Revisiting VAE for Unsupervised Time Series Anomaly Detection: A Frequency Perspective

Zexin Wang[*]
wangzexin@cnic.cn
Computer Network Information
Center, Chinese Academy of Sciences
Beijing, China

Changhua Pei[†]
chpei@cnic.cn
Computer Network Information
Center, Chinese Academy of Sciences
Beijing, China

Minghua Ma
minghuama@microsoft.com
Microsoft
Beijing, China

Xin Wang
x.wang@stonybrook.edu
Stony Brook University
New York, United States

Zhihan Li
lizhihan03@kuaishou.com
Kuaishou Technology
Beijing, China

Dan Pei
peidan@tsinghua.edu.cn
Tsinghua University
Beijing, China

Saravan Rajmohan
saravan.rajmohan@microsoft.com
Microsoft
Redmond, United States

Dongmei Zhang
dongmeiz@microsoft.com
Microsoft
Beijing, China

Qingwei Lin
qlin@microsoft.com
Microsoft
Beijing, China

Haiming Zhang
hai@cnic.cn
Computer Network Information
Center, Chinese Academy of Sciences
Beijing, China

Jianhui Li
lijh@cnic.cn
Computer Network Information
Center, Chinese Academy of Sciences
Beijing, China

Gaogang Xie
xie@cnic.cn
Computer Network Information
Center, Chinese Academy of Sciences
Beijing, China

ABSTRACT

Time series Anomaly Detection (AD) plays a crucial role for web systems. Various web systems rely on time series data to monitor and identify anomalies in real time, as well as to initiate diagnosis and remediation procedures. Variational Autoencoders (VAEs) have gained popularity in recent decades due to their superior denoising capabilities, which are useful for anomaly detection. However, our study reveals that VAE-based methods face challenges in capturing long-periodic heterogeneous patterns and detailed short-periodic trends simultaneously. To address these challenges, we propose Frequency-enhanced Conditional Variational Autoencoder (FCVAE), a novel unsupervised AD method for univariate time series. To ensure an accurate AD, FCVAE exploits an innovative approach to concurrently integrate both the global and local frequency features into the condition of Conditional Variational Autoencoder (CVAE) to significantly increase the accuracy of reconstructing the normal data. Together with a carefully designed "target attention" mechanism, our approach allows the model to pick the most useful information from the frequency domain for better short-periodic trend construction. Our FCVAE has been evaluated on public datasets and a large-scale cloud system, and the results demonstrate that it outperforms state-of-the-art methods. This confirms the practical applicability of our approach in addressing the limitations of current VAE-based anomaly detection models.

CCS CONCEPTS

• **Mathematics of computing → Time series analysis**; • **Computing methodologies → Anomaly detection**.

KEYWORDS

Univariate time series, Anomaly detection, Conditional variational autoencoder

[*]Also with University of Chinese Academy of Sciences.
[†]Corresponding author.

1 INTRODUCTION

Time series anomaly detection (AD) is ubiquitous for web systems [8, 10, 13, 14, 16, 20, 32, 35, 49, 50]. Numerous web systems, such as online advertising systems, are monitored using a vast array of time series data (*e.g.,* conversion rate) [52]. Deploying time series AD algorithms is essential for timely detecting anomalies and initiating subsequent diagnosis and remediation processes.

Anomalies are rare in real-world time series data [40], making it difficult to label them and train a supervised model for anomaly detection [50]. Instead, unsupervised machine learning techniques are commonly used [5, 7, 15, 17, 26, 33, 50, 55, 58–60]. These techniques can be divided into two categories: *prediction-based* [15, 17, 60] and *construction-based* [5, 26, 33, 50, 55]. Both types aim to identify normal values and compare them to actual values to detect anomalies. Prediction-based methods were originally developed for forecasting future data points, regardless of whether they were normal or anomalous. However, these methods may overfit to anomalous patterns and underperform. On the other hand, Variational AutoEncoders (VAEs) [22], the leading construction-based approach, encode raw time series into a lower-dimensional latent space and then reconstruct them back to their original dimensions. VAEs are well-suited for detecting anomalies, but existing VAE-based anomaly detection models have not yet reached theoretically optimal performance. In this paper, we aim to re-examine the VAE model and improve its effectiveness in anomaly detection.

In order to more effectively demonstrate the challenges associated with VAE-based techniques, we provide an example in Figure 1. The original curve is displayed in the first sub-figure, with anomalies highlighted in red ③. The subsequent four sub-figures represent curves reconstructed by four distinct VAE-methods, including our proposed method (referred to as FCVAE). The reconstruction error is indicated by the green shaded area ⑤. To achieve superior AD performance, the reconstructed result should closely resemble the original curve for normal points, while deviating significantly for anomalous points ③. As evident in the figure, all VAE-based methods successfully disregard the anomalies during reconstruction. However, the reconstruction results for some normal points, particularly those marked by a blue rectangle ① and ellipse ④, are not satisfactory. This substantially impacts the overall performance, leading us to identify three key challenges that we address in the subsequent sections.

Challenge 1: Capturing similar yet heterogeneous periodic patterns. From Figure 1, periodic patterns can be observed in the curves, with one such period emphasized by the red shaded area ②. However, the shapes across different periods vary. As demonstrated by the blue ellipse, existing VAE-based methods (as shown in the second sub-figure) are unable to capture these heterogeneous patterns effectively. This observation naturally leads to the idea of utilizing conditional VAE to map data into distinct Gaussian spaces by considering the timestamp as a condition. Unfortunately, as illustrated in the fourth sub-figure (Time CVAE [26]), the results are unsatisfactory, which we will further discuss below.

Challenge 2: Capturing detailed trends. Reconstructing monotonous patterns (*i.e.,* trends) might appear straightforward at first glance. However, upon a closer examination of the local window (highlighted in a blue rectangle ① in Figure 1 and magnified in Figure 2(a), it becomes evident that existing methods fail to restore detailed patterns within this time frame. In Figure 2(a), the two green lines initially overestimate the ground truth (purple curve) but subsequently underestimate it for the remainder of the window. This is primarily because existing methods aim to minimize the overall reconstruction error without focusing on
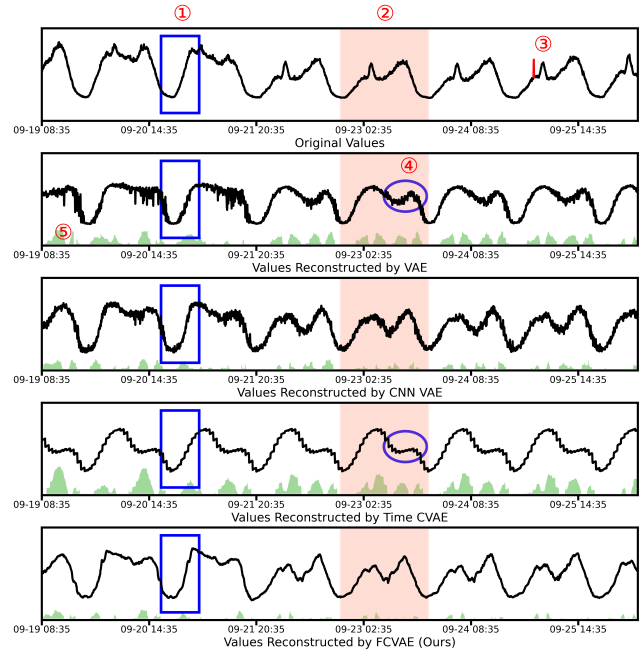

Figure 1: Comparison of four KPI reconstruction methods presented in our paper, highlighting anomalies in red ③. The green shade ⑤ represents the difference between the reconstructed values and the original values, the red shade ② represents a long period, and the blue ellipse ④ indicates peaks and valleys that are not properly reconstructed, the blue rectangle ① will be magnified in Figure 2 for detailed comparison.

(a) Original Values (b) Transformed Spectrogram
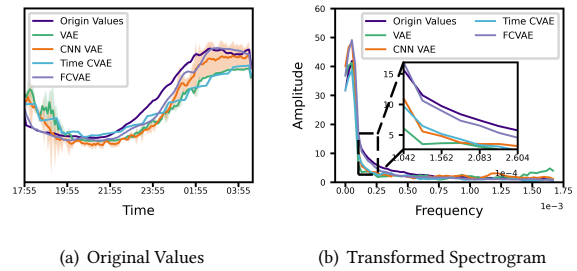
Figure 2: A detailed view of the region enclosed by a blue rectangle ① in Figure 1, where the shaded area represents the value range before applying a sliding window average.

"point-to-point" dependencies, *e.g.,* the precise upward and downward ranges following a specific point. This omission results in fluctuating reconstruction outcomes (as seen in the second sub-figure). Although Convolutional Neural Network (CNN) attempts to model point-to-point dependencies within the window, it still produces coarse-grained fluctuations (visible in the third sub-figure in Figure 1). The reason of unsatisfactory result of CNN-CVAE lies

in Figure 2(b). Upon converting the curves reconstructed by various methods (Figure 2(b)), it becomes evident that the primary cause of these phenomena is the **absence of some frequencies** (smaller amplitude of certain frequencies) in existing methods, hindering the reconstruction of detailed patterns. This observation logically suggests the possibility of employing frequency as the conditional factor in a Conditional Variational Autoencoder (CVAE). Nonetheless, employing frequency as the condition in CVAE presented a new challenge.

Challenge 3: A large number of sub-frequencies make the signal in condition of CVAE noisy and difficult to use. Directly converting the entire window into the frequency-domain results in numerous sub-frequencies, adding noise and obstructing effective VAE-based reconstruction. To address these challenges, we sub-divide the entire window into smaller ones and propose a **target attention** method to select the most useful sub-window frequencies.

In this paper, we introduce a novel unsupervised anomaly detection algorithm, named FCVAE (Frequency-enhanced Conditional Variational AutoEncoder). Different from current VAE-based anomaly detection methods, FCVAE innovatively incorporates both global and local frequency information to guide the encoding-decoding procedures, that both heterogeneous periodic and detailed trend patterns can be effectively captured. This in turn enables more accurate anomaly detection. Our paper's contributions can be summarized as follows:

- Our analysis of the widely-used VAE model for anomaly detection reveals that existing VAE-based models fail to capture both heterogeneous periodic patterns and detailed trend patterns. We attribute this failure to the missing of some frequency-domain information, which current methods fail to reconstruct.
- Our study systematically improves the long-standing VAE by focusing on frequency. Our proposed FCVAE makes the VAE-based approach the state-of-the-art in anomaly detection once more. This is significant because VAE-based methods can inherently handle mixed anomaly-normal training data, while prediction-based methods cannot.
- Evaluations demonstrate that our FCVAE substantially surpasses state-of-the-art methods (∼40% on public datasets and 10% in **a real-world web system** in terms of F1 score). Comprehensive ablation studies provide an in-depth analysis of the model, revealing the reasons behind its superior performance.

The replication package for this paper, including all our data, source code, and documentation, is publicly available online at **https://github.com/CSTCloudOps/FCVAE**.

2 PRELIMINARIES

2.1 Problem Statement

Given a UTS $\mathbf{x} = [x_0, x_1, x_2, \cdots, x_t]$ and label series $\mathbf{L} = [l_0, l_1, l_2, \cdots, l_t]$, where $x_i \in \mathbb{R}$, $l_i \in \{0, 1\}$, and $t \in \mathbb{N}$. $\mathbf{x}$ represents the entire time series data array, while x_i signifies the metric value at time i. $\mathbf{L}$ denotes the label of time series $\mathbf{x}$. We define the UTS anomaly detection task as follows:

Given a UTS $\mathbf{x} = [x_0, x_1, x_2, \cdots, x_t]$, *the objective of UTS anomaly detection is to utilize the data* $[x_0, x_1, \cdots, x_{i-1}]$ *preceding each point* x_i *to predict* l_i.

2.2 VAEs and CVAEs

VAE is composed of an encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder $p_\theta(\mathbf{z}|\mathbf{x})$. VAE can be trained by using the reparameterization trick. SGVB [38] is a commonly used training method for VAE because of its simplicity and effectiveness. It maximizes the evidence lower bound (ELBO) to simultaneously train the reconstruction and generation capabilities of VAE.

DONUT [50] proposed the modified ELBO (M-ELBO) to weaken the impact of abnormal and missing data in the window on the reconstruction. M-ELBO is defined in (1), α_w is defined as an indicator, where $\alpha_w = 1$ indicates x_w being not anomalous or missing, and $\alpha_w = 0$ otherwise. β is defined as $(\sum_{w=1}^{W} \alpha_w)/W$.

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \Big[\sum_{w=1}^{W} \alpha_w \log p_\theta(x_w|\mathbf{z}) + \beta \log p_\theta(\mathbf{z}|\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x}) \Big] \quad (1)$$

The overall structure of CVAE [43] is similar to VAE, and it combines conditional generative models with VAE to achieve stronger control over the generated data. The training objective of CVAE is defined as (2), where $\mathbf{c}$ is the condition, similar to that of VAE. FCVAE which will be elaborated on later extends the CVAE framework by incorporating frequency information.

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})} [\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c}) + \log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})] \quad (2)$$

3 METHODOLOGY

3.1 Framework Overview
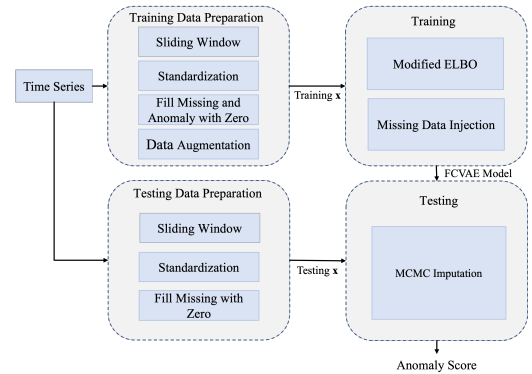
Figure 3: Overall Framework.

The proposed algorithm for anomaly detection is illustrated in Figure 3 and comprises three main components: data preprocessing, training, and testing.

3.2 Data Preprocessing

Data preprocessing encompasses standardization, filling missing and anomaly points, and the newly introduced method of **data augmentation**. The efficacy of data standardization and filling missing and anomaly points has been substantiated in prior studies
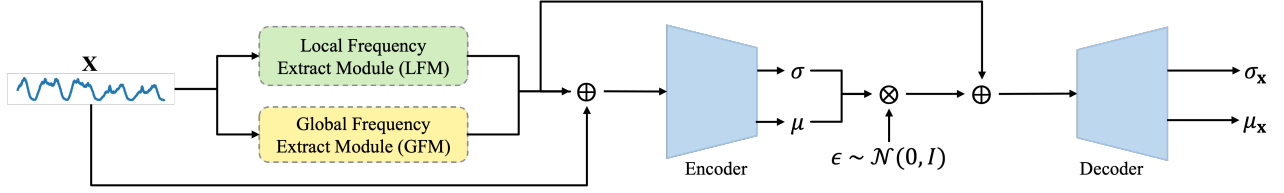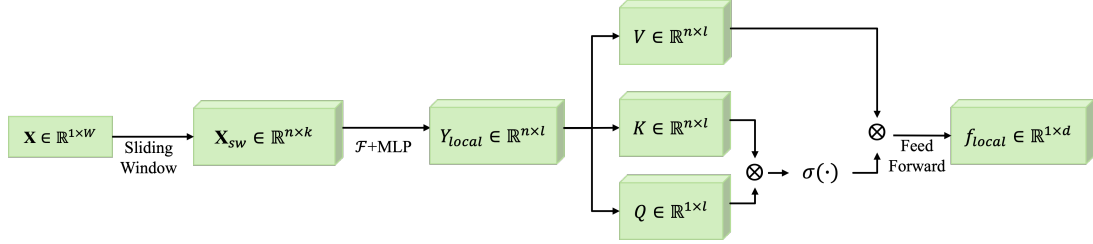
Figure 4: FCVAE Model Architecture.

Figure 5: Architecure of LFM.

[26, 29, 50]. Therefore, we directly incorporate these techniques into our approach.

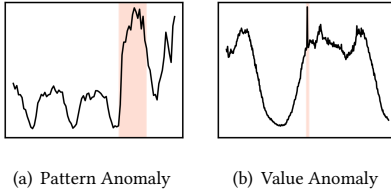(a) Pattern Anomaly   (b) Value Anomaly

Figure 6: Examples of the two most frequent anomalies, where the red shaded area denotes the abnormal segments.

Previous data augmentation methods [25, 47, 54] often added normal samples, such as variations of data from the time domain or frequency domain. However, for our method, we train the model by incorporating all the time series from the dataset together, which provides sufficient pattern diversity. Furthermore, FCVAE has the ability to extract pattern information due to the addition of frequency information, so it can handle new patterns well. Nonetheless, even with the introduction of frequency information, anomalies are often challenging to be effectively addressed. For the model to learn how to handle anomalies, we primarily focus on abnormal data augmentation. In time series data, anomalies are mostly manifested as pattern mutations or value mutations (shown in Figure 6), so our data augmentation mainly targets to these two aspects. The augmentation on the pattern mutation is generated by combining two windows from different curves, with the junction acting as the anomaly. Value mutation refers to changing some points in the window to randomly assigned abnormal values. With the augmented anomaly data, M-ELBO in CVAE, which will be introduced in detail later, can perform well even in an unsupervised setting without true labels.

3.3 Network Architecture

The proposed FCVAE model is illustrated in Figure 4. It comprises three main components: encoder, decoder, and a condition extraction block that includes a global frequency information extraction module (GFM) and a local frequency information extraction module (LFM). Equation (3) illustrates how our model works.

$$\mu, \sigma = \text{Encoder}(\mathbf{x}, \text{LFM}(\mathbf{x}), \text{GFM}(\mathbf{x}))$$
$$\mathbf{z} = \text{Sample}(\mu, \sigma) \tag{3}$$
$$\mu_{\mathbf{x}}, \sigma_{\mathbf{x}} = \text{Decoder}(\mathbf{z}, \text{LFM}(\mathbf{x}), \text{GFM}(\mathbf{x}))$$
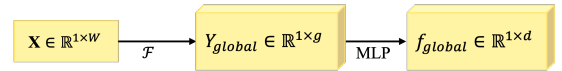
Figure 7: Architecure of GFM.

3.3.1 GFM. The GFM module (Figure 7) extracts the global frequency information using the FFT transformation ($\mathcal{F}$). However, not all frequency information is useful. The frequencies resulted from the noise and anomalies in the time series data appear as long tails in the frequency domain. Therefore, we employ a linear layer after the FFT to filter out the useful frequency information that can represent the current window pattern. Moreover, we incorporate a dropout layer following Fedformer [61] to enhance the model's ability to learn the missing frequency information.

The $f_{global} \in \mathbb{R}^{1 \times d}$ is calculated as (4), where d is the embedding dimension of the global frequency information and $\mathcal{F}$ means FFT.

$$f_{global} = \text{Dropout}(\text{Dense}(\mathcal{F}(\mathbf{x}))) \tag{4}$$

3.3.2 LFM. The attention mechanism [46] has been widely adopted in time series data processing due to its ability to dynamically process dependencies between different time steps and focus on important ones. Target attention, which is developed based on attention,

is widely used in the field of recommendation [4]. Specifically, target attention can weigh the features of the target domain, leading to more accurate domain adaptation.

The GFM module extracts the frequency information from the entire window, proving to be effective in reconstructing the data within the whole window. However, we use a window to detect whether the last point is abnormal, which poses a challenge because the GFM module does not provide sufficient attention to the last point. This can result in a situation where the reconstruction is satisfactory for part of the window but not for another part, especially when changes in system services lead to the concept drift in the time series data. Even in the absence of concept drift, GFM cannot capture local changes as it extracts the average frequency information from the entire window; hence, the reconstruction of the last key point may be unsatisfactory. Nonetheless, as previously mentioned, target attention can effectively address this issue, as it captures the frequency information of the entire window while paying a greater attention to the latest time point. Therefore, we propose the LFM that incorporates the target attention.

As depicted in Figure 5, the LFM module operates by sliding the entire window $\mathbf{x}$ to obtain several small windows $\mathbf{x}_{sw}$. Subsequently, FFT and frequency information extraction are applied to each small window. The most recent small window is used as the query Q because it contains the last point that we want to detect. The remaining small windows are utilized as keys K and values V for target attention. Finally, a linear layer is employed to facilitate the model in learning to extract the most important and useful part of the local frequency information, and dropout is also applied to enhance the model's ability to reconstruct some of the local frequency information like GFM.

$$
\begin{aligned}
\mathbf{x}_{sw} &= \mathrm{SlidingWindow}(\mathbf{x}) \\
Q &= \mathrm{Select}(\mathrm{Dense}(\mathcal{F}(\mathbf{x}_{sw}))) \\
K, V &= \mathrm{Dense}(\mathcal{F}(\mathbf{x}_{sw})) \\
f_{local} &= \mathrm{Dropout}(\mathrm{FeedFawrd}((\sigma(Q \cdot K^{\top}) \cdot V))
\end{aligned}
\tag{5}
$$

The calculation of $f_{local} \in \mathbb{R}^{1 \times d}$ in LFM is given by (5), where d is the embedding dimension of the local frequency information, which is the same as that of GFM. Here, $\mathbf{x}_{sw} \in \mathbb{R}^{n \times k}$ represents a group of small windows extracted from the original window, where k is the dimension of the small windows and n is the number of small windows. The Select function is employed to select the latest window as the query Q and the Dense function means dense neural network. The softmax function σ is used to calculate the attention weights for the small windows.

3.4 Training and Testing

The training process of FCVAE incorporates three key technologies: CVAE-based modified evidence lower bound (CM-ELBO), missing data injection, as well as the newly proposed **masking the last point**. As shown in (6), **CM-ELBO** is obtained by applying M-ELBO [50] to CVAE. Missing data injection [29, 50] is a commonly used technique in VAE that we directly apply. We observed that an anomalous point in time series data manifests as an outlier value in the time domain. However, when the data are transformed into the frequency domain, all frequency information is shifted, leading to a challenge. The impact of this issue will be amplified when the last point is abnormal, as we specifically aim to detect the last point

given the whole window. While we use the frequency enhancement method and frequency selection to mitigate this problem to some extent, we mask the last point as zero during the extraction of the frequency condition to address this issue further.

$$
\mathcal{L} = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x},\mathbf{c})} \Big[\sum_{w=1}^{W} \alpha_w \log p_{\theta}(x_w|\mathbf{z},\mathbf{c}) + \beta \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x},\mathbf{c}) \Big]
\tag{6}
$$

While testing, FCVAE adopts the Markov Chain Monte Carlo (MCMC)-based missing imputation algorithm proposed in [38] and applied in [29] to mitigate the impact of missing data. Since our goal is to detect the last point of a window, the last point is set to missing for MCMC to obtain a normal value. This also allows for a better adaptation to the last point mask mentioned earlier. FCVAE further utilizes reconstruction probabilities as anomaly scores, which are defined in equation (7).

$$
\mathrm{AnomalyScore} = -\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x},\mathbf{c})} \big[\log p_{\theta}(\mathbf{x}|\mathbf{z},\mathbf{c}) \big]
\tag{7}
$$

4 EXPERIMENTS

4.1 Experiment Settings

4.1.1 Datasets. To evaluate the effectiveness of our proposed algorithm, we conducted experiments on four datasets. **Yahoo** [2] is an open data set for anomaly detection released by Yahoo lab. **KPI** [27] KPI is collected from five large Internet companies (Sougo, eBay, Baidu, Tencent, and Ali). **WSD** [1] Web service dataset (WSD) contains real-world KPIs collected from three top-tier Internet companies, Baidu, Sogou, and eBay, providing large-scale Web services. **NAB** [24] The Numenta Anomaly Benchmark (NAB) is an open dataset created by Numenta company for evaluating the performance of time series anomaly detection algorithms.

4.1.2 Baseline Methods. To benchmark our model FCVAE against existing methods, we chose the following approaches for evaluation: SPOT [42], SRCNN [37], TFAD [54], DONUT [50], Informer [60], Anomaly-Transformer [51], AnoTransfer [55], VQRAE [21]. SPOT represents a traditional statistical method rooted in extreme value theory. SRCNN and TFAD are supervised methods relying on high-quality labels. Donut, VQRAE, and AnoTransfer are unsupervised reconstruction-based methods utilizing VAE for normal value reconstruction. Informer is an unsupervised prediction-based method that endeavors to predict normal values using an attention mechanism. Anomaly-Transformer is an unsupervised anomaly detection method leveraging the transformer architecture.

4.1.3 Evaluation Metrics. In practical applications, operators tend to be less concerned with point-wise anomaly detection, i.e., whether each individual point is classified as anomalous or not, and focus more on detecting continuous anomalous segments in time series data. Moreover, due to the substantial impact of anomalous segments, operators aim to identify such segments as early as possible. To address these requirements, we adopt two metrics, best F1 and delay F1, which are based on the works of DONUT [50] and SRCNN [37], respectively.

Best F1 is obtained by traversing all possible thresholds for anomaly scores, and subsequently applying a point adjustment strategy to the prediction in order to compute the F1 score. Delay F1 is similar to best F1 but employs a delay point adjustment strategy to

Table 1: Performance on test data. P means precison, R means recall, F1 means best F1 and F1* means delay F1.

Method	Yahoo						KPI						WSD						NAB					
	P	R	F1	P	R	F1*	P	R	F1	P	R	F1*	P	R	F1	P	R	F1*	P	R	F1	P	R	F1*
SPOT[42]	0.572	0.328	0.417	0.572	0.328	0.417	0.966	0.221	0.360	0.911	0.077	0.143	0.947	0.315	0.472	0.887	0.137	0.237	0.992	0.713	0.829	0.992	0.713	0.829
SRCNN[37]	0.268	0.236	0.251	0.219	0.181	0.198	0.673	0.944	0.786	0.617	0.753	0.678	0.093	0.903	0.170	0.028	0.361	0.053	0.825	0.832	0.828	0.460	0.766	0.575
DONUT[50]	0.381	0.150	0.215	0.381	0.150	0.215	0.378	0.569	0.454	0.328	0.407	0.364	0.263	0.195	0.224	0.199	0.131	0.158	0.933	0.937	0.935	0.821	0.774	0.797
VQRAE[21]	0.706	0.399	0.510	0.691	0.381	0.492	0.202	0.418	0.272	0.167	0.117	0.137	0.518	0.233	0.312	0.241	0.066	0.103	0.990	0.882	0.933	0.806	1.000	0.893
Anotransfer[55]	0.902	0.413	0.567	0.575	0.437	0.496	0.815	0.591	0.685	0.557	0.394	0.461	0.695	0.654	0.674	0.331	0.444	0.379	0.962	0.968	0.965	0.837	0.908	0.871
Informer[60]	0.747	0.671	0.707	0.731	0.619	0.671	0.927	0.910	**_0.918_**	0.801	0.845	**_0.822_**	0.532	0.583	0.557	0.402	0.385	0.393	0.971	0.974	**_0.973_**	0.878	0.907	0.892
TFAD[54]	0.884	0.739	**_0.805_**	0.883	0.734	**_0.802_**	0.684	0.834	0.752	0.650	0.714	0.680	0.541	0.750	0.628	0.431	0.482	**_0.455_**	0.749	0.719	0.734	0.265	0.233	0.248
Anomaly-Transformer[51]	0.588	0.179	0.274	0.054	0.020	0.029	0.930	0.814	0.868	0.622	0.240	0.346	0.861	0.630	**_0.728_**	0.144	0.129	0.137	0.944	1.000	0.971	0.891	0.932	**_0.911_**
FCVAE	0.897	0.821	**0.857**	0.897	0.792	**0.842**	0.930	0.924	**0.927**	0.906	0.772	**0.835**	0.786	0.881	**0.831**	0.705	0.571	**0.631**	0.953	1.000	**0.976**	0.925	0.909	**0.917**

Figure 8: Illustration of the adjustment strategy.

transform the prediction. The adjustment strategies are illustrated in Figure 8, with a delay set to 1 as an example. The detector misses the second anomalous segment because it takes two-time intervals to detect this segment, exceeding the maximum delay threshold we established. We configure the delay for all datasets to be 7, except for Yahoo, where it is set to 3, and NAB, where it is set to 150. This is because the anomaly segments in Yahoo are very short, while in NAB, they are typically much longer, often spanning several hundred data points.

4.1.4 Implementation Details. To guarantee the widespread applicability, all the experiments described below were conducted under entirely unsupervised conditions, without employing any actual labels (all labels are set to zero). For consistency across all methods, we trained a single model for all curves within a dataset. Regarding hyperparameters, we conducted a grid search to identify the most effective parameters for different datasets. Additionally, we later evaluated the sensitivity of these parameters to ensure robust performance.

4.2 Overall Performance

The performance of FCVAE and baseline methods across the four datasets is depicted in Table 1. Our method surpasses all baselines on four datasets regarding best F1 by 6.45%, 0.98%, 14.14% and 0.31%. In terms of delay F1, our method outperforms all baselines on four datasets by 4.98%, 1.58%, 38.68% and 0.65%.

The performance of various baseline methods on the datasets exhibits considerable variation. For instance, SPOT [42] does not excel on most datasets, as it erroneously treats extreme values as anomalies, whereas anomalies are not always manifested as such. SRCNN [37] is a reasonably proficient classifier, yet its performance falls short compared to most other models. This underscores the fact that implicitly extracting abnormal features is challenging. Informer [60] outperforms most other baselines across different datasets, as many anomalies exhibit notable value jumps, and prediction-based methods can effectively manage this situation. However, it

struggles with anomalies induced by frequency changes. Anomaly-Transformer [51] attains commendable results on most datasets in terms of best F1 but demonstrates a low delay F1. It detects anomalies based on the relationships with nearby points, and only when the anomalous point is relatively central within the window can it easily capture the correlation. Conversely, TFAD [54] achieves favorable results on various datasets but exhibits a certain delay in detection.

Moreover, our method surpasses DONUT [50] and VQRAE [21] in terms of reconstruction-based methods. Although VQRAE [21] introduces numerous modifications to the VAE, employing RNN to capture temporal relationships, our method still outperforms it. This finding implies that for UTS anomaly detection, it is imperative to incorporate only key information while avoiding overloading the model with superfluous data.

4.3 Different Types of Conditions in CVAE

We conduct experiments under identical settings to evaluate different types of conditions. The chosen conditions encompass information potentially useful for time series anomaly detection within the scope of our understanding, including timestamps [55], time domain information, and frequency domain information. To ensure consistency, we apply the same operation on the time domain information as we do on the frequency domain information.

As illustrated in Figure 9(a), the performance of employing the frequency information as a condition surpasses that using the timestamp or time domain information. This can be readily comprehended, as timestamps carry limited information and typically require one-hot encoding, resulting in sparse data representation. Time domain information is already incorporated in VAE, and utilizing it as a condition may lead to redundant information without significantly benefiting the reconstruction. Conversely, **frequency information, as a valuable and complementary prior, rendering it a more effective condition for anomaly detection.**

4.4 Frequency VAE and FACVAE

Is CVAE the optimal strategy for harnessing the frequency information in anomaly detection? In this study, we compare FCVAE with an improved frequency-based VAE (FVAE) model, in which the frequency information is integrated into VAE along with the input to reconstruct the original time series. As depicted in Figure 9(b), FCVAE surpasses FVAE. This outcome can be attributed to two primary reasons. Firstly, CVAE, due to its unique architecture that incorporates conditional information, intrinsically outperforms VAE
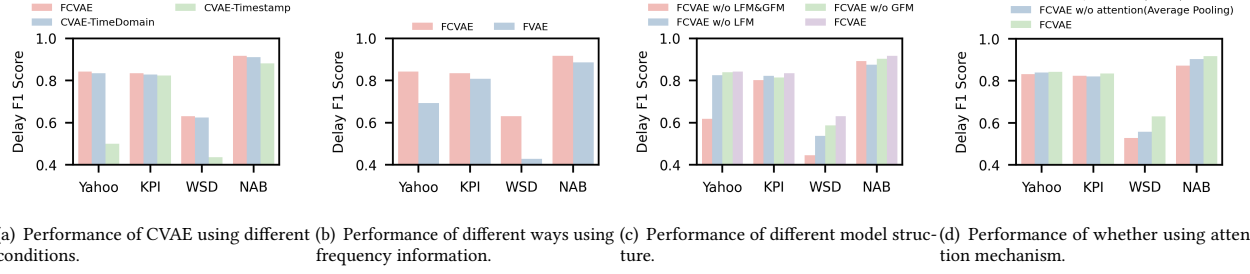
(a) Performance of CVAE using different conditions.

(b) Performance of different ways using frequency information.

(c) Performance of different model structure.

(d) Performance of whether using attention mechanism.

Figure 9: Delay F1 score of different settings.

in numerous applications. Secondly, FVAE does not fully exploit frequency information. Although it incorporates this additional information, it still lacks efficient utilization in practice, particularly in the decoder. Consequently, **the CVAE that incorporates the frequency information as a condition represents the most effective structure known to date.**

4.5 GFM and LFM

We propose GFM and LFM to extract global and local frequency information, respectively. However, do these two modules achieve our intended effects through their designs? Additionally, it is worth noting that GFM and LFM may overlap to some degree. Thus, we would like to determine if combining the two can further enhance the performance.

We conduct experiments and the results are depicted in Figure 9(c). It can be observed that, across the four datasets, employing either LFM or GFM in FCVAE outperforms the VAE model under identical conditions of other settings except for NAB, where the frequent oscillation of data results in inconsistency between the information extracted from GFM and the data value of the current time. For all datasets, when both LFM and GFM modules are utilized concurrently, they synergistically enhance each other, resulting in superior performance. Consequently, **both global and local frequency information play a crucial role in detecting anomalies.**

4.6 Attention Mechanism

It is crucial to discern whether the enhancement in LFM stems from the reduced window size or the attention mechanism. Thus, we perform experiments by excluding the attention operation from LFM while keeping GFM unaltered. Specifically, we utilized frequency information either from the latest small window in LFM (Latest) or from the average pooling of frequency information across all small windows in LFM (Average Pooling).

The findings in Figure 9(d) demonstrate that without attention, it is impossible to attain the original performance of FCVAE since it is not feasible to determine the specific weight of each small window in advance. However, **the attention mechanism effectively addresses this issue by assigning higher weights to more informative windows.**

We present a comprehensive explanation of the attention mechanism in LFM using a case. A specific data segment, denoted by the
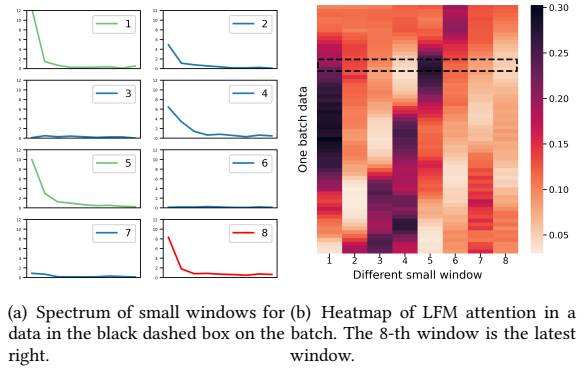

(a) Spectrum of small windows for data in the black dashed box on the right.

(b) Heatmap of LFM attention in a batch. The 8-th window is the latest window.

Figure 10: An example of attention mechanism in LFM.

black dashed box in Figure 10(b), is selected and all small windows produced by LFM's sliding window module are transformed into the frequency domain to obtain their spectra. As illustrated in Figure 10(a), the 5-th (green) and the 8-th (red) windows exhibit the highest similarity, where the 8-th window serves as the query Q for our attention. Upon examining Figure 10(b), it can be observed that the heat value of the 5-th window is the highest, which corresponds with the findings in Figure 10(a).

4.7 Key Techniques in Framework

In this section, we evaluate the effectiveness of our novel data augmentation technique, masking the last point, and the application of CM-ELBO on four distinct datasets. The results are presented in Table 2. Based on the results, it is clear that CM-ELBO plays the most crucial role in most datasets, which aligns with our expectations. This is because it can tolerate abnormal or missing data to a certain extent. Furthermore, masking the last point has a substantial impact on the results, as when an anomaly occurs at the last point of the window, it affects the entire frequency information. Effectively masking this point resolves the issue and improves the detection accuracy. Data augmentation, on the other hand, introduces some artificial anomalies to boost the performance of CM-ELBO, particularly in unsupervised settings.
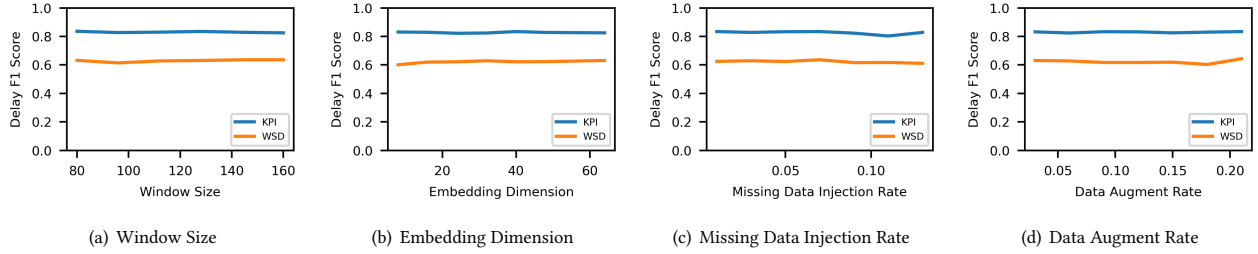
(a) Window Size   (b) Embedding Dimension   (c) Missing Data Injection Rate   (d) Data Augment Rate

Figure 11: Delay F1 score of different settings.

Table 2: Delay F1 of different settings.

Variants	Yahoo	KPI	WSD	NAB
w/o data augment	0.841	0.825	0.626	0.904
w/o mask last point	0.835	0.830	0.534	0.877
w/o CM-ELBO	0.690	0.757	0.435	0.897
FCVAE	**0.842**	**0.835**	**0.631**	**0.917**

4.8 Parameter Sensitivity

The stability of a model to different parameters is an important aspect to consider, and therefore we test the sensitivity of our model parameters on two datasets, KPI and WSD. We examine four aspects: the dimension of the condition, the window size, the proportion of missing data injection, and the proportion of data augmentation. The results, shown in Figure 11, indicate that our model can achieve stable and excellent results under different parameter settings.

5 PRODUCTION IMPACT AND EFFICIENCY

FCVAE has been incorporated as a crucial component in a large-scale cloud system that caters to millions of users globally [6, 18, 19]. The system generates billions of time series data points on a daily basis. The FCVAE detects anomalies in the cloud system, with the primary goal of identifying any potential regressions in the system that may indicate the occurrence of an incident[11, 13, 53, 58].

Table 3: Online performance of FCVAE in production compared to legacy detector. F1 and F1* are defined in Table 1.

Baseline		FCVAE		Improvement		Inference efficiency
F1	F1*	F1	F1*	F1	F1*	[points/second]
0.66	0.63	0.73	0.69	10.9%	11.1%	1195.7

Table 3 presents the online performance improvement achieved by employing FCVAE over a period of one year. The experiments were conducted on a 24GB memory 3090 GPU. The results demonstrate substantial enhancements in both Best F1 and Delay F1 compared to the legacy detector. This underscores the effectiveness and robustness of our proposed method. Furthermore, our model is lightweight and highly efficient, capable of processing over 1000 data points within 1 second. This far exceeds the speed at which the system generates new temporal points.

6 RELATED WORK

Traditional statistical methods [31, 34, 36, 39, 44, 45, 57] are widely used in time series anomaly detection because of their great advantages in time series data processing. For example, SPOT [42] considers that some extreme values are abnormal, therefore, detects them through Extreme Value Theory [9].

Supervised methods [23, 30, 37, 58] mostly learn the features of anomalies and identify them through classifiers based on the features learned. Opprentice [30] efficiently combines the results of many detectors through random forest. Some methods [3, 54] obtain pseudo-labels through data augmentation to enhance the learning ability.

Unsupervised methods are mainly divided into reconstruction-based and prediction-based methods. Reconstruction-based methods [5, 7, 21, 26, 28, 50] learn low-dimensional representations and reconstruct the "normal patterns" of data and detect anomalies according to reconstruction error. Prediction-based methods [17, 41, 60] try to predict the normal values of metrics based on historical data and detect anomalies according to the prediction error. In recent years, transformer-based methods[51] have been widely proposed. Some methods [48, 61] have begun to solve some practical problems from the frequency domain. Moerover, many transfer learning methods have been proposed[12, 26, 55, 56].

7 CONCLUSION

Our paper presents a novel unsupervised method for detecting anomalies in UTS, termed FCVAE. At the model level, we introduce the frequency domain information as a condition to work with CVAE. To capture the frequency information more accurately, we propose utilizing both GFM and LFM to concurrently capture the features from global and local frequency domains, and employing the target attention to more effectively extract local information. At the architecture level, we propose several new technologies. We carry out experiments on four dataset and an online cloud system to evaluate our approach's accuracy, and comprehensive ablation experiments to demonstrate the effectiveness of each module.

8 ACKNOWLEDGMENTS

REFERENCES

[1] [n. d.]. WSD dataset. Available: https://github.com/anotransfer/AnoTransfer-data/.

[2] [n. d.]. Yahoo dataset. Available: https://webscope.sandbox.yahoo.com/.

[3] Chris U Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. 2021. Neural contextual anomaly detection for time series. *arXiv preprint arXiv:2107.07702* (2021).

[4] Qiwei Chen, Changhua Pei, Shanshan Lv, Chao Li, Junfeng Ge, and Wenwu Ou. 2021. End-to-end user behavior retrieval in click-through rateprediction model. *arXiv preprint arXiv:2108.04468* (2021).

[5] Wenxiao Chen, Haowen Xu, Zeyan Li, Dan Pei, Jie Chen, Honglin Qiao, Yang Feng, and Zhaogang Wang. 2019. Unsupervised anomaly detection for intricate kpis via adversarial training of vae. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1891–1899.

[6] Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, et al. 2024. Automatic Root Cause Analysis via Large Language Models for Cloud Incidents. (2024).

[7] Yuhang Chen, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Imdiffusion: Imputed diffusion models for multivariate time series anomaly detection. *VLDB* (2023).

[8] Liang Dai, Tao Lin, Chang Liu, Bo Jiang, Yanwei Liu, Zhen Xu, and Zhi-Li Zhang. 2021. SDFVAE: Static and Dynamic Factorized VAE for Anomaly Detection of Multivariate CDN KPIs. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3076–3086. https://doi.org/10.1145/3442381.3450013

[9] L de Haan and A Ferreira. 2006. Extreme Value Theory: an Introduction Springer Science+ Business Media. *LLC, New York* (2006).

[10] Shohreh Deldari, Daniel V. Smith, Hao Xue, and Flora D. Salim. 2021. Time Series Change Point Detection with Self-Supervised Contrastive Predictive Coding. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3124–3135. https://doi.org/10.1145/3442381.3449903

[11] Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Xiaomin Wu, Meng Zhang, Qingjun Chen, Xin Gao, Xuedong Gao, et al. 2023. TraceDiag: Adaptive, Interpretable, and Efficient Root Cause Analysis on Large-Scale Microservice Systems. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1762–1773.

[12] XiaoYan Duan, NingJiang Chen, and YongSheng Xie. 2019. Intelligent detection of large-scale KPI streams anomaly based on transfer learning. In *Big Data: 7th CCF Conference, BigData 2019, Wuhan, China, September 26–28, 2019, Proceedings 7*. Springer, 366–379.

[13] Vaibhav Ganatra, Anjaly Parayil, Supriyo Ghosh, Yu Kang, Minghua Ma, Chetan Bansal, Suman Nath, and Jonathan Mace. 2023. Detection Is Better Than Cure: A Cloud Incidents Perspective. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1891–1902.

[14] Nikou Günnemann, Stephan Günnemann, and Christos Faloutsos. 2014. Robust Multivariate Autoregression for Anomaly Detection in Dynamic Product Ratings. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 361–372. https://doi.org/10.1145/2566486.2568008

[15] Siteng Huang, Donglin Wang, Xuehan Wu, and Ao Tang. 2019. Dsanet: Dual self-attention network for multivariate time series forecasting. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2129–2132.

[16] Tao Huang, Pengfei Chen, and Ruipeng Li. 2022. A Semi-Supervised VAE Based Active Anomaly Detection Framework in Multivariate Time Series for Online Systems. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1797–1806. https://doi.org/10.1145/3485447.3511984

[17] 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.

[18] Yuxuan Jiang, Chaoyun Zhang, Shilin He, Zhihao Yang, Minghua Ma, Si Qin, Yu Kang, Yingnong Dang, Saravan Rajmohan, Qingwei Lin, et al. 2024. Xpert: Empowering Incident Management with Query Recommendations via Large Language Models. (2024).

[19] Pengxiang Jin, Shenglin Zhang, Minghua Ma, Haozhe Li, Yu Kang, Liqun Li, Yudong Liu, Bo Qiao, Chaoyun Zhang, Pu Zhao, et al. 2023. Assess and Summarize: Improve Outage Understanding with Large Language Models. (2023).

[20] Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodriguez, Chao Zhang, and B Aditya Prakash. 2022. CAMul: Calibrated and Accurate Multi-View Time-Series Forecasting. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA,

3174–3185. https://doi.org/10.1145/3485447.3512037

[21] Tung Kieu, Bin Yang, Chenjuan Guo, Razvan-Gabriel Cirstea, Yan Zhao, Yale Song, and Christian S Jensen. 2022. Anomaly detection in time series with robust variational quasi-recurrent autoencoders. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 1342–1354.

[22] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[23] Nikolay Laptev, Saeed Amizadeh, and Ian Flint. 2015. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1939–1947.

[24] Alexander Lavin and Subutai Ahmad. 2015. Evaluating real-time anomaly detection algorithms–the Numenta anomaly benchmark. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE, 38–44.

[25] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. 2016. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*.

[26] Zeyan Li, Wenxiao Chen, and Dan Pei. 2018. Robust and unsupervised kpi anomaly detection based on conditional variational autoencoder. In *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*. IEEE, 1–9.

[27] Zeyan Li, Nengwen Zhao, Shenglin Zhang, Yongqian Sun, Pengfei Chen, Xidao Wen, Minghua Ma, and Dan Pei. 2022. Constructing Large-Scale Real-World Benchmark Datasets for AIOps. *arXiv preprint arXiv:2208.03938* (2022).

[28] Zhihan Li, Youjian Zhao, Yitong Geng, Zhanxiang Zhao, Hanzhang Wang, Wenxiao Chen, Huai Jiang, Amber Vaidya, Liangfei Su, and Dan Pei. 2022. Situation-Aware Multivariate Time Series Anomaly Detection Through Active Learning and Contrast VAE-Based Models in Large Distributed Systems. *IEEE Journal on Selected Areas in Communications* 40, 9 (2022), 2746–2765.

[29] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. 2021. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 3220–3230.

[30] Dapeng Liu, Youjian Zhao, Haowen Xu, Yongqian Sun, Dan Pei, Jiao Luo, Xiaowei Jing, and Mei Feng. 2015. Opprentice: Towards practical and automatic anomaly detection through machine learning. In *Proceedings of the 2015 internet measurement conference*. 211–224.

[31] Wei Lu and Ali A Ghorbani. 2008. Network anomaly detection based on wavelet analysis. *EURASIP Journal on Advances in Signal processing* 2009 (2008), 1–16.

[32] Xiaofeng Lu, Xiaoyu Zhang, and Pietro Lio. 2023. GAT-DNS: DNS Multivariate Time Series Prediction Model Based on Graph Attention Network. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 127–131. https://doi.org/10.1145/3543873.3587329

[33] Minghua Ma, Shenglin Zhang, Junjie Chen, Jim Xu, Haozhe Li, Yongliang Lin, Xiaohui Nie, Bo Zhou, Yong Wang, and Dan Pei. 2021. {Jump-Starting} Multivariate Time Series Anomaly Detection for Online Service Systems. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*. 413–426.

[34] Ajay Mahimkar, Zihui Ge, Jia Wang, Jennifer Yates, Yin Zhang, Joanne Emmons, Brian Huntley, and Mark Stockert. 2011. Rapid detection of maintenance induced changes in service performance. In *Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies*. 1–12.

[35] Oded Ovadia, Oren Elisha, and Elad Yom-Tov. 2022. Detection of Infectious Disease Outbreaks in Search Engine Time Series Using Non-Specific Syndromic Surveillance with Effect-Size Filtering. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 924–929. https://doi.org/10.1145/3487553.3524672

[36] Faraz Rasheed, Peter Peng, Reda Alhajj, and Jon Rokne. 2009. Fourier transform based spatial outlier mining. In *Intelligent Data Engineering and Automated Learning-IDEAL 2009: 10th International Conference, Burgos, Spain, September 23-26, 2009. Proceedings 10*. Springer, 317–324.

[37] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 3009–3017.

[38] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*. PMLR, 1278–1286.

[39] Bernard Rosner. 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 25, 2 (1983), 165–172.

[40] 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.

[41] Haotian Si, Changhua Pei, Zhihan Li, Yadong Zhao, Jingjing Li, Haiming Zhang, Zulong Diao, Jianhui Li, Gaogang Xie, and Dan Pei. 2023. Beyond Sharing: Conflict-Aware Multivariate Time Series Anomaly Detection. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on*

the Foundations of Software Engineering. 1635–1645.

[42] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. 2017. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1067–1075.

[43] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[44] Owen Vallis, Jordan Hochenbaum, and Arun Kejariwal. 2014. A novel technique for long-term anomaly detection in the cloud. In *6th {USENIX} workshop on hot topics in cloud computing (HotCloud 14)*.

[45] Charles Van Loan. 1992. *Computational frameworks for the fast Fourier transform*. SIAM.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[47] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. 2020. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478* (2020).

[48] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

[49] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. 2012. Review Spam Detection via Time Series Pattern Discovery. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 635–636. https://doi.org/10.1145/2187980.2188164

[50] 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*. 187–196.

[51] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *International Conference on Learning Representations*. https://openreview.net/forum?id=LzQQ89U1qm_

[52] Zhiqiang Xu, Dong Li, Weijie Zhao, Xing Shen, Tianbo Huang, Xiaoyun Li, and Ping Li. 2021. Agile and Accurate CTR Prediction Model Training for Massive-Scale Online Advertising Systems. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD*

'21). Association for Computing Machinery, New York, NY, USA, 2404–2409. https://doi.org/10.1145/3448016.3457236

[53] Zhengran Zeng, Yuqun Zhang, Yong Xu, Minghua Ma, Bo Qiao, Wentao Zou, Qingjun Chen, Meng Zhang, Xu Zhang, Hongyu Zhang, et al. 2023. Traceark: Towards actionable performance anomaly alerting for online service systems. In *IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 258–269.

[54] Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2022. TFAD: A Decomposition Time Series Anomaly Detection Architecture with Time-Frequency Analysis. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2497–2507.

[55] Shenglin Zhang, Zhenyu Zhong, Dongwen Li, Qiliang Fan, Yongqian Sun, Man Zhu, Yuzhi Zhang, Dan Pei, Jiyan Sun, Yinlong Liu, et al. 2022. Efficient kpi anomaly detection through transfer learning for large-scale web services. *IEEE Journal on Selected Areas in Communications* 40, 8 (2022), 2440–2455.

[56] Xu Zhang, Qingwei Lin, Yong Xu, Si Qin, Hongyu Zhang, Bo Qiao, Yingnong Dang, Xinsheng Yang, Qian Cheng, Murali Chintalapati, et al. 2019. Cross-dataset Time Series Anomaly Detection for Cloud Systems.. In *USENIX Annual Technical Conference*. 1063–1076.

[57] Yin Zhang, Zihui Ge, Albert Greenberg, and Matthew Roughan. 2005. Network anomography. In *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. 30–30.

[58] Chenyu Zhao, Minghua Ma, Zhenyu Zhong, Shenglin Zhang, Zhiyuan Tan, Xiao Xiong, LuLu Yu, Jiayi Feng, Yongqian Sun, Yuzhi Zhang, et al. 2023. Robust Multimodal Failure Detection for Microservice Systems. *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (2023).

[59] Nengwen Zhao, Jing Zhu, Yao Wang, Minghua Ma, Wenchi Zhang, Dapeng Liu, Ming Zhang, and Dan Pei. 2019. Automatic and generic periodicity adaptation for kpi anomaly detection. *IEEE Transactions on Network and Service Management* 16, 3 (2019), 1170–1183.

[60] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*. 11106–11115.

[61] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*. PMLR, 27268–27286.