

# GenAD: General unsupervised anomaly detection using multivariate time series for large-scale wireless base stations

Xiaolei Hua,<sup>1</sup> Lin Zhu,<sup>1,✉</sup> Shenglin Zhang,<sup>2</sup> Zeyan Li,<sup>3</sup> Su Wang,<sup>4</sup> Chao Deng,<sup>1</sup> Junlan Feng,<sup>1</sup> Zhao Zhang,<sup>4</sup> and Wei Wu<sup>5</sup>

<sup>1</sup>JIUTIAN Team, China Mobile Research Institute, Beijing, China

<sup>2</sup>College of Software, Nankai University, Tianjin, China

<sup>3</sup>Dept. of Computer Science and Technology, Tsinghua University, Beijing, China

<sup>4</sup>School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China

<sup>5</sup>Dept. of Network, China Mobile Group Guangdong Company Limited, Guangdong, China

✉ Email: zhulinyj@chinamobile.com

The reliability of wireless base stations is essential to guarantee the user experiences in wireless networks, thereby employing the anomaly detection on multivariate time series is indispensable for network operators to monitor the behaviours of large-scale wireless base stations. In this paper, a general unsupervised anomaly detection model is proposed using multivariate time series for large-scale wireless base stations, called GenAD. Firstly, multi-correlation attention and time-series attention are employed to learn the representations of the complex correlations and various temporal patterns of multivariate series. Secondly, a general model on large-scale wireless base stations is pre-trained with self-supervision, which can be easily transferred to a specific station with a small amount of training data. Experimentals show that GenAD boosts F1-score by total 9% on real-world datasets.

**Introduction:** With the large-scale commercialization of 5G, there are millions of wireless base stations (WBS) deployed in mobile networks, serving the connected users. Once an anomaly occurs within WBS, the connected users will suffer from the issues of dropped calls and slow access to the Internet. Thus, monitoring the behaviours of the large-scale WBSes is of vital importance to guarantee the user experiences. The behaviours of a WBS can be characterized by multiple performance metrics, e.g. wireless connection rate, etc. If one of these metrics becomes anomalous, the WBS is likely to suffer from performance degradation. These metrics are continuously collected at a predefined time interval and form the multivariate time series (MTS). Due to anomaly diversity and lack of training labels, the detection of WBSes' anomalous behaviours is converted to the unsupervised anomaly detection using MTS.

Recently, several deep learning-based methods [1–3] have been proposed for unsupervised anomaly detection using MTS. But there are two challenges in applying the methods in wireless networks: 1) The complex correlations and dynamic patterns of MTS. There exists complex inter-dependencies (correlations among time series) and various intra-dependencies (temporal patterns within one time series) for the MTS of each WBS. Additionally, these inter-dependencies and intra-dependencies of each WBS are dynamic, e.g. some WBSes behave in a pattern on weekdays, while the pattern can be dynamically changed during the holidays. These existing anomaly detection methods [4–6], either do not learn the inter-dependencies or intra-dependencies well, or lose the dynamic characteristics. 2) The heterogeneity of large-scale WBSes. Owing to various surrounding environments and different manufacturers, millions of WBSes behave in various patterns. It is infeasible to train one anomaly detection model for all WBSes.

To address the abovementioned challenges, we propose a general unsupervised anomaly detection model using multivariate time series, called GenAD. Firstly, GenAD proposes multi-correlation attention and time-series attention to represent the complex correlations and temporal patterns of MTS simultaneously. Specifically, the attention mechanism is introduced to capture the dynamics among time series and within one time series, and the multi-head and hidden-layers are utilized to capture non-linear, coupling, higher-order correlations, and trend, delay, periodicity of  $N$ -dimensional series. Secondly, GenAD pre-trains a general

Table 1. The 18 types of key performance metrics collected in WBS

No.	Performance metric	No.	Performance metric
1	Wireless connection rate	10	Interference level
2	Wireless drop rate	11	CCE utilization rate in PDCCH
3	Handover success rate	12	eNodeB paging congestion rate
4	E-RAB setup success rate	13	PRB uplink utilization rate
5	RRC connection success rate	14	PRB downlink utilization rate
6	VoLTE handover success rate	15	Uplink traffic (GB)
7	eSRVCC handover success rate	16	Downlink traffic (GB)
8	E-RAB drop rate	17	Average RRC connected users number
9	Teletraffic	18	Max RRC connected users number

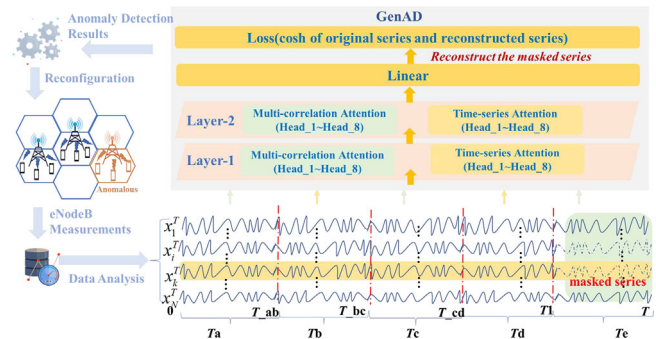


Fig. 1 The overall architecture of GenAD

model on large-scale WBSes with self-supervision. To adapt for each station-specific pattern, we fine-tune the model with only a small quantity of data. Lastly, we propose the two-level dynamic threshold setting strategy to achieve effective anomaly detection.

**Dataset description and problem setup:** In this work, we create the WBS dataset that has been collected from the real-world mobile networks of China Mobile (one of the major China mobile operators). The wireless base stations gather 18 types of key performance metrics with a 15-min sampling interval over a period of 10 days in 2021. For each WBS, these collected data constitute the 18-dimensional MTS, each MTS has 960 data points. More details of the WBS dataset can be found in Table 1. We first randomly select 3000 WBSes (unlabelled data) with 18-dimensional MTS in each station. Except for 3000 WBSes, we randomly pick three areas marked as WBS-Area1, WBS-Area2, WBS-Area3, respectively, and then choose 10 WBSes (labelled data) in each area at random. The MTS of each station is divided into two parts with the same data length: the training set size is 480 points (insufficient to train the existing deep models) and the testing set size is 480 points.

Given the multivariate time series  $X$  of a WBS, it contains  $N$  dimensions and  $T$  time points,

$$X = (x_1^T, x_2^T, \dots, x_N^T) \in \mathbb{R}^{N \times T} \quad (1)$$

Our goal is to determine whether there exist anomaly segments of  $X$  in future time points by learning the complex correlations and various temporal patterns of  $X$ .

**Overall architecture:** Figure 1 depicts the overall architecture of the proposed general unsupervised anomaly detection model using multivariate time series for large-scale wireless base stations. The wireless base stations collect the performance metrics for anomaly detection. The proposed GenAD model implements the anomaly detection by assessing the reconstruction error between the original series and the reconstructed series. It consists of three parts, including multiple hidden-layers with time-series attention and multi-correlation attention in each layer, linear-layer and loss function. Consequently, the anomaly detection results can be utilized to reconfigure the network and ensure the user experiences in mobile networks.

**Random masking for input series:** Due to the limited storage resources of WBS, only a small quantity of data can be leveraged to train the

anomaly detection model for each WBS. To reconstruct time series well on the limited training data, GenAD randomly selects series for masking and reconstruct them with the unmasked series. Considering that GenAD is unaware of which series have been picked for masking or have been employed for reconstruction during model training, it can force the model to automatically learn correlations and temporal patterns of MTS and minimize reconstruction loss. Specifically, we choose 20% of  $N$ -dimensions to be masked with a fixed random time series in  $T_e$ . As demonstrated in Figure 1, the entire time series is divided into five segments.  $T_a$ ,  $T_b$ ,  $T_c$ , and  $T_d$  have the same length that is equivalent to that of  $T_e$ . The remaining 80% original series and 20% masked series in  $T_e$ , along with all series in  $T_a$  to  $T_d$ , input to GenAD.

*Multi-correlation attention and time-series attention:* Existing deep learning-based methods employ convolutional neural networks and auto-encoders to represent correlations of MTS, after the offline model training, the correlations will not change during online inference. However, the MTS of WBSes are not identically distributed. For instance, there may exist strong correlation between  $x_i$  and  $x_j$  of  $X$  in  $(0, T_1)$ , but when distribution of  $x_i$  and  $x_j$  has changed, there will be weak or no correlation between  $x_i$  and  $x_j$  in  $(T_1, T)$ . Hence, GenAD designs multi-correlation attention to capture the dynamic correlation, along with non-linear, coupling, and high-order correlations.

The input of multi-correlation attention is all the MTS in  $T_e$  (including the masked and unmasked series), depicted in Figure 1. The original and masked series in  $T_e$  is denoted as  $X^{T_e}$ , and the reconstructed series by surroundings is  $\tilde{X}^{T_e}$  (only the masked series are reconstructed, not the whole series). Given one of the masked series  $x_i^{T_e}$  in  $X^{T_e}$  as an input, the reconstructed series of  $x_i^{T_e}$  is  $\tilde{x}_i^{T_e}$ ,

$$\begin{aligned}\tilde{x}_i^{T_e} &= \text{Multi correlation attention}[x_i^{T_e}] \\ &= \sum_{j \in [N] \setminus i} \text{softmax} \left( \frac{(Q_i^{T_e})(K_j^{T_e})^\top}{\sqrt{d}} \right) V_j^{T_e}\end{aligned}\quad (2)$$

where  $Q_i^{T_e}$  is the transformation of  $x_i^{T_e}$  in  $(T_1, T)$ ,  $Q_i^{T_e} = W^Q x_i^{T_e}$ , and the transition matrix  $W^Q$  can be learned through model training.  $K_j^{T_e}$  and  $V_j^{T_e}$  are also the transformations of the surrounding series of  $x_i^{T_e}$  in  $(T_1, T)$ .  $\top$  is the transpose of a matrix,  $d$  is the length of  $x_i^{T_e}$ . The softmax function is to calculate the correlations between the real-time dot-product of  $Q_i^{T_e}$  and  $K_j^{T_e}$ . Although the transition matrixes of  $Q_i^{T_e}$  and  $K_j^{T_e}$  keep unchanged after offline training,  $x_i^{T_e}$  and the surrounding series of  $x_i^{T_e}$  still change online, the result of  $Q_i^{T_e}$  and  $K_j^{T_e}$  also remain changing, such that the dynamic correlations between  $x_i^{T_e}$  and all series is captured, and the non-linear and coupling correlations are captured by ReLU activation function and dot-product operation. Furthermore, we introduce Multi-head to strengthen the ability of representing correlations, and stack multiple layers of multi-correlation attention to capture the higher-order correlations.

Existing deep learning-based methods employ gated recurrent unit and long short-term memory to represent temporal patterns of MTS, but they are troublesome to capture the various temporal patterns (including dynamic, trend, delay, and periodicity) of  $N$ -dimensional series simultaneously, especially when  $N$  is large. To this end, GenAD employs time-series attention to represent the various temporal patterns of MTS.

Given  $x_k$  in  $X$  as an input,  $x_k$  in time  $(T_1, T)$  is masked with a fixed random series, denoted as  $x_k^{T_e}$ . The input of time-series attention is  $x_k$  in  $(0, T)$ , including  $x_k^{T_a}$ ,  $x_k^{T_b}$ ,  $x_k^{T_c}$ ,  $x_k^{T_d}$ , and  $x_k^{T_e}$ , as shown in Figure 1. The reconstructed series of  $x_k^{T_e}$  is  $\tilde{x}_k^{T_e}$  and time-series attention is given,

$$\begin{aligned}\tilde{x}_k^{T_e} &= \text{Time series attention}[x_k^{T_e}] \\ &= \sum_{t \in \{T_a, T_b, T_c, T_d\}} \text{softmax} \left( \frac{(Q_k^{T_e})(K_k^t)^\top}{\sqrt{d}} \right) V_k^t\end{aligned}\quad (3)$$

where  $Q_k^{T_e}$  is the transformation of  $x_k^{T_e}$ ,  $K_k^t$ , and  $V_k^t$  represent the transformations of  $x_k^{T_a}$ ,  $x_k^{T_b}$ ,  $x_k^{T_c}$ , or  $x_k^{T_d}$ ;  $\top$  is the transpose of a matrix;  $d$  is the length of  $x_k^{T_e}$ . Equation (3) means that GenAD reconstructs  $x_k^{T_e}$  by learning the various temporal patterns of  $x_k^{T_a}$ ,  $x_k^{T_b}$ ,  $x_k^{T_c}$ , and  $x_k^{T_d}$ .

Ultimately, we adopt the fusion of time-series attention and multi-correlation attention to capture the temporal patterns and correlations of MTS simultaneously. Each layer captures both the temporal patterns of series and the correlations among series. The fusion is achieved by learning the general attention representations that can establish two types of attention meanwhile.

*Pretraining strategy for large-scale WBSes:* To tackle the challenge regarding to heterogeneity of large scale WBSes, we firstly pre-train GenAD with the unlabelled MTS data from 3000 real-world WBSes to learn the general representations. Then, to adapt for the specific patterns of each WBS, we finetune GenAD with only a small quantity of data from each WBS. Concretely, only 480 data points of each WBS are required to finetune the general model, the remaining 480 data points are utilized to test the model performances. Note that the pretraining strategy enables the model to obtain the universal feature representations and thus reduces the underfitting problem on a small dataset.

*Anomaly detection based on two-level dynamic threshold:* As GenAD detects anomalies by evaluating the reconstruction error between the original series and the reconstructed series, we set a two-level dynamic threshold, including the metric-level threshold  $\lambda_m$  and the entity-level threshold  $\lambda_e$ . For the metric-level, if reconstruction error of one time series at time  $t$  is greater than  $\lambda_m$ , the time series is declared as anomalous. For an  $N$ -dimensional entity at time  $t'$ , if there are  $M$  time series ( $M > \lambda_m$ ) that are anomalous, then the entity is declared as anomalous.

GenAD determines the dynamic anomaly threshold derived from the anomaly rate. The anomaly rate is marked as  $A_R$ , the anomaly threshold is denoted as  $\lambda$ , and the reconstruction error of single-dimensional or multi-dimensional in  $T_e$  are  $E^{T_e}$ , then

$$\mathbf{P}(E^{T_e} \leq \lambda) = 1 - A_R \quad (4)$$

Given that the probability density function of  $E^{T_e}$  is  $\mathbf{f}(e)$ , the probability distribution function  $\mathbf{F}(e) = \int_{-\infty}^e \mathbf{f}(t)dt$ , then Equation (4) is written as,

$$\mathbf{F}(\lambda) = \int_{-\infty}^{\lambda} \mathbf{f}(t)dt = 1 - A_R \quad (5)$$

To determine the dynamic anomaly threshold  $\lambda$ , we need to obtain the probability distribution function  $\mathbf{F}(e)$  of  $E^{T_e}$ . The intuitive idea for getting  $\mathbf{F}(e)$  is to analyse the statistical features of  $E^{T_e}$ , e.g.  $E^{T_e}$  obeys Gaussian-distribution or  $t$ -distribution. But it is not available in large-scale service or equipment scenarios, as the statistical features of each service are various and there are few data for each service to analyse. Besides, OmniAnomaly [6] applies extreme value theory (EVT) [7] to estimate the parameters of the distribution of  $E^{T_e}$ . Nonetheless, its high complexity leads to obtaining the anomaly threshold  $\lambda$  inefficiently. Unlike the abovementioned methods, we estimate the probability density function  $\mathbf{f}(e)$  of  $E^{T_e}$  based on the data of  $E^{T_e}$ ,

$$\mathbf{f}(e) = \sum_{i=1}^{\text{Num}_E} e_{\lambda_i}^{T_e}, e_{\lambda_i}^{T_e} = \begin{cases} 1, & \text{if } (e - \Delta e \leq E_i^{T_e} < e) \\ 0, & \text{else} \end{cases} \quad (6)$$

$\text{Num}_E$  is the number of  $E^{T_e}$ ,  $E_i^{T_e}$  is  $i$ th sample data of  $E^{T_e}$ , and  $\Delta e$  is the sample interval of  $E^{T_e}$  for  $\mathbf{f}(e)$ . Then we integrate  $\mathbf{f}(e)$  to get the probability distribution function  $\mathbf{F}(e) = \sum_{\min E^{T_e}}^e \mathbf{f}(t)$ , and get the dynamic threshold  $\lambda$  by Equation (5). Note that, in order to reduce the error of parameter estimation through sampled data, we set  $A_R = \eta + A_R$  and Equation (5) becomes  $\mathbf{F}(\lambda) = \int_{-\infty}^{\lambda} \mathbf{f}(t)dt = 1 - A_R'$ , where  $\eta \in [-0.01, 0.01]$  is fixed to maximize the F1-score during validation.

*Experimental results and analysis:* To evaluate the model performance, we compare the proposed GenAD model with the following baseline models: LSTM-NDT [8], OmniAnomaly [6], and MSCRED [5]. Table 2 reports the precision, recall, and F1-score on WBS dataset, where GenAD(G) represents GenAD with the pretraining strategy and the best scores are highlighted in bold. From Table 2, GenAD(G) outperforms all baseline methods and boosts its F1-score by up to 9% on the total dataset. It is worth mentioning that the recall of MSCRED,

Table 2. Results of GenAD(G) and baselines (WBS datasets)

Method	WBS-Area1			WBS-Area2			WBS-Area3			Total		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
LSTM-NDT	0.500	0.200	0.286	1.000	0.154	0.267	0.500	0.200	0.286	0.667	0.185	0.289
MSCRED	0.599	1.000	0.749	0.603	1.000	0.753	0.608	1.000	0.757	0.603	1.000	0.753
OmniAnomaly	0.768	1.000	0.868	0.762	1.000	0.865	0.747	1.000	0.855	0.759	1.000	0.863
GenAD (G)	<b>0.926</b>	1.000	<b>0.962</b>	<b>0.932</b>	1.000	<b>0.965</b>	<b>0.872</b>	1.000	<b>0.931</b>	<b>0.910</b>	1.000	<b>0.953</b>

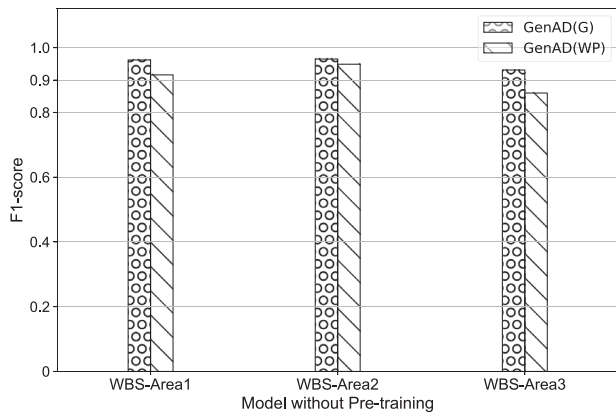


Fig. 2 Effect of pre-training

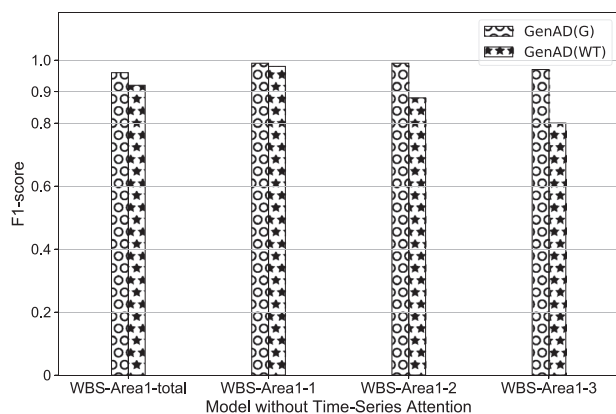


Fig. 3 Effect of time-series attention

OmniAnomaly, and GenAD(G) can achieve 1.0, since we adopt the point-adjust way [6] to calculate the metric. But the recall of the prediction-based model(LSTM-NDT) is lower than 1.0. It is due to the fact the prediction-based method is more sensitive to noise and some uncontrollable factors (e.g. the network environment change) make some series less predictable. Additionally, compared with MSCRED and OmniAnomaly, GenAD(G) performs better by virtue of the pre-training and fine-tuning framework.

Furthermore, to investigate the influence of the pretraining strategy, we train GenAD without pre-training (GenAD(WP)) on WBS datasets. From Figure 2, GenAD(G) achieves a higher F1-score on the three areas, verifying the effectiveness of pre-training strategy on heterogeneous WBSes. We also evaluate the performance of the GenAD without time-series attention (GenAD(WT)) on WBS-Area1, which are WBS-Area1-total and three WBSes selected in WBS-Area1 datasets. Figure 3 shows that the total performance (WBS-Area1-total) of GenAD(WT) decreases 4% compared with GenAD(G). This is because GenAD(WT) cannot capture the various temporal patterns of MTS in WBSes well. Besides, as for the effect of multi-correlation attention, it should be pointed out that GenAD without multi-correlation attention will detect each time series in isolation and lose the correlations among multivariate series.

**Conclusion:** In this paper, we propose a general unsupervised anomaly detection model using multivariate time series for large-scale WBSes, called GenAD. We adopt multi-correlation attention to represent the

complex correlations among the MTS and employ time-series attention to represent the various temporal patterns of each time series. Moreover, we employ the pretraining strategy to adapt to large-scale and heterogeneous WBSes. Extensive experiments reveal GenAD improves F1-score by up to 9% on real-world datasets in mobile networks. By leveraging the Jiutian artificial intelligence platform, we have applied GenAD in monitoring large-scale WBS behaviours in the realistic mobile networks of China Mobile, which has boosted the operation efficiency by 30–40% than these actually deployed anomaly detection techniques.

**Author contributions:** Xiaolei Hua: Conceptualization, methodology, writing - original draft. Lin Zhu: Project administration, writing - review and editing. Shenglin Zhang: Resources, supervision. Zeyan Li: Conceptualization. Su Wang: Investigation, validation. Chao Deng: Project administration. Junlan Feng: Formal analysis. Zhao Zhang: Writing - review and editing. Wei Wu: Data curation.

**Conflict of interest:** The authors declare no conflict of interest.

**Data availability statement:** Research data are not shared.

© 2022 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Received: 19 September 2022 Accepted: 4 November 2022

doi: 10.1049/ell2.12683

## References

- Zhao, P., Chang, X., Wang, M.: A novel multivariate time-series anomaly detection approach using an unsupervised deep neural network. *IEEE Access* 9, 109025–109041 (2021)
- Aytekini, C., Ni, X., Cricri, F., et al.: Clustering and unsupervised anomaly detection with 12 normalized deep auto-encoder representations. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE, Piscataway, NJ (2018)
- Zhang, Y., Chen, Y., Wang, J., et al.: Unsupervised deep anomaly detection for multi-sensor time-series signals. *IEEE Trans. Knowl. Data Eng.* (2021). <https://doi.org/10.1109/TKDE.2021.3102110>
- Zong, B., Song, Q., Min, M.R., et al.: Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations, pp. 1–19. Curran Associates, Red Hook, NY (2018)
- Zhang, C., Song, D., Chen, Y., et al.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1409–1416. AAAI, Washington, D.C (2019)
- Zhang, C., Song, D., Chen, Y., et al.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1409–1416. AAAI, Washington, D.C (2019)
- Xu, H., Chen, W., Zhao, N., et al.: Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In: Proceedings of the 2018 World Wide Web Conference, pp. 187–196. ACM Press, New York (2018)
- Hundman, K., Constantinou, V., Laporte, C., et al.: Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 387–395. ACM Press, New York (2018)