Computers in Industry 146 (2023) 103852

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Evaluation of deep unsupervised anomaly detection methods with a data-centric approach for on-line inspection

Alexander Zeiser [a,b,*], Bekir Özcan [a], Bas van Stein [b], Thomas Bäck [b]

[a] Bayerische Motorenwerke AG, Ohmstrasse 2, Landshut, 84030, Germany
[b] Leiden Institute of Advanced Computer Science, Niels Bohrweg 1, Leiden, 2333 CA, The Netherlands

ARTICLE INFO

Keywords:
Additive manufacturing
Deep unsupervised anomaly detection
Domain randomisation
Synthetic data
In-situ monitoring

ABSTRACT

Anomaly detection methods are used to find abnormal states, instances or data points that differ from a normal sample from the data domain space. Industrial processes are a domain where predictive models are needed for finding anomalous data instances for quality enhancement and to advance zero defect manufacturing objectives. A main challenge in this domain, however, is absence of labels for training supervised models. The potentials of domain randomisation and synthetic data in-the-loop are illustrated by a use case from additive manufacturing for automotive components. This paper contributes to a data-centric way of approaching artificial intelligence in industrial production by unsupervised anomaly detection. Based on a combination of WGAN and encoder CNN, adapted from f-AnoGAN, a solution with high potential for on-line anomaly detection is analysed. In an subsequent evaluation, we compare different metrics and neural network features for an automatic differentiation between normal or anomalous samples. We prove that using clustering methods with features generated by the discriminator yields better results than computing an anomaly score solely.

1. Introduction

Manufacturing efficiency and sustainability gained particular significance over recent years. Besides environmental-friendly consumables, the careful usage of resources for production is one major key performance indicator and more important than ever (Hristov and Chirico, 2019). In this context, high production quality and avoidance of scrap is still a current issue as production complexity and technological requirements increase. Even if already introduced in the 60s, zero defect manufacturing (ZDM) is still the hypernym for strategies and methods for quality improvement. With industry 4.0 new methods found their way into production management, now considered as key enabling technologies for ZDM, like artificial intelligence (AI), simulation or internet of things (IoT) (Powell et al., 2022), all of which are based on data from production.

Manufacturing complexity can be addressed through the development of predictive models that enable assessment of multidimensional relationships. By uncovering machining dependencies and root-causes of problems, they facilitate optimisation of processes and products. However, data quality and effortful pre-processing are major issues in industrial production because they make it difficult to construct reliable data-based solutions. Data drift, missing labels, severely unbalanced datasets, noise, and quickly changing conditions are a few examples of hurdles to be overcome. Surveys and literature on industrial applications indicate that there are still unresolved problems with integrating and utilising the potentials of machine learning (ML) in actual series manufacturing (Dogan and Birant, 2021).

Our work addresses the development of data-driven techniques in industrial production when access to labelled data is a huge difficulty. Thereby, we contribute to the still existing gap between advances in ML research and its direct applicability in industrial, real-world problems. We introduce a data-centric way of approaching artificial intelligence for production quality improvements by in-situ monitoring and anomaly detection (AD). We propose the concept of domain randomisation and synthetic data as an alternative when no labels are available for the real domain dataset. We then adapt f-AnoGAN from Schlegl et al. (2019) to a use case of additive manufacturing for automotive components. Especially Wasserstein Generative Adversarial Networks (WGAN) demonstrate good representation capabilities for data of this kind. The combination with a CNN-based encoder leads to a well performing solution for on-line anomaly detection. We further investigate the anomaly scores, proposed by Schlegl et al. (2019) and features extracted from the GAN components and compare different methods of clustering healthy and anomalous data.

* Corresponding author at: Bayerische Motorenwerke AG, Ohmstrasse 2, Landshut, 84030, Germany.
E-mail address: alexander.az.zeiser@bmw.de (A. Zeiser).

https://doi.org/10.1016/j.compind.2023.103852
Received 29 September 2022; Received in revised form 21 December 2022; Accepted 4 January 2023
Available online 19 January 2023

2. Data mining for zero defect manufacturing

After the third industrial revolution that computerised and automated production, the fourth industrial revolution was launched by digitised systems and network integration. Terms like 'smart manufacturing' or 'smart factory' describe concepts that enable a certain intelligence and autonomy in production. With the goal of greater efficiency, advances in computer science and robotics are increasingly being incorporated into the planning, management and execution of production systems. Connectivity and IoT, autonomous robotics, cyber physical systems (CPS), simulation, cloud computing and artificial intelligence are all pillars of industry 4.0, that provide and consume production data to operate (Erboz, 2017). Big data is therefore considered as one of the cross-sectional, technological pillars for industry 4.0 (Machado et al., 2020). On the contrary, new challenges for effective data management arise as manufacturing data comprised more than twice the amount of data than healthcare, media and financial services in 2018 altogether (Reinsel et al., 2018). So-called "dark" data is a result. It describes data that is not used in any analysis due to reasons, such as un-structuredness or loss of timely significance, and was quantified at 90% of generated manufacturing data in 2020 (Corallo et al., 2022). The 5 V's (velocity, volume, value, variety and veracity) that describe big data, are likewise hurdles of exploiting the informative value of data and sufficient data management tools and qualified personnel are needed (Nagorny et al., 2017; Abedjan, 2022).

Manufacturing output objectives and key performance indicators (KPI) aim for high productivity with little wastage and downtime of machines. Data-driven automation and process control improvements are aided by newly formed Cyber–physical production systems (CPPS) as processes are digitised and smart sensors allow (near) real time monitoring (Tao et al., 2019). Within the Industry 4.0 framework, the synchronisation of quality management techniques with the evolving capabilities of CPPS opens up new opportunities for quality and organisational performance. As advancement of traditional quality management practices, like Total Quality Management or Six Sigma, ZDM integrates logistics, maintenance, process control and manufacturing technology and forms a more holistic quality management approach (Christou et al., 2022). Traditional practices are based on detection and post-process correction, whereas in ZDM defects are completely prevented through prediction and prescriptive measures (Powell et al., 2022). Therefore, predictive modelling for quality enhancement is one major sub-technology of AI to address ZDM in the context of steadily increasing complexity. For advanced process monitoring and prediction methods to be pervasively used in series production, however, they must overcome constraints, such as adaptability to data shifts, robustness and reliability (Lepenioti et al., 2020; Webb et al., 2018).

The current attention to data-centric AI addresses a certain gap between advances of AI-based algorithms and models, mostly driven by academia, and a general applicability of those developments to real-world problems. In contrast to benchmark datasets or a laboratory setup, the industrial production system is characterised by numerous external and uncontrolled variables, which cause noise and errors in data gathering and poor data quality. It still needs high domain-specific knowledge, not only for interpretation of results but especially for meaningful, semantic data preparation (Abedjan, 2022). Data-centric AI puts data quality and model independent pre-processing in focus before fine-tuning the predictive model itself. With the 'Data-Centric AI Competition' (Ng, 2021), Andrew Ng emphasised the significance of context-based data preparation. Contributors like Motamedi et al. (2021) followed this paradigm, focusing on steps of data quality improvement prior to training and fine-tuning a predictive model (data-centric before model-centric). It was demonstrated that improving data quality may have a far bigger impact on prediction accuracy than simply optimising a machine learning model's hyperparameters. This approach is encouraging in the context of real-world industrial applications in order to advance AI maturity levels from descriptive to prescriptive analytics.

3. Machine learning for anomaly detection

Anomaly detection methods are used to find abnormal states, instances or data points that differ from a sample within the normal data domain space. The significance of being harmful is defined by the domain individually and problem specific (Goldstein and Uchida, 2016). According to Chandola et al. (2009), there are three categories of anomalies. Firstly, point anomalies: one instance is outside the normal value space. Secondly, contextual anomalies: a data instance is anomalous only in one context but not in another. Thirdly, collective anomalies: a collection of related data instances that are anomalous together even if their individuals are normal, e.g. as a consequence of the frequency or duration that a certain value appears. The term anomaly detection is mainly used to refer to very imbalanced problems in terms of sporadic occurrence of abnormal samples between the majority of normal samples. Commonly, there are no labels available, hence, it is unknown which samples belong to which states (Chandola et al., 2012). In many domains, a major challenge is access to reliably labelled data for training of ML algorithms. In the industrial context, quality inspection is often based on sampling and performed by changing operators of different experience, which makes supervised ML methods hardly applicable.

Unsupervised ML approaches have been applied successfully in several domains, such as medicine (e.g. detection of critical cardiac arrhythmia, tumor detection with computed tomography), banking (e.g. fraudulent financial transactions, payments with stolen credit cards), security (e.g. surveillance, document forgery, network intrusion) (Goldstein and Uchida, 2016) but also engineering (e.g. critical machine state detection) (Nassif et al., 2021). With ML complex distributions and dependencies are modelled to distinguish between classes, clusters and sub-cohorts. In anomaly detection related to manufacturing quality this characteristic is utilised similarly to detect defective (NOK) samples that differ from the bigger cohort of healthy (OK) samples (Nassif et al., 2021).

For more complex data with higher order, intrinsic features, like in images and videos, deep learning techniques become evermore popular to solve the anomaly detection task due to their precision and robustness. In contrast to more traditional image processing methods combined with machine learning, deep learning methods, like Convolutional Neural Networks (CNNs), do not need prior knowledge or manual feature engineering. They extract relevant features directly from raw input by filtering and assembling patterns on subsequent layers (Balzategui et al., 2021). CNNs are particularly designed for time series, image and video data as filters process grid-like data by convolution instead of general matrix multiplication (Goodfellow and Bengio, 2016). Examples of application are (Sabokrou et al., 2018; Caggiano et al., 2019; Scime et al., 2020). With the goal of anomaly detection Autoencoders and Generative Adversarial Networks (GANs) present a valuable framework. In a GAN architecture, a generator and a discriminator compete against each other, where the generator tries to build replicas from input data and the discriminator tries to distinguish between real input and generated data (Goodfellow et al., 2014). This approach can be utilised for anomaly detection in terms of a measured distance metric from healthy to anomalous as the GAN is trained only on healthy samples.

4. Domain randomisation for synthetic data in-the-loop

To maintain consistent processing and high-quality components in rapid and large-scale production, (real-time) process monitoring becomes even more crucial for on-line inspection and defect prediction. Unsupervised approaches are unavoidable in environments where unique identifiers, timestamps and labels from quality inspection are lacking, to still gain insights from data, nevertheless. As an additional approach domain randomisation can be adapted to create synthetic

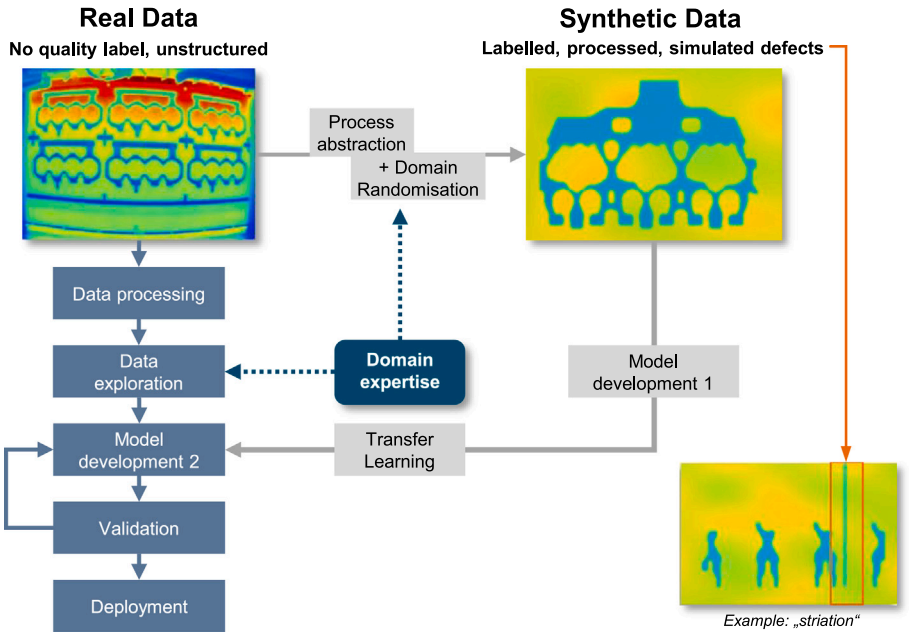

Fig. 1. Synthetic data in-the-loop (in false-colour representation for better visualisation): Data from the real process is abstracted and synthetic data is created by domain randomisation. Domain randomisation follows the hypothesis that if distributions are well projected from real to synthetic data and variability is significantly high in the synthetic dataset, then DL models, trained only on synthetic data, will generalise well on the real-world data.
Source: Figure adapted from (Zeiser et al., 2023).

data, either through simulation or through abstraction of the real domain. In particular when tasks include highly under-represented classes as in anomaly detection scenarios, the generation of labelled data and the vast volume of data at minimal cost are major advantages. As a result, domain randomisation enables evaluation in terms of precise labels and it may be used as a supplement to unsupervised learning methods. It is based on the hypothesis that DL models, trained only on synthetic data, will generalise well on the real-world data if distributions are well projected from real to synthetic data and variability is significantly high in the synthetic dataset (Tobin et al., 2017; Valtchev and Wu, 2021). Domain randomisation has been applied in robotics (Inoue et al., 2018; Tobin et al., 2017), image classification (Valtchev and Wu, 2021), pose estimation (Khirodkar et al., 2019) and traffic sign recognition (Villalonga et al., 2020) to enrich synthetic data, e.g. from simulation, with variability so that real-world data can be perceived as just another domain variation when synthetic and real data is mixed (Villalonga et al., 2020). Also in industrial processes, we see great potential for domain randomisation and synthetic data for predictive quality model development when access to labels cannot be established (see Fig. 1 for the whole concept of synthetic data in-the-loop). Following, we apply the approach to an additive manufacturing process.

4.1. Process data understanding

As use case we refer to data from additive manufacturing (AM) at BMW plant Landshut. AM comprises of several sub-technologies, defined by ISO 17296 (ISO/TC 261, 2015), and is mostly based on layer-wise workpiece creation, either by energy, binder or material deposition. Manufacturing free form geometries in a tool-independent way is a significant advantage compared to traditional manufacturing technologies. This allows quick adaption to revision indices or completely new products. As technological advancements could be achieved AM is not only used in rapid prototyping anymore but also applied in medium to high scale production. Especially binder jetting has a number of advantages for larger components and is applied successfully in series production (Gibson et al., 2021). Together with conventional technologies like casting it allows new geometric possibilities while

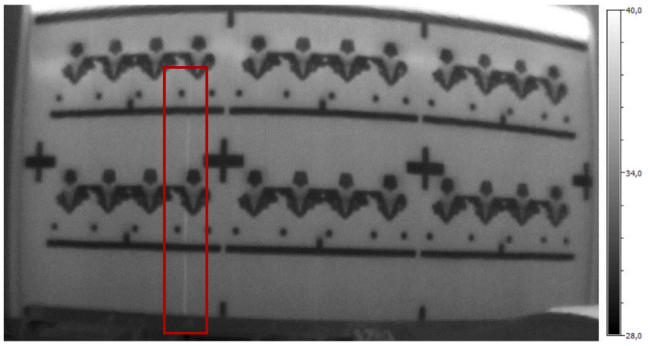

Fig. 2. Example of Type 2 anomaly (defect type “striation) that is related to product quality.

maintaining product stability and experience of established processes. Examples are automotive cylinder heads of BMW straight-four engines or structural parts of the car body.

Our research focuses on image data obtained from an infrared (IR) camera installed within the printing chamber of a binder jetting machine. As the process includes energy deposition by IR lamps for binder activation, thermal imaging allows a visualisation of the temperature distribution over the print bed. In order to support process optimisation effectively, a systematic image data analysis is required since binder and energy deposition have a significant impact on dimensional accuracy. Additionally, the aim is to detect anomalies that can become sources of product defects in (near-)real time. A print job consists of multiple hundred layers, of which images are taken, respectively. To allow comparison, the greyscale correspond to a fixed range of temperatures. Thus, the resulting dataset is comprised of $n \times j$ number of unlabelled images, with n being the number of jobs and j being the number of layers (see Fig. 2).

Two different types of anomalies must be differentiated for the binder jetting process. The first, Type 1, is connected to potential sensor failure or human interference that results in inaccurate data gathering, hence, a process flow outliers and a data quality issue. Contrarily, we

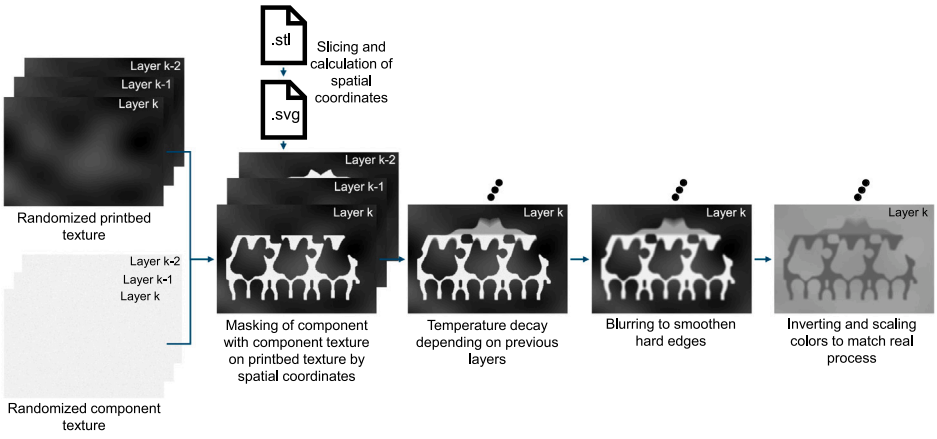

Fig. 3. Synthetic data creation process: Firstly, we slice a 3d-model and mask printbed and component with the created randomised textures. Then, we create a correlation of layer k to previous layers to model temperature decay over time. In a subsequent step, we blur, invert and scale the images to achieve an closer approximation to the data of the real AM process.
Source: Figure adapted from (Zeiser et al., 2023).

characterise Type 2 anomalies as abnormal variation of the normal production process that may result in defects and quality issues of the final product. These defects may also be only internally and not visible upon visual inspection. Typical types are porosities, agglomerates, striations an foreign objects. These cause layer disintegration which result in low material strength and broken sand cores during later casting operations. Dimensional errors are caused by other faults such layer shifts and sand agglomerates. The focus of an online process monitoring for anomaly detection is on Type 2 anomalies. Therefore Type 1 anomalies are filtered out in a series of cleaning and data preparation steps.

4.2. Synthetic data creation process

In AM the print job file is typically based on a Standard Tessellation Language file. We can reuse this format as basis for domain randomisation. Similar to how the AM print job is set up, we divide the STL file into layers and generate Support Vector Graphic (svg) files that specify the workpiece's layer-by-layer contours using spatial coordinates. We apply domain randomisation to closely resemble the changing temperature distribution of the printbed and component by dispersing the background and workpiece greyscale on texture maps (see Fig. 3).

We produce random layer-by-layer temperature distributions without regard to prior layers by masking the shapes defined by the edge coordinates and the parameterised work-piece location with the textures. To further simulate the actual AM process, parameters are modified to control the layer-to-layer temperature decrease and the opacification of the sharp work-piece edges. The minimum and maximum values of a randomly selected subsample of real images serve as the basis for the randomisation intervals for these job parameters. To improve variety, this method is carried out per layer as well as per job to be generated.

Next, aforementioned defects (Type 2 anomalies) are visually specified and positioned on particular layers per job. To avoid bias in the data, these anomalous jobs each have their own unique properties, such as geometry, location, and duration of anomaly in terms of the number of layers impacted. In order to exploit the resultant dataset of clean and anomalous jobs for further study and the validation of advancements, synthetic images, as shown in Fig. 4, are labelled respectively.

5. Generative unsupervised anomaly detection for AM

The real AM process under consideration may still have unknown anomaly types that are not represented in the synthetic data. Moreover, as mentioned previously, highly imbalanced problems and the lack of labels make the training and evaluation of supervised models difficult or impossible. Therefore, we consider the use of unsupervised methods for anomaly detection in our research. Furthermore, we use generative models such as DCGANs and VAEs due to their ability to handle complex and real data. Moreover, the presence of labels for our synthetic data allows us to evaluate the unsupervised models.

5.1. Experimental setup

The generative models were only trained on synthetic healthy data to solely learn the distribution of healthy samples and normalised depending on model requirements. We generated a total of 5000 healthy images and excluded 500 for evaluation and created 500 anomalous samples of each anomaly type shown in Fig. 4. The training parameters are shown in Table 1. The convolutional units of the VAE consist of convolutional layers with ReLU activations. The convolutional units of the DCGAN include batch normalisation and leaky ReLU activation functions, while the last convolution of the generator has a tanh activation to fit the normalised image pixel value bounds of [0,1]. The discriminator additionally features dropout layers to prevent overfitting of the model. However, the DCGANs trained by us still suffered from mode collapse, causing the generator to produce similar output images seemingly independent of the latent input. This issue is addressed by WGANs, where the discriminator not only discriminates real and fake images but gives a rating by the Wasserstein distance between the distributions of generator generated data and training data and therefore improves training (Arjovsky et al., 2017). Thus, we implemented an improved WGAN with gradient penalty (Gulrajani et al., 2017) to stabilise training and were able to generate realistic synthetic samples with dimensions of 128 by 128. The generator convolutional units consist of convolutional and batch normalisation layers paired with leaky ReLU activation functions. The discriminator convolutional units only feature dropout layers and leaky ReLU activations. Generated images for both GAN structures are shown in Fig. 5.

For training the VAE, we utilise the binary cross entropy because of our output bounds [0,1] and better training performance as well as the Kullback–Leibler divergence (KLD) for the latent space as specified in Kingma and Welling (2014) to fit the generative distribution to the training distribution. In comparison to standard Auto Encoders, the latent space is also defined as probability distributions. For the WGAN training, we use the Wasserstein-Loss approximation with gradient penalty as in Gulrajani et al. (2017). In addition, we train 3 discriminator iterations for each generator iteration as opposed to 5 in Gulrajani et al. (2017) to speed up the training process.

For the purpose of detecting anomalies, in this case data outside of the training data distribution, such as anomalies in Fig. 4, with the

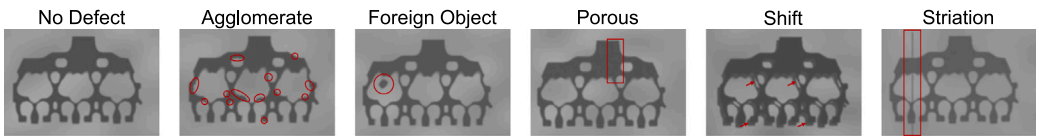

Fig. 4. Anomaly types included in the synthetic data (red: highlighted location of anomaly). Agglomerate: accumulation of unwanted loose components into a solid compound. Foreign Object: unknown objects on the printbed. Porous: Temperature difference in the direction of print head movement. Shift: Sudden change of the component location in the printbed. Striation: Vertical striations from excess binder, sand or nozzle-clogging. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Source: Figure adapted from (Zeiser et al., 2023).

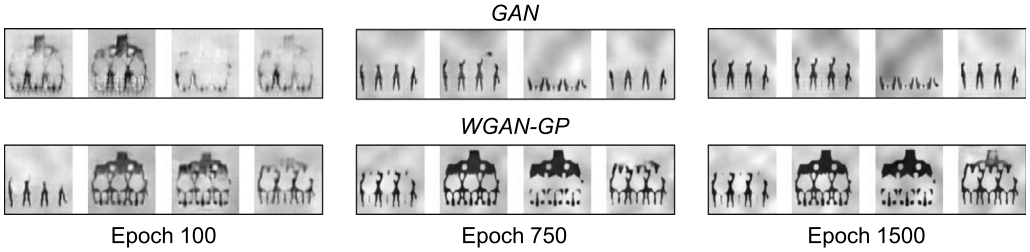

Fig. 5. Generated images of the DCGAN and the WGAN-GP during training with constant latent space vector. While WGAN successfully trains the distribution of the training data, GAN suffers of mode collapse and generates visually similar images.

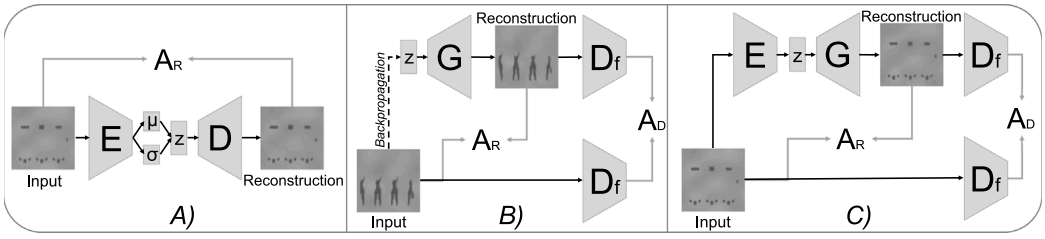

Fig. 6. Generative unsupervised anomaly detection architectures: (A) VAE with Reconstruction Error, (B) AnoGAN (Schlegl et al., 2017), (C) F-AnoGAN (Schlegl et al., 2019).

Table 1
Model and training parameters of generative methods.

	VAE	DCGAN	WGAN
Train data dimension	(4 500, 64, 64)	(4 500, 64, 64)	(4 500, 128, 128)
Batchsize	64	128	256
# Epochs	300	1 500	1 500
Convolutional units	3 for encoder 4 for decoder	4 for generator 2 for discriminator	5 for generator 5 for discriminator
Latent space dimension	16	100	128
# Trainable parameters	1 430 209	3 039 041	5 697 347

VAE we use the reconstruction error calculated by the mean squared error (MSE) between input image and VAE output image as shown in Fig. 6. Ideally, the VAE yields high reconstruction errors of anomalies, as it cannot reproduce image features outside the training distribution, while healthy samples obtain small reconstruction errors.

For anomaly detection with the use of GANs, we evaluate AnoGAN (Schlegl et al., 2017) and f-AnoGAN (Schlegl et al., 2019). For a better comparison, we use the trained WGAN for both methods, as it produces more realistic synthetic images than the DCGAN. As shown in Fig. 6, both methods compute an anomaly score consisting of two components:

$$A(X) = (1-\lambda) \cdot A_R(X, G(z(X))) + \lambda \cdot A_D(D_f(X), D_f(G(z(X)))) \quad (1)$$

While A_R describes the image reconstruction error between the input image X and generator output image $G(z(X))$, A_D describes the residual error of the discriminator feature of the same components. The discriminator features D_f are the outputs of the last convolutional layer as those features are most discriminative for the information trained by the discriminator (Salimans et al., 2016). For both error components, we compute the MSE and set the weighting factor $\lambda = 0.1$ as in Schlegl et al. (2017). The main difference between the methods is the approach to find $z(X)$ that minimises the error. AnoGAN finds $X \mapsto z$ (image to latent space representation) by utilising backpropagation steps for each input image, whereas f-AnoGAN separately trains an encoder solely for the task of mapping input images to the latent space $z(X) = E(X)$ (Schlegl et al., 2017, 2019).

In our AnoGAN approach, we use 3 500 backpropagation steps for the iterative reconstruction of test samples with the aim to minimising the anomaly score. Our f-AnoGAN encoder consists of 3 convolutional units with leaky ReLU activation functions and dropout layers to prevent overfitting. Training of the encoder is done in a subsequent step after WGAN training, using the same healthy input images X and resulting discriminator feature outputs $D_f(X)$ as labels for the loss function, which is defined as the anomaly score.

Our technical setup for the implementation of the methods and the evaluation is based on open source resources, namely Python running on version 3.7.8, TensorFlow on version 2.3, scikit-learn on version 0.21.2. As computational resource we use a Tesla V100 GPU on a shared NVIDIA DGX-1 platform.

5.2. Evaluation

Firstly, given the architectures shown in Fig. 6, the computation of the anomaly score components for VAE and f-AnoGAN requires only a

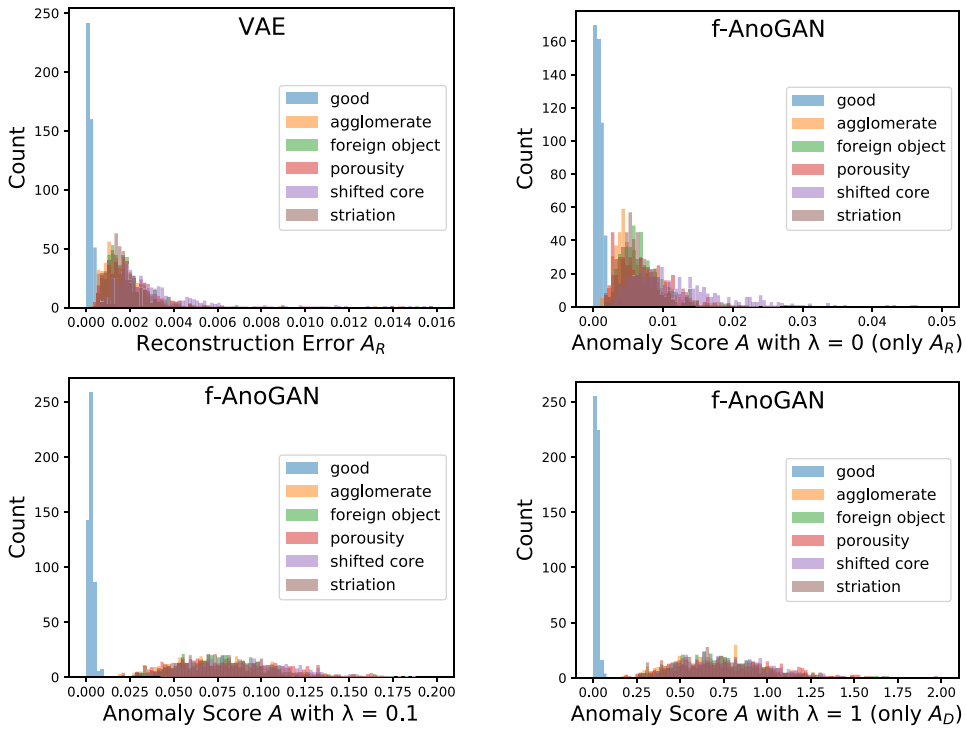

Fig. 7. Distribution of anomaly scores depending on anomaly detection architectures on the test dataset, coloured by anomaly types. For f-AnoGAN anomaly score, results are shown for $\lambda = \{0, 0.1, 1\}$ as weighing factors defined in Eq. (1). The discriminator feature error term A_D shows a clear differentiation capability between good and anomalous samples. For the reconstruction error A_R the distributions are partly overlapping which makes a clear differentiation and setting a threshold difficult without creating too many false negatives. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

forward pass of the input image. However, the AnoGAN architecture requires more computationally intensive backpropagation steps to find the latent space z for each input image. Moreover, the chosen 3 500 backpropagation steps do not converge to a suitable reconstruction for some input images. By choosing an even higher number of steps, the computation time increases proportionally. Depending on the current load of our shared GPU resources, the AnoGAN approach requires ~ 45 sec/image, making it infeasible for real world in-situ monitoring, which is why we will not inspect it further and focus on the VAE and f-AnoGAN approaches. For the computation of the test image outputs, the methods examined in more detail only require fragments of a second per image, thus enabling them to be implemented in a real world application.

The Histograms of the resulting anomaly scores for the test dataset consisting of 500 images for each class are shown in Fig. 7. For the f-AnoGAN approach, we tested $\lambda = \{0, 0.1, 1\}$, which essentially represents A as the combined anomaly score, as well as its components A_R and A_D separately. Each approach shows that the distribution of healthy/no defect synthetic images has smaller anomaly scores than these of anomalous samples, implying that the trained models can reconstruct healthy samples better than anomalous samples. The distributions of the reconstruction error A_R for the VAE and f-AnoGAN for good and anomalous samples have an overlapping range, which means that setting an absolute threshold for classification will result on the cost of having false positives or false negatives. Inclusion or sole consideration of the discriminator feature residual error A_D for the f-AnoGAN method shows a promising gap between the distributions of the good and anomalous samples. This confirms the statement that taking discriminator features into account yield a reliable indicator for anomalies (Schlegl et al., 2019).

However, to accomplish the task of finding anomalies in data, we still need to develop a method that assigns good or bad labels to the input data. We will follow the unsupervised approach to fit real world applications and use the labels only for evaluation. Essentially, we aim to cluster healthy and anomalous samples and as for the clustering methods considered, we want to compare an anomaly detection algorithm: Isolation Forest (Liu et al., 2008), a clustering method: k-Means, an automatic thresholding method: Otsu's (Otsu, 1979) and a method to split mixed Gaussian distributions: a non-Bayesian Gaussian mixture model (GMM) using the expectation–maximisation algorithm defined in scikit-learn platform. For training these methods, we split the test dataset defined in Section 5.1 and create a balanced training set of 300 healthy samples and 300 anomalous samples, equally divided with samples of the 5 anomalous classes shown in Fig. 4. The test dataset consisted of the remaining 200 healthy samples and a collection of 200 anomalous images. The evaluation of these methods using the anomaly scores of VAE and f-AnoGAN based on our synthetic labels is presented in Table 2. We used accuracy as the evaluation metric for our balanced binary data. While the reconstruction error of the VAE is insufficient for separation, the combined anomaly score $A(X)$ of the f-AnoGAN delivers promising results. In particular, the GMM is able to separate the distributions shown in Fig. 7 well.

Since there is a significant gap between the distributions in A_D, we believe there is room for improvement. Furthermore, we believe that compressing an image to a one-dimensional anomaly score reduces the information value. Therefore, we further inspect the feature map $D_f(X)$ of the input image itself and the feature map $D_f(G(E(X)))$ of the generated image by generator G. Each feature map has a dimension of $4 \times 4 \times 512 = 8192$ values. For comparison purposes, we train the same models once on the feature map $D_f(X)$, once on a stack of feature maps $[D_f(X), D_f(G(E(X)))]$, yielding an input dimension of 16384 values, and once on the element-wise subtraction of the feature maps $D_f(X) - D_f(G(E(X)))$. The subtracted feature maps using Isolation Forest deliver the most promising results as shown in Table 2. Isolation Forest is able to detect anomalies even for high-dimensional datasets with a large number of irrelevant attributes (Liu et al., 2008). Other

Table 2
Evaluation heatmap of different combinations from input data and clustering method. Best results achieved with elementwise subtracted features and Isolation Forest. Otsu's method was only used on 1d-input to calculate a threshold for clustering. Gaussian Mixture Model resulted in low memory error due to high input dimensions with stacked feature maps. (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

			Cluster method			
	Input data	Input dimension	Isolation Forest	K-Means	Otsu-Threshold	Gaussian Mixture Model
VAE	Reconstruction Error A_R	1	0.725	0.688	0.695	0.523
	Anomaly Score A	1	0.880	0.953	0.953	0.995
F-AnoGAN	Discriminator Features D_f	8192	0.525	0.520	–	0.515
	PCA on D_f	188	0.793	0.508	–	0.578
	poly kPCA on D_f	18	0.803	0.510	–	0.503
	rbf kPCA on D_f	29	0.863	0.520	–	0.543
	Stack of features $[D_f, D_f(G(E(X)))]$	16384	0.630	0.513	–	–
	PCA on feature stack	17	0.853	0.513	–	0.528
	poly kPCA on feature stack	29	0.870	0.510	–	0.513
	rbf kPCA on feature stack	27	0.855	0.530	–	0.503
	Elementwise subtraction of features $[D_f - D_f(G(E(X)))]$	8192	1.000	0.688	–	0.500
	PCA on subtracted features	4	1.000	0.675	–	0.995
	poly kPCA on subtracted features	7	1.000	0.508	–	0.998
	rbf kPCA on subtracted features	25	0.778	0.993	–	1.000

methods might suffer from the curse of dimensionality, meaning that larger dimensions result in a lower relevant information density and therefore lower scores. Thus, we use principal component analysis (PCA) to reduce dimension of the model input on the aforementioned feature maps directly and in other spaces transformed by polynomial and radial basis function kernels (Schölkopf et al., 1997). The shown number of used PCA components was incrementally identified by evaluating the Isolation Forest's accuracy. All in all, as shown in Table 2, this increases the performance of Isolation Forest for the input image feature map and the stack of feature maps, while the performance of k-Means and GMM do not differ. The dimensions of the model input can be reduced in the double digits space.

As for the subtracted feature maps, which are the direct input for the discriminator feature residual error A_D, the dimension reduction also greatly increased performances of k-Means and GMM. With only 4 linear PCA components, the GMM achieved 99.5% accuracy and with 25 kernelised PCA components with the radial basis function, the k-Means algorithm achieved 99.3% accuracy.

Furthermore, we increased the number of clusters in the hyperparameters of the clustering methods k-Means and the GMM to essentially separate different types of anomalies. As for the training data, we used 300 images of each class and computed the anomaly scores and feature maps to have a balanced dataset. The remaining 200 images for each class were used as validation data. However, even using the different input transformations as in Table 2, the trained models were not able to significantly cluster different anomaly types. The trained features and used models were only able to differentiate between healthy and anomalous samples but not classify multiple class.

All things considered, we believe that using subtractive F-AnoGAN features in combination with Isolation Forest is a better approach for locating unknown anomalies than simply computing an anomaly score. Especially, when real world data contains a lot of noise with small geometrical anomalies, the distributions of anomaly scores will overlap, making them insufficient for monitoring applications. We believe that collecting healthy samples and training of the WGAN and encoder is the most important step in this pipeline for anomaly detection, because the features of the discriminator heavily depend on the learned distribution.

6. Conclusion

Psarommatis et al. emphasise missing collaboration of academia and industry as a main problem in research on ZDM and advancing its methodologies (Psarommatis et al., 2020). With our work we counteract this shortcoming. One major challenge to advance potentials of deep learning for industrial processes, especially to further implement zero defect manufacturing, is access to relevant, clean and labelled data for model development. By synthesising data with domain randomisation and bringing it into the loop for development of anomaly detection pipelines, we identified great potential: Economically due to low cost of data creation without any physical production, as well as technically due to exact labelling. In the ZDM context this provides a possibility to develop data-based inspection methods already in parallel with manufacturing process and quality planning phases. Hence, in prototype production prior to product launch and start of serial production (SOP). This is an extension to what Powell et al. call "first-time-right and quality ramp-up minimization" as main research direction to advance ZDM (Powell et al., 2022).

Anomalies go beyond previously known defects and can occur as unknown, harmful deviations of the normal processing. As a consequence we chose a generative unsupervised anomaly detection approach over a simple supervised model. Based on f-AnoGAN, we implemented a method capable of detecting the anomalous samples in an unsupervised manner. In an evaluative comparison by utilising the synthetic labels, we proved that using clustering methods with features generated by the deep networks yields better results than computing an anomaly score solely. The presented, data-centric approach to artificial intelligence shifts attention towards data pre-processing and data quality. With this input optimisation, deep learning, like anomaly detection techniques, will likely perform better in real-world scenarios. Besides binder jetting and other AM techniques also different manufacturing technologies with in-line imaging inspection methods can benefit from the presented approach, e.g. automated welding, casting or coating. In cases where the starting point may not be a CAD file with the expected geometry, modifications must be made at synthetic data creation. However, the approach of domain randomisation and synthetic anomaly inclusion is independent of the manufacturing process. Future work will direct

towards enhancing domain randomisation in order to explore effects of different variability levels in synthetic data as well as analysing mixed datasets, i.e. real and synthetic for a comprehensive transfer learning approach.

Present work includes only object-like defects as anomalies. Though, trends, patterns and higher order features may be present, related to an uneven temperature distribution and indicating quality issues in the final part. Further analysis on unknown, more implicit anomalies is required. As generative models, especially WGAN, present good representation results for the AM dataset we will extend our experiments to non-geometric anomalies. These generative models are also further to be investigated by combining unlabelled big data with labelled small data or using combinations of real and synthetic data.

CRediT authorship contribution statement

Alexander Zeiser: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Visualization, Investigation. **Bekir Özcan:** Methodology, Software, Writing – original draft, Data curation, Validation, Visualization, Investigation. **Bas van Stein:** Supervision. **Thomas Bäck:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alexander Zeiser reports financial support and article publishing charges were provided by Bayerische Motoren Werke AG. Bekir Oezcan reports financial support and article publishing charges were provided by Bayerische Motoren Werke AG.

Data availability

The authors do not have permission to share data.

References

Abedjan, Z., 2022. Enabling data-centric AI through data quality management and data literacy. IT - Inf. Technol. 64 (1–2), 67–70. http://dx.doi.org/10.1515/itit-2021-0048.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, pp. 214–223.

Balzategui, J., Eciolaza, L., Maestro-Watson, D., 2021. Anomaly detection and automatic labeling for solar cell quality inspection based on generative adversarial network. Sensors 21 (13), http://dx.doi.org/10.3390/s21134361.

Caggiano, A., Zhang, J., Alfieri, V., Caiazzo, F., Gao, R., Teti, R., 2019. Machine learning-based image processing for on-line defect recognition in additive manufacturing. http://dx.doi.org/10.1016/j.cirp.2019.03.021.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41 (15), http://dx.doi.org/10.1145/1541880.1541882.

Chandola, V., Banerjee, A., Kumar, V., 2012. Anomaly detection for discrete sequences: A survey. IEEE Trans. Knowl. Data Eng. 24 (5), 823–839. http://dx.doi.org/10.1109/TKDE.2010.235.

Christou, I.T., Kefalakis, N., Soldatos, J.K., Despotopoulou, A.M., 2022. End-to-end industrial IoT platform for Quality 4.0 applications. Comput. Ind. 137, 103591. http://dx.doi.org/10.1016/j.compind.2021.103591.

Corallo, A., Crespino, A.M., Vecchio, V.D., Lazoi, M., Marra, M., 2022. Understanding and defining dark data for the manufacturing industry. IEEE Trans. Eng. Manage. 1–13. http://dx.doi.org/10.1109/TEM.2021.3051981.

Dogan, A., Birant, D., 2021. Machine learning and data mining in manufacturing. Expert Syst. Appl. 166, 114060. http://dx.doi.org/10.1016/j.eswa.2020.114060.

Erboz, G., 2017. How to define industry 4.0: Main pillars of industry 4.0. Manag. Trends Dev. Enterp. Glob. Era (November 2017), 761–767, URL https://spu.fem.uniag.sk/fem/ICoM_2017/files/international_scientific_conference_icom_2017.pdf.

Gibson, I., Rosen, D.W., Stucker, B., 2021. Additive Manufacturing Technologies. Springer US, Boston, MA, http://dx.doi.org/10.1007/978-1-4419-1120-9.

Goldstein, M., Uchida, S., 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS One 11 (4), http://dx.doi.org/10.1371/journal.pone.0152173.

Goodfellow, I.J., Bengio, Y., 2016. Convolutional networks. In: Goodfellow, I., Bengio, Y., Courville, A. (Eds.), Deep Learning. Cambridge University Press, Ch. 9.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 3, (January), pp. 2672–2680. http://dx.doi.org/10.3156/jsoft.29.5{_}177{_}2.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of Wasserstein GANs. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5769–5779.

Hristov, I., Chirico, A., 2019. The role of sustainability key performance indicators (KPIs) in implementing sustainable strategies. Sustainability (Switzerland) 11 (20), http://dx.doi.org/10.3390/su11205742.

Inoue, T., Choudhury, S., De Magistris, G., Dasgupta, S., 2018. Transfer learning from synthetic to real images using variational autoencoders for precise position detection. In: IEEE International Conference on Image Processing, no. October. ICIP, IEEE, pp. 2725–2729. http://dx.doi.org/10.1109/ICIP.2018.8451064.

ISO/TC 261, 2015. ISO 17296-2:2015 Additive manufacturing — General principles — Part 2: Overview of process categories and feedstock.

Khirodkar, R., Yoo, D., Kitani, K., 2019. Domain randomization for scene-specific car detection and pose estimation. In: 2019 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 1932–1940. http://dx.doi.org/10.1109/WACV.2019.00210.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational Bayes auto-encoding variational Bayes. In: International Conference on Learning Representations. ICLR.

Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. Int. J. Inf. Manag. 50, 57–70. http://dx.doi.org/10.1016/j.ijinfomgt.2019.04.003.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. In: Eighth IEEE International Conference on Data Mining, no. October 2018. IEEE, pp. 413–422. http://dx.doi.org/10.1109/ICDM.2008.17.

Machado, C.G., Winroth, P., Hans, E., 2020. Sustainable manufacturing in Industry 4.0: An emerging research agenda. Int. J. Prod. Res. 58 (5), 1462–1484. http://dx.doi.org/10.1080/00207543.2019.1652777.

Motamedi, M., Sakharnykh, N., Kaldewey, T., 2021. A data-centric approach for training deep neural networks with less data. In: 35th Conference on Neural Information Processing Systems. NeurIPS 2021.

Nagorny, K., Lima-Monteiro, P., Barata, J., Colombo, A.W., 2017. Big data analysis in smart manufacturing: A review. Int. J. Commun., Network Syst. Sci. 10 (03), 31–58. http://dx.doi.org/10.4236/ijcns.2017.103003.

Nassif, A.B., Talib, M.A., Nasir, Q., 2021. Machine learning for anomaly detection: A systematic review. IEEE Access 9, 78658–78700. http://dx.doi.org/10.1109/ACCESS.2021.3083060.

Ng, A., 2021. MLOps: From model-centric to data-centric AI. URL https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 20 (1), 62–66.

Powell, D., Magnanini, M.C., Colledani, M., Myklebust, O., 2022. Advancing zero defect manufacturing: A state-of-the-art perspective and future research directions. Comput. Ind. 136, 103596. http://dx.doi.org/10.1016/j.compind.2021.103596.

Psarommatis, F., May, G., Dreyfus, P.-A., 2020. Zero defect manufacturing: State-of-the-art review, shortcomings and future directions in research. Int. J. Prod. Res. 58 (1), 1–17. http://dx.doi.org/10.1080/00207543.2019.1605228.

Reinsel, D., Gantz, J., Rydning, J., 2018. The Digitization of the World From Edge to Core. Tech. Rep., IDC, Seagate.

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R., 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vis. Image Underst. 172 (October 2017), 88–97. http://dx.doi.org/10.1016/j.cviu.2018.02.006.

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.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54 (May), 30–44. http://dx.doi.org/10.1016/j.media.2019.01.010.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Information Processing in Medical Imaging, Vol. 10265. IPMI 2017, pp. 146–157. http://dx.doi.org/10.1007/978-3-319-59050-9{_}12.

Schölkopf, B., Smola, A., Müller, K.-R., 1997. Kernel principal component analysis. In: Artificial Neural Networks. ICANN'97, pp. 583–588. http://dx.doi.org/10.1007/BFb0020217.

Scime, L., Siddel, D., Baird, S., 2020. Layer-wise anomaly detection and classification for powder bed additive manufacturing processes. Addit. Manuf. 36 (March), 101453. http://dx.doi.org/10.1016/j.addma.2020.101453.

Tao, F., Qi, Q., Wang, L., Nee, A., 2019. Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison. Engineering 5 (4), 653–661. http://dx.doi.org/10.1016/j.eng.2019.01.014.

Tobin, J., Fong, R., Ray, A., 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems. IROS, IEEE, pp. 23–30. http://dx.doi.org/10.1109/IROS.2017.8202133.

Valtchev, S.Z., Wu, J., 2021. Domain randomization for neural network classification. J. Big Data 8 (1), 94. http://dx.doi.org/10.1186/s40537-021-00455-5.

Villalonga, G., Van de Weijer, J., López, A.M., 2020. Recognizing new classes with synthetic data in the loop: Application to traffic sign recognition. Sensors 20 (3), 583. http://dx.doi.org/10.3390/s20030583.

Webb, G.I., Lee, L.K., Goethals, B., Petitjean, F., 2018. Analyzing concept drift and shift from sample data. Data Min. Knowl. Discov. 32 (5), 1179–1199. http://dx.doi.org/10.1007/s10618-018-0554-1.

Zeiser, A., Özcan, B., Kracke, C., van Stein, B., Bäck, T., 2023. A data-centric approach to anomaly detection in layer-based additive manufacturing. at - Automatisierungstechnik 71 (1), 81–89. http://dx.doi.org/10.1515/auto-2022-0104.