CSCAD: Correlation Structure-based Collective Anomaly Detection in Complex System

Huiling Qin[1,2,3], Xianyuan Zhan[2,3], Yu Zheng[1,2,3]
[1]Xidian University, Xi'an, China
[2]JD Intelligent Cities Research, Beijing, China
[3]JD Intelligent Cities Business Unit, JD Digits, Beijing, China
orekinana@gmail.com;zhanxianyuan@jd.com;msyuzheng@outlook.com

ABSTRACT

Detecting anomalies in large complex systems is a critical and challenging task. The difficulties arise from several aspects. First, collecting ground truth labels or prior knowledge for anomalies is hard in real-world systems, which often leads to limited or no anomaly labels in the dataset. Second, anomalies in large systems usually occur in a collective manner due to the underlying dependency structure among devices or sensors. Lastly, real-time anomaly detection for high-dimensional data requires efficient algorithms that are capable of handling different types of data (i.e. continuous and discrete). We propose a correlation structure-based collective anomaly detection (CSCAD) model for high-dimensional anomaly detection problems in large systems, which is also generalizable to semi-supervised or supervised settings. Our framework utilize graph convolutional network combining a variational autoencoder to jointly exploit the feature space correlation and reconstruction deficiency of samples to perform anomaly detection. We propose an extended mutual information (EMI) metric to mine the internal correlation structure among different data features, which enhances the data reconstruction capability of CSCAD. The reconstruction loss and latent standard deviation vector of a sample obtained from the reconstruction network can be perceived as two natural anomalous degree measures. An anomaly discriminating network can then be trained using low anomalous degree samples as positive samples, and high anomalous degree samples as negative samples. Experimental results on five public datasets demonstrate that our approach consistently outperforms all the competing baselines.

CCS CONCEPTS

• **Information systems** → **Embedded systems**; *Data mining*; • **Computing methodologies** → Neural networks.

KEYWORDS

Anomaly Detection, Neural Networks, Deep Learning, Generative Model, Complex System, Correlation Structure

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '21, August 15, 2021, Virtual
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Huiling Qin[1,2,3], Xianyuan Zhan[2,3], Yu Zheng[1,2,3]. 2021. CSCAD: Correlation Structure-based Collective Anomaly Detection in Complex System. In *KDD '21: OUTLIER DETECTION AND DESCRIPTION WORKSHOP, August 15, 2021, Virtual.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

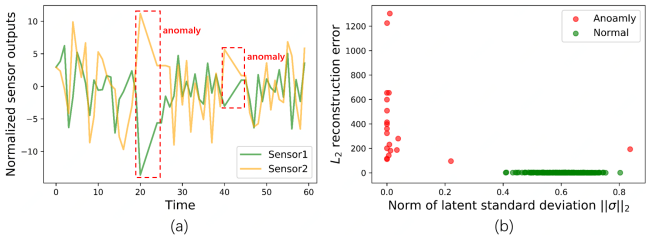


Figure 1: Sensory data and anomaly status from an human activity dataset: (a) Normalized readings of two sensors over time. (b) Reconstruction error and the norm of latent standard deviation vector of samples evaluated using a VAE.

Efficiently detecting faults or anomalies is vital to safeguard the daily operation of many large complex systems. Typical applications including manufacturing system fault detection, network intrusion detection, abnormal bioactivities discovery and so on [1, 21]. Although extensive anomaly detection studies have been conducted in the past, developing a robust anomaly detection technique for large systems with complex internal structures is still a challenging task to solve. The challenges arise from several aspects.

First, unlike other data-driven tasks, collecting ground truth labels or prior knowledge for anomalies in complex systems is much harder. Anomalies are typically rare in the population, leading to highly unbalanced datasets. In many real-world systems, anomaly labels are collected through manual inspection, leading to very limited or even no anomaly labels being collected. In extreme cases, it is even impossible to know the basic information about the anomalies in the system, which forbids the use of many anomaly detection algorithms with a predetermined threshold (e.g. anomalous degree threshold or the proportion of anomalies in the data) [7, 27, 31]. Due to these facts, unsupervised anomaly detection algorithms [25] that can adaptively learn the discriminating boundaries of normal and anomaly samples are particularly desirable.

Second, most sensors in large complex systems do not work independently. Many anomaly detection scenarios involve complicated internal dependency structures among sensors or devices.

For example, sensors monitoring different parts of a manufacturing system typically have a complex interdependent relationship. Fault from a sensor can propagate to other dependent sensors, causing cascading failures in parts or the entire system. This is also referred to as *collective anomaly* [4, 5, 28, 29]. Under such circumstances, an anomaly might not be that anomalous by checking a single data feature but could be identified as an anomaly when checking multiple related features simultaneously. Considering the feature space correlation to analyze the change in the underlying correlation structural pattern facilitates better modeling the operation characteristics of the system, leading to more accurate and robust results. This is particularly important in detecting early-stage anomalies when the anomalous pattern is not significant.

Lastly, sensory data collected in large systems typically involve data with different properties (e.g. static and time-series) or types, such as continuous (e.g. temperature reading from a sensor) and discrete (e.g. on/off status of a switch) data. A unified approach is needed to handle different types of data and generalizable to time-series settings.

In this paper, we propose a highly adaptive correlation structure-based anomaly detection (CSCAD) framework to address the aforementioned challenges. We consider the behavior of anomalies from three aspects: (1) break down of the original feature space correlation patterns (e.g. correlated features exhibit different correlation pattern); (2) anomalous samples are harder to reconstruct compared with normal samples, as they possess different statistical patterns; and (3) anomalous samples have larger variance when explaining using a model trained with the complete dataset. We illustrate the above anomalous behavior using a simple human activity dataset [12]. In Figure 1(a), anomalies occur when two positively correlated sensors suddenly become uncorrelated; and in Figure 1(b), the reconstruction error and latent standard deviation evaluated using a variational autoencoder (VAE) are observed to be highly discriminative features for the anomaly detection task.

Based on above observations, we use a graph convolutional variational autoencoder as the reconstruction network to capture the feature space correlation and reconstruction deficiency of samples, and further perform detection using a feedforward neural network as the discriminating network. Graph convolutional layers are used in the reconstruction network to mine the hidden structure in the feature space of data, which is constructed using a new correlation evaluation metric, the extended mutual information (EMI). It is capable of handling different types of data (continuous and categorical) and generalizable to time-series data. The feature space correlation provides extra information about the internal correlation structure among different features, thus enhances the data reconstruction capability of the model. Two anomalous degree measures, the reconstruction loss (measures reconstruction difficulty) and the latent standard deviation (measures internal variance) of samples can be obtained from the trained reconstruction network. The discriminating network is then trained using low anomalous degree samples as positive samples, and high anomalous degree samples or samples with anomaly labels as negative samples. This design allows the training process of the framework can be performed in an unsupervised manner without the need of introducing any predetermined anomaly threshold. Moreover, the proposed framework is highly flexible, which is also applicable to semi-supervised or supervised settings when parts or complete anomaly labels are available, and can be easily extended to model high-dimensional time-series data. Experiments on five public datasets show that the proposed framework, as well as its time-series extension, consistently outperform all the baselines in terms of precision, recall and F_1 score, which demonstrate the effectiveness and extensibility of our framework.

2 RELATED WORK

There is extensive literature related to anomaly detection. Our focus is mainly restricted to high-dimensional collective anomaly detection problem with few or no anomaly labels (see [2, 8, 23] for a wider scope survey). In this scope, most recent works tend to utilize a reconstruction-based approach or evaluate an anomaly score to solve the anomaly detection problem.

Reconstruction-based anomaly detection. The main assumption of reconstruction-based approach is that anomalous samples have different patterns compared with normal samples, hence are more difficult to be reconstructed. The anomalous degree of a data sample can be reflected by the loss or distance between the original and reconstructed data samples generated by some statistical or neural models. Classic methods include principal component analysis (PCA) with explicit linear projections [13], and the improved version, robust principal component analysis (RPCA) [7], [15], which makes PCA more robust by enforcing sparse structures. Inspired by RPCA and deep learning techniques, Zhou and Paffenroth [30] introduce a robust deep autoencoder (RDA) model which split the input data into reconstructed part and noise to improve the robustness of standard deep autoencoders. Other methods [10], [24], [19] detect anomalies using generative adversarial networks (GANs) [14]. The idea is that anomalies differ from the distribution of normal samples, which makes it difficult to generate similar non-anomalous samples through GANs. The weakness of this reconstruction-based approach lies in the need of a reliable data reconstruction model. A low-capacity model is unable to capture the complex patterns in the data, resulting in model-induced reconstruction deficiency.

Anoamly score-based anomaly detection. Another class of methods detect anomaly based on some anomaly representation score calculated by conventional or deep learning models. Data samples are considered as anomalous when they are located in a low density/probability region of the training data. Traditional methods include kernel density estimation [22] and Robust-KDE [17]. These methods are known to be problematic dealing with high-dimensional data due to the curse of dimensionality. To mitigate this problem, some studies [7] adopt a two-step process, which first compresses the high-dimensional data into low-dimensional latent representations using deep autoencoders, and then applies a density or distance-based model on the low-dimensional space to detect the anomaly. Some recent works combine these two steps and directly learning an anomaly score that perform density/probability estimation. Zhai et al. [27] utilize an energy-based autoencoder model to map each data sample to an energy score. Zong et al. [31] use a compression network combined with the Gaussian mixture model to estimate the density of each sample and further detect the anomaly. Other methods [3], [26], [16] use a variant of VAE

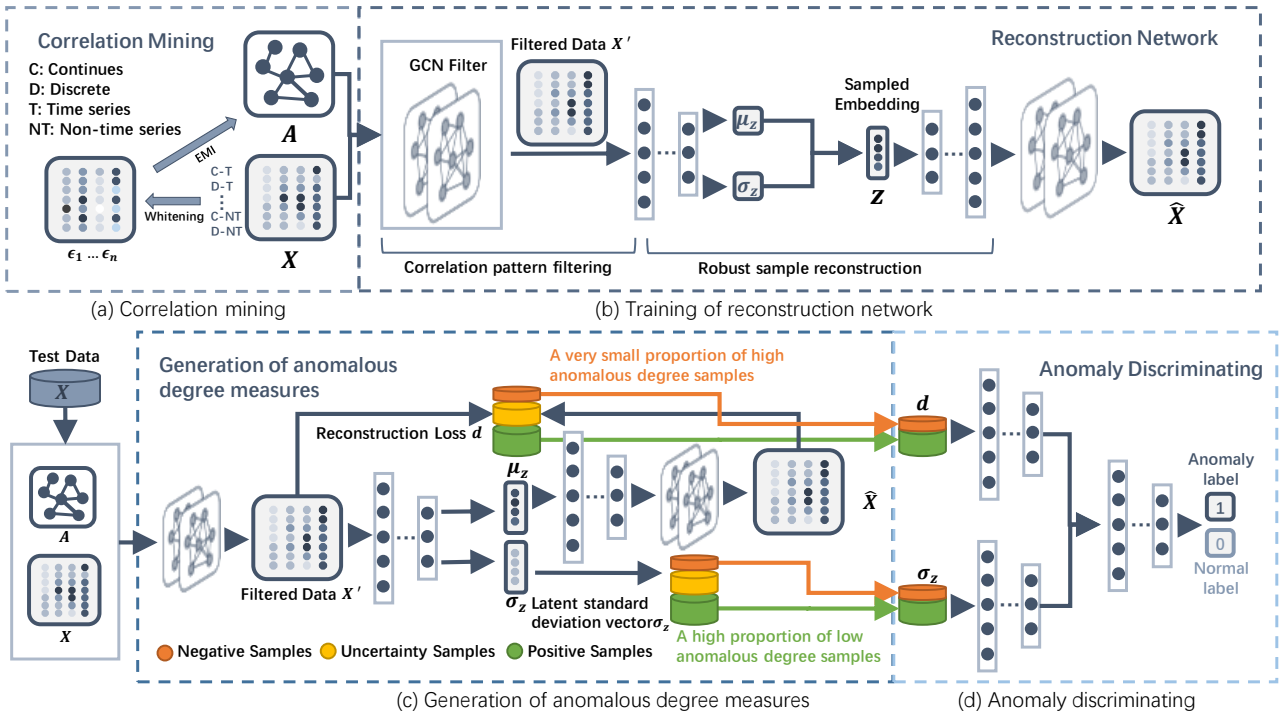


Figure 2: Proposed framework for collective anomaly detection

to learn the probability density of each sample and identify the low-probability samples as anomalies. A major drawback of many anomaly score-based methods is the need to specify some thresholds for the anomaly score or proportion of anomalies in the data to discriminate normal or anomalous samples, which involves additional assumptions on the data.

In this work, we combine the merits of both aforementioned approaches while overcomes their drawbacks. We exploit the internal correlation structure in data features and use a high-capacity model for data reconstruction. We further train a small discriminative network to perform detection using the data reconstruction loss and latent standard deviation. The proposed framework is highly adaptive, which can be used with limited or no anomaly labels, and easily generalizable to high-dimensional time-series data.

3 OVERALL FRAMEWORK

We consider the problem of collective anomaly detection in large complex systems. Let $\mathcal{X}$ be the space of all samples in a system, with the ith sample denoted as $X_i = (x_{i1}, \cdots, x_{im})^T$, with x_{ij} as the feature or sensor of sample X_i in dimension j. We judge a sample X_i to be normal or anomalous based on three criteria: (1) break down of usual feature space correlation pattern; (2) difficulty in reconstructing using a model trained using the complete dataset, and (3) latent space embedding exhibits high variance.

Our framework adopts the following design to incorporate the three criteria. First, the hidden structure of the feature space is mined using a new extended mutual information (EMI) metric, which constructs a correlation structure graph among features. A reconstruction network modeled as graph convolutional variational autoencoder is then trained to generate two sample anomalous degree measures (reconstruction loss d and latent standard deviation σ_z) by capturing the feature space correlation and perform robust sample reconstruction. The two anomaly degree measures are used as the inputs of a discriminative network to predict the final anomalous probability. The training of the discriminative network adopts a self-learning mechanism, which uses low anomalous degree samples as positive samples, and high anomalous degree samples or samples with anomaly labels as negative samples. An illustration of the proposed framework can be found in Figure 2.

3.1 Correlation Structure Mining

Large complex systems typically have large amount of sensors involving both continuous (e.g. temperature readings from a thermometer) and discrete (on/off status of a switch) data; some may be time varying and others are static. Most existing correlation measures, such as Pearson correlation coefficient, cross-variance and mutual information (MI), only work for data with identical types. In order to mine the correlation structure across multiple types and properties of data features, a generic metric is needed. We proposed the extended mutual information (EMI) metric extending the work of Galka et al. [13] to measure the correlation between features with different characteristics.

EMI is based on information theory and directly works with probabilistic distributions of variables, which can be applied to

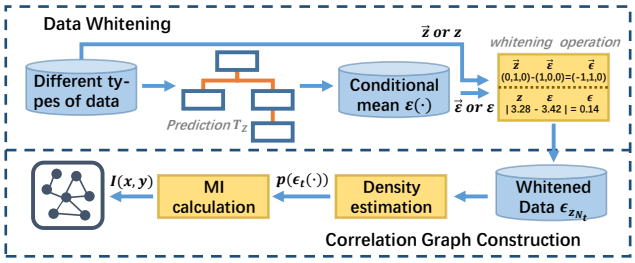


Figure 3: Illustration of the correlation mining process

both continuous and categorical data, and generalizable to time-series data. A llustration of the correlation mining process using EMI see Fig. 3

Time-series MI. The MI between two random variables $\mathbf{x}$ and $\mathbf{y}$ is defined as the Kullback-Leibler (KL) divergence between the joint distribution $p(\mathbf{x}, \mathbf{y})$ and the product of their marginals $p(\mathbf{x})$ and $p(\mathbf{y})$: $I(\mathbf{x}, \mathbf{y}) = KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y}))$. Galka et al. [13] proposed an improved version of MI, which generalizes the MI for time-series data. Let x_t and y_t be two time-series random variables at time step t, the generalized MI can be written as:

$$I(\mathbf{x}, \mathbf{y}) = \log p((x_1, y_1), \cdots, (x_{N_t}, y_{N_t})) - \log(p(x_1, \cdots, x_{N_t})p(y_1, \cdots, y_{N_t})) \quad (1)$$

As the high-dimensional joint distributions $p((x_1, y_1), \cdots, (x_{N_t}, y_{N_t}))$, $p(x_1, \cdots, x_{N_t})$ and $p(y_1, \cdots, y_{N_t})$ are complicated to estimate, Galka et al. [13] suggests removing the self-temporal correlation of the data by a whitening operation which utilizing a predictive model $\mathcal{E}(\cdot)$ to approximate the conditional means of variables, such that the whitened data satisfy the I.I.D. property (see [13] for details). Let $\epsilon_{(\cdot)}$ represent the whitened form of data, which is the residual between the original and predictive conditional mean of data. The time-series MI can thus be evaluated as:

$$I(\mathbf{x}, \mathbf{y}) = \log(p(\epsilon_{x_1}, \epsilon_{y_1}) \cdots p(\epsilon_{x_{N_t}}, \epsilon_{y_{N_t}})) - \log(p(\epsilon_{x_1}) \cdots p(\epsilon_{x_{N_t}})p(\epsilon_{y_1}) \cdots p(\epsilon_{y_{N_t}})) \quad (2)$$

Extended MI (EMI). With the nice statistical property of the whitened residuals, the key to compute the mutual information between two variables with different types lies in finding a unified predictive model to approximate the conditional means of the variable $\mathcal{E}(\cdot)$, as well as defining the residual for discrete variables.

Based on time-series MI, we propose an extended version of MI to cover different types of data. There are two key ingredients of EMI: (1) a unified predictive model to perform temporal whitening for different types of data; and (2) a new scheme to properly define the "residual" of whitened discrete variables.

For both continuous and discrete variables, we can perceive the conditional mean $\mathcal{E}(\cdot)$ as a temporal predictor $x_t = f(x_{t-1}, x_{t-2}, \cdots)$ that predict the data value at time t given historical time-series information. In order to unify the predictive model for different types of data, we calculate the conditional mean using a regression tree (for continuous data) or a decision tree (for discrete data) given historical time-series data from a sliding time window. And for the conditional mean of the joint of two variables $\mathcal{E}((x_t, y_t)|(x_{t-1}, y_{t-1}), (x_{t-2}, y_{t-2}), \cdots)$, we can still use the regression tree regardless the types of x_t and y_t, as the regression tree can handle both the continues and discrete data. The reasons that we choose the tree-based model to estimate the conditional means are because: (1) tree-based model can deal with different type of variable; (2) it can capture both the linear and non-linear relationship in the temporal data; (3) it is effective and simple, which help to reduce the computational cost during correlation evaluation for high-dimensional data.

Another challenge of evaluating mutual information between two variables across different data types is to properly define the residual for discrete variables, as simple subtraction is no longer well-defined for discrete variables. To address this issue while guarantee the linear transformation for the variable, we encode the discrete variables as one-hot encoded vector (m-dimensional vector if has m discrete states), and calculate vector subtraction results $\vec{y}_i - \vec{y}_j$. There are a total of $m^2 - m + 1$ possible outcomes of $\vec{y}_i - \vec{y}_j$ (each element in the m-dimensional vector $\vec{y}_i - \vec{y}_j$ takes value of -1, 0, 1), therefore, we encode different outcomes of the vector subtraction as a new $m^2 - m + 1$ dimensional one-hot encoded vector $\vec{z}$ and use it as the residual for discrete variables. The residual is a lossless recording of the transition between different discrete states of the variables and is well-defined.

Finally, we can unify the whitening operation as follows:

$$\epsilon_{z_{N_t}} = z_{N_t} \ominus T_Z(z_{N_t}|z_{N_t} \cdots z_1) \quad (3)$$

where $T_Z(\cdot)$ is the trained tree-based predictive model for variable Z; $\ominus$ is the subtraction operation when z is continuous and is the previously defined discrete residual operation if z is discrete.

After obtain the whitened residuals for the time-series data of the system, the marginal distribution of $p(\epsilon_t(x|x))$ and $p(\epsilon_t(x|(x, y)))$ can be estimated either by non-parametric statistics or methods such as kernel density estimation(KDE) (for continuous residuals) or simply count the proportion of each discrete residual state (for discrete residuals). For the joint distribution $p(\epsilon_t(x|x, y), \epsilon_t(y|x, y))$ involving both continuous and discrete variables, it is calculate as:

$$p(\epsilon_t(x|x, y), \epsilon_t(y|x, y)) = p(\epsilon_t(x|x, y)|\epsilon_t(y|x, y))p(\epsilon_t(y|x, y)) \quad (4)$$

where x is a continuous variable, and y is a discrete variable. The conditional probability of $p(\epsilon_t(x|x, y)|\epsilon_t(y|x, y))$ can be estimated by a set of non-parametric statistics conditioned on each discrete state of y. Finally, we can use Eq.1, 2 to obtain the MI-based correlation for two time-series variables with arbitrary types.

With the extended data whitening scheme, the MI of data with different types can be evaluated in a similar procedure as discussed in [13]. The proposed EMI allows evaluating correlation for both continuous and categorical data, and generalizable to time-series data. We use EMI to evaluate the pairwise correlation between different features of data, from which a correlation structure graph can be constructed, with nodes represent data features, and edges represent the pairwise correlation of the two features.

3.2 Reconstruction Network

A notable characteristic of anomalies is that they are harder to reconstruct compared with normal samples [31]. Inspired by this observation, we design a graph convolutional variational autoencoder as the reconstruction network for sample reconstruction, which is a combination of graph convolutional neural network (GCN) and variational autoencoder (VAE) [18] (see Figure 2 (b)). We use GCN layers to capture the correlation structure of features mined using the EMI. Hence when data features exhibit distinct correlation patterns, the outputs filtered by the GCN layers will facilitate enlarging the reconstruction error evaluated by the VAE. Furthermore, we use the VAE rather than a conventional autoencoder. As VAE can learn a latent mean and a latent standard deviation vector, which can be perceived as the denoised mean and the internal variance of the sample embedding, thus extracts more detailed information about the anomalous behavior of a sample.

We use the GCN layer proposed in Defferrard et al. [11] to model the correlation structure in the feature space. Consider a spectral convolution on graph defined as the multiplication of a graph signal $X \in \mathbb{R}^{m \times c}$ with a filter g_θ parameterized by θ in the Fourier domain:

$$g_\theta \star_{\mathcal{G}} X = g_\theta(L)X = g_\theta(U\Lambda U^T)X = U g_\theta(\Lambda)U^T X \tag{5}$$

where $U \in \mathbb{R}^{m \times m}$ is the matrix of eigenvectors, and $\Lambda \in \mathbb{R}^{m \times m}$ is the diagonal matrix of eigenvalues of the normalized graph Laplacian $L = I_N - D^{\frac{1}{2}}\Lambda D^{\frac{1}{2}} = U\Lambda U^T$, where I_N is the identity matrix, $D \in \mathbb{R}^{m \times m}$ is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$ and A is the adjacency matrix of the correlation structure graph. Directly perform convolution operation using above formulation is computationally expensive. Defferrard et al. [11] used the Kth-order polynomial in the Laplacian to enable fast evaluation, which restricts the GCN to capture the information at maximum K step away from the central node (K-localized). The corresponding graph convolutional operator and the lth layer output of GCN $H^{(l)} \in \mathbb{R}^{m \times d}$ given activation function $f(\cdot)$ are given as:

$$g_\theta(L)X = \sum_{k=0}^{K-1} \theta_k L^k X, \quad H^{(l+1)} = f\Big(\sum_{k=0}^{K-1} \theta_k L^k H^{(l)}\Big) \tag{6}$$

The subsequent VAE component of the reconstruction network takes the outputs from the GCN layer to perform a robust reconstruction. VAE forces its encoder to generate a latent vector z that roughly follow a Gaussian distribution, which is parameterized by a latent mean vector μ_z, and a latent standard deviation vector σ_z. μ_z and σ_z can be perceived as embeddings of the denoised mean and the internal uncertainty level of a sample, which provides important information about the anomalous degree of the sample. Given a dataset of N samples, the reconstruction network can be trained in a similar manner as VAE using variational inference [18] by minimizing the following objective function:

$$J(\theta, \phi) = \frac{1}{N}\sum_{i=1}^{N} L(X_i, \hat{X}_i) + \frac{\lambda}{N}\sum_{i=1}^{N} D_{KL}(q_\phi(z_i|X_i)||p_\theta(z_i)) \tag{7}$$

where $L(X_i, \hat{X}_i) = ||X_i - \hat{X}_i||_2^2$ is the loss between original sample X_i and reconstructed sample $\hat{X}_i$. The KL divergence $D_{KL}(q_\phi(z|X)||p_\theta(z))$ forces the approximated posterior distribution $q_\phi(z|X)$ to be similar to the prior distribution $p_\theta(z)$, which improves the robustness of reconstruction when separating the internal noise from input data.

The proposed reconstruction network can be easily extended to model high-dimensional time-series data. This is done by introducing the recurrent neural network layer (e.g. LSTM, GRU) after GCN layer to capture temporal information across different time steps.

3.3 Discriminating Network and Detection Strategy

Generation of anoamlous degree measures. Utilizing the trained reconstruction network, we derive two anomalous degree measures to support anomaly detection:

- The reconstruction loss $\vec{d_i} = \left[(x_{i1} - \hat{x}_{i1})^2, \cdots, (x_{im} - \hat{x}_{im})^2\right]^T$ is the element-wise reconstruction loss calculated by original sample X_i and reconstructed sample $\hat{X}_i$. We generate $\hat{X}_i$ by applying the decoder of GCVAE on the latent mean vector μ_z rather than the sampled latent vector z. As $\hat{X}_i$ generated from μ_z is deterministic and can be perceived as a denoised version of X_i, which helps reconstruction loss d more accurately reflect the reconstruction difficulty of a sample. As a result, a sample with a larger d indicates a higher anomalous degree.
- The latent standard deviation σ_z represents the internal variance level of a sample. It is another anomalous degree measure which reflects the uncertainty of the input data. A sample with high uncertainty or significantly deviate from the regular patterns of the data is likely to be an anomaly.

These two anomalous degree measures provide more detailed anomalous information of a sample to allow users to "interpret" the results. The reconstruction loss and latent standard deviation are given as a vector indicating the reconstruction difficulty and uncertainty level of each feature (e.g. each sensor), which can also help locate the most problematic sensors in the system.

Anomaly discriminating. We construct a discriminating network using reconstruction loss d and latent standard deviation σ_z as inputs (see Figure 2 (c), (d)). In the discriminating network, d and σ_z are first fed into two sets of fully connected layers. Their outputs are then concatenated and fed into two fully connected layers to output the final anomalous probability of a sample.

Training the discriminating network requires anomaly labels in the data, which can be difficult to acquire. We take an alternative approach by utilizing the already obtained anomalous degree measures d and σ_z. We first evaluate d and σ_z of all samples in the training set, and rank the samples according to their norms $(||d||_2, ||\sigma_z||_2)$. Considering that the anomalies are typically rare in the population, and a sample with large d or σ_z might be anomalous. We label 50% (the proportion can be flexibly adjusted according to different real-world datasets) of low anomalous degree samples as positive samples. The negative samples are selected based on a very conservative criterion, that we only select a very small proportion of high anomalous degree sample (e.g. top 2.5%) as negative samples. The discriminating network is trained only using selected positive and negative samples. When parts or full anomaly labels are available, the actual anomalies can be used as negative samples, which enables the proposed framework adaptable to semi-supervised or

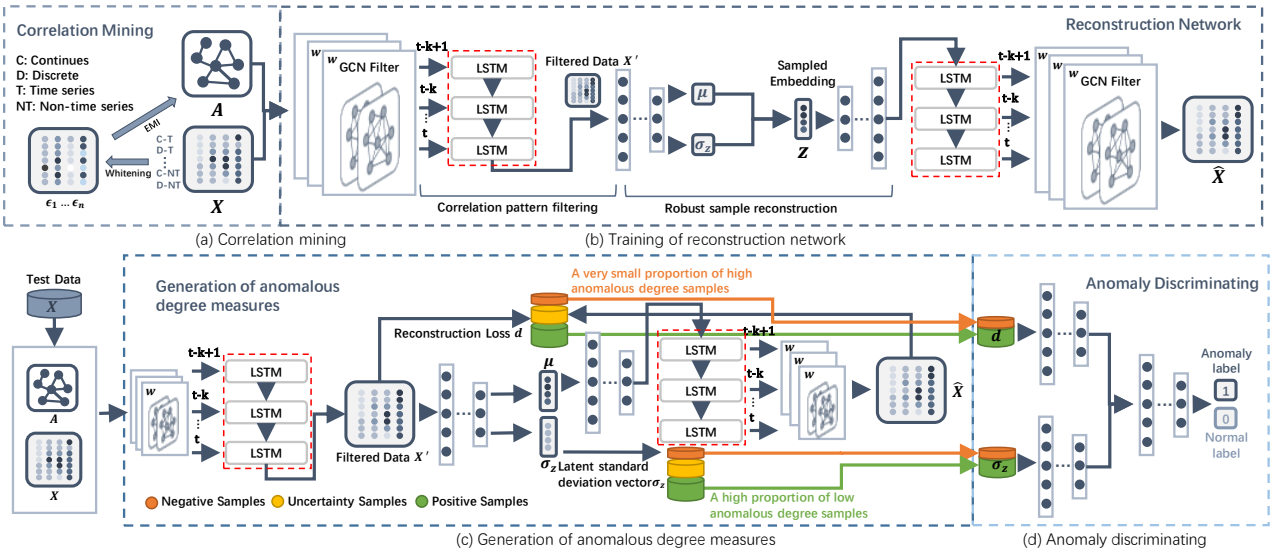


Figure 4: Proposed framework for collective anomaly detection in time-series settings

supervised settings. The proposed strategy does not need to specify a predetermined anomaly threshold that typically used in anomaly score-based methods [3, 16, 26, 27, 31], and capable of learning complex high-dimensional separation boundaries between normal and anomalous samples.

At the detection stage, the reconstruction loss d and latent standard deviation vector σ_z are obtained using the trained reconstruction network. They are fed into the trained discriminating network to evaluate the final anomalous probability of the given sample (see Figure 2 (d) for detailed illustration).

3.4 Time-Series Extension

The proposed CSCAD framework can also be generalized to time-series settings. We focus on the collective anomaly detection problem that finds anomalies at time step $t+1$ based on previous k time steps' data $(t-k+1, \cdots, t)$. Let the sample at time step t be $X_t = (x_{t1}, x_{t2}, \cdots, x_{tm})^T$. We use EMI to construct the feature space correlation structure graph of the time-series data. The reconstruction network is modified to include an LSTM layer to capture the temporal characteristics of the data. The time-series data X_t at each time step is first fed into the GCN layer, and the sequence of the GCN outputs are then modeled using the LSTM layer. The VAE component takes the final output of the LSTM layer and generates the latent mean and standard deviation of samples to support anomaly detection. The discriminating network and detection strategy remain the same as in the previous section. A illustration of the time-series extension framework see Fig. 4

4 EXPERIMENTAL RESULT

In this section, we use public benchmark datasets to demonstrate the effectiveness of the hidden structure-based collective anomaly detection (CSCAD) framework against several competing baselines.

4.1 Dataset

We use five public datasets from UCI machine learning repository [12] in our experiments, including KDDCUP, Thyroid, MoCap, UJIIndoorLoc for the static version of the framework, and Heterogeneity for the time-series extension of the framework. Detailed statistics of each dataset is presented in Table 1.

Table 1: Statistics of the four public datasets

	Dimensions	Instances	Anomaly ratio
KDDCUP99	121	494,021	20%
Thyroid	13	19,016	10%
MoCap	36	15,963	8.5%
UJIIndoorLoc	522	10,000	7.5%
Heterogeneity	6	54,000	20%

- **KDDCUP**. The KDDCUP99 dataset consists of 34 continuous and 7 categorical features. The categorical features are encoded using one-hot encoding, resulting in a dataset of 121 dimensions. As only 20% of samples are labeled as "normal" and 80% are labeled as "attack", therefore, we treat "normal" samples as anomalies in this task.
- **Thyroid**. The Thyroid dataset is a disease dataset in which "negative" samples are treated as normal and others are anomalies. 7 continuous and 2 categorical relevant features are used in this

Table 2: Experiment results of our framework and the baseline methods for four static datasets

Methods	KDDCUP			Thyroid			MoCap			UJIIndoorLoc		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
DAGMM	0.830	0.839	0.835	0.273	0.257	0.265	0.431	**1**	0.603	0.720	0.709	0.714
AE-LOF	0.456	0.461	0.458	0.269	0.507	0.351	0.122	0.283	0.171	0.067	0.264	0.107
LOF	-	-	-	0.248	0.467	0.324	0.123	0.286	0.172	0.134	0.528	0.214
IF	0.449	0.454	0.451	0.280	0.526	0.365	0	0	0	0.508	0.998	0.673
OC-SVM	-	-	-	0.331	0.340	0.335	0.001	0.002	0.001	-	-	-
VAE-DN	0.852	0.954	0.9	0.268	0.345	0.302	0.256	0.522	0.344	0.112	0.701	0.211
CSCAD(no σ)	0.734	**1**	0.846	0.440	0.586	0.503	**0.557**	0.684	0.614	0.059	**1**	0.112
CSCAD(7.5%)	0.857	0.994	0.920	**0.496**	**0.795**	**0.611**	0.486	**0.991**	0.652	**0.924**	0.856	**0.889**
CSCAD(5%)	0.862	0.921	0.890	0.480	0.662	0.557	0.496	0.956	**0.653**	0.916	0.858	0.886
CSCAD(2.5%)	**0.881**	0.996	**0.934**	0.495	0.553	0.522	0.507	0.792	0.618	0.706	0.894	0.789

task. Again, the categorical features are encoded using one-hot encoding.

- **MoCap**. The MoCap dataset consists of 36 continuous features. Representing the hand postures by 12 measuring points in the three-dimensional space. Here, we consider the hand posture "5" as anomalies and "1" as normal samples.
- **UJIIndoorLoc**. The UJIIndoorLoc is a multi-building indoor localization dataset containing 522 WiFi fingerprints as attributes. We consider "BUILDING ID" labeled "2" as normal samples and "0" as anomaly.
- **Heterogeneity**. The Heterogeneity is a human activity recognition dataset which contains records of gyro sensor and accelerometer from smartphones and smartwatches, including 6 continuous time-series features. We consider "Stand" activities as normal samples, and "Stair Up" activities as anomalies.

4.2 Baseline

We consider several state-of-the-art deep learning approaches and a few widely used methods as baselines:

- **DAGMM**. The deep autoencoding Gaussian mixture model (DAGMM) [31] uses latent low-dimensional representation and a Gaussian mixture model to perform density estimation of data, and further predict anomalies using a predetermined threshold.
- **LOF**. The local outlier factor (LOF) [6] finds anomalous data points by measuring the local deviation of a given data point with respect to its neighbors.
- **AE-LOF**. This method adopts a two-step approach. It first trains a deep autoencoder to generate a latent representation of a sample, then uses LOF to detect the anomaly.
- **IF**. Isolation Forest (IF) [20] is a decision tree-based ensemble method. It partitions the sample points by randomly selecting a split value between the maximum and minimum values of a feature. Anomalies are more easily separated under random splits. This baseline is also used in the experiment on time-series data.
- **OC-SVM**. One-class support vector machine (OC-SVM) [9] is a kernel-based method which learns a decision function between normal and anomalous samples.
- **VAE-DN**. This is a variant of the proposed CSCAD framework, which uses VAE as the reconstruction network without considering the feature space correlation.
- **CSCAD(no σ)**. This variant of the proposed framework only uses the reconstruction loss d in the discriminating network, without considering the latent standard deviation vector σ_z.
- **CSCAD(p%)**. These models are the proposed framework of which the discriminating network is trained by selecting different percentages (i.e. 2.5%, 5%, 7.5%, smaller than the actual anomaly ratio) of high anomalous degree samples as negative samples.
- **Baselines for time-series data**. We consider two baselines in addition to IF, which are AE(LSTM)-IF and VAE+DN(LSTM). AE(LSTM)-IF introduces LSTM layer in the autoencoder, and uses IF to perform detection on the encoded representation. VAE+DN(LSTM) is a variant of the time-series extension of CSCAD, which removes GCN layers in the reconstruction network.

4.3 Model Configurations

The detailed model configurations used on each individual datasets are summarized below. Let m be the dimension of data features.

- **Reconstruction Network**. For all datasets, the reconstruction network runs with GCN(m, k=2, ReLU)-FC(m, 60, ReLU)-FC(60, 30, ReLU)-FC(30, 10, ReLU)-2 FC(10, 5, none)-reparameterize-FC(5, 10, ReLU)-FC(10, 30, ReLU)-FC(30, 60, ReLU)-FC(60, m, ReLU)-GCN(m, k=2, none).
- **Discriminating Network**. The discriminating network contains 2 compression sub-networks that encode d and σ_z. The outputs of the two compression sub-networks are concatenated and fed into a sub-discriminating network to output the final anomalous probability of a sample. For all datasets, the compression sub-network that encodes the d runs with BN(m, elu)-FC(m, m/2, elu)-FC(m/2, m/4, elu)-FC(m/4, 10, elu)-FC(10, 5, elu); and the compression sub-network for σ_z runs with BN(m, elu)-FC(m, m/2, elu)-FC(m/2, 2, elu). Finally, the sub-discriminating network runs with BN(7, elu)-FC(7, 4, elu)-FC(4, 2, softmax).

4.4 Result

Detection accrancy. Table 2 presents the precision, recall and F_1 score of the baseline methods and our proposed framework for different datasets. It is shown that CSCAD demonstrates superior performance over all the baseline methods. Several baseline

methods such as LOF and OC-SVM even failed within an acceptable running time due to the large data size (e.g., KDDCUP, UJIndoorLoc). It is observed that even training by selecting only the most conservative 2.5% high anomalous degree samples as negative samples, the proposed framework (CSCAD(2.5%)) achieves 10.2%, 15.7%, 1.5% and 7.5% improvement on F_1 score on different datasets over the best baseline methods (excluding VAE+DN and other CSCAD variant models). Moreover, comparing the results of CSCAD(2.5%, 5% and 7.5%), it is observed that conservatively selecting a very small proportion (much smaller than the anomaly ratio of the dataset) of high anomalous degree samples as negative samples for training still enables the discriminating network to maintain reasonable anomaly detection accuracy. This is important, as in many real-world scenarios, the actual anomaly ratio in the dataset is not known. It is desired to have a model that works with a very conservative estimate of the anomaly ratio while still achieves reasonable accuracy.

Several interesting observations can be made by analyzing the results in Table 2. It is observed that considering feature space correlation clearly improves anomaly detection accuracy. More specifically, compared with the VAE-DN (without GCN to perform correlation pattern filtering), the proposed framework achieved almost 30-60% improvement in terms of F_1 score on Thyroid, MoCap and UJIIndoorLoc datasets. This shows that considering feature space correlation makes a significant improvement on the detection accuracy when data features have strong internal correlation structure (e.g. interdependency of the physiological features under disease for Thyroid dataset, the correlated relative movement of the measuring points in the hand posture dataset MoCap and the underlying pattern of the relative position of the WiFi fingerprint signals in UJIIndoorLoc dataset.).

Impact of the generated anomalous degree measures. It is found that the CSCAD consistently achieves higher F_1 score compared with the variant method CSCAD(no σ) that does not consider the latent standard deviation vector σ_z. This shows that considering only reconstruction loss may not fully reflect the anomalous behavior of a sample. Incorporating the internal uncertainty level information reflected in σ_z is also important. This observation is also confirmed in Figure. 5, which presents the visualization of detection results with respect to the L2-norms of reconstruction loss d and latent standard deviation vector σ_z. It is observed that there does not exist simple boundary values to separate the normal and anomalous samples by solely inspecting d or σ_z. However, by joint modeling both d and σ_z, we can learn a more robust discriminative boundary to detect anomalies.

Table 3: Experiment results with limited anomaly label available for KDDCUP dataset.

Labeled Rate	Precision	Recall	F_1
0%	0.881	0.996	0.934
0.50%	0.883	0.990	0.934
0.75%	0.883	0.997	0.936
1.00%	0.881	1	0.937

Results under semi-supervised settings. CSCAD can be easily adapted to semi-supervised or supervised settings by using actual anomalous or normal samples in the positive and negative sample set during training the discriminative network. We conduct an experiment on the KDDCUP dataset to test the performance of our framework when limited amount of anomaly labels are known. Table 3 presents the results of CSCAD(2.5%) model when replacing 0% to 1% of the 2.5% high anomalous degree negative samples with the actual anomalies. It is observed that with the introduction of higher amount of labeled anomalies, the F_1 score increases from 0.934 to 0.937. Even without labeled data, the framework trained under unsupervised setting already achieved reasonable accuracy (0.934), with only 0.003 decrease in F_1 score compared with the case introducing 1% of actual anomalies during training. This demonstrates the robustness of the proposed framework.

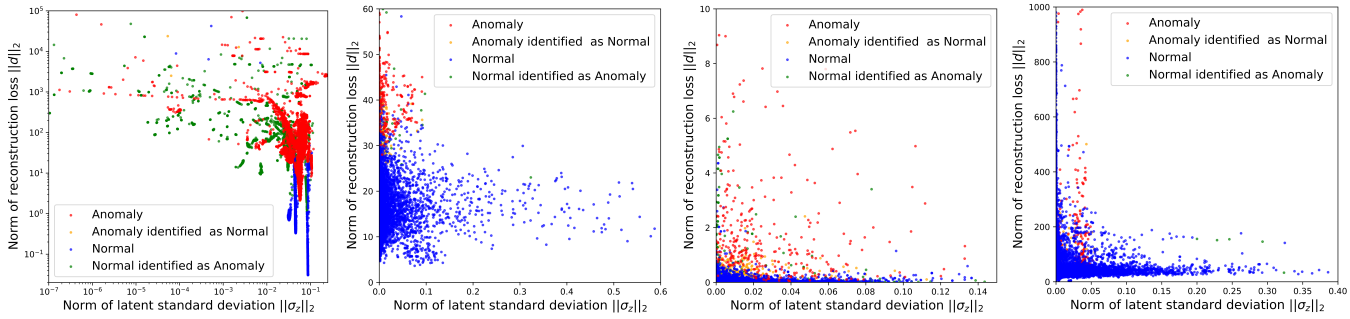
Time-series extension. To demonstrate the extensibility of our framework, we also present the experiment results on the time-series dataset Heterogeneity in Table 4. We compare our framework (CSCAD(LSTM)) with a classic method (IF) and two deep learning-based approaches (AE(LSTM)-IF and VAE+DN(LSTM)). Our framework outperforms all the baseline methods in terms of precision, recall and F_1 score. Compared with IF and AE(LSTM)-IF, the proposed framework achieved 8-9% improvement on F_1 score. Again, we observe the feature space correlation helps to improve detection accuracy, which accounts for 1.4% increase on F_1 score compared with the VAE-DN (LSTM) model. These results show the effectiveness and extensibility of our framework.

Table 4: Experiment results of our framework and the baseline methods for Heterogeneity dataset

Methods	Heterogeneity		
	Precision	Recall	F_1
IF	0.830	**1**	0.907
AE(LSTM)-IF	0.825	0.994	0.902
VAE+DN(LSTM)	0.953	**1**	0.976
CSCAD(LSTM)	**0.980**	**1**	**0.990**

5 CONCLUSION

We propose a new adaptive framework (CSCAD) for collective anomaly detection for a large complex system with limited or no anomaly labels. Unlike the state-of-the-art approaches so far, CSCAD jointly considers the correlation structure in the feature space and robust sample reconstruction, which leads to superior performance in high-dimensional collective anomaly detection tasks. In this framework, we propose a new EMI metric which is capable of evaluating the correlation of data with different types (continuous and categorical) and properties (static and time-series). A reconstruction network is developed to perform sample reconstruction and evaluates two natural anomalous degree measures of each sample: the reconstruction loss and the latent standard deviation. These two anomalous degree measures serve as inputs to a discriminating network to perform final anomaly detection, which is trained using high anomalous degree samples as positive samples, and



(a) KDDCUP99 (b) MoCap (c) Thyroid (d) UJIndoorLoc

Figure 5: Visualization of the detection results with respect to norms of reconstruction loss d and latent standard deviation σ_z

low anomalous degree samples or samples with anomaly labels as negative samples. This scheme allows the discriminating network can be trained without a predetermined threshold and adaptable to semi- or supervised settings, which is a perfect fit for many real-world applications. Moreover, CSCAD is very flexible and can be easily generalized to time-series collective anomaly detection tasks. Numerical results on five public datasets show that our framework consistently outperforms the baseline methods, which proves the effectiveness of CSCAD.

REFERENCES

[1] Naoki Abe, Bianca Zadrozny, and John Langford. 2006. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 504–509.

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. 2016. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* 60 (2016), 19–31.

[3] Jinwon An and Sungzoon Cho. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE* 2 (2015), 1–18.

[4] Daniel B Araya, Katarina Grolinger, Hany F ElYamany, Miriam AM Capretz, and G Bitsuamlak. 2016. Collective contextual anomaly detection framework for smart buildings. In *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 511–518.

[5] Loïc Bontemps, James McDermott, Nhien-An Le-Khac, et al. 2016. Collective anomaly detection based on long short-term memory recurrent neural networks. In *International Conference on Future Data and Security Engineering*. Springer, 141–152.

[6] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *ACM sigmod record*, Vol. 29. ACM, 93–104.

[7] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58, 3 (2011), 11.

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.

[9] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. 2001. One-class SVM for learning in image retrieval.. In *ICIP (1)*. Citeseer, 34–37.

[10] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. 2018. Anomaly Detection with Generative Adversarial Networks. (2018).

[11] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.

[12] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[13] Andreas Galka, Tohru Ozaki, Jorge Bosch Bayard, and Okito Yamashita. 2006. Whitening as a tool for estimating mutual information in spatiotemporal data sets. *Journal of statistical physics* 124, 5 (2006), 1275–1315.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[15] Peter J Huber. 2004. *Robust statistics*. Vol. 523. John Wiley & Sons.

[16] Yasuhiro Ikeda, Kengo Tajiri, Yuusuke Nakano, Keishiro Watanabe, and Keisuke Ishibashi. 2018. Estimation of Dimensions Contributing to Detected Anomalies with Variational Autoencoders. *arXiv preprint arXiv:1811.04576* (2018).

[17] JooSeuk Kim and Clayton D Scott. 2012. Robust kernel density estimation. *Journal of Machine Learning Research* 13, Sep (2012), 2529–2565.

[18] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2014).

[19] Dan Li, Dacheng Chen, Jonathan Goh, and See-kiong Ng. 2018. Anomaly detection with generative adversarial networks for multivariate time series. *arXiv preprint arXiv:1809.04758* (2018).

[20] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

[21] Biswanath Mukherjee, L Todd Heberlein, and Karl N Levitt. 1994. Network intrusion detection. *IEEE network* 8, 3 (1994), 26–41.

[22] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33, 3 (1962), 1065–1076.

[23] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing* 99 (2014), 215–249.

[24] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*. Springer, 146–157.

[25] Kai Tian, Shuigeng Zhou, Jianping Fan, and Jihong Guan. 2019. Learning Competitive and Discriminative Reconstructions for Anomaly Detection. *arXiv preprint arXiv:1903.07058* (2019).

[26] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 187–196.

[27] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. 2016. Deep Structured Energy Based Models for Anomaly Detection. In *International Conference on Machine Learning*. 1100–1109.

[28] Huichu Zhang, Yu Zheng, and Yong Yu. 2018. Detecting Urban Anomalies Using Multiple Spatio-Temporal Data Sources. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 54.

[29] Yu Zheng, Huichu Zhang, and Yong Yu. 2015. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2.

[30] Chong Zhou and Randy C Paffenroth. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 665–674.

[31] 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. (2018).