A Review of Cryptocurrency Data Mining and Fraud Detection

Zejing Chen*

Bashu Secondary School, Chongqing, 400013, China

*Corresponding author's e-mail: 616365344@qq.com

Abstract. Anomaly detection in Industrial IoT (IIoT) is critical, yet current data-driven methods suffer from high false positives due to data imbalance, poor generalization, and the limited applicability of single models. To address these challenges, we propose a multi-stage framework for robust IIoT anomaly detection. Our system first employs a pre-processing module that uniquely supports both supervised and unsupervised learning by using K-Means clustering to generate pseudo-labels for unlabeled data, reducing the reliance on extensive manual labeling. Subsequently, an XGBoost-based module selects the most salient features. To enhance model input, a deep learning module then extracts advanced features: a Convolutional Neural Network (CNN) captures spatial patterns from raw packets, while a Long Short-Term Memory (LSTM) network models temporal correlations, with its output augmenting the feature set for a CART model. Finally, an optimally weighted ensemble of diverse classifiers—including Decision Tree, Random Forest, SVM, ANN, and Bayesian models—performs the final anomaly type classification. This integrated framework is designed to overcome the limitations of traditional methods by improving detection accuracy, reducing false positives, and enhancing generalization for dynamic IIoT environments.

Keywords: IIoT, Anomaly Detection, Traffic Analysis

1 Introduction

The development of the Industrial Internet Internet of Things(IIoT) is reshaping global industry with unprecedented depth and breadth. Traditionally, operational technology (OT) networks were closed and isolated environments, using proprietary protocols and focusing primarily on production stability and safety. Today, however, in order to enable data-driven decision-making, remote monitoring, and predictive maintenance, OT networks are increasingly connected to enterprise IT networks and even the public Internet. This convergence breaks the original physical isolation and dramatically expands the "attack surface." Conventional industrial protocols such as Modbus and Profibus, ares not originally designed with cybersecurity in mind. With the widespread adoption of industrial Ethernet and TCP/IP protocol stacks (e.g., OPC UA, MQTT), industrial systems have become more interconnected, but they are also

J. Rak et al. (eds.), *Proceedings of the 2026 11th International Conference on Social Sciences and Economic Development (ICSSED 2026)*, Advances in Economics, Business and Management Research 391,
https://doi.org/10.2991/978-94-6239-701-9_78

more exposed to attacks from the standard IT world, such as DDoS attacks, virus scanning, and malware propagation. Furthermore, many traditional industrial devices—including sensors, actuators, and programmable logic controllers (PLCs)—are becoming "smart" and network-connected. These devices generally have limited computing power, weak security defenses, and long lifecycles (up to several decades), making frequent security updates difficult and rendering them the most vulnerable points in the network.

Therefore, detecting anomalous traffic during industrial device communications has become crucial. In August 2025, Colt Technology Services suffered a ransomware attack launched by the WarLock group, in which the attackers are suspected to have exploited a SharePoint Server vulnerability (CVE-2025-53770) [1]. This flaw allows remote code execution and enables attackers to implant backdoors on unpatched servers, leading to key leakage and privilege escalation. Anomaly-based traffic detection methods can promptly identify hidden data exfiltration paths created by such backdoors and trace the attackers' origin. Similarly, in 2016, the distributed denial-of-service (DDoS) attack on the domain name system provider Dyn caused major internet platforms and services in Europe and North America to become inaccessible [2]. In such cases, anomalous traffic detection can quickly pinpoint the compromised servers and enable the timely application of defense strategies.

Existing anomaly detection methods for Industrial Internet of Things (IIoT) networks partially rely on signature-based rules and finite state machine (FSM) constructions. The primary advantage of these fingerprint-based approaches lies in their high detection accuracy and low false-positive rates when handling known threats, coupled with their strong interpretability. However, their fundamental limitation is the inability to detect novel, unknown attacks, commonly referred to as zero-day attacks. Since the behavioral patterns of such new attacks have not yet been discovered or integrated into the signature database, the system effectively remains "blind" to them.

To address this shortcoming, many studies have explored data-driven, probabilistic modeling approaches. Instead of relying on predefined attack signatures, these methods learn from large-scale network data to build mathematical models that characterize the system's "normal" behavior. Typically, machine learning and deep learning techniques are employed to construct these models. However, these approaches face several limitations: **1. High false-positive rates caused by data imbalance.** Network traffic datasets inherently suffer from severe class imbalance, where normal traffic samples vastly outnumber malicious ones. In supervised learning scenarios, this imbalance causes models to overfit the majority class (normal traffic) during training, resulting in poor detection performance for minority attack classes. **2. Poor generalization and computational inefficiency.** High-performance deep neural networks (DNNs) are computationally intensive, demanding significant CPU, memory, and power resources, which poses challenges when deploying them on resource-constrained IIoT devices such as edge gateways or embedded controllers. While traditional machine learning algorithms are more lightweight, redundant features can lead them to perform well on training datasets but fail to generalize to unseen traffic or novel attack scenarios. **3. Limited adaptability of single-model**

approaches. Different algorithms exhibit varying sensitivities to features, resulting in discrepancies in accurately identifying different types of attacks. Consequently, a single model often struggles to simultaneously achieve both high recall and low false-positive rates, limiting its applicability across diverse detection scenarios.

To address these challenges, we propose a framework for anomaly detection in Industrial Internet of Things (IIoT) network traffic. In the traffic collection module, we adopt a rule-based and clustering-assisted labeling approach to balance and annotate the traffic data. We design an XGBoost-based feature selection method to reduce the high dimensionality of IIoT network traffic features, thereby mitigating the risk of the "curse of dimensionality," which can otherwise lead to model overfitting, longer training times, and reduced detection accuracy due to the noise introduced by irrelevant or redundant features. Furthermore, we propose a spatiotemporal feature extraction method based on Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, enabling subsequent classification models to achieve a better understanding of traffic data patterns. Finally, we design an ensemble learning model to integrate the classification results from multiple anomaly detection models, improving overall detection performance.

2 Background

In this section, we first summarize the anomaly traffic types in IIoT. Then we describe Next, we describe various types of anomaly detection methods.

2.1 Industrial Traffic Anomalies

Abnormal traffic in the Industrial Internet can be categorized into four typical patterns based on their causes and characteristics.

Network attack–type anomalies typically involve a clear malicious intent, aimed at disrupting, controlling, or exfiltrating industrial assets. These include reconnaissance and scanning attacks, denial-of-service (DoS) attacks, malicious command injection or logic tampering, as well as malware and data theft attacks. Reconnaissance and scanning attacks occur during the preparatory phase before launching a substantive attack, in which attackers probe the industrial network to identify active hosts, open ports, and running services. Examples include massive scan packets targeting common industrial control device ports—such as Modbus/TCP port 502 or S7 protocol port 102—and traffic that probes for known vulnerabilities in specific PLC or HMI models. Denial-of-service attacks aim to exhaust the network or computing resources of the target device, rendering it unable to respond to legitimate control commands and potentially causing production interruptions. Common forms include SYN floods and UDP floods. Malicious command injection or logic tampering represents the most dangerous type of anomaly in the industrial Internet, as it directly interferes with and disrupts physical production processes. Some attackers send reboot or reset commands to halt device operation. Even more critically, many industrial IoT devices have resource-constrained firmware that is easily compromised.

Attackers may upload malicious firmware during legitimate firmware update processes to take control of industrial control devices and issue harmful commands. Another form of this attack is a man-in-the-middle scenario, in which the attacker intercepts and manipulates normal communications between the engineering workstation and PLCs to issue false commands.

Operational fault–type anomalies are not caused by malicious attacks, but stem from failures within the devices, network, or software itself. These can be divided into two categories: device component failures and network configuration errors. Most component failures in industrial control devices manifest as a sensor or controller repeatedly sending a large number of duplicate or erroneous messages, creating a "broadcast storm." In other cases, a device that is supposed to report data periodically—such as an RTU—may suddenly go "silent," with its heartbeat or data packets disappearing. Network configuration errors are mostly caused by human mistakes from device engineers. Examples include IP address conflicts, unauthorized connection of personal laptops by maintenance personnel, critical control traffic being mistakenly routed to the office network, or communications occurring between devices in different security zones that should never be able to communicate.

Abnormal behavioral patterns aren't necessarily clear-cut attacks or malfunctions, but they deviate from established baselines of normal behavior and are important indicators of potential risk. These anomalies are well-suited for detection using model-based machine learning methods, including changes in communication patterns and protocol violations. Changes in communication patterns typically refer to unexpected alterations in the way network devices interact. For example, two PLCs that normally should not communicate may suddenly start exchanging data, there may be a surge of remote logins or data downloads during non-production periods, or baseline traffic patterns may drift abnormally. Another category of issues arises when developers misuse industrial control protocols in ways that deviate from standard specifications or intended business logic, such as using unauthorized function codes, engaging in cross-protocol communication, or generating anomalous packet structures.

2.2 Traffic Anomaly Detection

Industrial Internet anomaly traffic detection methods can be summarized into two major categories: rule/feature-driven and model/data-driven. Rule/feature-driven detection relies on predefined feature signature libraries and rule libraries to discover anomalies by comparing known attack features.

Rule/Character Driven Method. Hee-Yong et al. Kwon[9] proposed a hybrid intrusion detection method that combines signature-based statistical filtering and a behavior-based composite autoencoder (CAE). This approach effectively detects anomalies in Industrial Control Systems (ICS), improving detection performance and reducing computational time compared to autoencoder-only methods, as validated on the SWaT dataset. Jesús Díaz-Verdejo et al. [10] proposed a method to evaluate the performance of signature-based intrusion detection systems (SIDS) like Snort, ModSecurity, and Nemesida, using predefined rulesets for web attack detection. They

found that these off-the-shelf configurations either achieved insufficient detection rates (as low as 6-8%) or produced high false alarms, highlighting the critical need for manual ruleset tuning. To address this, they introduced an efficient rule-filtering algorithm that increased precision from 0.015 to 1.0 in real-world environments, significantly reducing false positives. Yang et al. [11] proposed a stateful intrusion detection system (IDS) using a Deep Packet Inspection (DPI) method and a novel Detection State Machine (DSM) to monitor and secure IEC 60870-5-104 SCADA networks. This approach validates protocol state transitions to detect deviations and misbehaviors, effectively identifying all injected abnormal packets with zero false positives in experiments. The solution addresses the protocol's lack of authentication and enhances SCADA security against cyber threats like packet injection and replay attacks.

Model/Data Driven Method. Chen et al. [3] proposes an optimized deep learning framework for anomaly detection in Industrial Internet of Things (IIoT) environments. They employs Extreme Gradient Boosting (XGBoost) to evaluate the importance of each feature by applying different importance thresholds to select the most informative features, and apply a customized loss function to handle highly imbalanced datasets and closely related anomaly classes. Finally, LSTM-based model called MIX_LSTM are designed for anomaly detection. Rodriguez et al. [4] proposes an IIoT Anomaly Classification Framework to address challenges of system complexity, security threats, and operational failures in industrial environments. The framework combines unsupervised anomaly detection with machine learning-based classification, enhanced by contextual information to improve accuracy and adaptability. Evaluated on a realistic IIoT testbed, it achieves high performance, demonstrating strong potential for improving reliability, security, and efficiency in industrial systems. Anton et al. [12] evaluated several machine learning algorithms (SVM, Random Forest, k-NN, k-means) for anomaly detection on a synthetic Modbus/TCP dataset. They found that SVM and Random Forest performed exceptionally well at detecting malicious traffic, while k-NN and k-means clustering yielded unsatisfactory results.

3 Design Details

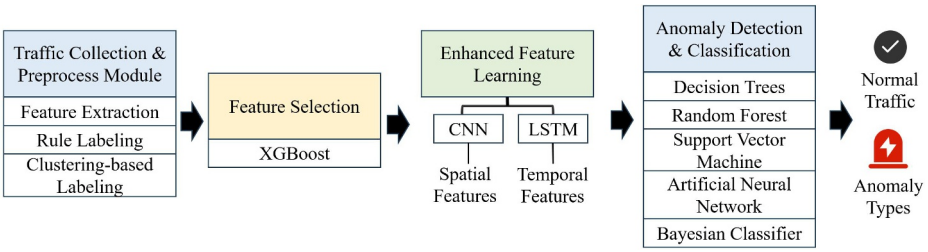

Fig. 1. The architecture of anomaly traffic detection system.

In this section, we introduce an anomaly detection framework for Industrial Internet of Things (IIoT) network traffic. As shown in Fig.1, the framework consists of four main components: a traffic collection and preprocessing module, a feature selection module, a deep learning–based feature extraction module, and a model-based anomaly classification module. The traffic collection and preprocessing module includes both traffic feature extraction and anomaly labeling. For datasets without predefined labels, we employ K-means clustering to generate predicted labels, thereby supporting both supervised and unsupervised classification scenarios. The feature selection module leverages the XGBoost algorithm to identify the most relevant features with the highest correlation to anomaly patterns. The deep learning feature extraction module integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks: the CNN extracts spatial features from raw packets, while the LSTM captures temporal dependencies in network traffic and outputs temporal context vectors. These vectors are then incorporated as new CART-based features, effectively expanding the feature space and enriching the inputs for downstream classification models. Finally, the model-based anomaly classification module adopts an ensemble learning approach that combines multiple classifiers, including Decision Trees, Random Forests, Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Bayesian classifiers. By applying optimized weight allocation to integrate the outputs of these models, the framework produces the final classification results for different anomaly types.

3.1 Traffic Collection \& Preprocess

Feature Extraction. To comprehensively capture network behavior, feature engineering primarily focuses on extracting two types of features: packet-level features and flow-level features. **Packet-level features:** These are extracted directly from individual packet headers, such as source/destination IP addresses, ports, protocol types, and packet lengths. These features serve as fundamental indicators but are often insufficient on their own for handling complex detection tasks.

Data Labeling. A core challenge in the field of Industrial Internet of Things (IIoT) security is the severe lack of labeled datasets representing anomalous traffic. The vast majority of network traffic is normal, while novel attacks, by definition, are unlabeled. Our framework addresses this challenge through a dual-mode approach, making it more practical than purely supervised systems. When labeled datasets are available (e.g., historical incident data from testbeds such as WUSTL-IIoT-2021 or the UNSW dataset), the system directly leverages these ground truth labels for model training and evaluation. This path maximizes the use of existing knowledge. For unlabeled real-world traffic, the framework switches to an unsupervised mode. In this mode, the system automatically generates labels for subsequent supervised classifiers. To achieve this, we designed a KMeans clustering algorithm to infer the intrinsic structure of the data. The core assumption is that normal traffic forms dense, well-defined clusters, whereas anomalous traffic either forms its own small, distinct clusters or appears as outliers[5]. We adopt the Mahalanobis distance as the distance metric, as it offers significant advantages for high-dimensional, correlated network

data. By taking into account the covariance between features, this method becomes less sensitive to variations in scale and inter-feature correlations.

3.2 Feature Selection

After feature engineering, Industrial IoT (IIoT) network traffic often results in datasets with a large number of features (high dimensionality). This "curse of dimensionality" can lead to model overfitting, increased training time, and reduced detection accuracy, as irrelevant or redundant features introduce noise. Therefore, feature selection is not just an optimization step but a necessity. It enhances model performance by allowing learning algorithms to focus on the most relevant variables, improves interpretability by revealing which factors are most indicative of attacks, and reduces the computational resources required for both training and inference[6].

In our framework, we adopt XGBoost, a powerful tree-based ensemble algorithm renowned for its outstanding performance and scalability. We apply Recursive Feature Elimination (RFE) techniques, where the model is iteratively retrained after removing the least important features, ranking all features from most to least significant based on their contribution. The output of XGBoost is a "semantically compressed" feature vector. When this vector is fed into subsequent classification models and feature learning modules, the models no longer need to waste learning capacity on discovering basic correlations. Instead, they can directly focus on extracting complex, higher-order temporal patterns from features already known to have high predictive value, making the entire processing pipeline more efficient and effective.

3.3 Enhanced Feature Learning

This is the most innovative component of the framework, employing a hybrid deep learning architecture to extract features that traditional machine learning models cannot capture. It treats network traffic not as static numerical tables but as a dynamic, multi-dimensional data stream with both spatial structure and temporal narrative. Traditional models struggle to capture the complex nonlinear relationships present in network traffic. In contrast, deep learning models—especially hybrid architectures—excel in this regard by automatically learning hierarchical feature representations from raw data [7].

The combination of CNNs and LSTMs represents a state-of-the-art approach for this task, as they capture complementary aspects of the data [8]. The CNN acts as a spatial feature extractor, analyzing the internal structure of individual data points (packets or flows), while the LSTM serves as a temporal feature extractor, analyzing the sequence of these data points over time. This hybrid approach enables the model to simultaneously understand both the "content" and the "temporal/behavioral patterns" of network activity. The specific output of this module is a "temporal context vector", which enhances the feature set for the final classifier. The final hidden state of the LSTM (or the output at the last time step) serves as this temporal context vector. It is a dense, fixed-size vector that encapsulates the learned temporal

features of the entire input sequence. This newly generated temporal context vector is then concatenated with the original feature vector (the XGBoost-selected features of the last flow in the sequence) and the spatial feature vector from the CNN.

3.4 Anomaly Detection & Classification

The final decision component of the framework leverages the collective power of multiple models to achieve higher accuracy and robustness than any single model alone. Ensemble methods combine multiple "weak learners" to create a more accurate, robust, and less overfitting-prone "strong learner." In our framework, we employ five types of ensemble learning models. **Decision Trees and Random Forests:** Tree-based models are highly interpretable and can effectively capture nonlinear relationships. Random Forests, as ensembles of decision trees, are particularly robust and resistant to overfitting. **Support Vector Machines (SVMs):** Effective in high-dimensional spaces, SVMs classify by finding an optimal hyperplane that separates classes. **Artificial Neural Networks (ANNs):** Multi-layer perceptrons (MLPs) can learn highly complex nonlinear decision boundaries, complementing the capabilities of other models.

A simple approach is to combine base classifiers using majority voting, but a more sophisticated method is to assign weights to each classifier's predictions based on its performance, giving more influence to more reliable models. Here, we adopt a meta-learning (Stacking) approach, where a meta-learner (e.g., logistic regression) is trained on the predictions of the base classifiers. The base classifiers' predictions on a validation set serve as the meta-learner's training data. This allows the system to learn the optimal way to combine base model outputs, effectively learning the weights itself. The final output of the optimally weighted ensemble model is a probability distribution over possible anomaly classes. A high-confidence prediction (e.g., 98% probability of 'DDoS') can trigger an automatic response, while a low-confidence prediction (e.g., 45% 'DDoS', 40% 'Reconnaissance') can be flagged for immediate human analyst review. The weight optimization process is essentially a way of learning which "expert" (base classifier) to trust under different circumstances, thereby minimizing the risk of erroneous decisions.

4 Conclusion

In this work, we have presented a comprehensive framework designed to overcome critical limitations in current data-driven anomaly detection for Industrial IoT. The experimental results validate that our multi-stage approach effectively addresses the challenges of data imbalance, poor generalization, and the narrow applicability of single models. By integrating a K-Means-based semi-supervised pre-processing stage, we significantly reduce the dependency on labeled data. The fusion of deep features, extracted via a hybrid CNN-LSTM architecture, with core statistical features selected by XGBoost, was shown to provide a richer and more discriminative input for classification. This addresses the poor generalization and efficiency of traditional

models by capturing both spatial and temporal characteristics of traffic data. Furthermore, the final classification, derived from an optimally weighted ensemble of diverse machine learning models, demonstrates superior stability and accuracy compared to any single classifier.

In conclusion, this research confirms that a hybrid, modular framework that synergizes unsupervised clustering, deep learning feature extraction, and ensemble methods provides a robust, adaptable, and high-performance solution. This approach represents a significant step towards creating more secure and resilient Industrial IoT systems in the face of complex and evolving cyber threats.

References

1. Colt technology services hit by warlock ransomware gang attack. https://www.itpro.com/security/cyber-attacks/uk-telecoms-firm-takes-systems-offline-after-cyber-attack.
2. Ddos attacks on dyn. https://en.wikipedia.org/wiki/DDoS attacks on Dyn.Coinmarketcap accessed, 2024. https://coinmarketcap.com/.
3. Zhen Chen, ZhenWan Li, Jia Huang, ShengZheng Liu, and HaiXia Long. An effective method for anomaly detection in industrial internet of things using xgboost and lstm. Scientific Reports, 14(1):23969, 2024.
4. Martha Rodriguez, Diana P Tobon, and Danny Munera. A framework for anomaly classification in industrial internet of things systems. Internet of Things, 29:101446, 2025.
5. Nico G¨ornitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. Journal of Artificial Intelligence Research, 46:235–262, 2013.
6. Silpa Chalichalamala, Niranjana Govindan, and Ramani Kasarapu. Logistic regression ensemble classifier for intrusion detection system in internet of things. Sensors, 23(23):9583, 2023.
7. Jawad Hussain Kalwar and Sania Bhatti. Deep learning approaches for network traffic classification in the internet of things (iot): A survey. arXiv preprint arXiv:2402.00920, 2024.
8. Afrah Gueriani, Hamza Kheddar, and Ahmed Cherif Mazari. Adaptive cyber-attack detection in iiot using attention-based lstm-cnn models. In 2024 International Conference on Telecommunications and Intelligent Systems (ICTIS), pages 1–6. IEEE, 2024
9. Kwon H Y, Kim T, Lee M K. Advanced intrusion detection combining signature-based and behavior-based detection methods[J]. Electronics, 2022, 11(6): 867.
10. Díaz-Verdejo J, Muñoz-Calle J, Estepa Alonso A, et al. On the detection capabilities of signature-based intrusion detection systems in the context of web attacks[J]. Applied Sciences, 2022, 12(2): 852.
11. Yang Y, McLaughlin K, Sezer S, et al. Stateful intrusion detection for IEC 60870-5-104 SCADA security[C]//2014 IEEE PES General Meeting| Conference & Exposition. IEEE, 2014: 1-5.
12. Anton S D, Kanoor S, Fraunholz D, et al. Evaluation of machine learning-based anomaly detection algorithms on an industrial modbus/tcp data set[C]//Proceedings of the 13th international conference on availability, reliability and security. 2018: 1-9.