Anomaly detection in smart manufacturing: An Adaptive Adversarial Transformer-based model

Moussab Orabi, Kim-Phuc Tran, Philipp Egger, Sebastien Thomassey

Journal of Manufacturing Systems 77 (2024) 591–611

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Technical paper

Anomaly detection in smart manufacturing: An Adaptive Adversarial Transformer-based model

Moussab Orabi [a,b,c], Kim Phuc Tran [b], Philipp Egger [a], Sébastien Thomassey [b,*]

[a] Rosenberger Hochfrequenztechnik GmbH & Co. KG, Hauptstraße 1, 83413 Fridolfing, Germany
[b] Univ. Lille, ENSAIT, ULR 2461 - GEMTEX - Génie et Matériaux Textiles, F-59000 Lille, France
[c] International Chair in DS & XAI, International Research Institute for Artificial Intelligence and Data Science, Dong A University, Danang, Viet Nam

ARTICLE INFO

Keywords:
Industry 5.0
Anomaly detection
Artificial intelligence
Smart manufacturing
Transformer
Deep learning
Multivariate time series data

ABSTRACT

In Industry 5.0, smart manufacturing brings additional intricacies and novel data processing challenges. Given the evolving nature of manufacturing processes and the inherent complexity of data, including noise and missing entries, achieving accurate anomaly detection becomes even more intricate. Conventional methods often miss nuanced anomalies, especially when dealing with high-dimensional, multivariate, non-stationary data. These data types are typical of smart manufacturing environments. Hence, many recent approaches have embraced deep learning to confront these challenges, making use of diverse attention mechanisms to acquire data representations. However, in manufacturing, where the dynamics of time series data change over time, methods relying solely on pointwise or pairwise representations often fall short. Thus, ensuring product quality and operational integrity calls for even more advanced methodologies. The deficiency lies in the capability of state-of-the-art models to effectively capture abnormal patterns while considering both local and global contextual information. This challenge is compounded by the rarity of anomalies, making it exceedingly challenging to establish substantial associations between individual abnormal points and the entire time series. To tackle these challenges, we introduce the "**A**daptive **A**dversarial **T**ransformer" as a novel deep learning technique that combines Transformer architecture with an anomaly attention mechanism and Adversarial Learning. Our Model effectively captures intricate temporal patterns, distinguishes normal and anomalous behaviors, and dynamically adjusts thresholds to align with the evolving dynamics of time-series data. Empirical validation on four benchmark datasets and three real-world manufacturing datasets demonstrates our model's effectiveness compared to the state-of-the-art, as evidenced by the F1-Score.

1. Introduction

Industry 5.0 represents a pivotal shift in manufacturing, emphasizing the integration of Artificial Intelligence (AI) with human expertise to forge a new era of collaborative and intelligent systems. Building upon the interconnected frameworks established by Industry 4.0 [1], Industry 5.0 enhances the analytical capabilities of AI with the nuanced understanding of human operators, fostering a sophisticated interaction between technology and intuition [2]. At the core of this evolution is the symbiotic relationship between AI and human insight, where AI systems analyze vast datasets to identify potential anomalies and inefficiencies, and human operators apply their deep contextual knowledge to interpret and manage these insights. This approach not only enhances the responsiveness of manufacturing systems but also significantly improves the effectiveness of anomaly detection mechanisms. Human-in-the-loop (HITL) systems [3] exemplify this integration by embedding human judgment within automated processes, proving invaluable in complex scenarios where anomalies are subtle or ambiguous. The progression of manufacturing technologies has also necessitated more advanced monitoring and management methods. Traditional anomaly detection often relied on manual inspections, which are time-consuming and miss finer details. Recent advancements, including Digital Twins and Edge Intelligence [4], facilitate real-time monitoring and decision-making, representing a significant step forward in automating anomaly detection processes. Furthermore, Multivariate Time Series (MVTS) data in manufacturing presents unique challenges due to its complexity and dynamic nature. While traditional methods such as One-Class SVMs have been utilized, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [5] have shown greater potential in managing such data, capturing crucial temporal dependencies. However, the high computational costs and training difficulties

* Corresponding author.
E-mail address: sebastien.thomassey@ensait.fr (S. Thomassey).

https://doi.org/10.1016/j.jmsy.2024.09.021
Received 12 December 2023; Received in revised form 27 August 2024; Accepted 29 September 2024

associated with these networks on large datasets underscore the need for more efficient and robust approaches [6]. Recently, there has been a significant trend towards utilizing the emerging Transformer architecture in anomaly detection, benefiting from its robustness and ability to capture fine and long-term contextual anomalies [7,8]. However, much work still needs to be done in developing the proper attention mechanisms and the training style for this architecture [9]. Moreover, while Transformers are used as a reconstruction system, there is still a need to enhance them with a mechanism for classifying inputs into normal or abnormal categories. This method needs to be adaptive to the intricate changes in the dynamics of the input data as expected in an industrial setting.

In response to these challenges, the main contributions of our study are as follows:

- Our research introduces the Adaptive Adversarial Transformer (AAT), a novel architecture that combines the strengths of Transformers with a dual attention mechanism and adversarial learning to improve anomaly detection accuracy and efficiency.
- We integrate a Support Vector Data Description (SVDD) model enhanced with a Radial Basis Function (RBF) kernel. Our approach adapts dynamically to changes in data characteristics, significantly reducing the necessity for frequent model retraining.
- We distinguish between the input and reconstructed input using a vector approach rather than a scaled-based one. This method enables distinctions to be based not only on the scale of the values but also on their directional characteristics, providing a more nuanced understanding of anomalies. This dynamic adaption to changes in data characteristics, significantly reduces the need for frequent model retraining [10].
- Our results demonstrate AAT's superiority in the field of anomaly detection. It surpasses numerous well-established techniques, achieving improvements of up to 10% as measured by the F1 score metric across both public and internal datasets. This not only aligns with but also exceeds, Rosenberger's expectations regarding efficiency and scalability.

The remainder of the paper is organized as follows: Section 2 reviews pertinent literature to frame the technological and conceptual foundations of our research. Section 3.1 delves into the specific challenges and solutions for anomaly detection in manufacturing MVTS data. Section 3 elaborates on our proposed AAT framework, while Sections 4 and 5 provide a comprehensive evaluation of our method's performance against established benchmarks. The paper concludes with Section 6, discussing future research directions and the potential impact of our findings on the field of smart manufacturing.

2. Contextual overview and related work

This section provides a comprehensive overview of the foundational concepts and relevant research in the field of Anomaly Detection in Multivariate Time Series Data.

2.1. Anomalies: Understanding definitions and overcoming challenges

In the context of handling anomalies in MVTS, it is crucial to distinguish between anomalies and outliers while also recognizing their intersections. Fundamentally, anomalies can be seen as truly aberrant or unforeseen events, whereas outliers are primarily data points that deviate from the typical value range. For example, a data point that drastically surpasses or falls short of its counterparts in a dataset might be labeled an outlier. In contrast, a data point representing a measurement glitch or an illicit transaction exemplifies an anomaly. Additionally, outlier detection (OD) serves a descriptive purpose, while anomaly detection leans more towards predictive methodologies. Therefore, in many scenarios, OD is employed as a preliminary step to ensure data integrity before training robust anomaly detection models. The traditional definition of an anomaly, as described by Hawkins [11], is: "An anomaly is an observation which significantly differs from other data points, raising suspicion that it originates from a distinct mechanism". While anomalies are sometimes obvious and can be easily identified by data scientists using straightforward methods, the presence of noise often complicates anomaly detection. The primary challenge lies in differentiating between noise and actual anomalies, as illustrated in Fig. 1.

The distinction between anomalies and outliers is often ambiguous, with the terms frequently being used interchangeably. Another significant challenge in anomaly detection, particularly in the context of manufacturing processes, is the issue of imbalanced data. Class imbalance (CI) is a prevalent problem where the number of data samples for different classes (such as 'normal' and 'anomalous') is disproportionately distributed. In manufacturing, where production efficiency is high and anomalies are rare, this imbalance becomes particularly pronounced. This imbalance poses a challenge in developing effective machine learning models, as models tend to be biased towards the majority class, often leading to poor performance in detecting rare, yet critical, anomalous events. Furthermore, the cost of acquiring labeled data in such imbalanced scenarios is another significant hurdle. Labeling data requires expert knowledge, especially for complex manufacturing processes where understanding what constitutes an anomaly can be intricate. The expense and effort involved in obtaining accurately labeled data cannot be understated, making it a crucial factor in the feasibility and effectiveness of anomaly detection models. The issue of class imbalance in manufacturing and its implications on data modeling is extensively reviewed by de Giorgio et al. [12]. They systematically analyze various machine learning and deep learning solutions to address the class imbalance problem in the manufacturing domain. Their work underscores the importance of considering class imbalance and the challenges of labeled data in developing robust anomaly detection systems in manufacturing.

To counteract this imbalance, techniques such as the oversampling of the minority class and the downsampling of the majority class are employed. This ensures that during the model training phase, the presence of the majority class does not overshadow the critical minority class, allowing for an equitable representation of all classes within the decision-making algorithm. To address the class imbalance in anomaly detection, techniques like oversampling the minority class and downsampling the majority class are used. This ensures balanced representation during model training. Oversampling with the Synthetic Minority Over-sampling Technique (SMOTE) [13] generates synthetic samples for the minority class through interpolation. Downsampling using a cluster-based method reduces the majority class size by clustering and selecting representative instances from each cluster, typically using centroids calculated from the clusters. Additionally, overfitting-where the model learns both signal and noise from the training data is a common issue. To mitigate overfitting, methods such as cross-validation, reducing the number of features, pruning, and regularization are employed. Traditional techniques like cross-validation and step-wise regression are effective for small feature sets, while regularization is preferable for large feature sets.

Not only do we quickly encounter the problem of imbalanced datasets in anomaly detection, but we also face the issue of overfitting, especially when the data model has a large number of features. Overfitting occurs when the model learns both the signal and the noise in the training data, leading to poor performance on new, unseen data. There are several methods to avoid overfitting, such as cross-validation sampling, reducing the number of features, pruning, and regularization. Conventional methods like cross-validation and step-wise regression work well with a small set of features, while regularization techniques are more suitable when dealing with a large set of features.

Optimizing the threshold between normal data and anomalies is also very complex, particularly with high-dimensional, multivariate,

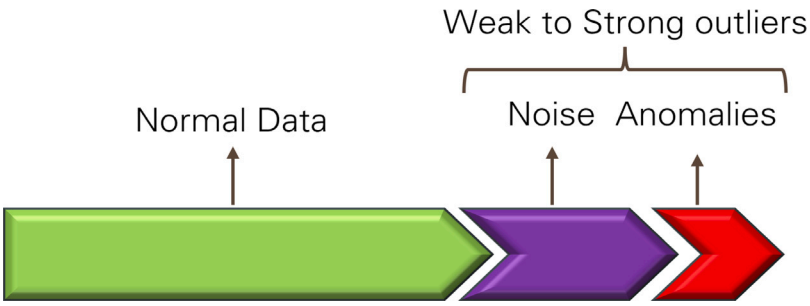


Fig. 1. Increasing outliers from left to right.

non-stationary data, such as that found in smart manufacturing. In this context, adaptive thresholding is essential for anomaly detection. This dynamic approach outperforms static methods by adjusting thresholds in response to data evolution, thereby accurately identifying anomalies indicative of operational issues or quality defects. Methods include Peak Over Threshold (POT), which identifies anomalies by setting a threshold for extreme values [14]; Kernel Quantile Estimation (KQE), which uses kernel-based techniques for quantile estimation to handle non-linear relationships [15]; Support Vector Data Description (SVDD), a supervised learning method that dynamically adjusts thresholds based on discrepancies between reconstructed and actual values, effectively reducing both False Positive Rate (FPR) and False Negative Rate (FNR) [16]; and Non-parametric Dynamic Threshold (NDT), which offers a dynamic, non-parametric approach without relying on assumed data distributions, making it ideal for changing manufacturing environments [17].

SVDD, in particular, provides a robust framework for adaptive anomaly detection by establishing a dynamic threshold for identifying outliers [16]. It constructs a hypersphere around the data distribution, aiming for the smallest volume that encloses the bulk of the data points, with allowances for anomalies. The SVDD optimization problem seeks to minimize the sphere's radius R of the hypersphere with center a and the slack variables ξ_i that account for outliers:

$$\min_{R,a} R^2 + C \sum_i \xi_i \quad \text{s.t.} \quad \|X_i - a\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \tag{1}$$

with C as the regularization parameter, X_i for the data points.

A new sample Z is classified as an outlier if its distance from the center a exceeds the radius R:

$$\|Z - a\|^2 > R^2 \Rightarrow \text{Outlier}, \tag{2}$$

The use of kernel functions further enhances the adaptivity of SVDD, allowing it to conform to non-linear data distributions and providing a flexible approach to setting thresholds in complex manufacturing environments [18].

2.2. Evolution of time series anomaly detection methods

The evolution of time series anomaly detection methods, particularly in the context of manufacturing-derived data, reflects a progressive integration of various techniques to handle the increasing complexity of data dynamics. Traditional statistical methods like ARIMA (Auto Regressive Integrated Moving Average) and Exponential Smoothing initially laid the groundwork, yet struggled to capture the intricate patterns in manufacturing data due to their deterministic nature [19,20]. Stochastic learning methods, such as Gaussian Processes and Hidden Markov Models, introduced probabilistic modeling, offering greater flexibility, though often at the cost of increased computational intensity [21]. The focus then shifted to outlier detection techniques, including Isolation Forest and Local Outlier Factor (LOF), which effectively addressed high-dimensional data but occasionally led to false alarms [22]. These limitations, as highlighted by Choi et al. [23], became more evident when dealing with complex sensor data irregularities, such as point, contextual, and collective anomalies. This led to advocacy for deep learning models, given their ability to identify intricate nonlinear relationships. Consequently, modern approaches often combine traditional and machine learning techniques, such as the Isolation Forest-LOF (IF-LOF) hybrid model, to enhance anomaly detection capabilities. With the advent of machine learning, algorithms like Support Vector Machines (SVM) and Decision Trees were adapted for anomaly detection, offering the advantage of handling large datasets and high dimensionality. However, these models often required extensive feature engineering and were sensitive to hyperparameters [10]. The limitations of classical and conventional methods highlight the need to harness deep learning models in anomaly detection applications. Long Short-Term Memory (LSTM) networks, being auto-regressive, captured the essence of sequential data dependencies. The *LSTM-NDT* model by Hundman et al. [17] employed this approach, predicting future data based on past feedback. However, its limitations arise from inefficiencies in modeling elongated temporal patterns, more so when faced with noisy data, as pointed out by Su et al. [24]. Deep Autoencoding Gaussian Mixture Model (DAGMM) introduced by Zong et al. [25], combines the strengths of deep autoencoding with a Gaussian mixture model. Despite the advantages of its decoupled training, it grapples with limitations such as slow computational performance and challenges in exploiting inter-modal correlations. In contrast, Li et al. [26] put forth an unsupervised anomaly detection technique tailored for rotating machinery, leveraging a memory-augmented temporal convolutional autoencoder. Further adding to the evolution of deep autoencoders, Unsupervised Anomaly Detection (USAD) [27] integrates an autoencoder with dual-decoders, carving a niche by significantly streamlining the training process. OmniAnomaly by Su et al. [24] marks a significant stride with its stochastic recurrent neural network, integrating elements from LSTM-VAE [28]. It proposes the Peak Over Threshold (POT) method [14], ushering in remarkable performance improvements. However, it comes at the expense of heightened training times. Anomaly detection has increasingly embraced attention-based architectures, especially in the context of smart manufacturing. Initially, models like *HitAnomaly* [9] integrated conventional transformers within an encoder–decoder framework, primarily targeting natural language log data. However, this approach showed limitations in handling continuous time-series datasets typical in manufacturing. Responding to these limitations, newer models such as *TranAd* (Transformer for Anomaly Detection) [8] and *Anomaly Transformer* [7] have evolved. These models shift focus from intensive transformer designs to more specialized attention mechanisms, better suited for the nuanced needs of time-series data in industrial settings. Further diversifying the application of transformers in anomaly detection, the *MLPT* (Multi-Layer Parallel Transformer) model by Leng et al. [29] represents a significant step forward. It demonstrates the versatility of transformer architectures in addressing the complexities of smart manufacturing systems, particularly in product quality issue detection and rapid anomaly localization.

Fig. 2 illustrates the recent benchmark anomaly detection models across three different learning approaches: supervised, unsupervised,

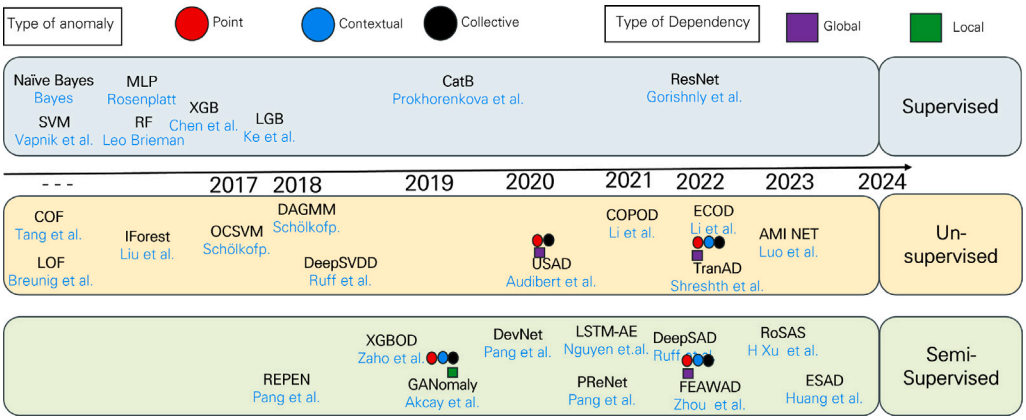


Fig. 2. A State-of-the-art perspective.

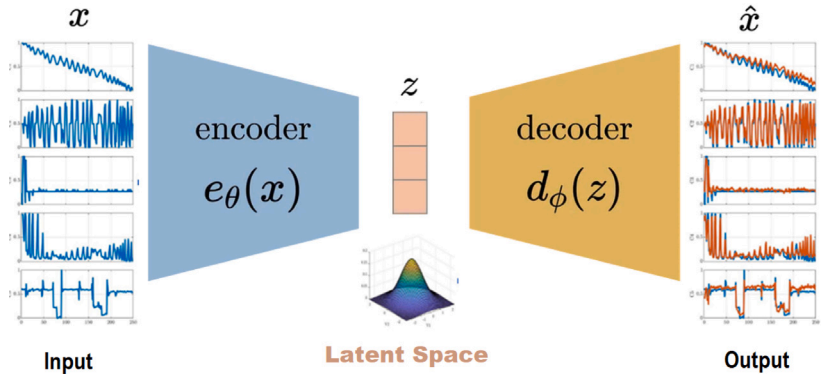


Fig. 3. Schematic of AE with latent space.

and semi-supervised. The figure demonstrates a trend towards using deep learning models in unsupervised and semi-supervised contexts. Additionally, it showcases the capability of some models to address anomalies in Multivariate Time Series (MVTS) from two distinct perspectives: the type of anomalies (point, contextual, and collective) and the type of dependency these anomalies have on the rest of the data (global and local). This visualization helps in understanding the evolving landscape of anomaly detection methodologies and their respective strengths in handling complex data interactions.

2.3. Emerging techniques for anomaly detection

In light of the evolution of the developed models in the last decade, three techniques emerged for their promising capacity to deal with issues related to anomaly detection in the manufacturing context. Our framework synergizes Autoencoder-Based Anomaly Detection, Transformer-based Architectures, and Adversarial Learning Frameworks, leveraging the strengths of each to enhance the detection of anomalous patterns in manufacturing data.

2.3.1. Autoencoder-based anomaly detection

Anomaly detection is often cast as a binary classification task, aimed at differentiating between normal and anomalous data instances. Given the rarity of anomalous samples, unsupervised learning approaches are favored, where models are trained solely on normal data. A prominent unsupervised method is the Autoencoder (AE), a type of neural network that excels in data compression and denoising tasks [30,31]. AE has two core components: the encoder, which compresses the input into a latent space, and the decoder, which attempts to reconstruct the input from the compressed form, as illustrated in Fig. 3.

The encoder maps the input X to a latent representation Z using a parametric function e_θ, which includes a linear transformation followed by a non-linear activation:

$$Z = e_\theta(X), \tag{3}$$

The decoder objective is to reconstruct the original input from Z using a similar parametric function d_ϕ:

$$\hat{x} = d_\phi(Z), \tag{4}$$

Training involves adjusting the AE's parameters to minimize the reconstruction loss, which measures the discrepancy between the original input x and the reconstructed output $\hat{x}$ during the training phase:

$$\text{loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2, \tag{5}$$

Once the AE is trained, its reconstruction error is utilized to calculate an anomaly score for new data points during the inference phase:

$$\text{Anomaly Score} = \|x - \hat{x}\|_2, \tag{6}$$

Despite the similarity in their calculation, the loss function is a training criterion, while the anomaly score is a post-training metric used to identify anomalies. While autoencoders form a foundational approach for anomaly detection, our AAT framework overcomes its limitations in handling high-dimensional, non-stationary data through its innovative architecture and adaptive learning mechanisms.

2.3.2. Transformer based architecture

The utilization of Transformer architecture for anomaly detection has witnessed a consistent rise in both academic and industrial spheres,

owing to its demonstrated effectiveness. This approach has found practical applications across various domains, including cybersecurity [32], telecommunications and networking [33], healthcare [34], and has recently garnered increased interest in industrial contexts, exemplified by its use in industrial quality monitoring [35] and real-time anomaly detection for time series data from industrial furnace [36].

The transformer, a neural network architecture, was first introduced in the paper "Attention is All You Need" by Vaswani et al. [37] . Since its inception, it has formed the basis for several significant projects, including Google's BERT and OpenAI's GPT series, all of which have delivered performance outcomes surpassing prior benchmarks. Although similar to Recurrent Neural Networks (RNNs) in their design for processing sequential data, transformers are unique in their structure, comprising an encoder and a decoder. Their defining feature is the utilization of self-attention mechanisms, enabling them to recognize dependencies throughout an entire sequence regardless of distance. This is in contrast to RNNs, which depend on recurrent connections for transmitting information through sequential time steps. Both the encoder and decoder in the transformer architecture consist of stacked layers that include multi-head self-attention (Fig. 4(a)), additive connections, layer normalization, and position-wise feed-forward layers. Additionally, positional encoding is applied to the input embeddings in a pre-processing step, ensuring that the model can retain the positions of elements within the sequence, as the transformer architecture does not inherently incorporate any information about sequence order. The core innovation of the transformer is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input sequence when processing each element. The formula for computing the attention scores for each element in the input sequence are as follows:

$$\text{Attention}(Q, K, V) = \text{Sofmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{7}$$

Here, Q represents the query matrix, encapsulating the input sequences in vector form. K denotes the keys, with sequences similarly represented as vectors, with d_k as its dimension and K^T the transpose of K, and V signifies the values, also depicted in vectorial sequence format. The attention mechanism, applied in a repeated manner with varying projections of Q, K, and V, is designed to enhance the model's interpretative capabilities. This process of parallel application is depicted in Fig. 4(b). Linear transformations are achieved by the multiplication of Q, K, and V with the training-evolved weight matrices W. The multi-head attention module, a pivotal element in connecting the encoder and decoder, integrates the encoder's input sequence with that of the decoder, considering up to a specified point in the sequence. Post the multi-head attention phase, both the encoder and decoder incorporate a feed-forward layer.

The utilization of transformer models, notably their independence from Recurrent Neural Networks (RNNs), has been pivotal in enhancing performance across a spectrum of domains. This includes significant advancements in language translation and natural language processing (NLP), as expounded in the works of Wolf et al. [38] and further contributions detailed in the HuggingFace's Transformers [39]. In the field of computer vision, notable developments are highlighted in studies such as those by Gao et al. [40] and Ayoub et al. [41]. Moreover, in the manufacturing sector, transformers have been instrumental in predictive maintenance, with models like Trans-Lighter discussed in [42]. Additionally, their application in human activity recognition (HAR) has been effectively explored, as delineated in [43]. However, Transformers are not without limitations. These include their high computational complexity, especially for long sequences, due to the self-attention mechanism which scales quadratically with the sequence length. Additionally, Transformers require large amounts of training data to achieve optimal performance, making them less effective for tasks with limited data availability. Training Transformers can be time-consuming, often requiring significant computational resources and extended training times. Furthermore, due to their high capacity, Transformers are prone to overfitting, especially when trained on small datasets without proper regularization techniques. These limitations can impact their applicability and efficiency in various scenarios, necessitating adaptations or alternative approaches for certain tasks.

2.3.3. Adversarial learning framework

Adversarial learning, introduced through the Generative Adversarial Network (GAN) [44] framework, consists of:

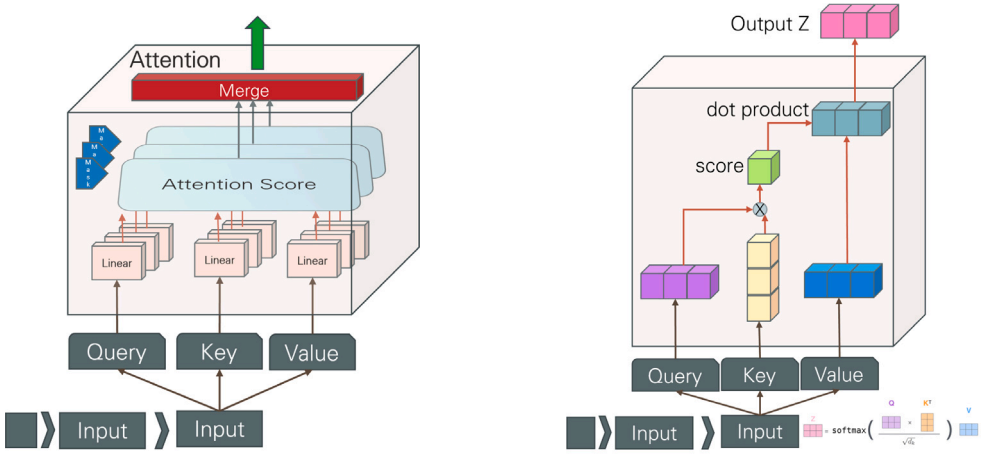
1. **The generator**: It crafts plausible data, which then act as negative training samples for the discriminator.
2. **The discriminator**: It discerns between genuine data and the generator's creations, penalizing the latter for implausible outputs.

Since their introduction in 2014, Generative Adversarial Networks (GANs) have established a strong foundation for adversarial learning. In this framework, a generator attempts to produce data that closely resembles real instances, while a discriminator works to distinguish between genuine and generated samples. This adversarial approach has been particularly influential in anomaly detection within Multivariate Time Series (MVTS), underscoring the importance of detailed and sophisticated feature extraction. Integrating adversarial learning into a transformer architecture designed for high-dimensional, multivariate, nonstationary manufacturing data can significantly improve the model's efficiency. We will further explore the integration of adversarial learning within this transformer paradigm, focusing on how it addresses the complexities presented by high-dimensional data—challenges that may hinder traditional transformer models. Furthermore, models like MAD-GAN [45], which employ LSTM-based GAN architectures, have proven effective in modeling data distributions. By incorporating prediction errors and discriminator loss into the anomaly detection process, these models offer a comprehensive analysis of time-series data, which is essential for identifying anomalies in complex manufacturing processes.

2.4. Remaining challenges

Despite the emergence of these promising techniques, a prevailing challenge remains: the developed models consistently confront difficulties in simultaneously handling both long-term and short-term dependencies (global and local), and in navigating the complex dynamics inherent to multivariate time series in manufacturing settings. Some of these challenges are:

1. **Limited Training Spectrum**: Solely training on normal data confines the model's ability to generalize.
2. **Feature Representations**: Current models' feature representations, whether they be pointwise, pairwise, or a combination, often lack the depth needed for capturing the full spectrum of normal patterns.
3. **Reconstruction Focus**: An encoder–decoder architecture's very focus on faithful reconstruction can sometimes blindside the model, causing it to overlook nuanced anomalies.
4. **Curse of Dimensionality**: As data dimensionality increases, the volume of its space surges exponentially. This escalation makes it an arduous task for models to capture relevant patterns, often resulting in suboptimal anomaly detection.
5. **Assumption of Normality**: These methods rest on the presumption that anomalous data points are the exception rather than the rule.
6. **Ambiguity in Thresholding**: One of the pivotal decisions in anomaly detection is determining the threshold that demarcates normal from anomalous. A lax threshold might overlook subtle deviations, whereas a stringent one could raise numerous false alarms.



(a) Multiple Heads Attention (b) scaled dot-product attention

Fig. 4. Principles of overall multiple heads attention (a) and Scaled Dot-Product Attention (b) in Transformers.

Solving these challenges in the manufacturing context requires the development of a new methodology for anomaly detection described in Section 3.

3. Proposed methods: Methodology and the AAT framework outline

This section provides a comprehensive overview of the proposed system, called AAT (Adversarial Adaptive Transformer), for AD. This system is composed of two modules, the Adversarial transformer (AT) and the Adaptive Support Vector Data Description (SVDD) classifier, to deal with the complexity of HO-MVTS data as depicted in Fig. 7, and to facilitate a clearer grasp of the AAT approach, we will present pseudocode along with a succinct description of the training procedure, Algorithm 2.

Before diving into the intricacies of our methodology, it is essential to grasp the real-world scenario that inspired its development and the specific challenges it aims to address. Understanding this context is crucial, as it not only grounds the abstract components of the AAT framework in a practical setting but also highlights its relevance and utility in real-world industrial applications. Additionally, in elaborating on the AAT framework, we include a comparative analysis with existing models, showcasing the distinct effectiveness of AAT in managing high-dimensional, multivariate time series data within the realm of smart manufacturing as illustrated in Section 5.

3.1. Problem formulation

Cable assembly manufacturing involves the production of cables and wire harnesses used in various applications, from electronics to automotive. Throughout the manufacturing process, numerous machines (M_i) and equipment are employed to assemble and test these cables. These machines continuously generate time series data, encompassing a sequence of data points collected at successive intervals (t).

Time series data in cable assembly manufacturing ($TS_{M_i,t}$) typically includes parameters such as voltage ($V_{M_i,t}$), current ($I_{M_i,t}$), temperature ($T_{M_i,t}$), pressure ($P_{M_i,t}$), and various sensor readings. These data points are collected in real-time, providing a detailed view of the manufacturing process's dynamics.

The time series data for each machine M_i and at each time t can be defined as follows:

$$TS_{M_i,t} = [V_{M_i,t}, I_{M_i,t}, T_{M_i,t}, P_{M_i,t}, \ldots] \tag{8}$$

Where: $TS_{M_i,t}$ represents the time series data for machine M_i at time t, $V_{M_i,t}$, $I_{M_i,t}$, $T_{M_i,t}$, $P_{M_i,t}$, and so on, represent specific parameters or sensor readings for machine M_i at time t.

We aim to harness this detailed data to pinpoint patterns and irregular behaviors that might result in a flawed product. This kind of detection is crucial during the manufacturing process, referred to as "online inspection", complemented by "offline inspection" post-production, utilizing specialized tools and advanced technologies like computer vision models powered by artificial intelligence. Both online and offline inspections collaborate to establish a robust quality barrier, as illustrated in Fig. 5.

The nature of high-order dependencies multivariate time series data, typical in manufacturing, presents numerous difficulties in detecting anomalies such as complexity, multi-modality, high dimensionality, noise and outliers, data drift, scalability, non-stationarity, and non-normality.

The case study in focus pertains to the identification of potential anomalies in the production of cable connectors at one of Rosenberger's assembly lines. This assembly process exemplifies the application of High-Order Multivariate Time Series (HO-MVTS) in Smart Manufacturing. The manufacturing protocol inherently follows a sequential series of steps, with each step building upon the previous one. Throughout this progression, the product undergoes various procedures, with multiple sensors monitoring specific parameters at each stage. These steps are executed either sequentially or in parallel, as depicted in Fig. 6(a). Any perturbation or irregularity in an early step can reverberate through subsequent steps, leading to cascading effects.

Each step utilizes various sophisticated machines (M_i), totaling eight in our context, equipped with numerous sensors. These sensors collect critical parameters such as voltage ($V_{M_i,t}$), current ($I_{M_i,t}$), temperature ($T_{M_i,t}$), and pressure ($P_{M_i,t}$), along with 122 additional sensor readings at successive time intervals (t). These interdependencies between steps, combined with the multivariate nature of the data from various sensors, lead to what we term High-Order dependencies Multivariate Time-Series (HO-MVTS), presenting a complex scenario for data analysis, as shown in Fig. 6(b).

The representation of the cable assembly process as a High-Order MTVS problem involves not only structuring the complex, multidimensional sensor data but also understanding the topology of the manufacturing steps. The topology—referring to the layout and interconnections of these steps—plays a crucial role in how data is collected, interconnected, and analyzed. This subsection outlines the essential components of this representation and the underlying process topology:

- **Step and Machine Representation:** The manufacturing process consists of several steps, where each step is associated with a machine M_i.

$$M = \{M_1, M_2, \ldots, M_i, \ldots, M_n\} \tag{9}$$

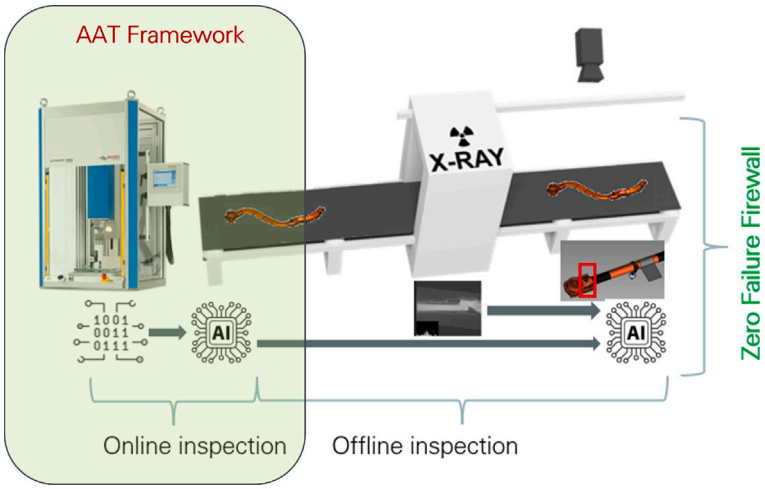


Fig. 5. Leverage AI for quality inspection for Power Connectors products.

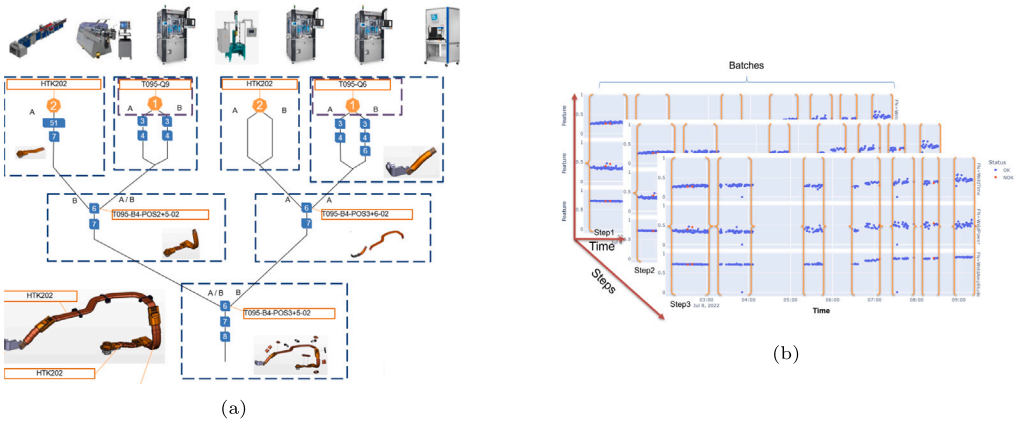


Fig. 6. Time series data in cable assembly manufacturing for E-mobility.

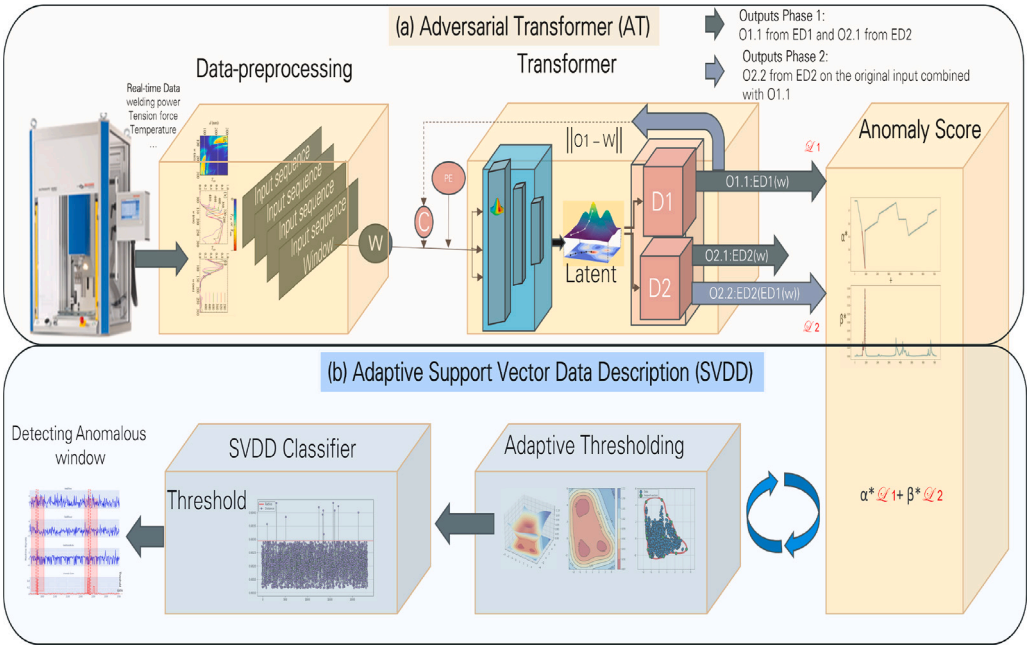


Fig. 7. Detailed illustration of two-phase training adaptive adversarial transformer for anomaly detection: Adversarial Transformer module (a) and Adaptive Support Vector Data Description module (b).

where M denotes the set of all machines involved in the process, where each M_i corresponds to a specific machine within the manufacturing process, and i ranges from 1 to n, with $n = 8$ in our case.

- **Sensor Data Representation:** For each machine M_i at step s, sensor readings at time t are represented as a vector:

$$\mathbf{X}_{M_i,t} = [V_{M_i,t}, I_{M_i,t}, T_{M_i,t}, P_{M_i,t}, \ldots], \tag{10}$$

encompassing all critical parameters and additional sensor data.

- **Time Series for Each Step:** The time series for each step involving machine M_i over a time interval T is captured as:

$$T_{M_i} = \{\mathbf{X}_{M_i,t_1}, \mathbf{X}_{M_i,t_2}, \ldots, \mathbf{X}_{M_i,t_T}\}, \tag{11}$$

providing a comprehensive machine dataset for each step in the manufacturing process.

- **Comprehensive High-Order MTVS Representation with High-Order Dependencies:** Aggregate the time series data across all steps into a collection $\mathcal{H} = \{T_{M_1}, T_{M_2}, \ldots, T_{M_n}\}$, with a mapping function $f : S \to \mathcal{H}$ that assigns each step to its corresponding time series data. This representation not only reflects the interconnected and temporal dynamics of the manufacturing process but also accounts for High-Order Dependencies. Given the sequential and, in some cases, parallel nature of manufacturing, parameters from one step can significantly influence subsequent steps. This interdependence results in data with high-order dependencies, emphasizing the complex relationships between different stages of the manufacturing process.
- **Matrix Representation of Time Series Data:** The comprehensive dataset for each machine M_i across all steps in the manufacturing process can be structured into a matrix representation. This matrix, denoted as $\mathbf{M}_{M_i}$, consolidates the time series data over the interval T for all observed parameters. Each row in $\mathbf{M}_{M_i}$ corresponds to a vector $\mathbf{X}_{M_i,t}$ at a specific time t:

$$\mathbf{M}_{M_i} = \begin{bmatrix} V_{M_i,t_1} & I_{M_i,t_1} & T_{M_i,t_1} & P_{M_i,t_1} & \cdots \\ V_{M_i,t_2} & I_{M_i,t_2} & T_{M_i,t_2} & P_{M_i,t_2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ V_{M_i,t_T} & I_{M_i,t_T} & T_{M_i,t_T} & P_{M_i,t_T} & \cdots \end{bmatrix} \tag{12}$$

where each column represents a specific parameter (e.g., voltage, current, temperature, pressure), and each row represents the sensor readings at a given time t.

- **Windowing:** To capture patterns and relationships over a period, the data is segmented into windows $W = \{w_1, \ldots, w_{T-r+1}\}$ with a stride of 1. Each window is defined as $w_t = \{u_t, \ldots, u_{t+r-1}\}$, where r is the window size.
- **Ground Truth Label:** Each window or segment of the time series has an associated label $y_t \in \{0, 1\}$. A label of 1 indicates that the window has an anomaly, whereas a label of 0 indicates a normal observation.

The approach to data representation in the cable assembly process highlights the complexity and dynamism of collecting, organizing, and preprocessing multivariate time series (MTVS) data. This process involves sophisticated technologies to handle the volume and velocity of data from numerous sensors, forming a matrix that encapsulates the intricate interdependencies of the manufacturing stages. Such a comprehensive representation is pivotal for understanding the nuanced interactions across time lags, which is essential for effective anomaly detection. Anomaly detection in this context is challenging due to the complex interplay between datasets from different stages of assembly, often termed "locality". The principle of locality underscores the correlation between temporally proximate data points, a critical factor in identifying relevant patterns and trends for precise anomaly detection. This analysis forms the basis for creating predictive models capable of identifying subtle fluctuations and potential issues early on, which is crucial for maintaining quality standards and minimizing downtime in cable assembly operations.

The analysis of HO-MVTS data, therefore, is not just about detecting anomalies but also about predicting potential disruptions by examining trends over time. Recognizing the significance of locality further refines this process, highlighting correlations within specific periods or operational contexts that might be missed in broader analyses. This detailed exploration of data interdependencies and temporal correlations underscores the challenges and opportunities in leveraging HO-MVTS data for smarter manufacturing practices.

An additional complexity in this context is the scarcity of labeled data, which is crucial for training accurate anomaly detection models. This scarcity poses a significant challenge, as labeled data are essential for guiding the learning process of AI models, ensuring they can distinguish between normal operations and potential anomalies effectively. In addressing these challenges, a particular focus is placed on scenarios where labels for training data may be scarce. This situation underscores the importance of combining AI technologies with human expertise. Domain experts play a crucial role by supplying a subset of labels to evaluate the model's performance, bridging the gap between automated systems and nuanced, real-world knowledge. Even though labels might be scarce, the synergy between AI and human expertise ensures that domain experts provide the model with a subset of labels to gauge its performance.

The ultimate aim is to define and implement an online anomaly detection (AD) approach capable of managing unlabeled multivariate time series, preprocessing them, training the AI model, and subsequently testing its efficacy on both standard trends and custom anomalies of interest. Building on this foundation, our aim extends to developing and deploying an online anomaly detection (AD) methodology. This innovative approach is designed to navigate the complexities of unlabeled multivariate time series data effectively. It involves preprocessing this data, training an AD model called adversarial adaptive transformer (AAT), and then rigorously testing its capability to identify both standard trends and bespoke anomalies of interest. Such an online AD system promises to leverage the strengths of AI while remaining flexible and responsive to the unique challenges presented by the cable assembly process, ensuring a robust solution to the problem of scarce labeled data.

3.2. Proposed transformer for anomaly detection

The Adversarial Adaptive Transformer (AAT) framework is meticulously crafted to tackle the nuanced challenges of anomaly detection in industrial environments. These challenges include managing high-order dependencies in multivariate time series data, navigating the complexities of high dimensionality, recognizing the importance of locality in data correlations, detecting subtle anomalies that may otherwise go unnoticed, and maintaining adaptability to continuously evolving data patterns. By integrating adversarial learning with adaptive mechanisms within a transformer architecture, the AAT framework is uniquely positioned to address these issues head-on. Its cutting-edge design not only enhances the accuracy of anomaly detection but also ensures the system's ability to adjust to new or unforeseen data patterns, as discussed in the previous section. Through this approach, the AAT framework aspires to set a new standard in anomaly detection, offering unprecedented precision and flexibility in handling the intricate dynamics of industrial data.

The following subsections delve into the components and functionalities of the AAT framework, elucidating its architecture, training methodology, and key features. By focusing on the model intricacies, this section aims to provide a thorough understanding of how the AAT framework operates and its potential applications in real-world industrial contexts.

3.2.1. Model overview

In this section, we will delve into the AAT Model architecture, emphasizing its unique attention mechanism and covering its adaptive thresholding. AAT is composed of two modules, the Adversarial transformer (AT) and the Adaptive Support Vector Data Description (SVDD) as a classifier (Fig. 7).

Anomalies in time series data pose a formidable challenge, often blending subtly with normal data and posing difficulties for methods solely dependent on local or global contexts. The self-attention mechanism offers a more expansive view, illuminating the nuanced differences between regular data points and anomalies. This insight forms the basis of 'contextual anomaly attending', emphasizing the pivotal role of encompassing temporal context in pinpointing anomalies, even those deeply intertwined with regular patterns.

High-order Multivariate Time Series (HO-MVTS) data further compounds the challenge, exhibiting intricate interdependencies among multiple variables. Such data often shows deviations from expected patterns across several variables and time intervals, amplifying the importance of capturing locality within the data.

Traditional anomaly detection methods, including conventional Transformer models, may struggle to effectively capture these high-order dependencies and locality issues, leading to suboptimal performance in anomaly detection tasks within HO-MVTS data.

To address these challenges and other challenges that have been introduced in Section 3.1, the Adversarial Transformer (AT) model is introduced. The AT employs a unique configuration with one encoder (E) and two decoders ($D1$ and $D2$), fostering a competitive dynamic between the encoder and decoders. This setup enhances the model's ability to discern regular patterns from anomalous ones and specifically targets the subtle blending of anomalies with normal data patterns.

The AT is designed to handle the complexities of HO-MVTS data and the locality issues inherent in anomaly detection tasks. By leveraging adversarial learning techniques and robust attention mechanisms, the AT effectively navigates the intricacies of time series data.

By incorporating adaptive mechanisms, the AT captures both local and global dependencies within the data, enabling it to effectively discern anomalies from normal patterns, even in the presence of subtle blending across multiple variables and time intervals. This capability positions AT as a promising solution for addressing the challenges posed by contextual anomalies and HO-MVTS data, setting a new standard for anomaly detection in complex industrial environments.

The use of a single encoder equipped with an anomaly attention mechanism in our AT model represents a balance between computational efficiency and functional capability, making it suitable for real-time and resource-constrained environments. The option of incorporating more decoders remains open for future exploration, particularly as computational capabilities evolve and more complex anomaly detection scenarios are considered.

Anchored on an adversarial basis, the AT boasts a singular configuration, composed of one encoder (E) and a pair of decoders ($D1$ and $D2$). This setup fosters a competitive dynamic between the encoder and decoders, bolstering the model's adeptness at discerning regular patterns from anomalous ones and addresses the intricate challenge of anomaly detection within complex HO-MVTS data by harnessing the advanced capabilities of the self-attention mechanism, particularly within the context of 'contextual anomaly attending'. This approach specifically targets the subtle blending of anomalies with normal data patterns, a common challenge in various industries, including manufacturing.

3.2.2. Attention mechanism in AT

In time series anomaly detection, recognizing both long-term and short-term dependencies is pivotal. Anomalies usually manifest as variations from immediate patterns and often correlate more with neighboring time points than with distant ones. By incorporating the anomaly attention mechanisms into the Transformer's Encoder block, the model can deftly identify standout features in time series data. This enhancement allows for more accurate and efficient anomaly detection, significantly boosting the identification and monitoring of potential malfunction alerts. This advancement provides a nuanced view that might elude traditional self-attention mechanisms, thereby bolstering the Anomaly Transformer's capabilities. Therefore, we introduce the dual attention system in AT:

1. Long-Term Dependency (LTD): Predominantly based on the transformer's self-attention mechanism, LTD captures correlations across the entire series, ensuring no long-term dependencies are missed. Formally, it is represented as:
$$\text{SelfAttn}_{ij} = \frac{\text{softmax}_k(QK^T)}{\sqrt{d_k}}V \tag{13}$$
with d_k the dimension of the key vectors and K^T the transpose of K. This mechanism is agnostic to the specific distance between points in a sequence, ensuring no inherent bias towards the spacing of data points. The Transformer model, which underpins this approach, is discussed in detail in Section 2.3.2, providing a comprehensive overview of its architecture and functionality.
2. Short-Term Dependency (STD): To cater to the immediate neighbors of a data point, AT utilizes an attention mechanism based on the Gaussian kernel. This attention is formulated as:
$$\text{Gaussian}_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \tag{14}$$
where d_{ij} represents the distance between positions i and j, and σ is a hyperparameter that dictates the width of the Gaussian kernel.

Balancing and tuning long-term and short-term attentions. The Adversarial Transformer (AT) achieves a harmonious integration of Long-Term Dynamics (LTD) and Short-Term Dynamics (STD) attentions through a learnable parameter, θ. This parameter is critical in dictating the balance between these two types of attention, with the combined attention mechanism, termed 'CombinedAttention (CA)', formulated as:
$$\text{CombinedAttn}_{ij} = \theta \times \text{SelfAttn}_{ij} + (1-\theta) \times \text{GaussianAttn}_{ij} \tag{15}$$
Initialized within the interval [0,1], θ is subject to updates during the training process via gradient descent, ensuring the model's adaptability to the varying significance of LTD and STD in different contexts.

To maintain θ within its intended range and ensure the adaptive nature of AAT, it is adjusted using a sigmoid activation function post-update:
$$\theta = \text{sigmoid}(\theta') \tag{16}$$
where θ' represents the interim value of θ prior to the application of the sigmoid function. This approach guarantees that θ dynamically finds the optimal equilibrium between long-term and short-term attention based on the training data's characteristics, enabling the model to effectively capture and respond to both immediate and evolving data patterns.

3.2.3. Training approach

The Adversarial Transformer (AT) is composed of three elements: Transformer Encoder E and two transformer decoders $D1$ and $D2$. As depicted in Fig. 7a in the AT module, the three elements are connected into an architecture composed of two transformer encoder/decoder blocks ED1 and ED2 sharing the same encoder network E: employs a two-phase training process for its encoder/decoder architecture in an adversarial style. These two phases are:

Phase 1: Standard Transformer Encoder/Decoder Training In the first phase, both transformer Encoder-Decoder blocks (ED1 and ED2) are trained to reconstruct the input data, where the encoder E compresses the input W into a latent space Z, and each decoder $D1$ and $D2$ attempts to reconstruct the input from this latent space. The objective of this phase is to minimize their reconstruction errors

on the original input data. This reconstructive role is fundamental during the first phase of training, where the primary goal is to learn to accurately replicate the normal operation data, ensuring that the model can identify what constitutes a non-anomalous state. The output of ED1 after Phase 1 training is denoted by $O1.1$, and the output of ED2 after Phase 1 training is denoted by $O2.1$:

$$O1.1 = D1(E(W)) \tag{17}$$

$$O2.1 = D2(E(W)) \tag{18}$$

The loss functions for this phase are:

$$L_{ED1} = \|W - O1.1\|^2 \tag{19}$$

$$L_{ED2} = \|W - O2.1\|^2 \tag{20}$$

Phase 2: Transition to Adversarial Training In the second phase, an adversarial relationship is established between ED1 and ED2. Here, ED_2 is trained to distinguish real data from the reconstructions produced by ED_1. and ED_1 is trained to fool ED_2, such that ED_2 cannot differentiate between real data and the reconstructions of ED_1. In other words, ED1 (Encoder + $D1$) tries to produce reconstructions $ED1(W) = D1(E(W))$ that, when passed through the encoder E again and then reconstructed by $D2$, are indistinguishable from the original inputs. ED2 (Encoder + $D2$) tries to distinguish between the real input W and the input reconstructed by ED1, i.e., $ED12(ED1(W)) = D2(E(D1(E(W))))$. ED1 tries to minimize the error when its output is reconstructed again by ED2 while ED2 tries to maximize this error, essentially trying to identify if the input is from ED1 or the original data. Therefore we can formulate the adversarial loss functions as follows:

$$\min_{ED1} L_{ED1} = \|W - ED2(ED1(W))\|^2 = \|W - D2(E(D1(E(W))))\|^2 \tag{21}$$

$$\max_{ED2} L_{ED2} = -\|W - ED2(ED1(W))\|^2 = -\|W - D2(E(D1(E(W))))\|^2 \tag{22}$$

During adversarial training, $D2$ processes both real inputs and reconstructions from $D1$, developing the ability to differentiate between normal and pseudo-anomalous inputs. Thus, $D2$ functions both as:

- **Traditional Decoder**: During Phase 1, $D2$ functions like a normal decoder, reconstructing the input from the latent space produced by the encoder.
- **Adversarial Decoder**: During Phase 2, $D2$ has a dual role. It reconstructs inputs from the latent space and also tries to maximize the reconstruction error of inputs that were passed through $D1$. Essentially, $D2$ is trained to detect when the input it receives has already been through another reconstruction process (from $D1$). This dual-input mechanism ensures that D2 develops a nuanced understanding of what constitutes normal versus anomalous patterns, bolstered by continuous feedback from its attempts to correctly classify the inputs (from $D1$).

This setup allows ED1 to learn better representations that can fool ED2, while ED2 becomes better at distinguishing real inputs from reconstructions, enhancing AT ability to distinguish between anomalies and regular data points in the adversarial phase and making it more adept at discerning between regular patterns and anomalies. This is achieved by ensuring that ED1 can produce reconstructions that are close to normal data, while ED2 learns to amplify the differences when the data is anomalous. This distinction makes it easier to detect anomalies during inference. This integration balances effective anomaly detection with computational efficiency. Introducing more than two decoders could enhance the model's ability to discern and classify multiple types of anomalies or operational states simultaneously, thereby increasing robustness and functional diversity. Each decoder could specialize in different aspects of the data or different types of anomalies, similar to ensemble methods that aggregate diverse perspectives. More decoders provide redundancy, reducing the risk of significant performance drops if one decoder pathway degrades or fails. Our methodology leverages a single-encoder structure, contrasting with some dual-encoder systems in similar applications. The integration of an anomaly attention mechanism within this single encoder enhances the model's ability to focus on pertinent features indicative of anomalies, without duplicative encoding layers. This approach reduces the number of trainable parameters, simplifying the training process and decreasing the computational burden during both the training and inference phases. With only one encoder to process inputs, data throughput speed increases, making the system more suitable for real-time anomaly detection applications where response time is critical. Compared to models employing dual encoders like the TranAD model [8], or multiple decoding pathways, our single-encoder model with an integrated anomaly attention mechanism is computationally less expensive and faster. This design choice addresses the need for efficient data processing in industrial settings, where delays can lead to increased operational risks and costs. While the single encoder design enhances computational efficiency, it might limit the model's ability to independently adapt to highly variable or disparate feature sets better handled by multiple encoders. However, the dual anomaly attention mechanism is specifically designed to mitigate this limitation by enhancing the encoder's focus on relevant features.

3.3. Proposed support vector data description (SVDD) module

For anomaly detection in manufacturing time series data, the integration of adaptive thresholding techniques, such as NDT, POT, KQE, and notably SVDD, which we adapt, offers a sophisticated approach, look at Section 2.1. These methods not only streamline the anomaly detection process but also boost the precision and promptness of pinpointing potential issues in high-dimensional multivariate nonstationary manufacturing data. Maintaining operational efficiency and meeting quality standards in today's manufacturing environments hinges on such proactive anomaly identification.

With the merger of adversarial learning and adaptive loss, the proposed Adversarial Adaptive Transformer (AAT) model capitalizes on the potency of these adaptive thresholding methods, Fig. 7b, the SVDD module. It promises a refined and robust framework tailored for intricate manufacturing contexts.

SVDD's prominence shines brightest in handling high-order Multivariate Nonstationary Manufacturing data. At its essence, SVDD encompasses the majority of data points within a hypersphere in the feature space, forming a robust representation of standard operational behaviors. This method is particularly apt for the complex, high-order nature of modern manufacturing.

SVDD's non-linear kernel trick excels in unraveling intricate, non-linear relationships intrinsic to multivariate manufacturing data, a boon in non-stationary settings where data dynamics shift constantly. When integrated with the transformer Encoder-Decoder structure, SVDD's unique prowess lies in its adaptive thresholding. Unlike conventional methods, SVDD dynamically adjusts based on the error vector, or anomaly score, generated by the transformer. This thresholding, honed by training on labeled test data's error vector, aims to optimize the F1 score. As a result, it effectively reduces both the False Positive Rate (FPR) and False Negative Rate (FNR), curbing the operational and quality costs linked to false alarms.

Moreover, SVDD's design allows the assimilation of domain expertise, which is invaluable in manufacturing. Such integration delivers a richer understanding of anomalies, which is important in the multifaceted world of modern manufacturing.

With these attributes, SVDD emerges as an adaptable and potent technique for anomaly detection in high-order Multivariate Nonstationary Manufacturing data. It forms the bedrock of our proposed Enable

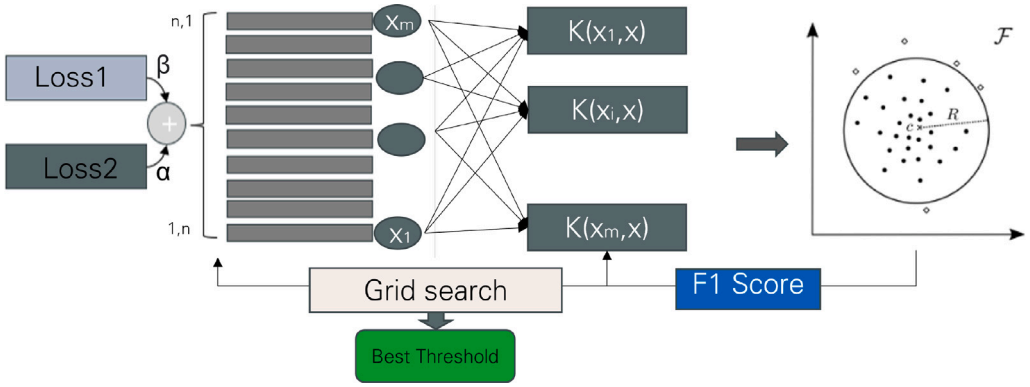


Fig. 8. Adaptive anomaly detection using SVDD.

Transformer model, which fuses Adversarial Learning and adaptive loss. We have employed the SVDD classifier with RBF kernel to sculpt a nonlinear SVDD as shown in Fig. 8.

First, we map the two loss vectors derived from the transformer model into a higher-dimensional feature space, denoted by matrix X. This is coupled with grid parameters essential for refining the SVDD model and a range determining the weights of two unique error types: 'Reconstruction_Loss' and 'Adversarial_Loss', then we generate a combined loss metric by systematically comparing these two errors, relying on a fluctuating weighting factor. With the accumulated losses, the dataset becomes richer, setting the stage for training the SVDD model. This training is facilitated through a grid search approach, guaranteeing the identification of the optimal model specifications. Once the best model is pinpointed, the algorithm embarks on an adaptive exploration of potential threshold values. The primary goal is to identify the threshold which optimally elevates the F1 score. This process is underpinned by the desire to harmoniously blend the two error types and recognize the ideal threshold, thereby optimizing the SVDD model's performance. Ultimately, the algorithm produces the threshold that ensures the peak F1 score as its output. More details are provided in Algorithm 1.

Algorithm 1 Determine Adaptive Threshold for AAT with SVDD

Require: Data matrix X, Grid parameters for SVDD `param_grid`, Range for α between 1 and 10 and β equals 10 - α.
Ensure: the Best Threshold for the Global F1 Score
for $\alpha = 1$ to 10 **do**
  $combined\_loss \leftarrow \alpha \times$ Reconstruction_Loss + $(10 - \alpha) \times$ Adversarial_Loss
  X[Loss-α] $\leftarrow combined\_loss$
end for
Split X into $X\_train, X\_test, y\_train, y\_test$
$svdd \leftarrow$ GridSearchCV(BaseSVDD(display='off'), param_grid, scoring="F1_score")
$svdd$.fit(X_train, y_train)
`best_model` $\leftarrow$ `svdd.best_estimator_`
$best\_F1 \leftarrow 0$
$best\_threshold \leftarrow 0$
for each $threshold$ in a predefined range **do**
  Predict labels using distances from `best_model` and the $threshold$
  Compute $F1\_score$ for the predicted labels
  if $F1\_score > best\_F1$ **then**
    $best\_F1 \leftarrow F1\_score$
    $best\_threshold \leftarrow threshold$
  end if
end for
  return $best\_threshold$

Algorithm 2 Adversarial Encoder Transformer Training Algorithm

Notations:

- E - Encoder
- D_1, D_2 - Decoders
- W - Dataset for training
- N - Iteration limit
- θ - Weight parameter for combined attention
- LTD - LongTermDependency
- STD - ShortTermDependency
- CA - CombinedAttention

Require:
Encoder E with two branches for LTD and STD.
Decoders D_1 and D_2.
Dataset for training W.
Adaptive hyperparameter σ.
Epochs N.
procedure TRAINING
  Initialization:
  Initialize weights for E, D_1, D_2.
  Initialize θ to 0.5.
  $n \leftarrow 0$
  while $n < N$ **do**
    // Process each window in the dataset
    for each window W_i in W **do**
      $Q, K, V \leftarrow LinearProjection(W_i)$ ▷ Linear projection of window
      $LTD \leftarrow AnomalyAttention(Q, K, V)$
      $STD \leftarrow GaussianKernel(W_i, \varsigma)$
      CA = $\theta \times$ LTD + $(1 - \theta) \times$ STD
      $O_1, O_2 \leftarrow D_1(E(W$ with CA)), $D_2(E(W$ with CA))
      // Calculate loss based on differences between outputs
      $L_1 \leftarrow \frac{1}{n}||O_1 - W_i|| + (\frac{1}{n})||O_1 - W_2||$ ▷ Reconstruction Loss
      $L_2 \leftarrow \frac{1}{n}||O_2 - W_1|| + (\frac{1}{n})||O_2 - W_i||$ ▷ Adversarial Loss
      Max$D$1Min$D$2(RescosError)
      // Update weights using calculated loss
      Update θ based on the gradient of the combined loss concerning θ.
      Update weights of E, D_1, D_2 using L_1, L_2
    end for
    $n \leftarrow n + 1$ Backpropagate using Combined Loss and updated weights, (Alog3)
  end while
end procedure

The 3 algorithm outlines an adaptive weight update process tailored for the Adversarial Transformer (AT) Model, focusing on the interplay between its encoder and dual decoders. Within each training iteration, the algorithm processes a minibatch of examples from the dataset, potentially supplemented by noise samples if applicable, embodying an adversarial training context. It iteratively updates the encoder and each decoder (focusing on Long-Term Dynamics (LTD) and Short-Term Dynamics (STD) respectively) based on their stochastic gradients, enhancing the model's ability to discern patterns and anomalies in the data.

A critical aspect of this approach is the dynamic adjustment of the θ parameter, which balances the combined attention mechanism between LTD and STD. This is achieved through gradient descent, minimizing the combined loss from both decoders, thereby refining the model's focus and improving anomaly detection accuracy. The use of standard gradient-based learning rules, including momentum, ensures efficient convergence and adaptability of the model to evolving data characteristics, making this algorithm a cornerstone of the AT model's training process.

Algorithm 3 Adaptive Weight Update for AAT Model Training

for each training iteration **do**

Sample minibatch of m examples $\{W_1, \ldots, W_m\}$ from the training dataset.

Sample minibatch of m noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p_g(z)$ ▷ Only if applicable.

for k steps **do**

Update the encoder E by ascending its stochastic gradient:

$$\nabla_{\theta_e} \frac{1}{m} \sum_{i=1}^{m} \log E(W^{(i)}).$$

Update the first decoder D_1 by ascending its stochastic gradient related to LTD:

$$\nabla_{\theta_{d1}} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_1(E(W^{(i)}))\right].$$

Update the second decoder D_2 by ascending its stochastic gradient related to STD:

$$\nabla_{\theta_{d2}} \frac{1}{m} \sum_{i=1}^{m} \left[\log D_2(E(W^{(i)}))\right].$$

Update the combined attention θ parameter by applying gradient descent to minimize the combined loss from both D_1 and D_2:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} \text{CombinedLoss}(D_1, D_2, \theta),$$

where η is the learning rate.

end for

This algorithmic approach ensures dynamic adjustment of the encoder and decoders, adapting to both LTD and STD features in the data. Gradient-based updates leverage standard learning rules, incorporating momentum for enhanced convergence.

3.4. Advantages and differentiators

AAT's design addresses several limitations of existing models: **The transformer-based architecture** excels in managing high-dimensional time series data, effectively capturing complex dependencies through its attention mechanisms. By incorporating Gaussian attention and multi-head attention, our model can focus on both local and global dependencies, which is essential for time series analysis in industrial applications. Traditional methods often struggle with the dimensionality of the data, performing well only on low-dimensional data and thus lacking scalability in manufacturing settings.

Unlike some approaches that necessitate two encoders, the AAT utilizes a single encoder with an integrated anomaly attention mechanism. This **dual attention mechanism** combines scaled-dot product attention with Gaussian attention, reducing model complexity while efficiently capturing both short-term and long-term dependencies. Multi-headed attention enhances the model's ability to capture complex interdependencies across different positions of the input sequence, improving global contextual understanding. In contrast, Gaussian attention focuses on local contextual dependencies by assigning higher weights to nearer elements, which is crucial for detailed anomaly detection where local patterns are significant.

The AAT's training process consists of **two phases of training, transformer training and adversarial training**. In the initial phase, the transformer model captures intricate patterns of normal data. In the adversarial phase, it enhances its ability to distinguish between anomalies and regular data points. This dual-phase approach mitigates overfitting commonly encountered in complex deep-learning models. By using a shared encoder and assigning distinct roles to each decoder, the model maintains stability and benefits from improved gradient flow and feature utilization across different training phases. In comparison with pure GAN approaches, excluding reconstruction errors might seem to simplify training and reduce costs, but it can lead to convergence issues due to the adversarial nature of training, which can oscillate without a balanced generator and discriminator. Additionally, pure GANs require careful tuning to ensure effective learning, and without reconstruction errors, they might not learn the detailed structure of the data, reducing anomaly detection accuracy.

By training the model on normal data and then using the reconstruction errors to train the SVDD, our approach avoids solely relying on the assumption that anomalous data points are the exception. The **dynamic thresholding mechanism** ensures that the model remains effective even when the input data distribution changes. It is important to note that the threshold set during the training of the SVDD is based on these reconstruction error vectors and not on scaled values. This approach ensures that when both normal and abnormal windows generate vectors of the same scale but with different slopes or directions, the SVDD can effectively distinguish between them, particularly by employing the Radial Basis Function (RBF) kernel to transfer these vectors to another space. In the first training phase, the outputs $O1$ and $O2$ are the results of the transformations by ED1 and ED2, as illustrated in Fig. 7a, the AT module. Our model can be adapted to other industrial processes by leveraging the adaptive thresholding mechanism. This adaptability allows for effective generalization across different processes with varying characteristics. To capture the specific normal and anomalous patterns of each unique environment, it is necessary to retrain only the SVDD component. This approach simplifies the adaptation process and ensures the model's effectiveness in diverse industrial settings.

Our choice to integrate a GAN-like mechanism with reconstruction errors is driven by the goal of leveraging the strengths of both approaches. This hybrid model is designed to achieve high sensitivity in anomaly detection with enhanced stability and reliability in training, qualities that are essential for practical applications, especially in complex industrial settings. Compared to state-of-the-art approaches and benchmarked models, our approach demonstrates significant enhancements in effectively addressing the complexities of data in manufacturing. However, our approach still has some limitations.

Inability to Efficiently Handle Tabular Data: Our model, which utilizes a transformer architecture, faces limitations in efficiently incorporating tabular data such as statistical process control (SPC) data and other categorical parameters from the machine or product. Our current model architecture is primarily designed for time series data and does not inherently accommodate tabular data, which could contain significant predictive information. However, recent techniques such as the tabular transformer have shown promise in overcoming this limitation. By adding another embedding block to incorporate tabular data, future work could integrate these features more effectively, potentially

using a hybrid model that combines elements of traditional machine learning techniques with deep learning. In summary, The AAT model further leverages its design to focus on variations in data correlations, employing adaptive thresholding techniques. This approach allows the AAT to effectively navigate the complexities inherent in anomaly detection within multivariate time series data. By concentrating on these variances and utilizing adaptive mechanisms, the AAT presents a robust and comprehensive solution, adept at accurately detecting both subtle and significant anomalies, thus addressing the multifaceted challenges of anomaly detection.

4. Experimental framework

This section outlines the datasets utilized, the performance metrics employed in the experimental evaluations, and the practical application study conducted in the real-world manufacturing context at Rosenberger.

4.1. Challenges with assessing model performance on public datasets

Numerous papers employ public datasets that have been repetitively used in research. Over time, researchers have come to understand the key features of these datasets, thereby accumulating experience that aids in training models. This recurrent use poses a potential pitfall: it becomes challenging to gauge a model's genuine performance as both the model and the researcher are overly familiar with the dataset. Consequently, relying solely on such datasets for performance assessment can be misleading. In essence, the researcher, not necessarily the model, knows where to look, which may not always reflect real-world applicability.

In our work, we aim to tackle this limitation head-on. Instead of solely relying on public datasets, our experiments extend to real manufacturing data, including one internal and two external datasets.[1]

This allows for a more comprehensive evaluation of our model's performance in realistic scenarios. Additionally, we do incorporate certain public datasets, ensuring a comparative evaluation with state-of-the-art models.

4.2. Datasets description

While public datasets offer a common ground for model comparison and benchmarking, one must approach them with a degree of caution. The familiarity of these datasets within the research community means that they can sometimes offer a skewed representation of a model's capabilities. It is essential to consider the broader context and not just the dataset in isolation.

The public datasets used for the evaluation of the performances of our model are described below.

SecureWater Treatment (SWaT) Dataset: This dataset originates from an actual industrial water treatment facility focused on producing filtered water. It covers 11 days of the plant's continuous operations. The first 7 days document normal functioning, while the remaining 4 days record scenarios where the system was under cyber–physical attacks. The dataset encompasses a variety of measurements, including sensor readings like water levels and flow rates, as well as actuator activities such as the operation of valves and pumps.

Soil Moisture Active Passive (SMAP) Dataset: The SMAP dataset is a collection of telemetry data associated with soil samples, gathered using the Mars rover. Provided by NASA, this dataset includes 55 separate traces, each encompassing 25 different dimensions. The data offers valuable insights into telemetry anomalies, as reported in NASA's Incident Surprise Anomaly (ISA) reports, which are part of the spacecraft monitoring system's data collection.

Server Machine (SMD) Dataset: SMD is one of the largest public datasets currently available for evaluating multivariate time-series anomaly detection. It contains metrics like CPU load, network usage, memory usage, etc, over 5 weeks long, monitors 28 server machines for a large Internet company with 33 sensors.

Mars Science Laboratory (MSL) Dataset: Analogous to the SMAP dataset, this dataset represents sensor and actuator data from the Mars rover. It contains complex sequences labeled as A4, C2, and T1. Notably, the MSL dataset encompasses telemetry anomaly data, originating from NASA's spacecraft monitoring systems, and covers a range of 55 dimensions.

In addition to the frequently used public datasets in this field, we evaluated our method on actual industrial data, where the proportion of anomalies does not exceed 3.5%. **E-Coating**: An industry case study dataset from the air filtration system of an electrophoresis paint shop. Can be used to predict process conditions as a basis for maintenance improvements. Inspection records over 7 years and sampled every 30 min.

Prognostics and Health Management (PHM): The data provided by PHM Data Challenge 18 was provided to investigate the fault behavior of ion mill etch tools in the wafer manufacturing process. It is a database that collects sensor data in time sequence from ion mill etching tools operating under various setting conditions.

Statistical details are summarized in Table 1.

4.3. Feasibility study on Rosenberger's dataset

We conducted our feasibility study using Rosenberger's internal dataset, which was sourced from various sensors involved in the cable assembly manufacturing process. Given that multiple machines function in tandem during this assembly process, the data can be classified as a high-order multivariate time series, as elaborated in Section 3.1.

This dataset encapsulates data collected during the initial half of 2023, comprising 126 variables in total. For our analysis, we utilized data from the first four months as our training set and data from the subsequent two months for testing.

Anomalous windows, which correspond to defective batches, were meticulously identified and excluded from the training data with invaluable assistance from domain experts. Comprehensive statistical details related to this dataset are presented in Table 1.

4.4. Implementation details

The Adaptive Adversarial Transformer (AAT) is optimized for real-time evaluation in smart manufacturing through a comprehensive implementation strategy that encompasses model training, data preprocessing, and lifecycle management

Model Training and Evaluation: Our proposed Transformer was optimized to achieve minimal reconstruction error within ten epochs. With a learning rate of 0.005 and a gradient clip value of 0.2 were set, employing the mean squared error (MSE) as the loss function. For hyperparameter tuning of the SVDD component, Hyperopt's optimizer was utilized for distributed asynchronous hyper-parameter optimization, exploring a Γ search space of (0.1, 0.2, 0.5) for the RBF kernel with empirically determined 30-fold cross-validation.

Data Infrastructure and Management:

1. **Data Storage and Scalability**: A lakehouse architecture is employed for data storage, offering a harmonized platform that combines the benefits of data lakes and data warehouses. This architecture supports extensive data volume and variety, providing a scalable foundation for advanced analytics in manufacturing settings.

[1] The two external manufacturing datasets PHM and ECoating can be found here Fraunhofer.

Table 1
Details of benchmark Dataset, with three datasets in manufacturing.

Dataset	Applications	Application Senario(s)	Train	Test	Dimensions	Anomalies (%)
SMAP	Space	Monitoring	135 183	427 617	55	13.13
MSL	Space	Monitoring	58 317	73 729	55	10.72
SWaT	Water	Monitoring	496 800	449 919	51	11.98
SMD	Server	Monitoring	708 405	708 420	38	4.16
E-Coating	Machinery Equipment	Monitoring	73 760	36 881	17	3.18
PHM	Ion mill etch tool	Production Quality	750 000	250 000	24	2.14
Rosen	Manufacturing	Production Quality	802 912	378 015	126	1.1

2. **Preprocessing Pipeline**: A systematic pipeline ensures data quality, encompassing normalization, imputation, feature extraction, and secure feature storage. These preprocessing steps are crucial for preparing the manufacturing data for subsequent analysis by AAT.
3. **Model Lifecycle Oversight**: Comprehensive management of the AAT model's lifecycle is implemented, encompassing stages from training and deployment to active monitoring, with a focus on version control and tracking of model iterations for consistent performance evaluation.
4. **Operator Interface**: An interface is devised for system operators, equipped with analytical and visualization tools, to effectively engage with the AAT system and leverage its insights for operational oversight.
5. **Human–Machine Interface (HMI) Development**: An intuitive HMI was designed for operator interaction with AAT, offering clear visualizations of anomalies and analysis tools, which improves the decision-making process in manufacturing settings.

Operational Integration and Continuous Improvement: Integration challenges, such as ensuring data compatibility and model adaptability to the evolving manufacturing processes, are addressed through methodical engineering and continuous optimization practices. Security measures are prioritized to protect data integrity in the deployed environment. Additionally, operator training is emphasized to ensure proficient use of the AAT system and its associated data management tools.

By addressing these key areas, the AAT system is rendered capable of robust performance and seamless integration into the manufacturing workflow, ensuring it remains responsive to the dynamic and complex nature of industrial data.

5. Evaluating the effectiveness of AAT

In this section, we conduct an extensive performance assessment of the proposed Adaptive Adversarial Transformer (AAT) framework, aiming to provide a comprehensive evaluation through various advanced metrics. Our analysis is meticulously structured to not only benchmark AAT against established state of art models in the anomaly detection field but also to explore the impact of different configurations, particularly focusing on the Peak-Over-Threshold (POT) approach and the Support Vector Data Description (SVDD). This exploration is crucial for understanding AAT's adaptability and determining the optimal setup that enhances its robustness across diverse operational scenarios. Additionally, by employing advanced metrics tailored to time series anomaly detection — including distance-based, range-based, and affiliation metrics — we delve into the precision, recall, and F1 score among other metrics, to assess AAT's ability to finely balance false positives and false negatives. This multifaceted analysis aims to elucidate AAT's operational effectiveness, adaptability, and comparative performance in the broader context of anomaly detection technologies.

5.1. Detection performance

For our evaluation approach, we utilize the point adjustment technique. Here, a window is deemed abnormal if any of its points are identified as anomalies. This methodology is particularly pertinent in manufacturing contexts where maintaining a near-perfect quality gate is imperative to ensure product integrity. Accepting a few false negatives within a window is tolerable given the relatively short window size, making subsequent manual inspections manageable. Precision, recall, and F1-score were used to measure the classification performance.

The Precision, in the context of model evaluation, is the proportion of true positive predictions in the set of all samples predicted as positive. It is calculated with the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{23}$$

where TP represents the count of true positives and FP denotes the count of false positives. Note that TN (true negatives) and FN (false negatives) are not directly involved in the calculation of precision, but are relevant in other performance metrics.

The Recall, also known as sensitivity, measures the fraction of actual positive samples that are correctly identified by the model. It is mathematically expressed as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{24}$$

The F1-score, a metric that harmonizes precision and recall, provided a single measure of a test's accuracy by calculating their harmonic mean. It is mathematically defined as:

$$\text{F1-Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{25}$$

5.2. Ablation study

In this subsection, we conduct a comprehensive ablation study to critically analyze the key components of our proposed Adaptive Adversarial Transformer (AAT) model for anomaly detection in smart manufacturing environments. we aim to understand the individual and collective contributions of various model parameters and mechanisms towards the overall performance.

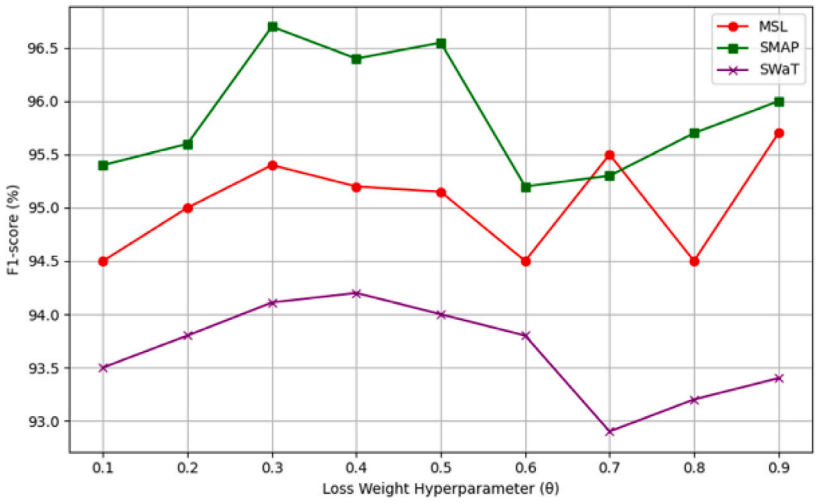


Fig. 9. Dual attention mechanism tuning parameter θ.

5.2.1. Dual attention mechanism, tuning parameter θ

We begin by dissecting the dual attention mechanism, focusing on θ, the parameter tuning the balance between Gaussian attention and multi-head attention within our proposed anomaly detection model. θ is defined within the range [0,1], where $\theta = 1$ fully activates the multi-head attention, and $\theta = 0$ fully employs Gaussian attention. The attention output is dynamically adjusted as follows:

$$\text{Final Attention} = \theta \times \text{Multi-head Attention} + (1-\theta) \times \text{Gaussian Attention} \tag{26}$$

We aim to assess how varying θ influences the model's sensitivity to short-term dependencies across different datasets, including MSL, SMAP, and SWaT, which are known for their diverse temporal characteristics. We will increment θ in steps (0.1 to 1.0) and evaluate each setting's impact using precision, recall, and F1-score. Fig. 9 indicates that the anomaly detection performance in datasets such as MSL, SMAP, and SWaT is significantly affected by how θ is tuned. These datasets typically feature anomalies that are more effectively detected through enhanced sensitivity to short-term dependencies, suggesting a higher weighting towards Gaussian attention in these scenarios. The dynamic tuning of θ strikes an optimal balance between the two attention mechanisms, influenced by the characteristics of the training data.

These observations will guide the development of algorithms for automatic θ adjustment, enhancing the model's adaptiveness and applicability in real-time anomaly detection environments.

5.2.2. Adversarial loss, tuning parameters α and β

As we combine the two loss vectors the reconstructed and the adversarial, we can tune the impact of the adversarial effect by tuning the α and β parameters. See Fig. 8. This analysis highlights how the interplay between adversarial and reconstruction loss influences the model's efficacy in detecting anomalies. By adjusting these parameters, we aim to find the optimal balance that enhances the model's robustness.

Our approach enables the possibility to control the contribution of each training phase through α and β, where

$$\text{combined\_loss} = \alpha \times \text{Adversarial\_Loss} + \beta \times \text{Reconstruction\_Loss} \tag{27}$$

with $\alpha + \beta = 1$.

We noticed that when $\alpha = 0.7$ and $\beta = 0.3$, we achieved the best results, indicating that the adversarial phase contribution is higher than the standard training, look at Fig. 10. This finding suggests that adversarial training plays a crucial role in enhancing the model's performance in anomaly detection within smart manufacturing systems. By allowing the flexibility to adjust these parameters, our approach can adapt to various datasets and conditions, maintaining robust and effective anomaly detection across different scenarios. This adaptability is critical in addressing data drift and changing data distributions over time, ensuring the model's continued efficacy in real-world applications.

5.2.3. Adaptive thresholding

The adaptive thresholding component, implemented via the Support Vector Data Description (SVDD) model, is critical for dynamically adjusting the threshold based on the anomaly scores produced by the model. This adaptation is key to managing varying degrees of anomaly severities and frequencies encountered in real-world data. Removing adaptive thresholding results in a less flexible model, potentially leading to increased false positives or false negatives, depending on the static threshold set. This diminishes the model's practical applicability, especially in environments where operational conditions change over time. Table 2 illustrates the impact of adaptive thresholding on model performance. SVDD is used due to its effectiveness in defining a boundary around the majority of data points, which represents normal conditions in high-dimensional space. This choice is particularly pertinent given the complexity and dimensionality of the data involved. Replacing SVDD with another classifier or using a different thresholding approach, such as Point Over Threshold (POT), results in a degradation in the model's sensitivity and specificity. POT, while simpler to implement, often fails to account for the intricate patterns present in high-dimensional data, leading to less accurate anomaly detection. For instance, simpler classifiers might not capture complex boundary shapes in high-dimensional space as effectively as SVDD, potentially leading to poorer performance in anomaly detection. Table 2 compares the performance of SVDD with other classifiers and thresholding approaches.

5.3. Comparative evaluation

After having a look at the AAT model performance on the benchmarked datasets and analyzing its performance against some ablations on its main components, we juxtapose the AAT model with renowned state-of-the-art methods [7,8,24,27,28,46], evaluating both their feature capabilities and overall detection accuracy. Table 2 presents the effectiveness of the Adaptive Adversarial Transformer (AAT) in anomaly detection, as assessed by our developed adaptive anomaly detection methodology, across standard benchmark datasets as well as proprietary internal data. Impressively, our method achieves cutting-edge performance across both data types.

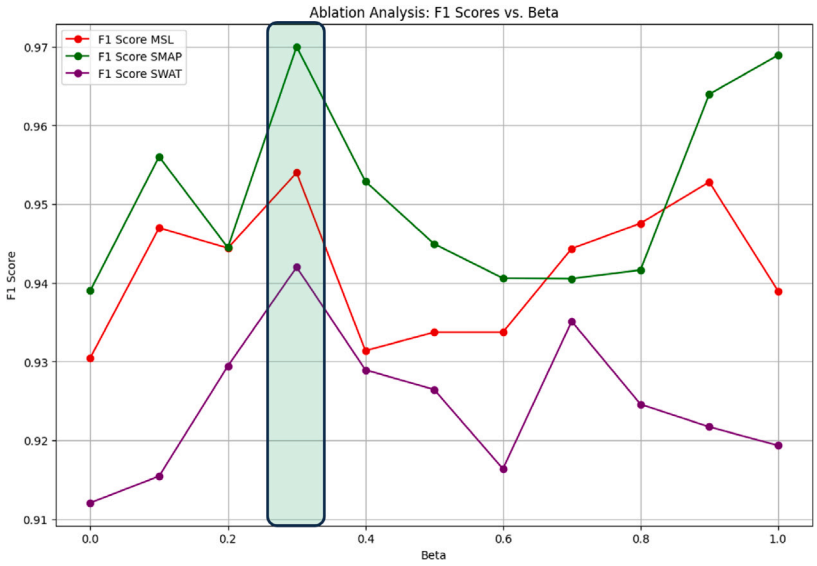


Fig. 10. Reconstruction loss vs Adversarial loss, α and β parameters.

Table 2
Performance comparison of AAT against the state of the art methods.

Method \ Dataset		SMAP			MSL			SWAT		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
LSTM-AE		0.852	0.733	0.788	0.629	1,000	0.772	0.778	0.511	0.617
OmniAnomaly		0.741	0.977	0.843	0.886	0.911	0.898	0.978	0.696	0.813
AD-GAN		0.816	0.922	0.865	0.852	0.993	0.917	0.959	0.696	0.807
USAD		0.769	0.983	0.863	0.881	0.978	0.927	0.987	0.740	0.846
TranAD		0.804	0.999	0.891	0.903	0.999	0.949	0.976	0.699	0.815
AnomalyTrans		0.941	0.994	0.966	0.920	0.951	0.935	0.915	0.967	0.941
AAT		0.884	0.990	0.934	0.901	0.963	0.930	0.936	0.892	0.913
Dynamic Thresholding	POT	0 *	0 *	0 *	0.924	0.983	0.952	0.968	0.889	0.926
	SVDD	0.946	0.996	**0.970**	0.920	0.991	**0.954**	0.947	0.938	**0.942**

In bold, the best F1 score for each dataset
(*)- In the case of the SMAP dataset, the threshold established through the POT method is excessive high, leading to both True Positives (TP) and False Positives (FP) being zero.

Our approach does not rely on pre-existing knowledge of the data distribution but is versatile, adapting to shifts in data distribution as new data emerges. The synergy of the adversarial framework's capacity to grasp intricate data distributions and the transformer's adeptness at managing temporal dependencies results in a robust model, especially proficient in high-dimensional scenarios.

In the comparative assessment, while each benchmarked technique has its strengths in anomaly detection, our *AAT* model integrates the best attributes of prior reconstruction-based models. It focuses on boosting anomaly detection accuracy and reducing training times. Embracing the blend of adversarial learning with transformer architecture holds promise for reliable anomaly detection in high-dimensional multivariate nonstationary manufacturing data.

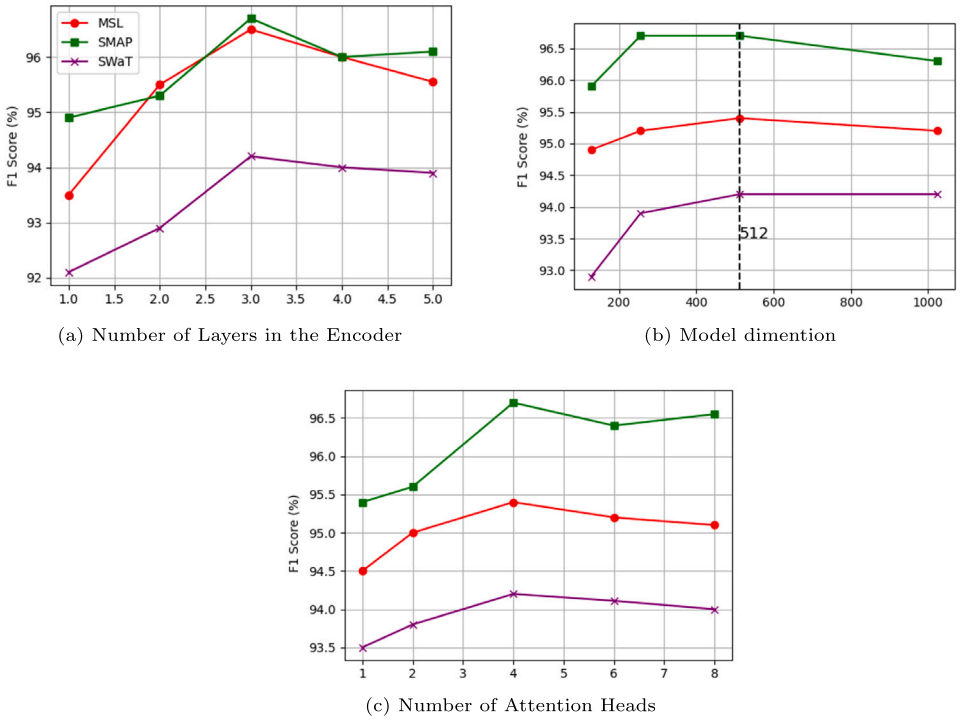
Fig. 11 presents a performance evaluation of the AAT model across various architectural configurations.

The upper-left chart 11(a), showcases the AAT model's proficiency in achieving high F1 scores with relatively few stacked layers. Such architectural simplicity not only streamlines the model but also aids in resisting overfitting. This remarkable performance can be attributed to the dual attention mechanisms employed by AAT.

The upper-right chart 11(b), provides insights into the relationship between the dimensionality of the attention mechanism (d_model) and the AAT model's performance in anomaly detection within multivariate time series data. While a larger d_model size can empower the model to encapsulate a richer set of features, and potentially discern more complex anomalous patterns, it may also introduce challenges. An increased d_model dimensionality can lead to a growth in computational requirements and memory overhead. Additionally, a more extensive representation space might elevate the risk of overfitting if not paired with adequate regularization or training data. In the presented results, a d_model size of 512 appears to offer the best trade-off between performance gains and the mentioned complexities. Thus, while there are evident performance enhancements with an expanding d_model, it is crucial to balance these gains against the inherent complexities and computational considerations.

The lower char 11(c), demonstrates the AAT model's adeptness in leveraging multiple attention heads for enhancing anomaly detection performance in multivariate time series data. The model's ability to achieve commendable results even with a limited number of attention heads underlines the efficiency of its attention mechanisms. This is particularly significant for time series data where both sparse and densely occurring anomalies are present. The observed trend suggests that AAT's attention design is optimized to capture diverse anomalous patterns without the necessity of an excessive number of heads, further emphasizing the model's architectural efficiency.

This comparative analysis not only showcases the Adaptive Adversarial Transformer (AAT) model's strengths but also affirms its position within the anomaly detection domain, signaling a notable progression of methodologies propelled by deep learning technologies in tackling complex data challenges. Our AAT model, characterized by its flexibility, excels without relying on pre-established data distribution knowledge. It adeptly adapts to data distribution changes, a testament to the adversarial framework's capacity to model intricate distributions coupled with the transformer's proficiency in temporal dependencies.



(a) Number of Layers in the Encoder

(b) Model dimention

(c) Number of Attention Heads

Fig. 11. Performances of the AAT model with various architectural configurations: (a) Number of Layers in the Encoder, (b) Model dimension, and (c) Number of Attention Heads.

Table 3
AAT Performance on Rosenberger Internal data and two other public manufacturing datasets.

Method \ Dataset	E-Coating			PHM			Rosenberger		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
TranAD	0.869	0.675	0.759	0.469	0.665	0.550	0.348	0.363	0.355
AnomalyTrans	0.671	0.845	0.748	0.535	0.706	0.608	0.528	0.608	0.565
AAT-SVDD	0.774	0.813	**0.793**	0.761	0.847	**0.801**	0.882	0.841	**0.861**

In bold, the best F1 score for each dataset.

This yields a model that is especially capable in high-dimensional, dynamic environments.

The experimental results in Table 2 highlight the superiority of the Support Vector Data Description (SVDD) approach in adaptive threshold selection over the traditional Peak Over Threshold (POT) method, as evidenced by the SVDD's impressive F1 score of 0.970 on the SMAP dataset, outperforming other models. Moreover, the AAT's overall performance, utilizing our adaptive anomaly detection methodology, excels on Rosenberger's real-world shop floor dataset. It achieves an F1 score of 0.861, which is indicative of its robustness and effectiveness in a practical manufacturing setting. Additionally, when evaluated against two other public industrial datasets, our AAT model not only competes well but often surpasses other state-of-the-art models, achieving an average F1 score of 0.818 across all manufacturing datasets. The real-world shop floor dataset from Rosenberger's cable assembly manufacturing. Moreover, additionally, on two other public industrial datasets. These results are encapsulated in Table 3, underscoring the AAT model's advanced detection capabilities and the efficacy of employing SVDD for dynamic thresholding in complex, multivariate, non-stationary manufacturing data scenarios and surpasses the results of state-of-the-art models.

In summary, the AAT configuration demonstrates a pronounced improvement in adapting the threshold selection mechanism, which is crucial for accurate anomaly detection. The framework's ability to maintain high performance across a variety of datasets, including high-dimensional, multivariate, and non-stationary manufacturing data, is evident from its superior F1 scores, heralding a significant advancement in the field of smart manufacturing anomaly detection.

Knowing that ATT is using the SVDD as the dynamic thresholding part, Fig. 12 illustrates the proficiency of the Adaptive Adversarial Transformer (AAT) in distinguishing between normal and anomalous test samples, by finding the smallest hypersphere that encloses most of the data points in the feature space, with the center of this hypersphere being point c and the radius R. In this depiction, test samples that extend beyond the predefined radius of the normal class's hypersphere are considered anomalies. These points represent data instances that significantly deviate from the learned distribution of normal data, based on their feature characteristics and the relational context captured by the AAT model using the radial basis function (RBF) kernel with gamma=0.3. Such deviations are flagged as anomalies rather than mere noise.

Anomalies are typically contextual or collective, showing significant deviations from normal patterns that are relevant to operational conditions. Noise, however, often consists of random variations or outliers that do not necessarily signify operational issues. These experimental results show that our model focuses on these contextual deviations by learning complex dependencies within the data, thereby reducing the likelihood of misclassifying noise as anomalies.

In Figs. 13(a) and 13(b), we plot the input and the reconstructed input from normal and anomalous samples, respectively. These reconstructions are based on an adaptive threshold set to 0.014, as depicted in Fig. 14.

Fig. 15 illustrates a time series example from the feasibility study, highlighting instances where anomalies were successfully identified by the AAT framework.

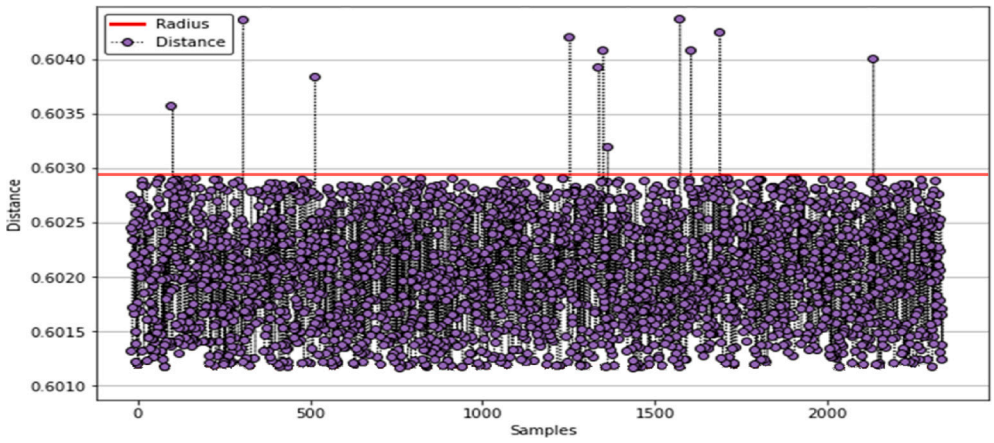
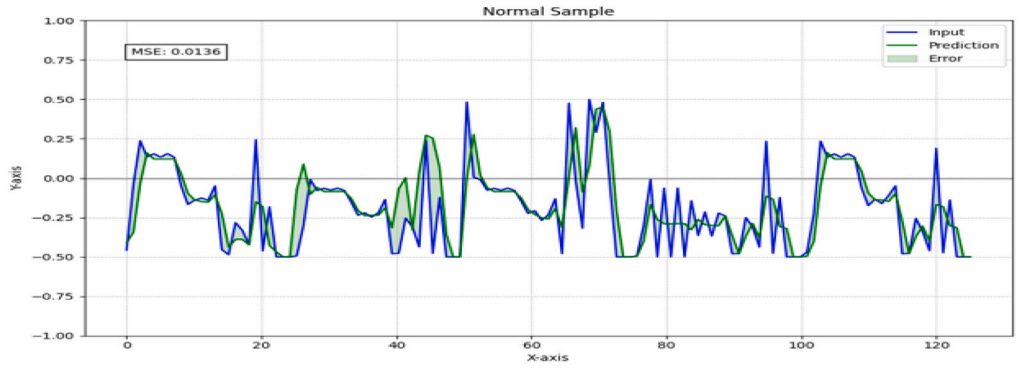
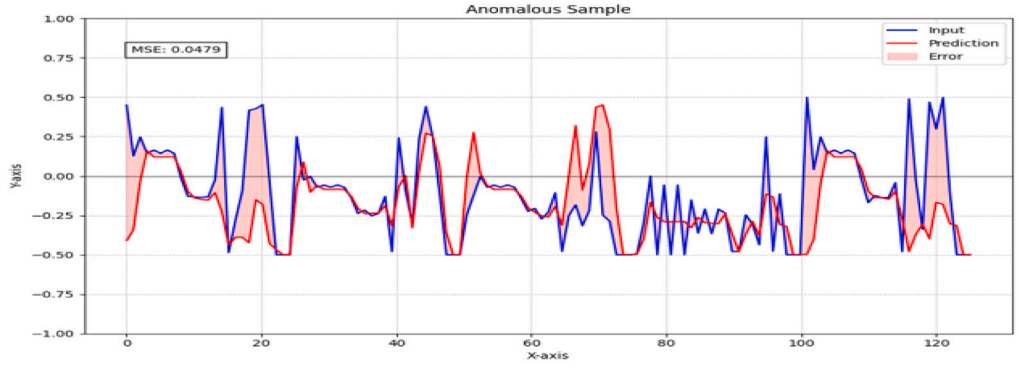


Fig. 12. Hyperspheres obtained using AAT-SVDD using rbf kernel, gamma=0.3 on Rosenberger data.



(a) Normal: Threshold= 0.014



(b) Anomal: Threshold= 0.014

Fig. 13. Input and reconstructed input from normal (a) and anomalous (b) samples on Rosenberger Data.

5.4. Training time and computational complexity

This subsection elucidates the time complexity associated with the Adaptive Adversarial Transformer (AAT) pipeline, alongside empirical training times observed during our experiments. The computational complexity of the AAT primarily revolves around the transformer architecture with a multi-head attention mechanism. As the number of these layers and the attention heads are fixed and invariant to the input size, the computation for each input with n features predominantly incurs a time complexity of $O(n)$. This stems from the dot product operations in the multi-head attention. Furthermore, each output is derived from the sum product of n input features, multiplied by a constant set of weights, ensuring that the computation time does not scale with n. The activation function applied subsequently also aligns with a linear time complexity. During our experimental analysis, the AAT-SVDD framework demonstrated a total training time of 64.3 s for one complete training cycle. This duration was further broken down into approximately 35.4 s for training the AAT-Transformer for one iteration and about 29.3 s for training the SVDD component. Testing a single sample took approximately 0.43 s. It is important to note that these training times are subject to variation based on several factors such as the computational capability of the hardware used, optimization of the implementation, and the volume of data each local model processes. As such, these figures should be considered as indicative benchmarks rather than absolute constants.

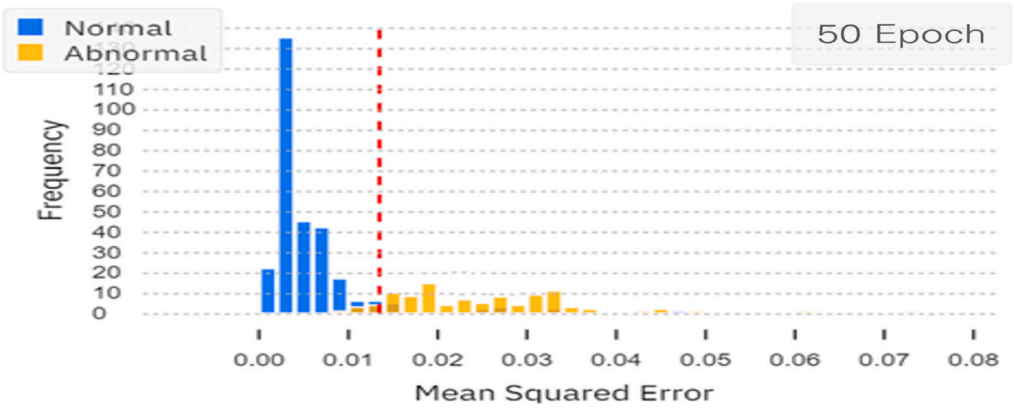


Fig. 14. Normal and abnormal samples distribution.

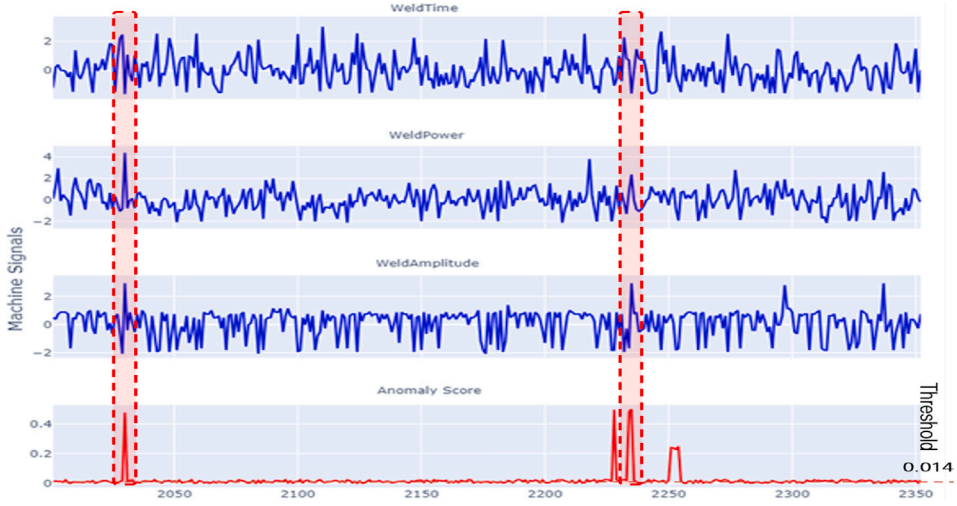


Fig. 15. Visualization of anomaly score on Rosenberger dataset.

6. Conclusions

In this work, we introduced the AAT, an Anomaly Detection model tailored for Smart Manufacturing, leveraging the prowess of transformer architectures and built upon the principles of adaptive adversarial transformers. By harnessing the capabilities of transformer-based encoder–decoders, AAT ensures expedited model training without compromising on the efficacy of detection. The model's unique fusion of adaptivity and adversarial techniques, reminiscent of the paradigm of Generative Adversarial Networks, accentuates its robustness and precision, especially in challenging environments such as the shop floor data for cable assembly manufacturing.

For the myriad datasets explored in this endeavor, AAT not only manifested superior detection benchmarks but also evinced a marked decrement in training durations relative to traditional paradigms. The stability inherent to AAT, coupled with its rapid training mechanics and adaptability, underscore its scalability and its aptitude for deployment in demanding industrial milieus. An additional feather in AAT's cap is its provision to modulate sensitivity, proffering a gamut of detection granularities from a singular model instantiation. Such versatility is a boon in industrial scenarios, enabling operative teams to calibrate detection sensitivities in tandem with the vicissitudes of operational exigencies, thus optimizing the fulcrum of operational efficiencies.

Our hands-on exploration with the shop floor data for cable assembly manufacturing stood a testament to AAT's industrial aptitude, underscoring its potential to herald a paradigm shift in the realm of anomaly detection in smart manufacturing arenas. Despite the many accolades, as with any nascent approach, challenges were not in short supply—thus delineating a roadmap replete with avenues ripe for future exploration and refinement.

Peering into the future, we envisage augmenting the AAT framework with avant-garde transformer paradigms, potentially drawing inspirations from bidirectional neural architectures, all in a bid to further the model's adaptability ante across diverse temporal data landscapes. Moreover, the odyssey to full-fledged deployment might unravel novel infrastructure prerequisites and other considerations, all crucial for the seamless integration and zenith performance of AAT in a plethora of manufacturing scenarios.

CRediT authorship contribution statement

Moussab Orabi: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kim Phuc Tran:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Conceptualization. **Philipp Egger:** Validation, Resources, Funding acquisition. **Sébastien Thomassey:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis.

Declaration of competing 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.

Acknowledgments

This research was carried out with the support of Rosenberger Hochfrequenztechnik GmbH & Co. KG, located at Hauptstraße 1, 83413 Fridolfing, Germany. The data employed in this study originates from a genuine production line at this location. All results, analyses, and codes are recognized as the intellectual property of Rosenberger.

References

[1] Dan Luo ST, Dolgui A. A state-of-the-art on production planning in industry 4.0. Int J Prod Res 2023;61(19):6602–32. http://dx.doi.org/10.1080/00207543.2022.2122622.

[2] Maddikunta PKR, Pham Q-V, B P, Deepa N, Dev K, Gadekallu TR, et al. Industry 5.0: A survey on enabling technologies and potential applications. J Ind Inform Integr 2022;26:100257. http://dx.doi.org/10.1016/j.jii.2021.100257, URL: https://www.sciencedirect.com/science/article/pii/S2452414X21000558.

[3] Eduardo M-R. Human-in-the-loop machine learning: A state of the art. Artif Intell Rev 2023;56. http://dx.doi.org/10.1007/s10462-022-10246-w.

[4] Huang H, Yang L, Wang Y, Xu X, Lu Y. Digital twin-driven online anomaly detection for an automation system based on edge intelligence. J Manuf Syst 2021;59:138–50. http://dx.doi.org/10.1016/j.jmsy.2021.02.010, URL: https://www.sciencedirect.com/science/article/pii/S0278612521000467.

[5] Wu D, Jiang Z, Xie X, Wei X, Yu W, Li R. LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT. IEEE Trans Ind Inf 2020;16(8):5244–53. http://dx.doi.org/10.1109/TII.2019.2952917.

[6] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: Proceedings of the 30th international conference on international conference on machine learning, vol. 28, JMLR.org; 2013, p. III–1310–III–1318.

[7] Xu J, Wu H, Wang J, Long M. Anomaly transformer: Time series anomaly detection with association discrepancy. In: International conference on learning representations. 2022, URL: https://openreview.net/forum?id=LzQQ89U1qm_.

[8] Tuli S, Casale G, Jennings NR. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. Proc VLDB Endow 2022;15(6):1201–14. http://dx.doi.org/10.14778/3514061.3514067.

[9] Shaohan H, Yi L, Carol F, Rong H, Yining Z, Hailong Y, et al. HitAnomaly: Hierarchical transformers for anomaly detection in system log. IEEE Trans Netw Serv Manag 2020;PP:1. http://dx.doi.org/10.1109/TNSM.2020.3034647.

[10] Ma Q, Sun C, Cui B, Jin X. A novel model for anomaly detection in network traffic based on kernel support vector machine. Comput Secur 2021;104:102215. http://dx.doi.org/10.1016/j.cose.2021.102215, URL: https://www.sciencedirect.com/science/article/pii/S0167404821000390.

[11] Hawkins DM. Identification of outliers. Springer Dordrecht; 1980, http://dx.doi.org/10.1007/978-94-015-3994-4.

[12] de Giorgio A, Cola G, Wang L. Systematic review of class imbalance problems in manufacturing. J Manuf Syst 2023;71:620–44. http://dx.doi.org/10.1016/j.jmsy.2023.10.014, URL: https://www.sciencedirect.com/science/article/pii/S0278612523002157.

[13] Ma X, Shi W. AESMOTE: Adversarial reinforcement learning with SMOTE for anomaly detection. IEEE Trans Netw Sci Eng 2021;8(2):943–56. http://dx.doi.org/10.1109/TNSE.2020.3004312.

[14] Pickands J. Statistical inference using extreme order statistics. Ann Statist 1975;119–31.

[15] Sheather SJ, Marron JS. Kernel quantile estimators. J Amer Statist Assoc 1990;85(410):410–6. http://dx.doi.org/10.1080/01621459.1990.10476214, arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1990.10476214, URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476214.

[16] Tax DMJ, Duin RPW. Support vector data description. Mach Learn 2004;54(1):45–66. http://dx.doi.org/10.1023/B:MACH.0000008084.60811.49.

[17] Hundman K, Constantinou V, Laporte C, Colwell I, Soderstrom T. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2018, p. 387–95. http://dx.doi.org/10.1145/3219819.3219845.

[18] Lee K, Kim D-W, Lee D, Lee KH. Improving support vector data description using local density degree. Pattern Recognit 2005;38(10):1768–71. http://dx.doi.org/10.1016/j.patcog.2005.03.020.

[19] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. 5th ed. Hoboken, New Jersey: John Wiley & Sons; 2015, p. 712.

[20] Alizadeh M, Rahimi S, Ma J. A hybrid ARIMA–WNN approach to model vehicle operating behavior and detect unhealthy states. Expert Syst Appl 2022;194:116515. http://dx.doi.org/10.1016/j.eswa.2022.116515, URL: https://www.sciencedirect.com/science/article/pii/S0957417422000161.

[21] Görnitz N, Braun M, Kloft M. Hidden Markov anomaly detection. In: Proceedings of the 32nd international conference on international conference on machine learning - vol. 37. JMLR.org; 2015, p. 1833–42.

[22] Fan L, Ma J, Tian J, Li T, Wang H. Comparative study of isolation forest and LOF algorithm in anomaly detection of data mining. In: 2021 international conference on big data, artificial intelligence and risk management. 2021, p. 1–5. http://dx.doi.org/10.1109/ICBAR55169.2021.00008.

[23] Choi K, Yi J, Park C, Yoon S. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. IEEE Access 2021;9:120043–65. http://dx.doi.org/10.1109/ACCESS.2021.3107975.

[24] Su Y, Zhao Y, Niu C, Liu R, Sun W, Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. New York, NY, USA: Association for Computing Machinery; 2019, http://dx.doi.org/10.1145/3292500.3330672.

[25] Zong B, Song Q, Min MR, Cheng W, Lumezanu C, ki Cho D, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations. 2018, URL: https://api.semanticscholar.org/CorpusID:51805340.

[26] Li W, Shang Z, Zhang J, Gao M, Qian S. A novel unsupervised anomaly detection method for rotating machinery based on memory augmented temporal convolutional autoencoder. Eng Appl Artif Intell 2023;123:106312. http://dx.doi.org/10.1016/j.engappai.2023.106312, URL: https://www.sciencedirect.com/science/article/pii/S0952197623004967.

[27] Audibert J, Michiardi P, Guyard F, Marti S, Zuluaga MA. USAD: UnSupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. New York, NY, USA: Association for Computing Machinery; 2020, p. 3395–404. http://dx.doi.org/10.1145/3394486.3403392.

[28] Park D, Hoshi Y, Kemp CC. A multimodal anomaly detector for robot-assisted feeding using an LSTM-Based variational autoencoder. IEEE Robot Autom Lett 2018;3(3):1544–51. http://dx.doi.org/10.1109/LRA.2018.2801475.

[29] Leng J, Lin Z, Zhou M, Liu Q, Zheng P, Liu Z, et al. Multi-layer parallel transformer model for detecting product quality issues and locating anomalies based on multiple time-series process data in Industry 4.0. J Manuf Syst 2023;70:501–13. http://dx.doi.org/10.1016/j.jmsy.2023.08.013, URL: https://www.sciencedirect.com/science/article/pii/S0278612523001632.

[30] Chiang H-T, Hsieh Y-Y, Fu S-W, Hung K-H, Tsao Y, Chien S-Y. Noise reduction in ECG signals using fully convolutional denoising autoencoders. IEEE Access 2019;7:60806–13. http://dx.doi.org/10.1109/ACCESS.2019.2912036.

[31] Hong S, Kang M, Kim J, Baek J. Investigation of denoising autoencoder-based deep learning model in noise-riding experimental data for reliable state-of-charge estimation. J Energy Storage 2023;72:108421. http://dx.doi.org/10.1016/j.est.2023.108421, URL: https://www.sciencedirect.com/science/article/pii/S2352152X23018182.

[32] Wu Z, Zhang H, Wang P, Sun Z. RTIDS: A robust transformer-based approach for intrusion detection system. IEEE Access 2022;10:64375–87. http://dx.doi.org/10.1109/ACCESS.2022.3182333.

[33] Ryndyuk VA, Varakin YS, Pisarenko EA. New architecture of transformer networks for generating natural dialogues. In: 2022 wave electronics and its application in information and telecommunication systems. 2022, p. 1–5. http://dx.doi.org/10.1109/WECONF55058.2022.9803724.

[34] Krichen M. Anomalies detection through smartphone sensors: A review. IEEE Sens J 2021;21(6):7207–17. http://dx.doi.org/10.1109/JSEN.2021.3051931.

[35] Cai X, Xiao R, Zeng Z, Gong P, Ni Y. Itran: A novel transformer-based approach for industrial anomaly detection and localization. Eng Appl Artif Intell 2023;125:106677. http://dx.doi.org/10.1016/j.engappai.2023.106677, URL: https://www.sciencedirect.com/science/article/pii/S0952197623008618.

[36] Pota M, De Pietro G, Esposito M. Real-time anomaly detection on time series of industrial furnaces: A comparison of autoencoder architectures. Eng Appl Artif Intell 2023;124:106597. http://dx.doi.org/10.1016/j.engappai.2023.106597, URL: https://www.sciencedirect.com/science/article/pii/S0952197623007819.

[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems 30 (NIPS 2017), 2017, p. 5998–6008.

[38] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. In: Liu Q, Schlangen D, editors. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Online: Association for Computational Linguistics; 2020, p. 38–45. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6, URL: https://aclanthology.org/2020.emnlp-demos.6.

[39] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: State-of-the-art natural language processing. 2019, CoRR abs/1910.03771, arXiv:1910.03771.

[40] Gao L, Zhang J, Yang C, Zhou Y. Cas-VSwin transformer: A variant swin transformer for surface-defect detection. Comput Ind 2022;140:103689. http://dx.doi.org/10.1016/j.compind.2022.103689, URL: https://www.sciencedirect.com/science/article/pii/S0166361522000860.

[41] Ayoub M, Liao Z, Hussain S, Li L, Zhang CW, Wong KK. End to end vision transformer architecture for brain stroke assessment based on multi-slice classification and localization using computed tomography. Comput Med Imaging Graph 2023;109:102294. http://dx.doi.org/10.1016/j.compmedimag.2023.102294, URL: https://www.sciencedirect.com/science/article/pii/S089561112300112X.

[42] Du NH, Long NH, Ha KN, Hoang NV, Huong TT, Tran KP. Trans-lighter: A light-weight federated learning-based architecture for remaining useful life-time prediction. Comput Ind 2023;148:103888. http://dx.doi.org/10.1016/j.compind.2023.103888, URL: https://www.sciencedirect.com/science/article/pii/S0166361523000386.

[43] Li B, Cui W, Wang W, Zhang L, Chen Z, Wu M. Two-stream convolution augmented transformer for human activity recognition. Proc AAAI Conf Artif Intell 2021;35(1):286–93.

[44] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. NIPS '14, Cambridge, MA, USA: MIT Press; 2014, p. 2672–80.

[45] Li D, Chen D, Jin B, Shi L, Goh J, Ng S-K. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In: Artificial neural networks and machine learning – ICANN 2019: text and time series: 28th international conference on artificial neural networks, Munich, Germany, September 17–19, 2019, proceedings, part IV. Berlin, Heidelberg: Springer-Verlag; 2019, p. 703–16. http://dx.doi.org/10.1007/978-3-030-30490-4_56.

[46] Deng A, Hooi B. Graph neural network-based anomaly detection in multivariate time series. 2021, ArXiv, abs/2106.06947 arXiv:2106.06947, URL: https://api.semanticscholar.org/CorpusID:231646473.