Computers and Artificial Intelligence; Vol.3, No.5; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n5.96102
Published by: Zeus Press

A Review of the Application of Data Science Methods in Communication Signal Anomaly Detection

Weihong Tang*

School of Information Engineering and Artificial Intelligence, Nanchang Hangkong University, Nanchang 330063, China

**Corresponding author: Weihong Tang.*

Abstract

With the large-scale deployment of 5G and industrial Internet of Things devices, the electromagnetic environment has become complex and diverse, and communication signal anomalies caused by jamming interference, spoofing interference, and equipment faults have become increasingly prominent. Traditional threshold and time-frequency analysis methods depend on manual experience, and their performance is limited in practical scenarios with unknown interference, low signal-to-noise ratios, and scarce samples. In recent years, data science technologies represented by few-shot learning, open-set recognition, lightweight deep models, and improved unsupervised frameworks have become research hotspots in communication signal anomaly detection. This study reviews recent domestic and international research results from three dimensions: few-shot optimization under supervised paradigms, architectural improvement of unsupervised reconstruction models, and open semi-supervised models. It compares the advantages and existing shortcomings of different technical routes, analyzes practical challenges related to samples, signal-to-noise ratios, and deployment in real electromagnetic environments, and looks ahead to future development directions.

Keywords

communication signals, anomaly detection, machine learning, data science

1. Introduction

Wireless communication application scenarios continue to expand, and industrial wireless systems, the Internet of Things, and aviation communication systems place higher requirements on spectrum security. Sources of communication signal anomalies include human jamming interference, signal spoofing attacks, aging and faults in hardware equipment, and signal distortion caused by channel multipath fading. These anomalies can directly cause communication interruption, an increase in bit error rate, and failure of key services. Traditional detection methods rely on fixed energy thresholds and manually designed time-frequency features, and they can only adapt to ideal conditions with stable environments and known interference types. When the electromagnetic environment changes dynamically, the false alarm rate and missed detection rate of traditional methods rise significantly, and it becomes difficult to identify new interference signals that have never appeared before. Benefiting from the rapid development of radio-frequency datasets and deep learning frameworks, data science methods no longer rely completely on manually constructed signal features, and they can directly learn the intrinsic patterns of signals from IQ sampling sequences and time-frequency

© 2026 The Author(s). Authors retain copyright in their articles.

spectrograms. In recent years, research has no longer been limited to simple classification tasks, but has become more oriented toward real engineering constraints, such as scarce abnormal samples, unknown interference, limited computing power of edge devices, and superimposed composite interference. These practical problems have promoted communication signal anomaly detection technology to gradually move from laboratory simulation toward real monitoring scenarios. Reviewing recent research progress in data science has practical reference value for grasping the development trend of this field.

In recent years, the research focus at home and abroad has changed significantly. Domestic studies have mainly focused on interference recognition under few-shot conditions, lightweight network deployment for embedded receivers, and composite interference detection. Foreign studies have paid more attention to open-set recognition, self-supervised pretraining, and model uncertainty estimation, and have applied them to real scenarios [1, 2, 3]. At present, large-scale public datasets of real radio-frequency anomalies are still lacking. Most algorithm verification still relies on simulation data, and measured studies in real scenarios remain relatively limited.

With recent research results as its core, this study introduces research progress in few-shot directions under supervised paradigms, improved unsupervised reconstruction models, and semi-supervised-open-set recognition. It compares the advantages and disadvantages of various frontier schemes, summarizes the practical challenges that have not yet been solved at the present stage, and finally looks ahead to research directions that deserve further exploration in this field.

2. Basic Overview of Communication Signal Anomaly Detection

Communication signal anomalies refer to the phenomenon in which, in a wireless communication system, the signal characteristics observed by the receiving end deviate from the expected normal transmission model, leading to information demodulation failure, reduced positioning accuracy, or impaired system availability. Typical anomaly types are mainly divided into jamming interference, spoofing attacks, and composite interference. Jamming interference is the most traditional and direct electromagnetic attack method. Its core logic is to significantly raise the background noise floor at the receiver front end by transmitting high-power noise signals or continuous waves (CW). When the power spectral density of the interference signal far exceeds that of the useful signal, the signal-to-noise ratio (SNR) at the receiving end rapidly falls below the demodulation threshold, causing a substantive interruption of the communication link. Different from the "brute force" nature of jamming interference, spoofing attacks show very high concealment and intelligent characteristics. Attackers generate forged waveforms that are highly similar to real satellite or base-station signals in carrier frequency, modulation mode, and even spreading-code structure, inducing the receiver to lock onto an incorrect timestamp, phase, or spatial position. This is especially common in global navigation satellite systems [4]. In addition, as the electromagnetic environment becomes crowded, composite interference signals are increasingly common. That is, multiple interference sources exist simultaneously in the same frequency band and are superimposed on each other, forming complex waveforms with multiple modulation characteristics. This superposition is not simple linear addition, but produces serious cross-term interference and nonstationary distortion in the time-frequency domain, which greatly increases the difficulty of signal separation and feature extraction [5].

3. Recent Mainstream Research Progress in Data Science

3.1 Few-Shot Supervised Learning for Sample Scarcity

Meta-learning [6] enables a model to learn general radio-frequency signal features through multi-task episodic training. When facing a new type of interference, the model needs only a small number of samples to complete rapid adaptation. Some studies have applied meta-prototypical networks to interference recognition, and they still maintain considerable detection accuracy when each type of interference has only a few samples. For example, for the problem that network intrusions are difficult to detect, a meta-learning framework based on the Meta-SGD+ANIL algorithm was proposed. It aims to improve the detection success rate for few-shot intrusions and enhance the model's ability to recognize minority-class attacks and unknown attacks [7].

© 2026 The Author(s). Authors retain copyright in their articles.

Data augmentation is also an important means in few-shot scenarios. Generative adversarial networks (GANs) [8] are used to generate simulated abnormal samples, make up for the shortage of real abnormal samples, expand the distribution of the training set, and reduce the risk of model overfitting. At the same time, researchers have carried out lightweight network reconstruction, adopting depthwise separable convolution and knowledge distillation to compress the parameter scale of CNN and CNN-LSTM models and adapt them for deployment in embedded receivers [9].

Although few-shot supervised learning has made progress, it still has obvious limitations. Most existing few-shot schemes still rely on simulation datasets. When there is a large difference between simulation and the real channel environment, the generalization ability of the model declines. In addition, this paradigm still cannot effectively identify completely new interference types that have never appeared during the training stage.

3.2 Improved Unsupervised Reconstruction Models

Unsupervised learning uses only a large number of normal signals for training and does not require abnormal labels, so it is more consistent with engineering practice. Early basic autoencoders and isolation forests had shortcomings: the false alarm rate was high in strong-noise environments, and the detection effect was limited when facing composite interference. In the past four years, research has mainly improved and optimized the architecture of reconstruction models.

Memory-enhanced autoencoders, LSTM-AE, and variational autoencoders (VAE) have been widely discussed. The memory module constrains the model so that it can reconstruct the input signal only by using normal patterns in the memory bank, suppressing the model's reconstruction ability for abnormal signals and thereby reducing misjudgment caused by noise [10]. Some studies introduce an attention mechanism to focus on capturing weak abnormal fragments in time-frequency spectra and improve the detection ability for weak interference [11]. In addition to autoencoders, unsupervised frameworks based on kernel density estimation and deep clustering have also been used for real-time anomaly screening of large-scale base-station signals [12].

Improved unsupervised models alleviate some defects of basic schemes, but inherent problems still exist. The interpretability of these models is generally weak, and they only output anomaly scores without locating the causes of anomalies. In low-SNR environments, noise still interferes with reconstruction errors and causes false alarms. Their recognition ability is also limited for spoofing anomalies whose features are highly close to those of normal signals [13].

3.3 Open Semi-Supervised Models

With the continuous development and improvement of the communication signal field, the limitations of traditional semi-supervised learning have gradually become exposed. When facing massive high-dimensional data, computational complexity grows exponentially, limiting its further application. Therefore, in recent years, related research has no longer been limited to basic pseudo-label generation, but has turned to deeper theoretical research. Centered on consistency regularization, it seeks more robust and scalable solutions. The Mean Teacher model is a milestone work in this period. It not only provides a complete theoretical basis for semi-supervised learning under deep convolutional neural networks, but also lays a foundation for later research on advanced algorithms such as FixMatch [14].

Although the above studies have achieved certain effects, pseudo-label quality is always the bottleneck that restricts the performance of semi-supervised learning. For this reason, researchers have developed various training methods to reduce the bias introduced by pseudo-labels as much as possible. For example, the BaCon method targets class-imbalanced semi-supervised learning (CISSL) and proposes a balanced feature-level contrastive learning mechanism. By enlarging inter-class distances and reducing intra-class distances in feature space, it effectively alleviates the bias introduced by unreliable pseudo-labels [15]. In addition, Cross Pseudo Supervision (CPS) and its generalized form, n-CPS, train multiple subnetworks and let them provide pseudo-label supervision to each other. A voting mechanism is used to further improve pseudo-label reliability and reduce the risk of overfitting by a single model [16].

© 2026 The Author(s). Authors retain copyright in their articles.

4. Comparison of Mainstream Methods

Table 1: Comparison of Different Data Science Methods for Signal Anomaly Detection

Method category	Representative schemes	Labels	Core advantages	Limits	Scenarios
Few-shot supervised learning	Meta-prototypes; lightweight CNN-LSTM; GAN augmentation	Few labeled samples	High known-anomaly accuracy; lightweight deployment	Cannot detect unknown interference	Small abnormal-sample monitoring
Improved unsupervised learning	Memory-enhanced autoencoder; attention LSTM-AE	Normal samples only	No anomaly labels; suitable for engineering use	Weak spoofing-anomaly detection	Spectrum monitoring with scarce abnormal samples
Semi-supervised open-set recognition	Dual-branch fusion; energy models; open-set pretraining	Few labeled + many unlabeled samples	Handles known and unknown interference; fits open electromagnetic settings	Complex model and difficult training	Complex environments with changing interference

The comparison of the three types of schemes is shown in Table 1, from which the advantages and disadvantages of the three research schemes can be clearly understood. Few-shot supervised learning performs well when known abnormal labels are available and has high sample recognition accuracy, but it has poor effects on completely unknown interference. Therefore, it is mainly applied to scenarios where a small number of abnormal samples can be collected. Unsupervised learning is limited by its weak ability to identify weak anomalies, and it is mainly applied to environments with a large number of normal samples and no need to identify abnormal labels. Compared with the former two, semi-supervised learning has more prominent anti-interference ability and abnormal-label processing ability, so it is more often applied to complex environments with new and changing interference types. Different technical routes have their own trade-offs and are applied to different environments.

5. Current Practical Challenges

In recent years, data science has made sufficient progress in communication signal detection, but many urgent problems still remain.

a) Confirmation bias and pseudo-label quality: In the self-training paradigm, the model tends to regard the wrong pseudo-labels it generates as real labels for retraining, thereby causing error accumulation and amplification. Although methods based on consistency regularization, such as FixMatch, improve performance by introducing weak and strong augmentation strategies, their fixed confidence thresholds often cannot adapt to the learning progress of different classes [17].

b) Scarcity of high-quality public anomaly datasets: In real environments, the cost of collecting effective datasets is high, which leads to a relatively small number of high-quality public datasets. Most experiments are based on simulation data. Because simulation data differ from real data, some models that perform well in simulation show degraded performance after being applied to real scenarios.

c) Implementation of semantic alignment for multimodal fusion and heterogeneous data: With the rise of multimodal large models, feature fusion has become another means of improving performance, but it has also brought new challenges. During multi-feature fusion, how to realize semantic alignment of different modalities in a shared latent space and avoid "modal collapse" or noise dominance is a highly difficult technical problem [18].

d) Privacy protection in distributed environments: In real scenarios, semi-supervised models face privacy protection constraints. How to realize collaborative use of unlabeled data across institutions and platforms while protecting user privacy, and how to prevent leakage of sensitive information through gradient inversion, are legal and technical bottlenecks that restrict the large-scale application of this technology.

© 2026 The Author(s). Authors retain copyright in their articles.

6. Future Research Prospects

At present, recognition methods for communication signal anomaly detection are shifting from previous passive signal recognition to active cognition and adaptive prevention. Facing the increasingly complex electromagnetic environment and the intelligent development of interference means, the core research direction in the future should focus on constructing intelligent systems with adaptive and self-evolving abilities. By embedding knowledge such as signal propagation models and Maxwell's equations into architectural design, the model can not only accurately classify data statistically, but also effectively process and analyze composite signals in a dynamically changing electromagnetic environment. Second, the deepened application of open-set semi-supervised learning will be a core means of coping with confirmation bias and class distribution mismatch. In the future, when facing different types of unknown interference, research can focus on the unknown signals themselves, excavate and use the structural information contained in them, and enable the model to have zero-shot or few-shot adaptability even without explicit labels. Furthermore, multimodal feature fusion is also an important direction for the future development of communication systems. As communication systems develop toward multiple frequency bands, multiple standards, and multiple sensors, single-modality signal analysis has become unable to fully capture the characteristics of complex electromagnetic environments. Therefore, future anomaly detection will rely heavily on multimodal fusion technology and integrate multi-source heterogeneous information such as time-frequency, spatial, polarization, and even text log information. Through learnable graph operator expansion technology, the system can deeply capture structured relationships among different modalities and make up for the deficiencies of traditional concatenation or element-level operations in feature interaction [19].

7. Conclusion

In recent years, communication signal anomaly detection research has gradually moved away from early simple closed-set simulation classification and has turned toward research under the constraints of real electromagnetic environments. Early studies mostly relied on closed datasets with balanced distributions, which made it difficult to cope with the dynamic evolution of interference types and the frequent occurrence of unknown threats in real scenarios. With the continuous deepening of research, new achievements have continued to emerge in directions such as few-shot supervised learning, improved unsupervised reconstruction models, semi-supervised open-set recognition, and multi-feature fusion for composite interference. Data science methods have significantly expanded the capability of communication signal anomaly detection, but they still face many severe challenges, such as a lack of high-quality datasets and a significant computing-power trade-off between lightweight deployment and open-set recognition ability. In the future, with the continuous development of lightweight open-set methods, self-supervised pretraining, and hybrid frameworks, models will have stronger adaptive and continuous-evolution abilities, and data science methods will play a greater role in wireless spectrum monitoring, communication security, and electronic warfare defense.

References

[1] Zhao, G., Li, G., Qin, Y., Zhang, J., Chai, Z., Wei, X., Lin, L., & Yu, Y. (2024). Exploration and exploitation of unlabeled data for open-set semi-supervised learning. International Journal of Computer Vision, 132, 5888-5904. https://doi.org/10.1007/s11263-024-02155-y

[2] Elango, A., & Landry, R. J. (2024). XAI GNSS—A Comprehensive Study on Signal Quality Assessment of GNSS Disruptions Using Explainable AI Technique. Sensors, 24(24), 8039. https://doi.org/10.3390/s24248039

[3] Feng, Y., Han, B., Wang, X., Shen, J., Guan, X., & Ding, H. (2024). Self-supervised transformers for unsupervised SAR complex interference detection using Canny edge detector. Remote Sensing, 16(2), 306. https://doi.org/10.3390/rs16020306

[4] Elango, A., & Landry, R. J. (2024). XAI GNSS—A Comprehensive Study on Signal Quality Assessment of GNSS Disruptions Using Explainable AI Technique. Sensors, 24(24), 8039. https://doi.org/10.3390/s24248039

© 2026 The Author(s). Authors retain copyright in their articles.

[5] Tu, S., Cheng, Q., Qian, F., Hui, P., & Guan, L. (2022). Composite interference signals recognition based on YOLOv5. In 2022 IEEE 21st International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS) (pp. 210-215). IEEE. https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS57392.2022.00042

[6] Chen, G., Wang, J., Liu, W., Lin, Y., & Xu, F. (2023). The concept, model, and practical inspiration of meta-learning: A study based on spatio-temporal theory. Chinese Journal of Management Science,1-17. https://doi.org/10.16381/j.cnki.issn1003-207x.2023.0498.

[7] Huang, X., Fu, C., & Zeng, Q. (2026). Research on few-shot intrusion detection based on a Meta-SGD+ANIL meta-learning framework.12(8),712-720.

[8] Yiding Wang. (2025). Evolution and Breakthroughs of Generative Adversarial Network Technology. ITM Web of Conferences (ITM WoC), 80(11): 01023. https://doi.org/10.1051/itmconf/20258001023.

[9] Wang, H., Liu, H., & Li, M. (2024). Lightweight communication signal recognition network for embedded deployment. Journal of Electronics & Information Technology, 46(3), 912-920.

[10] Yu, Y., & Zhang, J. (2024). Unsupervised communication signal anomaly detection method based on memory-enhanced autoencoder. Journal on Communications, 45(6), 182-193.

[11] Jin, Z., & Liu, C. (2023). Communication weak interference detection algorithm based on time-frequency channel attention. Journal of Signal Processing, 39(2), 345-354.

[12] Wang, H., & Li, M. (2024). Unsupervised signal anomaly screening framework for large-scale base-station monitoring. Journal of Electronics & Information Technology, 46(8), 2761-2769.

[13] Mehr, A., & Dovis, F. (2023). Limitations of reconstruction-based unsupervised models for GNSS spoofing detection. IEEE Transactions on Aerospace and Electronic Systems, 59(4), 4123-4136.

[14] Li, J., Pan, J., Tan, V.Y., Toh, K., & Zhou, P. (2024). Towards Understanding Why FixMatch Generalizes Better Than Supervised Learning. ArXiv, abs/2410.11206.

[15] Feng, Q., Fang, S., Lin, T., & Xie, L. (2024). BaCon: Boosting imbalanced semi-supervised learning via balanced feature-level contrastive learning. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.48448/wxgy-7s45

[16] Filipiak, D., Tempczyk, P., & Cygan, M. (2026). n-CPS: Generalising cross pseudo supervision to n networks. IEEE Access, 14, 101774-101789. https://doi.org/10.1109/ACCESS.2026.3704899

[17] Xie, S., Wang, H., Niu, Z., Sun, H., Ouyang, S., Chen, Y.-W., & Lin, L. (2024). SemSim: Revisiting weak-to-strong consistency from a semantic similarity perspective for semi-supervised medical image segmentation. arXiv. https://arxiv.org/abs/2410.13486

[18] Ding, D., Wang, L., Zhu, L., Gedeon, T., & Koniusz, P. (2024). Learnable expansion of graph operators for multi-modal feature fusion. arXiv. https://arxiv.org/abs/2410.01506

[19] Ding, D., Wang, L., Zhu, L., Gedeon, T., & Koniusz, P. (2025). Learnable expansion of graph operators for multi-modal feature fusion. International Conference on Learning Representations. arXiv. https://doi.org/10.48550/arXiv.2410.01506

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

© 2026 The Author(s). Authors retain copyright in their articles.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.