzaro et al. (2022), Faltings et al. (2022)
	Tabular features	Fathy et al. (2021), Li et al. (2022), Ranasinghe et al. (2019)
Medical	Time series	Liu et al. (2022)
	Tabular features	Li et al. (2022), Lim et al. (2018)
Traffic	Time series	Smolyak et al. (2020)
	Images	Mahenge et al. (2021)
Other	Time series	Quintana et al. (2020)
	Tabular features	Li et al. (2022)

predictive maintenance in manufacturing processes like casting steel billets (Faltings et al. 2022), polymerization reactions (Zhu et al. 2022), additive manufacturing (Cannizzaro et al. 2022), etc. and to detect failures or predict the RUL of components in machines. These components can be ball bearings (Lu et al. 2021c; Zheng et al. 2020; Sadoughi et al. 2019), gearboxes (Bui et al. 2021; Wu et al. 2020), induction motors (Pasqualotto et al. 2021; Shao et al. 2019), etc.

Other application domains found are medical and traffic. In the medical area generative algorithms are used to generate additional time series samples and to detect anomalies in ECG data (Liu et al. 2022). In the traffic domain, reference (Mahenge et al. 2021) creates additional images to train a road crack detection model and Smolyak et al. (2020) creates GPS trajectories to detect anomalous routes and behavior of drivers.

There are many more domains in which generative algorithms are used that are not covered by this review due to its focus on predictive maintenance tasks. Sabuhi et al. (2021) identified additional application areas in their literature review about general use cases and applications of GANs. To these areas belong surveillance, intrusion detection, image recognition, fraud detection, etc.

4.4 RQ4: which type of data and datasets are most commonly used for predictive maintenance?

Table 3 shows that for predictive maintenance tasks three types of data are used, times series data, images and tabular data. Most of the publications found use time series data. This is

not surprising due to the fact that most predictive maintenance applications rely on some form of sensor data which are usually recorded as time series. The most used datasets of this type are ball bearing datasets, like CWRU bearing fault dataset[2] (Lin et al. 2020; Jiang et al. 2021; Zheng et al. 2020; Cui et al. 2021; Hong and Suh 2021; Dong et al. 2022) and the PRONOSTIA FEMTO-ST bearing dataset[3] used in the IEEE PHM 2012 Data Challenge (Lu et al. 2021c; Sadoughi et al. 2019; Ding et al. 2022). Another publicly available dataset that was used is the NASA C-MAPPS aircraft simulation data[4] (Zhang et al. 2021; Mo et al. 2022; Behera and Misra 2021; Liu et al. 2023). The remaining papers that used time series data in the areas of induction motors (Shao et al. 2019), data captured in an oil refinery (Xu et al. 2019), data of a polymerization reaction (Zhu et al. 2022), or data from rotating machinery parts Martins et al. (2023) do not make their datasets public.

Six of the publications found use images as input data to train their predictive maintenance or data augmentation models. Li et al. (2021b) use the freely available MVTec anomaly detection dataset[5] from Bergmann et al. (2021, 2019) which is a benchmarking dataset for industrial inspection. Another public image dataset is the RDD2020 dataset used in the IEEE Global Road Damage Detection Challenge 2020.[6] This dataset is used in the paper from Mahenge et al. (2021) to detect cracks in roads. The other publications used non public datasets for tool wear classification (Molitor et al. 2022), anomaly detection in powder bed fusion additive manufacturing (Cannizzaro et al. 2022) and for data augmentation of images of stamps on casted steel billets (Faltings et al. 2022).

Four publications used tabular datasets. Li et al. (2022) used multiple datasets from the UCI Machine Learning Repository[7] and Lim et al. (2018) from the ODDS Repository.[8] Another public dataset is the Scania air pressure system[9] dataset used by Fathy et al. (2021) and Ranasinghe et al. (2019).

4.5 RQ5: which validation methods and metrics are used to evaluate the quality of generated data?

When data augmentation is used to create additional data of minority classes to balance a dataset, the quality of the generated data samples should be evaluated. In the case of time series data and generative algorithms it should be assured qualitatively and quantitatively that the generated samples are plausible and of high quality. Plausibility in this context means that the generated samples can occur in real data and are not physically or by any other restrictions impossible to occur in reality.

Table 4 provides an overview of the methods and metrics used to evaluate the quality of generated data samples. Most publications use either no validation at all (Pasqualotto et al. 2021; Mahenge et al. 2021; Huang et al. 2021; Lu et al. 2021b; Sadoughi et al. 2019;

[2] https://engineering.case.edu/bearingdatacenter/download-data-file.

[3] https://www.kaggle.com/datasets/alanhabrony/ieee-phm-2012-data-challenge.

[4] https://data.nasa.gov/dataset/C-MAPSS-Aircraft-Engine-Simulator-Data/xaut-bemq.

[5] https://www.mvtec.com/company/research/datasets/mvtec-ad.

[6] https://rdd2020.sekilab.global/data/.

[7] https://archive.ics.uci.edu/ml/index.php.

[8] http://odds.cs.stonybrook.edu/.

[9] https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks.

Table 4 Relevant papers grouped by quality validation method

Quality metric	Publication
Visual comparison	Molitor et al. (2022), Cannizzaro et al. (2022), Lin et al. (2020), Li et al. (2022), Jiang et al. (2021), Zhang et al. (2021), Xu et al. (2019), Quintana et al. (2020), Faltings et al. (2022), Smolyak et al. (2020), Cui et al. (2021), Lim et al. (2018), Liu et al. (2022), Dong et al. (2022), Behera and Misra (2021)
Time-regulated Hausdorff Distance	Li et al. (2021a)
Dynamic Time Warping	Liu et al. (2022), Mo et al. (2022)
Structural similarity index measure	Hong and Suh (2021)
Kolmogorov–Smirnov test	Fathy et al. (2021), Bui et al. (2021), Ranasinghe et al. (2019)
t-SNE visualization	Li et al. (2021b), Zheng et al. (2020), Smolyak et al. (2020)
Euclidean distance	Lin et al. (2020), Quintana et al. (2020), Zheng et al. (2020), Shao et al. (2019), Mo et al. (2022), Behera and Misra (2021)
Cosine distance	Zheng et al. (2020), Behera and Misra (2021)
Maximum mean discrepancy	Cui et al. (2021)
Pearson correlation coefficient	Lin et al. (2020), Zheng et al. (2020), Shao et al. (2019), Cui et al. (2021)
Inception score	Molitor et al. (2022)
Fréchet inception distance	Molitor et al. (2022), Cannizzaro et al. (2022)
Earth mover's distance	Zhu et al. (2022)
Time domain indicators	Jiang et al. (2021)
Kullback–Leibler divergence	Shao et al. (2019)
Variational Autoencode	Yan et al. (2022)
No validation	Pasqualotto et al. (2021), Mahenge et al. (2021), Huang et al. (2021), Lu et al. (2021b), Sadoughi et al. (2019), Ding et al. (2022), Martins et al. (2023), Liu et al. (2023), Kim et al. (2023)

Ding et al. 2022; Martins et al. 2023; Liu et al. 2023; Kim et al. 2023) or only do a visual comparison between generated and real samples (Li et al. 2022; Zhang et al. 2021; Xu et al. 2019; Faltings et al. 2022; Smolyak et al. 2020; Lim et al. 2018; Dong et al. 2022). This indicates a clear lack of a good quality metric for the evaluation of synthetic time series data.

Three of the papers evaluated the quality of generated data by using metrics that quantify the similarity between images. Molitor et al. (2022) and Cannizzaro et al. (2022) use the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) to quantify the quality of generated samples. These metrics can not directly be used to evaluate the quality of generated time series data samples. Therefore Hong and Suh

(2021) first transform their time series data into MEL spectrogram images and use the Structural Similarity Index Measure (SSIM) to quantify the similarity of generated and real data.

Three publications (Fathy et al. 2021; Bui et al. 2021; Ranasinghe et al. 2019), use the Kolmogorov–Smirnov (K–S) test to evaluate the quality of generated data. The K–S test is a measure for the similarity between the distribution of generated samples and the distribution of real samples. Two other metrics for the similarity between distributions, the Maximum Mean Discrepancy (MMD) and the Kullback–Leibler divergence (K-LD) are also used (Shao et al. 2019; Cui et al. 2021). Seven of the included papers (Lin et al. 2020; Quintana et al. 2020; Zheng et al. 2020; Shao et al. 2019; Liu et al. 2022; Mo et al. 2022; Behera and Misra 2021) use distance measures such as Euclidean Distance (ED), Dynamic Time Warping (DTW) and Cosine Distance to test if the synthetic time series samples are similar to the real samples. The Pearson Correlation Coefficient (PCC) was adopted by four papers (Lin et al. 2020; Zheng et al. 2020; Shao et al. 2019; Cui et al. 2021) to measure the linear correlation between generated and real time series data. Three papers (Li et al. 2021b; Zheng et al. 2020; Smolyak et al. 2020) use t-distributed Stochastic Neighbor Embedding (t-SNE) visualization. T-SNE is a dimensionality reduction technique that transforms high dimensional data into a low dimensional space of two or three dimensions where similar samples are presented by nearby points and unsimilar samples by distant points.

In their study, Yan et al. (2022) employ a VAE to identify high-quality generated time series samples. The VAE is trained with a randomly selected set of samples generated by a WCGAN and tested with real-world anomaly samples. This process is iteratively repeated until the reconstruction error of the VAE for all real-world test samples falls below a specified threshold. At this point, the generated samples used to train the VAE are deemed to be of high quality.

A novel method called TRH distance was introduced by Li et al. (2021a). TRH distance extends the Hausdorff distance by adding a time-regularized penalty that represents the temporal order difference between two points from different time series samples.

5 Research findings on the WaVe use case

As previously described, a hydrogen-based drive system is being developed in the WaVe project and tested in field trials in two demonstrators from the medium duty vehicle sector. Additionally, a predictive maintenance system is being developed for this drive system. Sensor data from engine test benches and field tests are available for this purpose. Since it can be assumed that these data mainly consist of time series from normal operation and therefore little abnormal data will be available, a suitable data augmentation method is needed to generate additional data of possible failure cases. For these reasons, the literature search was conducted and the results were examined to determine if they were suitable for the WaVe use case.

This section first summarizes the limitations according to the WaVe use case of the remaining papers and then summarizes the most suitable approaches for a data augmentation model to generate new data of fault cases.

5.1 Limitations according to the WaVe use case

It was found that most of the approaches are not suitable due to various limitations. Since the recorded data from the field tests and from the test bench are time series data, a method is needed that can generate such data. Image-based data augmentation methods such as rotation, scaling, color or brightness changes are not applicable to time series sensor data. The addition of Gaussian noise could be used to augment time series data, but would not incorporate the time dependencies. GANs that generate images as training data could be applied to the WaVe use case, but require modifications due to the different characteristics of time series data. Generative models that recreate time series data for RUL prediction or anomaly detection using the recreation error are of interest. However, since they are not used to augment the training data, it is not clear whether they would have a positive impact if used in this way. Time domain data augmentation methods may have the problem that the additional samples have a low variety and are therefore not suitable for the WaVe use case. Methods that use and generate tabular data are partially useful for time series data. However, the time dependencies are not taken into account, which means that the points in time are independent. This behavior is not a reflection of the real-world use case. Simulations to generate additional data are not feasible in the WaVe use case because access to simulations of the different parts and processes of the hydrogen combustion drive system is limited and creating such simulations from scratch is far too complex a task.

5.2 Suitable approaches for the WaVe use case

This section highlights the most suitable approaches for the WaVe use case and discusses how they can be used to create and evaluate additional training data.

The WGAN-GP architecture is the most appropriate approach for the WaVe use case based on the results of the literature review. The publications that have achieved the best results in the generation of time series data have based their approaches on the WGAN-GP architecture. Several methods can be used to extend this architecture. The existing fault time series data from the engine test bench and field trials can initially be clustered using approaches similar to the ones used by Zhu et al. (2022) and Li et al. (2022). A WGAN-GP can then be trained for each cluster to generate additional time series data of the specific failure cases. Depending on the type of nominal data and the performance of the predictive maintenance model, additional data can also be generated from rarely occurring nominal time series signals as suggested by Lim et al. (2018) to reduce the number of false positives. To train the WGAN-GP, the two-step training approach described by Cui et al. (2021) can be used. This means that the model will first be pre-trained with all available data, i.e. normal data and data from failure cases in an unsupervised step. After this step the model is fine-tuned with labeled data from the fault cases.

Both training phases can use the TRH distance metric proposed by Li et al. (2021a), to filter generated time series data samples with low quality, so that the discriminator network is trained only with high quality data. This should generally improve the quality of the generated data. To measure the quality of the generated data, the TRH distance can be used. Other metrics that can be used to evaluate the similarity of time series are DTW and ED. These can also be used to assess the quality of the generated data. To evaluate the plausibility of the synthetic data, a suitable metric is still needed. Plausibility here means, as already

described, that the generated data can actually occur in reality. In the selected publications of the literature research, this aspect was only rarely considered and, if at all, only visually tested on a few samples.

6 Conclusion

The WaVe research project aims to develop a hydrogen-based drive system based on an internal combustion engine. A digital twin and a predictive maintenance solution will be implemented for this drive system. It is expected that the data will be highly imbalanced, since mostly data from the normal operating conditions will be available and only a small amount of data from failure cases. This makes it difficult to train a predictive maintenance model that can detect specific faults reliably. Therefore, a suitable data augmentation strategy is needed to generate more data of the underrepresented failure cases.

In order to identify, or develop a suitable strategy, a literature review was conducted. This review highlights the current state of the art of data augmentation methods for predictive maintenance procedures and time series and answers the previously posed research questions. It has been shown that mainly generative algorithms, especially GANs are used for data augmentation in predictive maintenance. On the one hand, these are used to generate new images for defect detection in image-based applications, such as images of stamps on casted steel billets or images of stamped parts. On the other hand they are also used to generate time series data recorded by vibration sensors or accelerometers on machine components, for example. Sampling-based data augmentation methods are used rather rarely.

While the literature review indicated that GAN are the predominant approach for data augmentation of time series data, alternative generative techniques may also prove effective for data augmentation in predictive maintenance. VAE, initially developed by Kingma and Welling (2014), can also be employed to generate novel time series data samples. Another noteworthy approach is that of diffusion models, initially proposed by Sohl-Dickstein et al. (2015) and subsequently refined through the introduction of denoising diffusion probabilistic models by Ho et al. (2020). These models have demonstrated remarkable efficacy in the generation of high-quality synthetic images and, in comparison to GAN, exhibit a more stable training process. GAN, on the other hand, are less computationally complex, which leads to shorter training and inference times.

The approaches found in the literature are discussed about their limitations and suitability for the WaVe use case. Based on this discussion the most promising methods include a two-step training approach for generative models. A method to reduce the amount of false positives is the generation of rarely occurring data of the majority class instead of creating additional samples of minority classes. Also a novel distance metric, TRH distance to evaluate the similarity between time series samples is found appropriate.

The question of how to evaluate and ensure the quality and plausibility of generated data has shown that most publications only perform a visual inspection. In order to evaluate the quality of time series, the TRH distance can be used as well as other metrics such as ED or DTW are suitable. To evaluate the plausibility of the data, however, suitable metrics for a quantitative assessment are still missing. The minority of publications checked the plausibility visually, but the majority did not consider it at all.

The next steps now are the analysis of the hydrogen combustion engine data captured at the engine test bench and implementation as well as experimental evaluation of a suitable data augmentation strategy.

Acknowledgements This work was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), under grant No. (19I21028R), research project WaVe. The authors alone are responsible for the content of the paper.

Author contributions Alexander Schwarz: Conceptualization, Writing—original draft, Writing—review and editing, Investigation. Jhonny Rodriguez Rahal: Writing—review and editing, Investigation. Benjamín Sahelices: Supervision, Writing—review and editing Verónica Barroso-García: Supervision, Writing—review and editing Ronny Weis: Project administration, Writing—review and editing Simon Duque Antón: Supervision, Writing—review and editing.

Funding This work was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), under grant No. (19I21028R), research project WaVe. Verónica Barroso-García was supported by the projects CPP2022-009735 and PID2020-115468RB-I00, funded by MICIU/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR. Her research was also funded by the "CIBER-Consorcio Centro de Investigación Biomédica en Red" (CB19/01/00012) through "Instituto de Salud Carlos III", co-funded with European Regional Development Fund.
Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability Data sharing not applicable to this article as no datasets were generated during the current study.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Behera S, Misra R (2021) Generative adversarial networks based remaining useful life estimation for IIoT. Comput Electr Eng 92:107195. https://doi.org/10.1016/j.compeleceng.2021.107195

Bergmann P, Fauser M, Sattlegger D, Steger C (2019) MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 9584–9592

Bergmann P, Batzner K, Fauser M, Sattlegger D, Steger C (2021) The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. Int J Comput Vis 129:1038–1059. https://doi.org/10.1007/s11263-020-01400-4

Bui V, Pham TL, Nguyen H, Jang YM (2021) Data augmentation using generative adversarial network for automatic machine fault detection based on vibration signals. Appl Sci 11(5). https://doi.org/10.3390/app11052166

Cannizzaro D, Varrella AG, Paradiso S, Sampieri R, Chen Y, Macii A, Patti E, Cataldo SD (2022) In-situ defect detection of metal additive manufacturing: an integrated framework. IEEE Trans Emerg Top Comput 10(1):74–86. https://doi.org/10.1109/TETC.2021.3108844

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3). https://doi.org/10.1145/1541880.1541882
Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
Commercial Vehicle Cluster-Nutzfahrzeug GmbH (2021) Verbundvorhaben WaVe—Entwicklung und prototypische Erprobung von Wasserstoff-Verbrennungsmotoren
Cui L, Tian X, Shi X, Wang X, Cui Y (2021) A semi-supervised fault diagnosis method based on improved bidirectional generative adversarial network. Appl Sci 11(20). https://doi.org/10.3390/app11209401
Ding Y, Zhuang J, Ding P, Jia M (2022) Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. Reliab Eng Syst Saf 218. https://doi.org/10.1016/j.ress.2021.108126
Dong Y, Li Y, Zheng H, Wang R, Xu M (2022) A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: solving the small sample problem. ISA Trans 121:327–348. https://doi.org/10.1016/j.isatra.2021.03.042
Faltings U, Bettinger T, Barth S, Schäfer M (2022) Impact on inference model performance for ML tasks using real-life training data and synthetic training data from GANs. Information 13(1). https://doi.org/10.3390/info13010009
Fathy Y, Jaber M, Brintrup A (2021) Learning with imbalanced data in smart manufacturing: a comparative analysis. IEEE Access 9:2734–2757. https://doi.org/10.1109/ACCESS.2020.3047838
Fuller A, Fan Z, Day C, Barlow C (2020) Digital twin: enabling technologies, challenges and open research. IEEE Access 8:108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358
Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems—volume 2, NIPS'14. MIT Press, Cambridge, pp 2672–2680
Grieves M, Vickers J (2017) Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Transdisciplinary perspectives on complex systems: new findings and approaches. Springer, Cham, pp 85–113
He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International joint conference on neural networks (IEEE world congress on computational intelligence), pp 1322–1328
Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17. Curran Associates Inc, Red Hook, pp 6629–6640
Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Proceedings of the 34th international conference on neural information processing systems, NIPS'20. Curran Associates Inc, Red Hook
Hong G, Suh D (2021) Supervised-learning-based intelligent fault diagnosis for mechanical equipment. IEEE Access 9:116147–116162. https://doi.org/10.1109/ACCESS.2021.3104189
Huang H, Xu C, Yoo S (2021) Interpretable temporal GANs for industrial imbalanced multivariate time series simulation and classification. In: Proceedings of the 36th annual ACM symposium on applied computing, SAC'21. Association for Computing Machinery, New York, pp 924–933
Jiang W, Wang C, Zou J, Zhang S (2021) Application of deep learning in fault diagnosis of rotating machinery. Processes 9(6). https://doi.org/10.3390/pr9060919
Kim Y, Lee T, Hyun Y, Coatanea E, Mika S, Mo J, Yoo Y (2023) Self-supervised representation learning anomaly detection methodology based on boosting algorithms enhanced by data augmentation using stylegan for manufacturing imbalanced data. Comput Ind 153:104024. https://doi.org/10.1016/j.compind.2023.104024
Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: 2nd International conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, conference track proceedings
Li Y, Shi Z, Liu C, Tian W, Kong Z, Williams CB (2021a) Augmented time regularized generative adversarial network (ATR-GAN) for data augmentation in online process anomaly detection. IEEE Trans Autom Sci Eng 1–18. https://doi.org/10.1109/TASE.2021.3118635
Li CL, Sohn K, Yoon J, Pfister T (2021b) CutPaste: self-supervised learning for anomaly detection and localization. In: 2021 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 9659–9669
Li Z, Sun C, Liu C, Chen X, Wang M, Liu Y (2022) Dual-MGAN: an efficient approach for semi-supervised outlier detection with few identified anomalies. ACM Trans Knowl Discov Data 16(6). https://doi.org/10.1145/3522690
Lim SK, Loo Y, Tran NT, Cheung NM, Roig G, Elovici Y (2018) DOPING: generative data augmentation for unsupervised anomaly detection with GAN. In: 2018 IEEE International conference on data mining (ICDM), pp 1122–1127

Lin Q, Zhang Y, Yang S, Ma S, Zhang T, Xiao Q (2020) a self-learning and self-optimizing framework for the fault diagnosis knowledge base in a workshop. Robotics Comput Integr Manuf 65. https://doi.org/10.1016/j.rcim.2020.101975

Liu S, Zhou B, Ding Q, Hooi B, Zhang Z, Shen H, Cheng X (2022) Time series anomaly detection with adversarial reconstruction networks. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2021.3140058

Liu D, Zhong S, Lin L, Zhao M, Fu X, Liu X (2023) Deep attention smote: data augmentation with a learnable interpolation factor for imbalanced anomaly detection of gas turbines. Comput Ind 151(C). https://doi.org/10.1016/j.compind.2023.103972

Lu H, Barzegar V, Nemani VP, Hu C, Laflamme S, Zimmerman AT (2021a) GAN-LSTM predictor for failure prognostics of rolling element bearings. In: 2021 IEEE International conference on prognostics and health management (ICPHM), pp 1–8

Lu H, Du M, Qian K, He X, Wang K (2021b) GAN-based data augmentation strategy for sensor anomaly detection in industrial robots. IEEE Sens J 1–1. https://doi.org/10.1109/JSEN.2021.3069452

Lu BL, Liu ZH, Wei HL, Chen L, Zhang H, Li XH (2021c) A deep adversarial learning prognostics model for remaining useful life prediction of rolling bearing. IEEE Trans Artif Intell 2(4):329–340. https://doi.org/10.1109/TAI.2021.3097311

Mahenge SF, Wambura S, Jiao L (2021) Robust deep representation learning for road crack detection. In: 2021 The 5th international conference on video and image processing, ICVIP 2021. Association for Computing Machinery, New York, pp 117–125

Martins DH, de Lima AA, Pinto MF, Hemerly DD, Prego TD, e Silva FL, Tarrataca L, Monteiro UA, Gutiérrez RH, Haddad DB (2023) Hybrid data augmentation method for combined failure recognition in rotating machines. J Intell Manuf 34:1795–1813. https://doi.org/10.1007/s10845-021-01873-1

Mo Y, Li L, Huang B, Li X (2022) Few-shot RUL estimation based on model-agnostic meta-learning. J Intell Manuf. https://doi.org/10.1007/s10845-022-01929-w

Molitor DA, Kubik C, Becker M, Hetfleisch RH, Lyu F, Groche P (2022) Towards high-performance deep learning models in tool wear classification with generative adversarial networks. J Mater Process Technol 302. https://doi.org/10.1016/j.jmatprotec.2021.117484

Murphy KP (2012) Machine learning: a probabilistic perspective. The MIT Press

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021a) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n71

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, McKenzie JE (2021b) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372. https://doi.org/10.1136/bmj.n160

Pasqualotto D, Navarro AN, Zigliotto M, Antonino-Daviu JA, Biot-Monterde V (2021) Fault detection in soft-started induction motors using convolutional neural network enhanced by data augmentation techniques. In: IECON 2021—47th annual conference of the ieee industrial electronics society, pp 1–6

Quintana M, Schiavon S, Tham KW, Miller C (2020) Balancing thermal comfort datasets: we GAN, but should we? In: Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation, BuildSys'20. Association for Computing Machinery, New York, pp 120–129

Ranasinghe GD, Lindgren T, Girolami M, Parlikad AK (2019) A methodology for prognostics under the conditions of limited failure data availability. IEEE Access 7:183996–184007. https://doi.org/10.1109/ACCESS.2019.2960310

Sabuhi M, Zhou M, Bezemer CP, Musilek P (2021) Applications of generative adversarial networks in anomaly detection: a systematic literature review. IEEE Access 9:161003–161029. https://doi.org/10.1109/ACCESS.2021.3131949

Sadoughi M, Lu H, Hu C (2019) A deep learning approach for failure prognostics of rolling element bearings. In: 2019 IEEE International conference on prognostics and health management (ICPHM), pp 1–7

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training GANs. In: Proceedings of the 30th international conference on neural information processing systems, NIPS'16. Curran Associates Inc, Red Hook, pp 2234–2242

Shao S, Wang P, Yan R (2019) Generative adversarial networks for data augmentation in machine fault diagnosis. Comput Ind 106:85–93. https://doi.org/10.1016/j.compind.2019.01.001

Smolyak D, Gray K, Badirli S, Mohler G (2020). Coupled IGMM-GANs with applications to anomaly detection in human mobility data. ACM Trans Spatial Algorithms Syst 6(4). https://doi.org/10.1145/3385809

Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using non-equilibrium thermodynamics. In: Proceedings of the 32nd international conference on international conference on machine learning—volume 37, ICML'15. JMLR.org, pp 2256–2265

Thomas D (2018) The Costs and benefits of advanced maintenance in manufacturing. Technical report, Advanced Manufacturing Series (NIST AMS), National Institute of Standards and Technology, Gaithersburg, MD

Thomas D, Weiss B (2020) Economics of manufacturing machinery maintenance: a survey and analysis of U.S. costs and benefits. Technical report, Advanced Manufacturing Series (NIST AMS), National Institute of Standards and Technology, Gaithersburg, MD

Wen Y, Fashiar Rahman M, Xu H, Tseng TLB (2022) Recent advances and trends of predictive maintenance from data-driven machine prognostics perspective. Measurement 187. https://doi.org/10.1016/j.measurement.2021.110276

Wu J, Zhao Z, Sun C, Yan R, Chen X (2020) Fault-attention generative probabilistic adversarial autoencoder for machine anomaly detection. IEEE Trans Ind Inf 16(12):7479–7488. https://doi.org/10.1109/TII.2020.2976752

Xu P, Du R, Zhang Z (2019) Predicting pipeline leakage in petrochemical system through GAN and LSTM. Knowl Based Syst 175:50–61. https://doi.org/10.1016/j.knosys.2019.03.013

Yan K, Su J, Huang J, Mo Y (2022) Chiller fault diagnosis based on VAE-enabled generative adversarial networks. IEEE Trans Autom Sci Eng 19(1):387–395. https://doi.org/10.1109/TASE.2020.3035620

Zhang X, Qin Y, Yuen C, Jayasinghe L, Liu X (2021) Time-series regeneration with convolutional recurrent generative adversarial network for remaining useful life estimation. IEEE Trans Ind Inf 17(10):6820–6831. https://doi.org/10.1109/TII.2020.3046036

Zheng T, Song L, Wang J, Teng W, Xu X, Ma C (2020) Data synthesis using dual discriminator conditional generative adversarial networks for imbalanced fault diagnosis of rolling bearings. Measurement 158. https://doi.org/10.1016/j.measurement.2020.107741

Zhu QX, Xu T, Xu Y, He YL (2022) Improved virtual sample generation method using enhanced conditional generative adversarial networks with cycle structures for soft sensors with limited data. Ind Eng Chem Res 61(1):530–540. https://doi.org/10.1021/acs.iecr.1c03197

Authors and Affiliations

Alexander Schwarz[1,2] · Jhonny Rodriguez Rahal[1,2] · Benjamín Sahelices[1] · Verónica Barroso-García[3,4] · Ronny Weis[2] · Simon Duque Antón[2]

✉ Alexander Schwarz
alexander.schwarz@estudiantes.uva.es

Jhonny Rodriguez Rahal
jhonny.rodriguez@estudiantes.uva.es

Benjamín Sahelices
benjamin.sahelices@uva.es

Verónica Barroso-García
veronica.barroso@uva.es

Ronny Weis
ronny.weis@comlet.de

Simon Duque Antón
simon.duque-anton@comlet.de

[1] GCME Research Group, Department of Informatics, University of Valladolid, Paseo de Belén, 15, 47011 Valladolid, Spain

[2] comlet Verteilte Systeme GmbH, Amerikastraße 27, 66482 Zweibrücken, Germany

[3] Biomedical Engineering Group, University of Valladolid, Paseo de Belén, 15, 47011 Valladolid, Spain

[4] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Paseo de Belén, 15, 47011 Valladolid, Spain