Towards an Awareness of Time Series Anomaly Detection Models' Adversarial Vulnerability

Shahroz Tariq
Data61 CSIRO
Sydney, Australia
shahroz.tariq@csiro.au

Binh M. Le
College of Computing and Informatics
Sungkyunkwan University, S. Korea
bmle@g.skku.edu

Simon S. Woo*
Department of Artificial Intelligence
Sungkyunkwan University, S. Korea
swoo@g.skku.edu

ABSTRACT

Time series anomaly detection is extensively studied in statistics, economics, and computer science. Over the years, numerous methods have been proposed for time series anomaly detection using deep learning-based methods. Many of these methods demonstrate state-of-the-art performance on benchmark datasets, giving the false impression that these systems are robust and deployable in many practical and industrial real-world scenarios. In this paper, we demonstrate that the performance of state-of-the-art anomaly detection methods is degraded substantially by adding only small adversarial perturbations to the sensor data. We use different scoring metrics such as prediction errors, anomaly, and classification scores over several public and private datasets ranging from aerospace applications, server machines, to cyber-physical systems in power plants. Under well-known adversarial attacks from Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) methods, we demonstrate that state-of-the-art deep neural networks (DNNs) and graph neural networks (GNNs) methods, which claim to be robust against anomalies and have been possibly integrated in real-life systems, have their performance drop to as low as 0%. To the best of our understanding, we demonstrate, for the first time, the vulnerabilities of anomaly detection systems against adversarial attacks. The overarching goal of this research is to raise awareness towards the adversarial vulnerabilities of time series anomaly detectors.

CCS CONCEPTS

• **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Computing methodologies** → Adversarial learning.

KEYWORDS

Adversarial Attack, Anomaly Detection, Time Series, Classification

ACM Reference Format:
Shahroz Tariq, Binh M. Le, and Simon S. Woo. 2022. Towards an Awareness of Time Series Anomaly Detection Models' Adversarial Vulnerability. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3511808.3557073

*Corresponding author ✉

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557073

1 INTRODUCTION

Machine learning and deep learning have profoundly impacted numerous fields of research and society over the last decade [16, 31]. Medical imaging [33], speech recognition [27], environmental sciences [35, 57] and smart manufacturing systems [61] are a few of these areas. With the proliferation of smart sensors, massive advances in data collection and storage, and the ease with which data analytics and predictive modeling can be applied, multivariate time series data obtained from collections of sensors can be analyzed to identify particular patterns that can be interpreted and exploited. Numerous researchers have been interested in time series anomaly detection [9, 26, 39, 42, 45, 46, 53, 64]. For instance, time series anomaly detection methods are used in the aerospace industry for satellite health monitoring [46, 48, 53]. These deep neural network-based solutions outperform the competition on a variety of benchmark datasets. However, as deep learning became more prevalent, researchers began to investigate the vulnerability of deep neural networks, particularly to adversarial attacks. In the context of image recognition, an adversarial attack entails modifying an original image in such a way that the modifications are nearly imperceptible to the human eye [63]. The modified image is referred to as an adversarial image, as it will be classified incorrectly by the neural network, whereas the original image will be classified correctly. One of the most well-known real-world attacks involves manipulating the image of a traffic sign in such a way that it is misinterpreted by an autonomous vehicle [14]. The most common type of attack is gradient-based, in which the attacker modifies the image in the direction of the gradient of the loss function relative to the input image, thereby increasing the rate of misclassification [17, 37, 63].

While adversarial attacks have been extensively studied in the context of computer vision areas, they have not been extensively investigated for anomaly detection systems with time-series data. It is surprising to see much less research performed, given the increasing popularity of deep learning models for classifying time series [36, 62, 68]. Additionally, adversarial attacks are possible in a large number of applications that require the use of time series data. For instance, Figure 1 (top) depicts the original and perturbed time series for the Korean Aerospace Research Institute's KOMPSAT-5 satellite (KARI) [53]. The prediction error (see Figure 1, right) is generated by the Convolutional LSTM with Mixtures of Probabilistic Principal Component Analyzers (CLMPPCA) method [53], which is currently incorporated at KARI, to predict anomalies. While CLMPPCA accurately predicts the anomaly for the original time series, adding small perturbations in the form of FGSM and PGD attacks causes the entire input samples to be classified as an anomaly. This attack can have a severe impact on the satellite health monitoring.

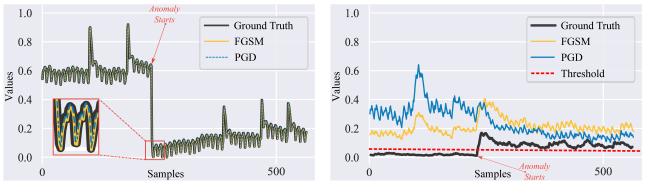


Figure 1: Example of ground truth and perturbed time series using FGSM and PGD attacks on CLMPPCA.

We present, transfer, and apply adversarial attacks that have been demonstrated to work well on images to time series data (containing anomalies) in this work. Additionally, we present an experimental study utilizing benchmark datasets from the aerospace and power plant industries and server machines, demonstrating that state-of-the-art anomaly detection methods are vulnerable to adversarial attacks. We highlight specific real-world use cases to emphasize the critical nature of such attacks in real-world scenarios. Our key findings indicate that deep networks for time series data, similar to their computer vision counterparts, are vulnerable to adversarial attacks. As a result, this paper emphasizes the importance of protecting against such attacks, particularly when anomaly detection systems are used in sensitive industries such as aerospace and power plants. Finally, we discuss some mechanisms for avoiding these attacks while strengthening the models' resistance to adversarial examples.

Aim, Scope and Contribution. In this work, we do not propose any novel adversarial attack method. However, we apply and demonstrate the threat of well-known existing adversarial attacks such as FGSM and PGD towards state-of-the-art anomaly detection methods for multivariate time-series data. In comparison to the computer vision domain, where adversarial attack has been extensively studied and investigated, the literature on novelty detection, and particularly on anomaly detection, is noticeably devoid of such studies. The purpose of this paper is to bring attention to this critical issue, whereas time series anomaly detection models also play pivotal roles in real-world scenarios as other vision tasks. Additionally, we hope to encourage researchers to consider robustness to adversarial attacks when evaluating time-series based future detectors. The paper's scope was limited to analyzing SOTA anomaly detectors. Finally, we successfully degraded the detection performance of deployed systems in the power plant and aerospace industries by employing adversarial attacks. It highlights how vulnerable the current generation of anomaly detectors is to adversarial attacks. Our source code and other implementation details are available here: https://github.com/shahroztariq/Adversarial-Attacks-on-Timeseries.

2 RELATED WORK

In this section, we present background information, notations, and related works, with a particular emphasis on time series anomaly detection and adversarial attacks.

Background and Notations. When performing a supervised learning task, we define $D = \{(s_i, y_i)|i = 1, \ldots, N\}$ to represent a dataset containing N data samples. Each data sample is composed of a m-dimensional multivariate time series s_i and a single target value y_i for classification. However, the majority of anomaly detection occurs in an unsupervised setting. As a result, we take a slightly different approach from the supervised task. Hence, for unsupervised learning, each data sample is again composed of a m-dimensional multivariate time series s_i however, y_i is an n-dimensional multivariate time series obtained from an autoregressive model, predicting the future. In most cases, $n = m$ however, they can be different as well. Moreover, we define any deep learning method as $\mathcal{F}(\cdot) \in f : \mathbb{R}^N \rightarrow \widehat{y}$ and loss function (e.g., cross entropy or mean squared error) as $\mathcal{L}_f(\cdot, \cdot)$. Finally, generating an adversarial instance s_i^{adv} can be described as an optimization problem given a trained deep learning model $\mathcal{F}$ and an original input time series s_i, as follows:

$$\min \left\| s_i - s_i^{adv} \right\| \; s.t. \; \mathcal{F}(s_i) = \widehat{y}_i, \;\; \mathcal{F}(s_i^{adv}) = \widehat{y_i^{adv}} \;\; and \;\; \widehat{y}_i \neq \widehat{y_i^{adv}} \quad (1)$$

Adversarial Attacks. In 2014, Szegedy et al. [49] introduced adversarial examples against deep neural networks for image recognition tasks for the first time. Following these inspiring discoveries, an enormous amount of research has been devoted to generating, understanding, and preventing adversarial attacks on deep neural networks [14, 17, 37]. Adversarial attacks can be broadly classified into two types: White-box and Black-box attacks. As White-box attacks presume access to the model's design and parameters, they can attack the model effectively and efficiently using gradient information. By contrast, Black-box attacks do not require access to the output probabilities or even the label, making them more practical in real-world situations. However, Black-box attacks frequently take hundreds, if not millions, of model queries to calculate a single adversarial case.

The majority of adversarial attack techniques have been proposed for use in image recognition. For instance, a Fast Gradient Sign Method attack was developed by Goodfellow et al. [17] as a substitute for expensive optimization techniques [49]. Madry et al. [37] proposed Projected Gradient Descent (PGD) in response to the success of FGSM. PGD seeks to find the perturbation that maximizes a model's loss on a particular input over a specified number of iterations while keeping the perturbation's size below a specified value called epsilon (ϵ). This constraint is typically expressed as the perturbation's L^2 or L^∞ norm. It is added to ensure that the content of the adversarial example is identical to that of the unperturbed sample — or even to ensure that the adversarial example is imperceptibly different from the unperturbed sample. Carlini-Wagner is another well-known attack [7]. However, it is primarily intended for L^2 norm-based attacks, whereas this study focuses exclusively on L^∞ norm-based attacks.

Adversarial Attacks on Time Series Anomaly Detectors. Surprisingly, limited efforts have been made to extend computer vision-based adversarial attacks to time series anomaly detection domain. However, a few adversarial attack approaches have been proposed recently for the time series classification task, which are tangentially related to our work. For instance, in their work on adopting a soft K Nearest Neighbors (KNN) classifier with Dynamic Time Warping (DTW), Oregi et al. [38] demonstrated that adversarial examples could trick the proposed nearest neighbors classifier on a single simulated synthetic control dataset from the UCR archive [12]. Given that the KNN classifier is no longer considered the state-of-the-art classifier for time series data [4], Fawaz et al. [15] extend this work by examining the effect of adversarial attack on the more recent and commonly used ResNet classifier [20]. Fawaz et al. [15],

on the other hand, focused mainly on univariate datasets from the UCR repository. As a result, Harford et al. [19] investigate the influence of adversarial attacks on multivariate time series classification using the multivariate dataset from UEA repository [3]. However, Harford et al. [19] only consider basic methods such as 1-Nearest Neighbor Dynamic Time Warping [43] (1-NN DTW) and a Fully Convolutional Network (FCN). Karim et al. [24] and Harford et al. [19] attacked models using Gradient Adversarial Transformation Networks (GATNs). However, they examined just transfer attacks, a relatively weak sort of Black-box attack. Only Siddiqui et al. [47] demonstrated the effectiveness of gradient-based adversarial attacks on time series classification and regression networks. However, they considered a very simple baseline for the attack, containing only 3 convolutional, 2 max-pooling, and 1 dense layer.

***Note*:** Our study differs from previous research in that we focus on time series anomaly detection rather than the broader classification problem. More precisely, we explore autoregressive models that have been mostly overlooked in prior works. Additionally, rather than targeting generic deep neural networks KNN with DTW or ResNet, we investigate state-of-the-art anomaly detection methods. For instance, when it comes to anomaly detection, we focus on the most contemporary and commonly used techniques, such as MSCRED [65], CLMPPCA [53], and MTAD-GAT [67]. Section 5 will cover these methods in further depth.

3 THREAT MODEL

To fully define the adversary, we divide the threat model into three subsections based on the adversary's capabilities, knowledge, and goals.

Adversary's Capabilities. We consider an adversary whose objective is to reduce the effectiveness of a victim model. The attacker can apply the perturbations by modifying the victim's test-time samples, for example, by compromising a sensor or the data link that collects the data for inference. We investigate a L^∞ norm threat model with a 0.1 epsilon. Due to the variable input range of time series data, there are no box constraints, in contrast to the visual image, where the pixels take on a definite value between $[0, 255]$. As a result, the data was standardized in our case using a zero-mean and unit standard deviation which justified the choice of 0.1 as the epsilon value.

Adversary's Knowledge. To evaluate the vulnerability of anomaly detection systems, we examine non-targeted White-box and Black-box scenarios. Typically, the attacker is given complete knowledge of the victim model, including its training data and the model's tunable parameters and weights. However, we believe it to be unpractical in our scenario. As most of the system in our analysis are behind some layer of firewall or defense protection and most of the model parameters are hidden. Therefore, we consider two types of adversary's knowledge as follows:

1) *Complete Knowledge*: The attacker understands how the model and its parameters works. We can consider a *White-box* attack to be the most appropriate method for this type of adversary.

2) *Partial or No Knowledge*: Given that the attacker has no or limited knowledge of the system, a *Black-box* attack is the most appropriate method in this case. As a result, strategies such as transfer-based priors [8] can be applied by the adversary.

Adversary's Goals. The adversary considers two cases: (i) normal to anomaly and (ii) anomaly to normal. In (i), the adversary creates a s_i^{adv} for each test sample s_i so that the models interpret it as an anomaly, thereby generating a false-positive. However, in (ii), the adversary fabricates s_i^{adv} to achieve the inverse effect, namely, to cause the model to predict an anomaly as normal, hence generating false-negative examples. As anomalies are rare events, even a few misclassifications caused by the adversary can have a detrimental effect on the model's performance.

4 ADVERSARIAL ATTACK GENERATION

The Fast Gradient Sign Method (FGSM) attack was proposed for the first time by Goodfellow et al. [17]. The training of neural networks entails minimizing a loss function by adjusting the network weights. FGSM, on the other hand, does the opposite by adjusting the input samples in the direction opposite to the loss function's minimum. Thus, the FGSM attack is concerned with the computation of optimal perturbation series η, which can be added/summed to an input sample pointwise (i.e., a point refers to a single timestep) in order to maximize the classification loss function, i.e., cause misclassifications. This is mathematically expressed as:

$$\eta = \epsilon \cdot \text{sign}\left(\nabla_s \, \mathcal{L}_f(s_i, y_i)\right) \tag{2}$$

where ∇_s denotes the derivative of the network's loss, $\mathcal{L}_f(\cdot, \cdot)$, with respect to each timestep in s_i (calculated for an input datapoint s_i and it's true output y_i). To control the magnitude of the perturbation (i.e., to keep it imperceptibly small), ϵ is used as a multiplier factor. After that, the perturbed sample s_i^{adv} can be computed as $s_i + \eta$. Note that FGSM requires the attacker to compute the loss function gradient with respect to a given input, which may not be possible directly. Due to the fact that FGSM requires knowledge of the internal workings of the network it is therefore referred to as a White-box attack. However, a surrogate model can be used to simulate the target model. An FGSM attack can be applied to the surrogate to generate adversarial examples [41], allowing for the use of such White-box attacks in practical scenarios [29].

Madry et al. [37] proposed a more robust adversarial attack called Projected Gradient Descent (PGD). This attack employs a multi-step procedure and a negative loss function. It overcomes the problem of network overfitting and the shortcomings of the FGSM attack. It is more robust than first-order network information-based FGSM, and it performs well under large-scale constraints. Gradient Descent is essentially identical to the Basic Iterative Method (BIM) [29] or the Iterative FGSM (IFGSM) [28] attacks. The only difference is that PGD initializes the example at a random location within the ball of interest (determined by the L^∞ norm) and performs random restarts, whereas BIM initializes at the original location.

$$\begin{gathered} s_{i,t+1}^{adv} = \Pi_{s+\delta}\left(s_{i,t}^{adv} + \alpha \ \text{sign}(\nabla_s \mathcal{L}_f(s_{i,t}^{adv}, y))\right) \\ s.t. \quad 1 \leq t \leq T \end{gathered} \tag{3}$$

where δ is a nonempty compact topological space, T is the total number of iterations, and α is the control rate. An illustration of the overall pipeline is provided in Figure 2.

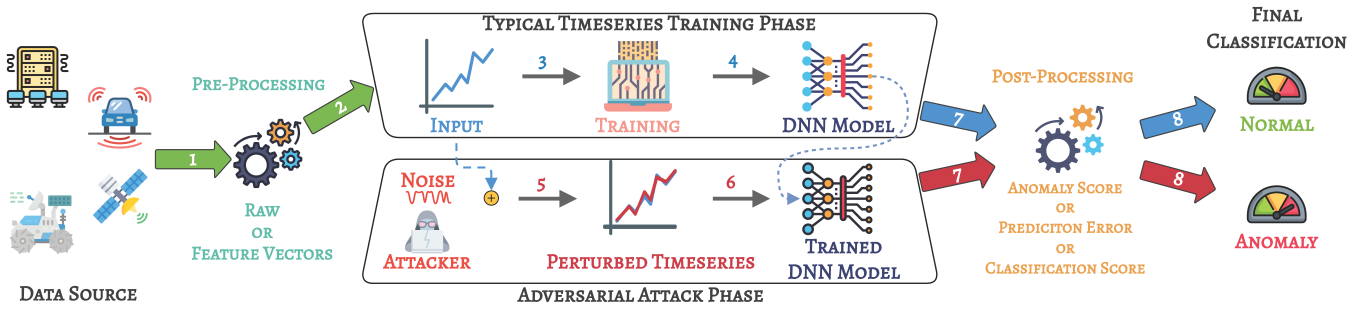


Figure 2: Pipeline of a typical time series anomaly detection's training phase and our adversarial attack phase.

Table 1: A summary of anomaly detection datasets.

Statistics	SMAP	MSL	SMD	KARI	Synthetic
Dimensions	55	27	28	4-35	30
Anomalies	13.13%	10.27%	4.16%	1.00%	1.10%
Train Size	135,183	58,317	708,405	4,405,636	8,000
Test Size	427,617	73,729	708,420	17,622,546	10,000

5 EXPERIMENTAL SETUP

This section contains information on the benchmark datasets, evaluation metrics, criteria for selecting baselines, and chosen baselines.

Datasets For anomaly detection we employ three public datasets: (i) Mars Science Laboratory rover (MSL) [22], (ii) Soil Moisture Active Passive satellite (SMAP) [22], and (iii) Server Machine Dataset (SMD) [48], as well as one private dataset: (vi) Korean Aerospace Research Institute KOMPSAT-5 satellite (KARI) [53] and one synthetic dataset: (v) from the MSCRED paper [65]. The datasets were chosen based on our baselines' shown ability to provide state-of-the-art performance on them. Table 1 summarize these datasets.

Evaluation Metrics To obtain the final classification result for anomaly detection methods, we observed that the majority of detectors use a thresholding method on top of the neural network's predictions, which are expressed as an anomaly score or prediction error. The precision, recall, and F1-score are then calculated using the results from thresholding methods. While these metrics are beneficial, the true impact of the adversarial attack is visible primarily in anomaly detectors' anomaly score and prediction errors. Therefore, we include Figure 1, 3b and 3a, as illustrations of this impact. Additionally, we include more related figures in Appendix B– D.

5.1 Anomaly Detection Baselines

5.1.1 ***Selection Criteria.*** We conduct experiments on the following baselines to demonstrate that the vulnerability to adversarial attacks is common among several state-of-the-art anomaly detection architectures. Anomaly detectors based on Deep Neural Networks (DNNs) are the most frequently used method. However, some methods based on Graph Neural Networks (GNNs) have also been proposed recently. As a result, we evaluated both DNNs- and GNNs-based anomaly detectors. We used the following criteria to determine the baseline:

(1) ***Diverse Architecture*:** To ensure that we cover a broad range of methods, we decide that the baselines should be diverse, i.e., no two baselines have similar model architecture.
(2) ***Diverse pre-processing*:** They should consider a different preprocessing technique (e.g., using raw data or feature vectors).
(3) ***Diverse post-processing*:** They should take into account various post-processing techniques for prediction (e.g., anomaly score, prediction error, or classification score).
(4) ***Peer-reviewed*:** The method is widely accepted and peer-reviewed. For this criterion, we take into account GitHub Forks, paper citations, and publication venues.
(5) ***Open-source*:** The source code is freely available or can be obtained upon request.

5.1.2 ***Selected Baselines.*** We choose the following baselines based on the aforementioned criteria:

***MSCRED* [65] [*AAAI'19*]:** Taking advantage of the temporal dependencies inherent in multivariate time series, Zhang et al. [65] proposed a Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) for anomaly detection on two datasets: (i) synthetic and (ii) power plant. Sidenote: Shen et al. [44] demonstrated that MSCRED outperforms all SOTA anomaly detection methods except Temporal Hierarchical One-Class (THOC), but we were unable to evaluate THOC as the code is not available (see more details below this list). As a result, we chose the second best method (i.e., MSCRED) among recently developed SOTA anomaly detection methods. Because the power plant dataset is not publicly available, we compare MSCRED with and without adversarial attack using the synthetic dataset used by Zhang et al. [65] in their work.

***CLMPPCA* [53] [*KDD'19*]:** Tariq et al. [53] proposed a hybrid approach for anomaly detection in multivariate satellite telemetry data. Based on the accomplishments of Convolutional LSTM-based networks in understanding spatiotemporal data in various domains [25, 50, 54, 56], they propose a Convolutional LSTM with Mixtures of Probabilistic Principal Component Analyzers (CLMP-PCA) method for transforming the time window containing several telemetry data samples into a feature vector that is used to train the model and to predict the future data instances. To make final classification, the prediction errors calculated from the prediction and ground truth are combined with a moving average-based threshold method. In their work Tariq et al. [53] used a private dataset from the Korean Aerospace Research Institute's (KARI) KOMPSAT-5 satellite for evaluation. We were able to obtain the same private dataset and demonstrate how adversarial attacks affect the performance of CLMPPCA. One of the primary reasons for selecting CLMPPCA is that it is currently deployed at KARI. Thus, successfully demonstrating an attack on this method will demonstrate its applicability in a practical scenario.

***MTAD-GAT* [67] [*ICDM'20*]:** Zhao et al. [67] proposed a multivariate time series anomaly detector based on Graph Attention

Networks. The authors treat each univariate time series as a separate feature and employ two parallel graph attention layers to learn the complex dependencies between multivariate time series in both temporal and feature dimensions by jointly optimizing a forecasting-based and reconstruction-based model. MTAD-GAT outperformed several recent time series anomaly detectors such as OmniAnomaly [48], MAD-GAN [32], and DAGMM [69] from ICLR 2018, on three publicly available anomaly datasets (SMAP, MSL, and SMD). As a result, MTAD-GAT is one of the best SOTA methods currently available. We evaluate MTAD-GAT with and without adversarial attacks on all three datasets (i.e., SMAP, MSL, and SMD).

***Note*:** We chose these three baselines based on their compliance with our defined criteria. Additionally, we were unable to evaluate some recent methods, such as Temporal Hierarchical One-Class (THOC) published at NeurIPS 2020 because the source code is not publicly available and our request to obtain the source code from the author was not answered. We discuss this further in Section A.

5.1.3 ***White- and Black-box Attack Settings.*** As the attacker will have complete knowledge of the underlying system in a White-box attack, we create attack vectors using the same selected baselines, namely MSCRED, CLMPPCA, and MTAD-GAT. Whereas for the Black-box attack, we build attack vectors using a model that is similar to but simpler than the victim model. For example, we utilise a vanilla recurrent autoencoder to create attack vectors for MSCRED, a simple CNN+LSTM model for CLMPPCA, and a vanilla GNN for MTAD-GAT.

6 EMPIRICAL EVALUATION

We present results for the L^∞ FGSM and PGD attacks against three SOTA anomaly detection methods—MSCRED, CLMPPCA, and MTAD-GAT. The Appendix includes additional details about the L^∞, L^1, and L^2 attacks results (Appendix B); more details on impact of adversarial attacks on MTAD-GAT (Appendix C); some original vs. perturbed time series samples (Appendix D. Moreover, results from the FGSM, PGD, BIM, Carlini-Wagner, and Momentum Iterative Method (MIM) [13] attacks on 71 datasets from the UCR repository are available on our GitHub repository. In general, we observe that perturbations that are L^∞-bounded are more effective. This could be explained by optimization challenges, as L^1 and L^2 attacks are typically more difficult to optimize [7, 59].

6.1 Adversarial Attack on MSCRED

6.1.1 ***MSCRED (White-box).*** We employ non-targeted FGSM and PGD methods to attack MSCRED. As a result, only s_i from the test set is made available to the attack methods. The ϵ is set to 0.1 for the FGSM attack, and α is set to 0.1 for the PGD attack with $T = 40$. The MSCRED method determines the appropriate threshold between normal and anomalous data points based on the training data. As a result, any modification to the test samples should not affect the threshold. As shown in Table 2, the victim model (MSCRED) has no efficacy on the samples perturbed by FGSM and PGD attacks and thus fails to detect all anomalies. Additionally, MSCRED classifies all instances of normal data as anomalies. We demonstrate in Figure 3a that MSCRED (No Attack) can accurately predict the majority of anomalies with an F_1 score of 0.890 (see Table 2). According to Figure 3a, the anomaly scores under FGSM (yellow) and PGD (blue) attacks are always higher than the threshold (red dashed line), which means that MSCRED is predicting everything as an anomaly, resulting in an F_1 score of less than 0.50. It is intriguing that such a small amount of change in the time series, which is primarily imperceptible to the naked eye, can greatly affect the MSCRED's anomaly scores, even when the perturbations are so minute. The results in Table 2 are demonstrating that MSCRED is not robust against adversarial attacks.

Table 2: MSCRED results (F_1 score) on synthetic dataset from original paper. Both White- and Black-box attacks show significant success with FGSM and PGD. We used surrogate model (vanilla recurrent autoencoder) to generate adversarial examples for Black-box attack.

	White-box			Black-box		
Method	***Pre.***	***Rec.***	F_1	***Pre.***	***Rec.***	F_1
No Attack	1.000	0.800	0.890	1.000	0.800	0.890
FGSM	0.487	0.500	0.493	0.651	0.693	0.671
PGD	**0.485**	**0.500**	**0.492**	**0.634**	**0.677**	**0.655**

6.1.2 ***MSCRED (Black-box).*** As with the White-box attack, the Black-box attack significantly reduced MSCRED's F1-score. This reduction, however, is slightly less than that caused by a White-box attack. Moreover, the PGD attack reduced F1-scores more than the FGSM attack, as shown in Table 2. This experiment demonstrates that even when we build the attack vector using a different backbone model, we can still achieve significant success by transferring the adversarial attack.

6.2 Adversarial Attack on MTAD-GAT

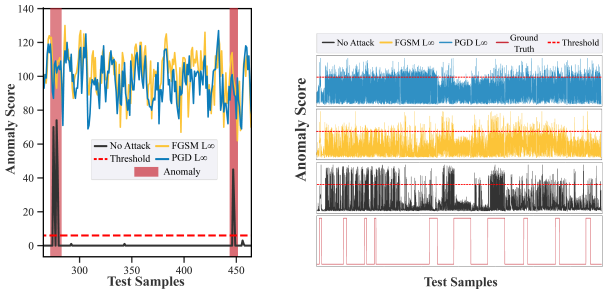
6.2.1 ***MTAD-GAT (White-box).*** As with MSCRED, we attack MTAD-GAT using a non-targeted FGSM and PGD method with $\epsilon = 0.1$, $\alpha = 0.1$, and $t = 40$. The results of adversarial attacks against MTAD-GAT trained on the MSL, SMAP, and SMD datasets are shown in Table 3. MTAD-GAT demonstrates state-of-the-art performance for anomaly detection in the absence of an adversarial attack (No Attack). However, when adversarial examples from FGSM and PGD are used to evaluate it, the detection performance drops to as low as 66%. The impact of the PGD attack is more significant than that of the FGSM attack, which is understandable given that PGD is a more powerful attack than FGSM. It leads us to ponder that if more sophisticated attacks are explicitly developed for time series data, they will have a significantly greater impact on SOTA anomaly detectors. As a result, future anomaly detection methods should take adversarial examples into account.

Additionally, Figure 3b illustrates the effect of adversarial examples from FGSM (yellow) and PGD (blue) attacks on the MTAD-GAT anomaly score for the MSL dataset. We can see that the anomaly scores for FGSM and PGD frequently exceed the threshold (red dashed line), resulting in a large number of false positives and lowering the F_1 score from 94.98% to 71.90% for FGSM and 68.69% for PGD.

6.2.2 ***MTAD-GAT (Black-box).*** Like the Black-box attack on MSCRED, the attack on MTAD-GAT has a similar effect, lowering

Table 3: MTAD-GAT results (F_1 score) on MSL, SMAP and SMD datasets. For all three datasets, both White- and Black-box attacks are highly effective. We used vanilla GNN as a surrogate model to generate adversarial examples for Black-box attack.

	White-box			Black-box		
Method	***MSL***	***SMAP***	***SMD***	***MSL***	***SMAP***	***SMD***
No Attack	0.950	0.894	0.999	0.950	0.894	0.999
FGSM	0.719	0.804	0.803	0.751	0.847	0.852
PGD	**0.687**	**0.775**	**0.665**	**0.727**	**0.815**	**0.749**



(a) No Attack, FGSM and PGD on MSCRED **(b) No Attack, FGSM and PGD on MTAD-GAT**

Figure 3: Anomaly score of No Attack, FGSM and PGD on MSCRED (a) and on MTAD-GAT for MSL dataset (b) The y-axis scale is between 0 and 1 for (b). See Appendix B and C for more details on (a) and (b), respectively.

the F1-score for MSL, SMAP, and SMD to 0.751, 0.847, and 0.852 with the FGSM attack and 0.727, 0.815, and 0.749 with the PGD attack, as shown in Table 3. As with the Autoencoder, this experiment indicates that it is possible to transfer adversarial attack to graph neural networks. Thus, demonstrating that the adversary may not require extensive knowledge of the backbone to launch a successful attack.

6.3 Adversarial Attack on CLMPPCA

6.3.1 ***CLMPPCA (White-box).*** The KARI dataset is divided into ten subsystems. As a result, we trained the CLMPPCA model on each subsystem separately, as described in the original paper. We then used FGSM and PGD attacks to evaluate each of these trained models. For FGSM, we use $\epsilon = 0.1$, for PGD, we use $\alpha = 0.1$, and $t = 40$. Table 4 summarizes the prediction errors for each subsystem prior to and following the attack. We can see that when adversarial attacks are used, the prediction error increases up to twentyfold. Note: For brevity and space constraints, we omit the F_1 score from Table 4, as it is 0.50 for all subsystems. CLMPPCA fails to predict any anomalies under FGSM and PGD attacks because the prediction error is always higher than the threshold (see Figure 1). We believe that by employing these straightforward yet effective attacks, an adversary can easily introduce false positives into CLMPPCA's predictions at will, posing significant difficulties for satellite operators.

6.3.2 ***CLMPPCA (Black-box).*** We generated the attack vector using a CNN+LSTM surrogate model and evaluated the CLMPPCA model in a Black-box scenario. As with the other two Black-box experiments (i.e., MSCRED and MTAD-GAT), we saw a similar trend. The CLMPPCA model's prediction error does increase consistently for all subsystems when the attack vector is constructed using the surrogate model, as shown in Table 4; in round brackets. We can deduce from the CLMPPCA Black-box results that all three types of models investigated in this work (i.e., Autoencoder, DNNs, GNNs) are relatively equally susceptible to transferable adversarial attacks via surrogate models.

6.4 Summary of Results

Our findings indicate that the majority of SOTA anomaly detectors prioritized performance over robustness. This could have dire consequences if such systems are deployed in real-world systems. CLMPPCA is one such example, which is currently being deployed at KARI. Please note that we have informed KARI of the vulnerability in CLMPPCA; additional information is available in our Ethics Statement (see Section 7). Additionally, leveraging a surrogate model to conduct a Black-box attack can have a severe effect on the performance of the victim model. However, there are several limitations to surrogate, which we shall discuss in Section 7.

7 DISCUSSION

Adversarial Time Series Defense. Adversarial training is one of the most commonly used defense methods against adversarial examples. However, as Kang et al. [23] suggest, training a network to withstand one type of attack may weaken it against others. Additionally, Tramer et al. [60] outline various methods to conduct an adaptive attack and demonstrate that none of the 13 recently developed defense methods can withstand all types of adaptive attacks. Recently, a few techniques for defending against adversarial time series have been proposed. For example, Goodge et al. [18] propose an Approximate Projection Autoencoder (APAE) resistant to IFGSM attacks. However, it only considers autoencoder-based anomaly detectors. Moreover, the performance of several SOTA baselines reported in the paper is significantly lower than that reported in their original paper using the same publicly available benchmark dataset. As a result, a thorough examination of the defense methods is required.

In order to encourage the studies of adversarial robustness for time series anomaly detection models, we will discuss here some possible approaches that are primarily motivated by computer vision areas. From the perspective of adversarial generation, perturbations created by attackers have mainly relied on gradients of model predictions *w.r.t* its invaded inputs. We can apply the input-output Jacobian regularization in order to agnostically silent the model's gradients regardless of its input as was shown in [10, 21]. On the other hand, when we have multiple classes in the training dataset, we can focus on aligning distributions of adversarial samples to be resembling to clean ones in the latent space, namely adversarial training [5, 6, 66]. In the one-class training manner, we expect our defense model to learn the intrinsic representative features from the training dataset and be more robust to adding noise in the test set. Therefore, regularizing the embedding space to be more compact is an appealing approach that so far has not been investigated in time-series anomaly detection areas thoroughly. This objective can

Table 4: CLMPPCA prediction errors for subsystems (SS) 1-10 on the KARI KOMPSAT-5 dataset from FGSM and PGD attack. The prediction errors enclosed in brackets are the result of Black-box attack, whereas those outside the brackets are from White-box attack. A higher error value indicates a more powerful attack.

Methods	SS1	SS2	SS3	SS4	SS5	SS6	SS7	SS8	SS9	SS10
No Attack	0.025	0.020	0.646	0.018	0.078	0.081	0.028	0.015	0.043	0.106
FGSM	0.306 (0.132)	0.327 (0.159)	5.657 (3.163)	0.153 (0.092)	1.744 (0.680)	1.708 (0.616)	0.246 (0.115)	0.201 (0.098)	1.303 (0.724)	0.314 (0.191)
PGD	**0.688 (0.333)**	**0.748 (0.382)**	**11.20 (5.216)**	**0.205 (0.135)**	**2.459 (1.301)**	**3.391 (1.630)**	**0.430 (0.206)**	**0.231 (0.139)**	**1.798 (1.105)**	**0.555 (0.249)**

be achieved via sparing the latent space with principal component analysis as demonstrated in [30, 34].

Limitations and Future Work. There are some limitations to our work, and future work will try to solve them. For instance, we could not evaluate all of the recent anomaly detectors in our work due to the following reasons: (i) The most important reason is that the codes are not publicly available in many cases or the code is outdated, making it hard to compare (we discuss this in detail in reproducibility section). (ii) It is hard to reproduce the same results as demonstrated by the paper, mainly when the codes are not from the original authors but developed by the community. Therefore, future work should look for more methods. Moreover, we have only applied FGSM, PGD, and SL1D (see Appendix B) attacks on the detectors. We do provide results from other attacks such as Carlini-Wagner L2 and MIM on the UCR dataset on our GitHub repository. Another future work will be to transfer these and new adversarial attacks to anomaly detectors.

We assumed that the training data for the surrogate model is either publicly available or obtained by probing the simulation results at multiple intelligently chosen places in the design parameter space. However, such an assumption may not hold true in a closed loop system. As a result, future research should focus on developing a more comprehensive strategy for acquiring training data for surrogate models. Finally, developing robust detectors should be considered in future studies.

New Adversarial Attacks. We have not developed a novel type of adversarial attack in this study and have instead utilised some of the more prevalent adversarial methods for the following reasons: (i) We believe that if a simple attack can demonstrate a system's vulnerability, then developing a new more complex attack solely to increase the novelty of the paper is futile, as the primary objective of this paper is to expose anomaly detector's vulnerabilities, not to develop new adversarial attacks. (ii) At the time the baselines reviewed in this study were published, the FGSM and PGD attacks were already well-known; thus, establishing that those baselines are not resilient against FGSM and PGD adversarial attacks provides a fair comparison.

Attack on Intrusion Detection System. Intrusion detection is frequently associated with anomaly detection and, more broadly, novelty detection systems. In contrast to the realm of anomaly detection, numerous attempts have been made to investigate adversarial attacks against intrusion detection systems [1, 2, 11]. As a proof of concept, we also deployed similar adversarial attacks (i.e., FGSM and PGD) to an intrusion detection system for Controller Area Networks and discovered that the attacks are just as effective against them. Owing to the fact that this experiment requires extensive background information, and due to a shortage of space, we provide further details on our GitHub[1] and more context here [58].

Ethics Statement. Our study, in our opinion, raises only one significant ethical issue (i.e., presenting the vulnerabilities of a deployed system). Now, we will describe how we deal with it. To begin, we downloaded the CLMPPCA code from GitHub. Second, we contacted the authors of the CLMPPCA paper and requested the dataset. Following KARI's security clearance. We were able to obtain access to the dataset and some code associated with the driver, which was kept private on purpose. We contacted the authors and informed them of our findings after identifying the vulnerabilities in CLMPPCA. The authors replicated our findings on the deployed system using the same attacks. For the time being, the system is offline, and the authors of the CLMPPCA paper and other KARI developers are investigating possible defense methods. We believe that adhering to this entire procedure resolves any ethical concerns regarding this matter.

8 CONCLUSION

The concept of adversarial attacks on deep learning models for time series anomaly detection was considered in this paper. We defined and adapted adversarial attacks initially proposed for image recognition to time series data. On several benchmark datasets, we demonstrated how adversarial perturbations could reduce the accuracy of state-of-the-art anomaly detectors. As data scientists and developers increasingly implement deep neural network-based solutions for time series related real-world critical decision-making systems (e.g., in aerospace industries), we shed light on several critical use cases where adversarial attacks could have severe and dangerous repercussions. Additionally, we demonstrate empirically that White- and Black-box attacks are both conceivable and can result in significant performance deterioration. Finally, we discuss several defense strategies and possible future directions for adversarially resilient anomaly detector development.

ACKNOWLEDGMENTS

This work was partially supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and ICT (MSIT) under No. 2020R1C1C1006004 and Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open

[1] https://github.com/shahroztariq/Adversarial-Attacks-on-Timeseries

domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. 2021-0-02309, Object Detection Research under Low Quality Video Condition).

A REPRODUCIBILITY

Issues in Baselines. According to our research, the majority of recent anomaly detection methods do not make their source code publicly available. Additionally, many methods whose source code was made publicly available by their authors (or implemented unofficially) were outdated. As a result, we were unable to run them directly on the most recent machines. For instance, in our experiment, we used an Nvidia RTX 3090 GPU. We discovered that, due to some issues with the CUDA version, we could not run an older version of TensorFlow optimally. As a result, the code either takes an eternity to execute or does not execute at all.

Our Solution. We chose to port the baselines to the latest versions of TensorFlow and PyTorch, respectively, which were 2.5.0 for TensorFlow and 1.9.0 for PyTorch at the time of our experiments. We used the CleverHans Library's [40] FGSM, PGD, BIM, Carlini-Wagner L2, SL1D, and MIM attacks, which were recently ported to TensorFlow2 and PyTorch in version 4.0.0. As a result, our workflows are compatible with the latest libraries. Additionally, after cleaning the code, we will include some tutorial attacks (similar to those included in the CleverHans library for image datasets) that can be used to assess future detectors to adversarial attacks.

Guidelines for Baseline. Note that it is difficult to port or implement all of the most recent methods on our own. Therefore, we tried our best with the limited resources that we had to make the baselines compatible with the latest version of libraries. We will provide some guidelines for creating new baselines and evaluating them against adversarial attacks on our GitHub page. We will leave it up to the community to add additional methods in the future.

Links to Baselines and Datasets. We will include the updated codes for each baseline in our repository as well. We obtained the code of the baselines from the following repositories:

- MSCRED [65]: https://github.com/Zhang-Zhi-Jie/Pytorch-MSCRED
- MTAD-GAT [67]: https://github.com/ML4ITS/mtad-gat-pytorch
- CLMPPCA [53]: https://github.com/shahroztariq/CL-MPPCA
- CAN-ADF [51, 52]: https://github.com/shahroztariq/CAN-ADF
- CANTransfer [55]: https://github.com/shahroztariq/CANTransfer

Note that we are unable to share the KARI dataset as it is proprietary and requires security clearance to access. The link to rest of the dataset used in our evaluation are as follows:

- SMAP and MSL: https://s3-us-west-2.amazonaws.com/telemanom/data.zip
- SMD: https://github.com/ML4ITS/mtad-gat-pytorch/tree/main/datasets
- Synthetic: https://github.com/Zhang-Zhi-Jie/Pytorch-MSCRED
- OTIDS: https://ocslab.hksecurity.net/Dataset/CAN-intrusion-dataset

B SUPPLEMENTARY RESULTS: MSCRED

Details on L^∞ FGSM and PGD Attacks. In Figure 4a, we detail MSCRED's performance against L^∞-norm FGSM and PGD attacks. Under normal conditions, we can see that the model correctly predicted three large anomalies but missed two minor ones. As a result, an F_1 score of 0.890 is obtained. However, when attacked with either FGSM or PGD, the MSCRED model produces no meaningful results because it predicts everything as an anomaly. Furthermore, the patterns of anomaly score under FGSM and PGD attack are very similar to those observed during non-anomalous (or normal) periods. As a result, adjusting the threshold to account for changes in the anomaly score will not be as effective.

SL1D and FGSM L^1 Attack. In Figure 4b, we present the results from two L^1 attacks: (i) FGSM L^1 and (ii) Sparse L^1 Descent (SL1D) attacks. As discussed previously in the main paper, optimizing L^1 and L^2-based attacks can be challenging. We can see an excellent illustration of this with the FGSM L^1 attack, where adversarial examples from the L^1-based FGSM attack produce nearly identical results to the No Attack data samples (with a few minor differences). However, the SL1D attack, also an L^1-based attack, performs similar to the L^∞ attack discussed previously. Although the range of anomaly scores produced by SL1D attacks is slightly less than that produced by L^∞ attacks, it is still significantly higher than the threshold making the MSCRED model to predict the whole input time series as an anomaly.

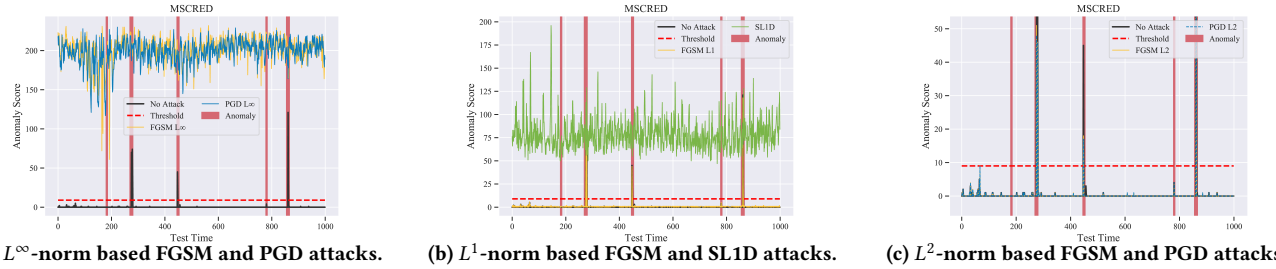
L^2 FGSM and PGD Attack. The results of the L^2-based FGSM and PGD attacks are shown in Figure 4c. Almost identical to the L^1-based FGSM attack, the L^2-based FGSM attack produces adversarial samples that have no effect on the anomaly score and are thus deemed ineffective. Similar results are obtained using the L^2-based PGD attack. As illustrated in Figure 4c, the Anomaly scores for No Attack, FGSM L^2, and PGD L^2 all overlap significantly.

C SUPPLEMENTARY RESULTS: MTAD-GAT

Detailed view of MTAD-GAT Results on MSL Dataset. In this section, we discuss the MTAD-GAT results on the MSL dataset in greater detail. Figure 5a– 5c show No Attack, FGSM attack, and PGD attack results on the entire test data, respectively. We can see that MTAD-GAT predicts fewer anomalies under FGSM and PGD attacks than normal conditions (i.e., No Attack), resulting in a higher rate of false negatives. We have now discussed both of these scenarios in detail in this work: (i) adversarial attack to generate false positives and (ii) adversarial attack to generate false negatives. Additionally, consistent with our previous findings, PGD performs better than FGSM and generates more false negatives than FGSM.

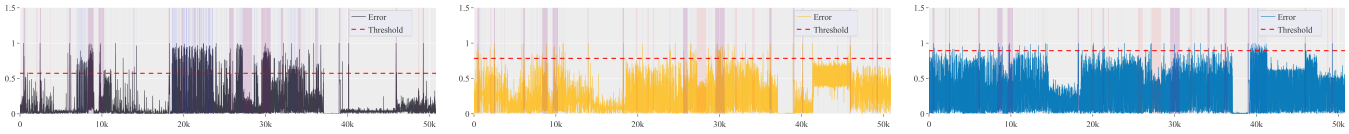
Results on SMD Dataset for MTAD-GAT. We present additional details on the MTAD-GAT results using the Server Machine Dataset (SMD) in Figure 6a, 6b and 6c . In the figures, the top row (in red) represents the Anomaly scores, the middle row (in brown) represents the MTAD-GAT predictions, and the bottom row (in blue) represents the ground truth. We can see that MTAD-GAT performs at a state-of-the-art level under normal conditions. However, when subjected to FGSM and PGD attacks, it generates a large number of false positives, resulting in a significant decrease in overall performance. Also, we can observe that when PGD is used, MTAD-GAT produces more false positives than when FGSM is used.

Effects of FGSM and PGD attacks on MTAD-GAT's Features. As previously stated, MTAD-GAT is composed of two components (i.e., forecasting and reconstruction). We demonstrate in Figure 7a–7c that both components become equally ineffective when subjected to adversarial attacks. For example, in normal circumstances (as illustrated in Figure 7a), the forecast and reconstruction are quite



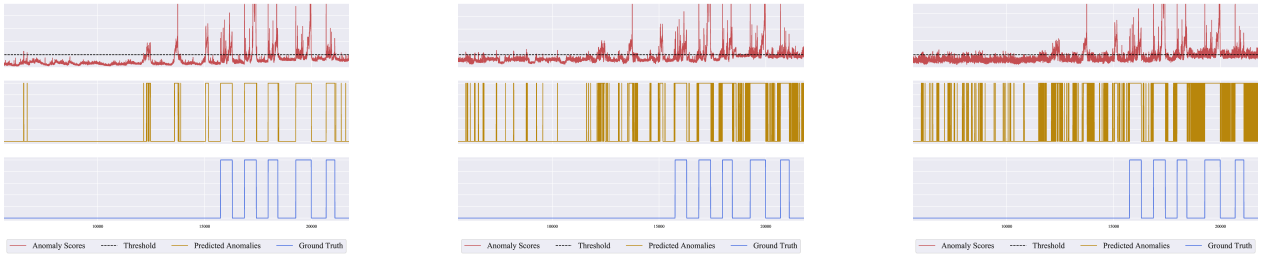
(a) L^∞-norm based FGSM and PGD attacks. (b) L^1-norm based FGSM and SL1D attacks. (c) L^2-norm based FGSM and PGD attacks.

Figure 4: Anomaly score comparison of MSCRED under No Attack, FGSM and PGD attacks.



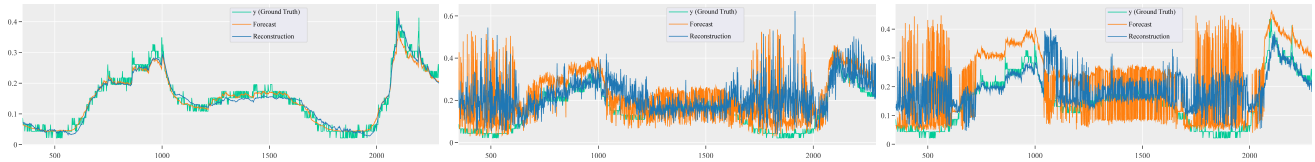
(a) Normal data (i.e., No Attack). (b) FGSM attack. (c) PGD attack.

Figure 5: MTAD-GAT's anomaly score.



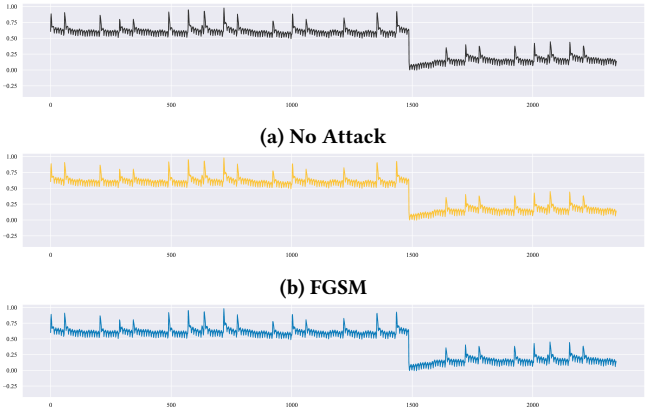
(a) Normal data (i.e., No Attack). (b) FGSM (c) PGD

Figure 6: The anomaly score and predicted anomalies for MTAD-GAT on SMD dataset.



(a) No Attack. (b) FGSM Attack. (c) PGD Attack.

Figure 7: Comparison of Forecast and Reconstruction with y_i



(a) No Attack

(b) FGSM

(c) PGD

Figure 8: A more detailed view of the same time series as in Figure 1.

close to the y_i (ground truth). However, when attacked by FGSM, they deviate from the ground truth, fooling the system into believing it is an anomaly. Additionally, forecast and reconstruction are more chaotic during a PGD attack. As a result, detection performance is even lower than that of a FGSM attack.

D SUPPLEMENTARY RESULTS: CLMPPCA

Original vs. Perturbed Samples. We compare some samples of original and perturbed time series in this section. The ground truth (in black), the FGSM (in yellow), and the PGD (in Blue) are depicted in Figure 8. We can easily see that all three of the time series overlap, rendering them largely indistinguishable to the naked eye. Additionally, Figure 8a– 8c show an expanded version of the time series depicted in Figure 1. Each of the three time series (i.e., No Attack, FGSM, and PGD) appears identical. Here, we demonstrate that even simpler adversarial attacks such as FGSM and PGD can be highly effective on time series data. Such perturbations will go unnoticed by a human observer.

REFERENCES

[1] Giovanni Apruzzese and Michele Colajanni. 2018. Evading botnet detectors based on flows and random forest with adversarial samples. In *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)*. IEEE, 1–8.

[2] Giovanni Apruzzese, Michele Colajanni, and Mirco Marchetti. 2019. Evaluating the effectiveness of adversarial attacks against botnet detectors. In *2019 IEEE 18th International Symposium on Network Computing and Applications (NCA)*. IEEE, 1–8.

[3] Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. 2018. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075* (2018).

[4] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery* 31, 3 (2017), 606–660.

[5] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. 2021. Improving adversarial robustness via channel-wise activation suppressing. *International Conference on Learning Representations (ICLR)* (2021).

[6] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. 2021. Optimal transport as a defense against adversarial attacks. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 5044–5051.

[7] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[8] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems* 32 (2019).

[9] Jinwoo Cho, Shahroz Tariq, Sangyup Lee, Young Geun Kim, Jeong-Han Yun, Jonguk Kim, Hyoung Chun Kim, and Simon S Woo. 2019. Robust Anomaly Detection in Cyber Physical System using Kullback-Leibler Divergence in Error Distributions. In *5th Workshop on Mining and Learning from Time Series (MileTS'19), Anchorage, Alaska, USA*.

[10] Kenneth T Co, David Martinez Rego, and Emil C Lupu. 2021. Jacobian regularization for mitigating universal adversarial perturbations. In *International Conference on Artificial Neural Networks*. Springer, 202–213.

[11] Igino Corona, Giorgio Giacinto, and Fabio Roli. 2013. Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. *Information Sciences* 239 (2013), 201–225.

[12] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica* 6, 6 (2019), 1293–1305.

[13] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9185–9193.

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634.

[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Adversarial attacks on deep neural networks for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT Press.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[18] Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. 2020. Robustness of Autoencoders for Anomaly Detection Under Adversarial Impact.. In *IJCAI*. 1244–1250.

[19] Samuel Harford, Fazle Karim, and Houshang Darabi. 2020. Adversarial attacks on multivariate time series. *arXiv preprint arXiv:2004.00410* (2020).

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Judy Hoffman, Daniel A Roberts, and Sho Yaida. 2019. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729* (2019).

[22] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.

[23] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. 2019. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016* (2019).

[24] Fazle Karim, Somshubra Majumdar, and Houshang Darabi. 2020. Adversarial attacks on time series. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[25] Seongchan Kim, Seungkyun Hong, Minsu Joh, and Sa-kwang Song. 2017. Deeprain: Convlstm network for precipitation prediction using multichannel radar data. *arXiv preprint arXiv:1711.02316* (2017).

[26] Young Geun Kim, Jeong-Han Yun, Siho Han, Hyoung Chun Kim, and Simon S Woo. 2021. Revitalizing Self-Organizing Map: Anomaly Detection Using Forecasting Error Patterns. In *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 382–397.

[27] Akshi Kumar, Sukriti Verma, and Himanshu Mangla. 2018. A survey of deep learning techniques in speech recognition. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 179–185.

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[29] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. 2016. Adversarial examples in the physical world.

[30] Chieh-Hsin Lai, Dongmian Zou, and Gilad Lerman. 2019. Robust subspace recovery layer for unsupervised anomaly detection. *arXiv preprint arXiv:1904.00152* (2019).

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[32] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks*. Springer, 703–716.

[33] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[34] Shao-Yuan Lo, Poojan Oza, and Vishal M Patel. 2021. Adversarially Robust One-class Novelty Detection. *arXiv preprint arXiv:2108.11168* (2021).

[35] Jorge Loy-Benitez, Shahzeb Tariq, Hai Tra Nguyen, Usman Safder, KiJeon Nam, and ChangKyoo Yoo. 2022. Neural circuit policies-based temporal flexible soft-sensor modeling of subway PM2. 5 with applications on indoor air quality management. *Building and Environment* 207 (2022), 108537.

[36] Tengfei Ma, Cao Xiao, and Fei Wang. 2018. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 261–269.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[38] Izaskun Oregi, Javier Del Ser, Aritz Perez, and Jose A Lozano. 2018. Adversarial sample crafting for time series classification with elastic similarity measures. In *International Symposium on Intelligent and Distributed Computing*. Springer, 26–39.

[39] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[40] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *arXiv preprint arXiv:1610.00768* (2018).

[41] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 506–519.

[42] Seoyoung Park, Siho Han, and Simon S Woo. 2020. Forecasting Error Pattern-Based Anomaly Detection in Multivariate Time Series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 157–172.

[43] Skyler Seto, Wenyu Zhang, and Yichen Zhou. 2015. Multivariate time series classification using dynamic time warping template selection for human activity recognition. In *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, 1399–1406.

[44] Lifeng Shen, Zhuocong Li, and James Kwok. 2020. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems* 33 (2020), 13016–13026.

[45] Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, Daewon Chung, Simon Woo, et al. 2019. Integrative Tensor-based Anomaly Detection System For Satellites. (2019).

[46] Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S. Woo. 2020. ITAD: Integrative Tensor-Based Anomaly Detection System for Reducing False Positives of Satellite Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New

York, NY, USA, 2733–2740. https://doi.org/10.1145/3340531.3412716

[47] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. 2019. Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access* 7 (2019), 67027–67040.

[48] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2828–2837.

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[50] Shahroz Tariq, Sowon Jeon, and Simon S Woo. 2022. Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack. In *Proceedings of the ACM Web Conference 2022*. 512–523.

[51] Shahroz Tariq, Sangyup Lee, Huy Kang Kim, and Simon S Woo. 2018. Detecting in-vehicle CAN message attacks using heuristics and RNNs. In *International Workshop on Information and Operational Technology Security Systems*. Springer, 39–45.

[52] Shahroz Tariq, Sangyup Lee, Huy Kang Kim, and Simon S Woo. 2020. CAN-ADF: The controller area network attack detection framework. *Computers & Security* 94 (2020), 101857.

[53] Shahroz Tariq, Sangyup Lee, Youjin Shin, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S Woo. 2019. Detecting anomalies in space using multivariate convolutional LSTM with mixtures of probabilistic PCA. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2123–2133.

[54] Shahroz Tariq, Sangyup Lee, and Simon Woo. 2021. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the Web Conference 2021*. 3625–3637.

[55] Shahroz Tariq, Sangyup Lee, and Simon S Woo. 2020. CANTransfer: transfer learning based intrusion detection on a controller area network using convolutional LSTM network. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 1048–1055.

[56] Shahroz Tariq, Sangyup Lee, and Simon S Woo. 2020. A convolutional LSTM based residual network for deepfake video detection. *arXiv preprint arXiv:2009.07480* (2020).

[57] Shahzeb Tariq, Jorge Loy-Benitez, KiJeon Nam, Gahye Lee, MinJeong Kim, Duck-Shin Park, and ChangKyoo Yoo. 2021. Transfer learning driven sequential forecasting and ventilation control of PM2. 5 associated health risk levels in underground public facilities. *Journal of Hazardous Materials* 406 (2021), 124753.

[58] Shahroz Tariq and Simon S. Woo. 2022. Evaluating the Robustness of Time Series Anomaly and Intrusion Detection Methods against Adversarial Attacks. https://openreview.net/forum?id=C5u6Z9voQ1

[59] Florian Tramer and Dan Boneh. 2019. Adversarial training and robustness for multiple perturbations. *arXiv preprint arXiv:1904.13000* (2019).

[60] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1633–1645. https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf

[61] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. 2018. Deep learning for smart manufacturing: Methods and applications. *Journal of manufacturing systems* 48 (2018), 144–156.

[62] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*. IEEE, 1578–1585.

[63] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems* 30, 9 (2019), 2805–2824.

[64] Jeong-Han Yun, Jonguk Kim, Won-Seok Hwang, Young Geun Kim, Simon S Woo, and Byung-Gil Min. 2022. Residual size is not enough for anomaly detection: improving detection performance using residual similarity in multivariate time series. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 87–96.

[65] Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. 2019. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1409–1416.

[66] Haichao Zhang and Jianyu Wang. 2019. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019), 1831–1841.

[67] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. 2020. Multivariate time-series anomaly detection via graph attention network. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 841–850.

[68] Zibin Zheng, Yatao Yang, Xiangdong Niu, Hong-Ning Dai, and Yuren Zhou. 2017. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics* 14, 4 (2017), 1606–1615.

[69] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.