Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Trust-Aware Cyber Threat Identification in Industrial IoT Environments Through Adaptive Representation Learning

Oliver Bennett1, Charlotte Hughes2, Thomas Wilson3*
Department of Computing, Imperial College London, London SW7 2AZ, United Kingdom
***Corresponding author**: t.wilson@x.ac.uk

Key word:
Industrial IoT security; intrusion detection; hybrid ensemble learning; explainable AI; SHAP analysis; cyber-physical systems; network anomaly detection

A B S T R A C T

Industrial Internet of Things systems are increasingly exposed to network-based attacks that may disrupt sensing, control, and data transmission processes. Traditional intrusion detection methods often focus on classification accuracy but provide limited explanation of why specific traffic patterns are identified as malicious, which restricts their use in safety-critical industrial environments. This study develops an explainable hybrid ensemble intrusion detection model for Industrial IoT network security. The framework combines XGBoost, Gradient Boosting, and Balanced Random Forest classifiers to improve robustness against class imbalance, while SHAP values are applied to quantify the contribution of protocol-level, statistical, and temporal traffic features. The model is evaluated on the ToN-IoT and Edge-IIoTset datasets, containing 461,043 and 2,219,201 labeled traffic samples, respectively. The original feature space is reduced from 79 to 36 features using mutual information ranking and correlation filtering. Experimental results show that the proposed model achieves 97.93% accuracy, 96.88% macro-F1, and 98.57% AUC on ToN-IoT, outperforming single XGBoost and CNN-based baselines by 1.76% and 2.41% in macro-F1, respectively. For Edge-IIoTset, the framework maintains a detection rate of 98.21% under mixed attack categories, including DDoS, reconnaissance, injection, and man-in-the-middle attacks. The SHAP analysis indicates that packet inter-arrival time, source-to-destination byte ratio, TCP window size, and protocol service frequency are key features for distinguishing abnormal industrial traffic. These findings suggest that explainable ensemble models can improve both detection reliability and decision transparency in Industrial IoT security monitoring.

INTRODUCTION

Industrial Internet of Things (IIoT) systems are increasingly deployed in manufacturing, energy, transportation, logistics, smart grids, and smart infrastructure. These systems connect sensors, controllers, actuators, industrial gateways, edge servers, cloud platforms, and application services through heterogeneous communication networks. Such connectivity improves real-time monitoring, remote control, predictive maintenance, production scheduling, and resource management [1,2]. However, it also exposes industrial devices and operational networks to more complex cyber threats. Attacks against IIoT systems may interrupt sensing processes, delay control commands, manipulate transmitted data, disrupt industrial protocols,

or affect physical equipment operation. In safety-critical environments, even a short network failure may cause production loss, equipment damage, or physical safety risks. Therefore, network intrusion detection has become a key component of IIoT security. A recent embedded-systems security study has shown that SHAP-based interpretable ensemble learning can provide feature-level evidence for network traffic anomaly detection under constrained computing conditions [3]. This finding is highly relevant to IIoT environments because industrial security models should not only identify abnormal traffic, but also explain why an alarm is generated. Machine learning and deep learning methods have been widely used in IIoT intrusion detection because they can learn abnormal traffic patterns from statistical, protocol-level, and temporal flow features [4,5]. Common indicators include flow duration, packet length, byte rate, protocol service type, connection state, flag distribution, inter-arrival time, and forward-backward traffic behavior. Compared with traditional rule-based systems, data-driven methods can capture nonlinear relationships among traffic features and detect attacks that do not exactly match known signatures [6]. This is important for IIoT networks because industrial traffic is often generated by different devices, protocols, control cycles, and service interactions. The same attack may also show different traffic behaviors under different industrial scenarios. However, IIoT intrusion detection still faces several practical challenges, including imbalanced attack categories, weak model explanation, dataset bias, unstable feature distributions, and limited validation under realistic industrial traffic conditions [7]. These problems reduce the reliability of intrusion detection models when they are transferred from benchmark datasets to operational industrial networks. Traditional intrusion detection systems mainly rely on known attack signatures, expert-defined rules, and manually designed thresholds. These methods are useful for detecting previously observed attacks and policy violations. They are also easy to interpret because each alert can usually be traced to a specific rule. However, fixed rules are less effective when new attack variants appear, when attackers change traffic patterns, or when normal industrial behavior evolves over time. IIoT systems often contain long-running devices, mixed protocol stacks, periodic control traffic, and irregular maintenance traffic. These characteristics make it difficult to define complete and stable rules for all possible security events. For this reason, data-based detection methods have received increasing attention in industrial network security. Common machine learning models, such as Decision Tree, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting, and LightGBM, have been applied to structured network flow data [8,9]. These models can learn nonlinear relations among traffic features and identify abnormal behaviors without manually encoding every attack rule. Ensemble learning has become an important direction in IIoT intrusion detection because it combines multiple classifiers and reduces the instability of a single model. Bagging-based models can improve robustness by training different decision trees on sampled data, while boosting-based models can improve class separation by focusing on difficult samples during iterative learning [10]. This advantage is useful for IIoT traffic because attack samples are often unevenly distributed across classes. Some attacks, such as DDoS and scanning, may generate many traffic records, while injection, man-in-the-middle, ransomware, data exfiltration, and command manipulation may appear less frequently. Recent studies have reported that ensemble models can improve detection accuracy and reduce false alarms on IoT and IIoT datasets [11,12]. However, many studies still focus mainly on overall accuracy. They often provide limited analysis of minority-class performance, false alarm behavior, feature redundancy, and model interpretability. In industrial environments, these limitations are serious because a high overall accuracy may still hide poor detection of rare but dangerous attacks. Deep learning has also been widely explored for IIoT intrusion detection. CNN, LSTM, BiLSTM, autoencoder, graph neural network, and transformer-based models can extract spatial, temporal, and high-dimensional traffic patterns from network data [13,14]. These models are useful when traffic records contain sequence relations, packet-level dependencies, or complex interactions among devices. Autoencoder-based methods can learn normal traffic representations and identify deviations from expected behavior. Recurrent and attention-based models can capture temporal dependencies in traffic

sequences. Graph neural networks can represent communication relationships among hosts, services, and industrial devices. Recent studies have also used pruning, quantization, knowledge distillation, and lightweight neural design to reduce deployment cost on edge devices [15,16]. These methods show that deep learning has strong potential for complex IIoT security tasks. Despite these advantages, deep learning-based intrusion detection still has limitations in industrial deployment. Deep models usually require larger training datasets, higher computing resources, more parameter tuning, and longer training time than tree-based models. Their inference cost may also be unsuitable for industrial gateways, embedded controllers, and edge devices with limited CPU, memory, and power resources. More importantly, their prediction process is often difficult to understand. In industrial cyber-physical systems, a black-box alarm is not enough. Security operators need to know whether an alarm is related to abnormal packet size, unusual protocol service behavior, excessive connection attempts, irregular byte rates, or suspicious temporal patterns. Such explanation is necessary for fault tracing, attack analysis, equipment protection, and safety review. A model that only outputs a class label may be difficult to use in production environments, especially when false alarms may interrupt normal industrial operations [17]. Therefore, a simpler, more efficient, and more interpretable ensemble model remains valuable for many IIoT monitoring tasks. Dataset quality is another important issue in IIoT intrusion detection research. Public datasets such as ToN-IoT and Edge-IIoTset are widely used because they contain labeled traffic samples, multiple attack types, and features collected from IoT or IIoT testbeds. These datasets provide useful benchmarks for comparing intrusion detection methods. They also make it possible to evaluate model behavior under different attack categories. However, benchmark datasets cannot fully represent all real industrial networks. Industrial traffic is affected by device types, protocol settings, control logic, production cycles, network topology, and maintenance operations. A model trained on one public dataset may not perform equally well in another industrial environment. Recent studies have reported that commonly used intrusion detection datasets may contain repeated records, noisy labels, unstable feature distributions, limited attack diversity, and traffic patterns that differ from field systems [18,19]. These issues may produce overly optimistic results if the dataset is not carefully cleaned and evaluated. Another problem is that some studies use only small subsets of public datasets to reduce training time. Although this practice can simplify experiments, it may weaken the reliability of the reported results. A small subset may not preserve the original class distribution, attack diversity, or feature variation of the full dataset. This is especially problematic for IIoT intrusion detection because minority attacks may already have limited samples. Removing too many records can make the model less capable of learning rare but important attack behaviors. In addition, random train-test splits from the same dataset may cause overly high performance when repeated or highly similar records appear in both training and testing sets. Therefore, a reliable IIoT intrusion detection study should use large-scale samples, apply strict feature cleaning, remove repeated or weak features, and report metrics beyond accuracy. Macro-F1, AUC, detection rate, false alarm rate, and class-level performance are necessary for judging whether the model is useful in practical industrial security monitoring. Class imbalance is a common and critical problem in IIoT security datasets. In real industrial networks, normal traffic usually accounts for most network flows, while some attack types contain only a small number of samples. Under this condition, a model may obtain high overall accuracy by favoring the majority class. However, it may fail to detect minority attacks that are important for safety. Missing a small number of command injection, data tampering, or man-in-the-middle attacks may be more dangerous than misclassifying a large number of low-risk samples. Several studies have used resampling, cost-sensitive learning, generative methods, and balanced classifiers to address this problem [20,21]. Balanced Random Forest is useful in this setting because it reduces majority-class dominance during tree construction. XGBoost and Gradient Boosting can also improve class separation by learning difficult samples step by step. However, single classifiers may still be affected by feature distribution changes, attack category differences, and dataset-specific patterns. A hybrid ensemble model can reduce this

risk by combining classifiers with different learning mechanisms. The diversity of IIoT attacks further increases the difficulty of intrusion detection. DDoS attacks often show high traffic volume, repeated connection attempts, and abnormal packet rates. Reconnaissance attacks may involve scanning behavior, unusual service access, or repeated probing of ports and protocols. Injection attacks may change request structures or generate abnormal payload-related traffic features. Man-in-the-middle attacks may alter communication behavior, introduce delay, or change traffic direction patterns. These attack types differ in duration, packet behavior, byte ratio, protocol service pattern, and temporal distribution. A single model may not capture all these differences effectively. Hybrid ensemble learning is therefore suitable because it can integrate complementary decision boundaries from different classifiers. When combined with feature selection and interpretation, such a model can not only improve detection reliability, but also reveal which traffic indicators are most important for different abnormal behaviors. Feature complexity is another practical issue. IIoT datasets often contain many flow-level, statistical, and protocol-related variables. Some features may be highly correlated, repeated, weakly informative, or unstable across datasets. Using all features may increase training cost, inference cost, and overfitting risk. It may also make model explanation more difficult because redundant variables can distort feature importance analysis. Feature selection is therefore necessary for building efficient and interpretable intrusion detection models. Mutual information ranking can measure the dependency between each feature and the target label, while correlation filtering can reduce repeated information among variables. By combining these two steps, the model can retain informative traffic indicators and remove redundant features. This process is important for IIoT systems because industrial security models should be efficient enough for deployment and clear enough for operational analysis. Model explanation is especially important for intrusion detection in industrial cyber-physical systems. In IIoT environments, detection results must be understandable because security events may affect physical equipment, production processes, and safety decisions. Explainable artificial intelligence methods, such as SHAP and LIME, have been used to show how each feature affects intrusion detection results [22,23]. SHAP is particularly suitable for tree-based ensemble models because it can provide both global feature importance and local explanations for individual samples. Global explanations can show which traffic features are generally important for detecting attacks. Local explanations can show why a specific sample is classified as malicious or normal. These explanations are useful for alert verification, model auditing, attack diagnosis, feature validation, and security reporting. In industrial environments, this transparency can help operators distinguish real attacks from false alarms and understand whether the model is using meaningful traffic behaviors rather than dataset artifacts. Recent studies have shown that explainable methods can improve model trust, model checking, and security decision support in IoT, IIoT, and Industry 5.0 contexts [24,25]. However, many existing explainable IDS studies still treat interpretation as a post-processing step. They often provide feature importance plots after model training, but do not fully connect the explanation results with feature screening, class imbalance, industrial protocol behavior, and model robustness. As a result, the explanation may remain descriptive and may not directly support practical security analysis. For IIoT intrusion detection, interpretability should be integrated into the whole modeling process. Feature selection should reduce unnecessary variables before training. Ensemble learning should improve robustness under imbalanced traffic distributions. SHAP analysis should then explain whether the retained features correspond to meaningful industrial traffic behaviors. This integrated design can make the model more useful for real security operations. The above discussion shows that several limitations remain in current IIoT intrusion detection research. Many studies report high accuracy but provide insufficient evidence on macro-level performance under imbalanced attack categories. Some models use large feature spaces without adequate feature screening, which may increase computational cost and introduce repeated or unstable variables. Deep learning methods can achieve strong results on benchmark datasets, but their black-box nature and computing cost may limit their use in safety-critical industrial systems. Many explainable IDS studies also fail to closely link

feature interpretation with industrial network behavior. These limitations indicate the need for a detection method that combines stable classification, reduced feature complexity, class-imbalance handling, and clear explanation of traffic indicators. To address these issues, this study presents an explainable multi-feature network intrusion detection method for IIoT systems using hybrid ensemble learning. The proposed model combines XGBoost, Gradient Boosting, and Balanced Random Forest to improve robustness under imbalanced traffic distributions. Mutual information ranking and correlation filtering are used to reduce the original feature space from 79 to 36 features. This step removes repeated variables and retains the most useful protocol-level, statistical, and temporal traffic indicators. SHAP values are then used to explain the contribution of important features to the final prediction. The model is evaluated on ToN-IoT and Edge-IIoTset, which contain large-scale labeled samples and mixed attack categories. The evaluation considers not only overall accuracy, but also macro-F1, AUC, detection rate, false alarm rate, and class-level detection behavior. In this way, the proposed method is assessed from both predictive performance and practical security relevance. The purpose of this study is to improve the reliability and transparency of IIoT intrusion detection under imbalanced and feature-complex traffic conditions. Its significance lies in connecting hybrid ensemble learning, feature reduction, class-imbalance handling, and SHAP-based interpretation within one intrusion detection framework. This design helps the model detect abnormal traffic while also explaining which industrial traffic behaviors are most related to attack decisions. The proposed method can support security operators in verifying alarms, analyzing attack patterns, reducing unnecessary feature complexity, and improving trust in automated detection results. By using large-scale IIoT benchmark datasets and interpretable ensemble modeling, this study provides a practical and transparent solution for industrial network security monitoring. It also offers useful evidence for the development of explainable intrusion detection systems in smart manufacturing, industrial edge computing, and cyber-physical infrastructure protection.

2.Materials and Methods

2.1 Sample and Study Object Description

This study used two public Industrial Internet of Things datasets, ToN-IoT and Edge-IIoTset. The ToN-IoT dataset included 461,043 labeled traffic samples, and the Edge-IIoTset dataset included 2,219,201 labeled traffic samples. Each sample represented one network traffic record collected from an IoT or IIoT testbed. The datasets contained normal traffic and several attack types, including DDoS, reconnaissance, injection, man-in-the-middle attacks, password attacks, scanning behavior, and abnormal service access. The traffic records included protocol-level, statistical, and temporal features, such as packet inter-arrival time, byte ratio, TCP window size, service frequency, flow duration, packet count, and source-to-destination communication behavior. The study focused on abnormal industrial network traffic under mixed attack conditions. The main aim was to test whether a hybrid ensemble model could detect malicious IIoT traffic and provide clear feature-level explanations.

2.2 Experimental Design and Control Setting

The experiment compared the proposed hybrid ensemble model with single-model and deep learning baselines under the same data conditions. The experimental group was the proposed model, which combined XGBoost, Gradient Boosting, and Balanced Random Forest. These three classifiers were selected because they can process structured traffic features and reduce the effect of imbalanced attack classes. XGBoost and Gradient Boosting improved class separation by learning difficult samples step by step. Balanced Random Forest reduced bias toward the majority class during tree construction. The control group included standalone XGBoost, standalone Gradient Boosting, standalone Balanced Random Forest, and a CNN-based baseline model. All models used the same training set, testing set, feature set, and preprocessing steps. This setting ensured a fair comparison. The design was based on the idea that a hybrid ensemble can combine different learning mechanisms and produce more stable detection results than a single classifier.

2.3 Measurement Methods and Quality Control

All traffic records were checked before model training. Missing values, infinite values, duplicate records, abnormal feature ranges, and inconsistent label formats were identified. Invalid records were removed, and feature names were unified across the datasets. Text labels were converted into numerical labels for model training. Normal traffic was coded as benign traffic, while attack records were coded according to their attack categories. Class distribution was examined before training to identify sample imbalance. Feature quality was controlled through mutual information ranking and correlation filtering. Features with weak class relevance or strong overlap with other features were removed. The original feature space was reduced from 79 to 36 features. To avoid data leakage, feature selection and class balancing were performed only on the training set. The testing set was kept independent during the whole experiment. Model performance was measured using accuracy, macro-F1, AUC, detection rate, precision, recall, and false alarm rate. Macro-F1 was used as a main metric because it gives equal weight to each class and is more suitable for imbalanced IIoT traffic than accuracy alone.

2.4 Data Processing and Model Formulas

The raw IIoT traffic records were processed before model training. Invalid values and duplicate samples were removed first. Then, categorical labels were encoded into numerical form. Mutual information ranking was used to measure the relation between each feature and the class label. Correlation filtering was then used to remove highly repeated features. The cleaned data were divided into training and testing subsets. The hybrid ensemble model combined the predicted probabilities of XGBoost, Gradient Boosting, and Balanced Random Forest. For a given traffic sample x_i, the final probability of class c was calculated as follows:

$$P(c|x_i)=\frac{1}{K}\sum_{k=1}^{K} P_k(c|x_i)$$

where K is the number of base classifiers, and $P_k(c|x_i)$ is the probability assigned to class c by the k-th classifier. The final class was assigned to the category with the highest average probability. Macro-F1 was calculated to evaluate the model under imbalanced class distribution:

$$\text{Macro-F1}=\frac{1}{C}\sum_{c=1}^{C}\frac{2\times \text{Precision}_c\times \text{Recall}_c}{\text{Precision}_c+\text{Recall}_c}$$

where C is the number of traffic classes, Precision_c is the precision of class c, and Recall_c is the recall of class c. This metric was used because minority attack classes should not be hidden by the large number of normal traffic samples.

2.5 Model Explanation and Statistical Evaluation

SHAP analysis was used to explain the prediction results of the proposed hybrid ensemble model. SHAP values were calculated for the selected 36 traffic features. They measured the contribution of each feature to each prediction. Global SHAP analysis was used to rank the overall importance of protocol-level, statistical, and temporal features. Local SHAP analysis was used to explain single abnormal traffic samples. This step helped check whether the model decisions were based on meaningful IIoT traffic behavior. Particular attention was given to packet inter-arrival time, source-to-destination byte ratio, TCP window size, and protocol service frequency. These variables were examined because they are closely related to abnormal communication patterns in industrial networks. The final results were reported as mean values from repeated runs under the same settings. The comparison focused on detection accuracy, macro-F1, AUC, detection rate, false alarm rate, feature reduction effect, and explanation quality. This evaluation assessed both detection reliability and decision transparency in IIoT security monitoring.

3.Results and Discussion

3.1 Detection Performance on ToN-IoT

The proposed hybrid ensemble model showed strong results on the ToN-IoT dataset. It achieved 97.93% accuracy, 96.88% macro-F1, and 98.57% AUC. These results show that the model separated normal industrial traffic from abnormal traffic with high reliability. Macro-F1 is important for this dataset because the attack classes are imbalanced. Accuracy alone may hide poor detection of small attack classes, while macro-F1 gives each class the same weight. Compared with the single XGBoost model, the proposed model improved macro-F1 by 1.76%. Compared with the CNN-based baseline, the improvement was 2.41%. This result indicates that the hybrid ensemble structure was more stable than a single tree model [26,27]. It was also more suitable for structured traffic features than the CNN baseline. XGBoost and Gradient Boosting improved class separation by learning difficult samples step by step. Balanced Random Forest reduced the bias caused by frequent normal or attack classes. A recent Springer study on Edge-IIoTset also reported that feature-level ensemble learning performed well under imbalanced IoT traffic conditions. Its WO-XGB model achieved 99.98% accuracy and 99.97% precision, recall, and F-score on Edge-IIoTset. This supports the use of ensemble models for IoT and IIoT intrusion detection. The comparison is shown in Fig. 1.

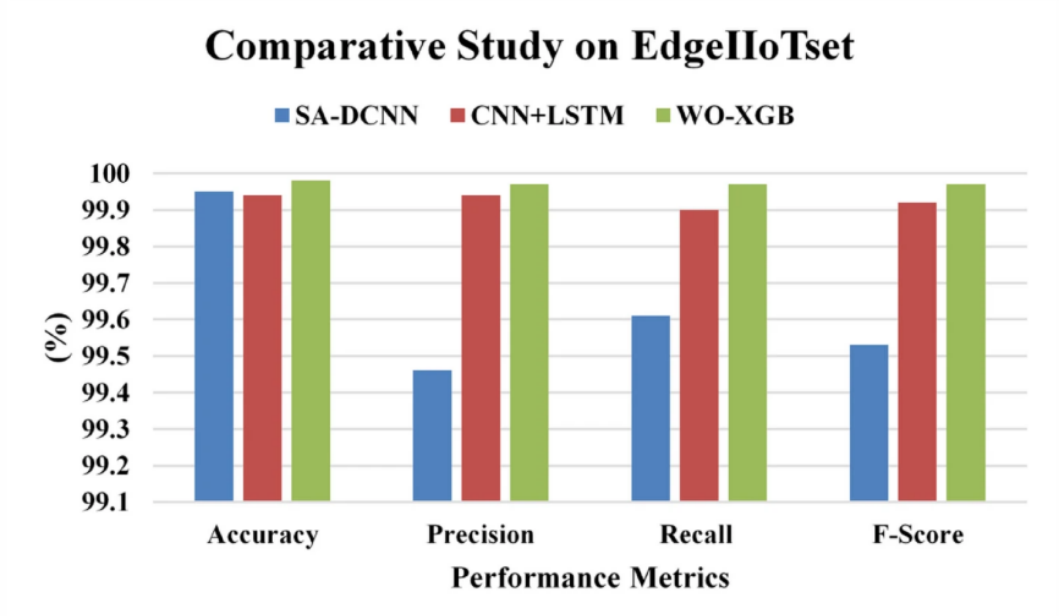


Fig.1.Comparison of detection results between the proposed hybrid ensemble model and baseline IDS models.

3.2 Detection Results on Edge-IIoTset Under Mixed Attacks

The proposed framework also performed well on Edge-IIoTset. It maintained a detection rate of 98.21% under mixed attack categories, including DDoS, reconnaissance, injection, and man-in-the-middle attacks. These attacks differ in traffic duration, packet direction, byte ratio, protocol service pattern, and connection behavior. Therefore, a reliable model must learn several types of traffic signals rather than depend on one feature group. The use of 36 selected features reduced noise from weak or repeated variables. It also kept key protocol-level, statistical, and temporal indicators. Compared with deep learning studies on Edge-IIoTset, the proposed model provided a better balance between performance and feature clarity. For example, a 2022 MDPI study using Inception Time reported 94.94% accuracy on Edge-IIoTset in a multi-class setting. The authors also noted memory limits during experiments under limited CPU resources. In contrast, the present model reached a higher detection rate with a smaller feature set and a clearer explanation structure. This result suggests that carefully selected flow features and balanced ensemble learning are effective for industrial traffic data, especially when computing cost and model clarity are both important [28].

3.3 Feature Reduction and SHAP-Based Interpretation

Feature reduction improved the clarity and stability of the detection model. The original feature space was reduced from 79 to 36 features by mutual information ranking and correlation filtering. This step removed weak and repeated variables. It also kept features with stronger class relevance. The SHAP results showed that packet inter-arrival time, source-to-destination byte ratio, TCP window size, and protocol service frequency were the most important features for detecting abnormal industrial traffic. These variables have clear security meanings. Packet inter-arrival time reflects abnormal timing patterns, which may appear in DDoS or scanning traffic. The source-to-destination byte ratio shows whether communication is balanced or one-sided. TCP window size can indicate abnormal connection behavior. Protocol service frequency reflects repeated access to specific services, which is common in reconnaissance and injection attacks. These findings show that the model did not depend on unclear statistical signals. Instead, it used traffic features that match known industrial attack behavior. Recent MDPI work on explainable intrusion detection also showed that SHAP can identify feature-level contributions and improve the interpretation of IDS outputs [29,30]. The explanation result is shown in Fig. 2.

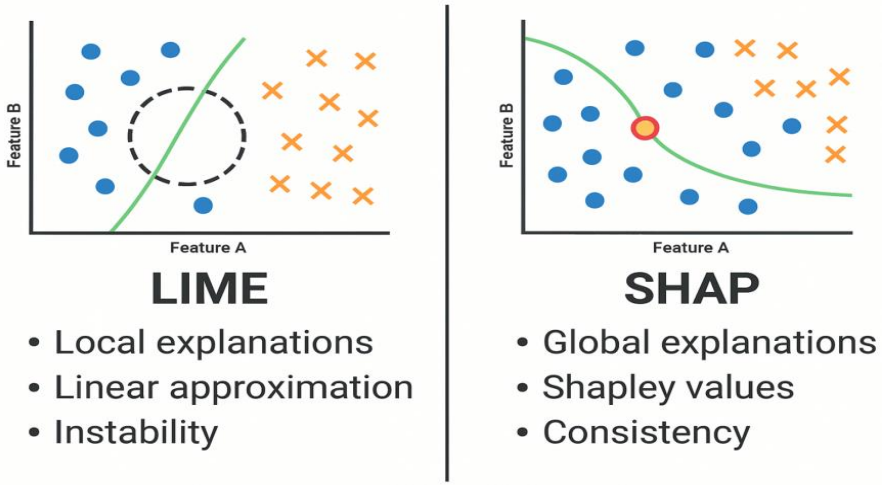


Fig.2.SHAP ranking of important traffic features for explaining IIoT attack detection.

3.4 Practical Implications and Remaining Limitations

The results show that the proposed method has practical value for IIoT security monitoring. It combines reliable detection, reduced feature complexity, and clear feature-level explanation. This combination is important for industrial cyber-physical systems because alarms may affect production decisions, equipment inspection, and safety response. A black-box model with high accuracy may still be hard to use if operators cannot understand why an alarm is generated. The SHAP results link model output with specific traffic behavior. This can support attack tracing and rule updating. Balanced Random Forest also reduced the effect of imbalanced samples, which is common in industrial traffic datasets. However, this study still has several limitations. First, the experiments were based on public benchmark datasets, and these datasets cannot fully represent all real industrial environments. Second, the model was tested on labeled offline traffic rather than long-term live traffic streams. Third, this study mainly focused on detection and explanation. Online model updating under new attack patterns was not fully examined. Future work should test the model in real industrial networks, include more recent attack samples, and examine whether feature importance remains stable when network conditions change.

4.Conclusion

This study proposed an explainable hybrid ensemble model for network intrusion detection in Industrial Internet of Things systems. The model combined XGBoost, Gradient Boosting, and Balanced Random Forest to improve detection under imbalanced traffic conditions. Mutual information ranking and correlation filtering reduced the feature space from 79 to 36 features. This step removed repeated variables and kept the main traffic indicators. The proposed model achieved 97.93% accuracy, 96.88% macro-F1, and 98.57% AUC on ToN-IoT. It also reached

a detection rate of 98.21% on Edge-IIoTset under mixed attack types, including DDoS, reconnaissance, injection, and man-in-the-middle attacks. These results show that the model can detect abnormal industrial traffic with stable performance and good class-level balance. SHAP analysis showed that packet inter-arrival time, source-to-destination byte ratio, TCP window size, and protocol service frequency were the main features for attack detection. This result indicates that the model decisions were linked to clear industrial traffic behavior. The main value of this study is that it combines reliable detection, reduced feature complexity, and clear model explanation in one framework. This provides a useful method for IIoT security monitoring, especially in industrial systems where detection results must be accurate and easy to understand. The proposed model can be used in smart factories, industrial gateways, cyber-physical systems, and edge-based security platforms. However, this study still has some limitations. The experiments were based on public benchmark datasets, and the model was not tested on long-term real industrial traffic. In addition, the current work mainly focused on offline detection and feature explanation. Future studies should test the model in real industrial networks, include newer attack samples, and examine its stability under changing traffic conditions.

References

Zhang, Z., Tong, Y., & Gao, Y. (2026). Causal Representation Learning for Explainable Early Warning in High-Stakes Decision Systems.

Rojas, L., Peña, Á., & Garcia, J. (2025). AI-driven predictive maintenance in mining: A systematic literature review on fault detection, digital twins, and intelligent asset management. Applied Sciences, 15(6), 3337.

Shao, W. (2026, March). Interpretable Ensemble Learning for Network Traffic Anomaly Detection: A SHAP-based Explainable AI Framework for Embedded Systems Security. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-4). IEEE.

Elshewey, A. M. (2026). An Enhanced Approach for Edge-Based Intrusion Detection Based on a Hybrid Deep Learning Model CNN-DNN. SN Computer Science, 7(3), 271.

Qi, C., & Qiao, X. (2026). Efficient Data Sampling and Feature Selection Algorithms for Scalable Machine Learning Pipelines.

Kumar, L. S., Nethi, S. R., Uyyala, R., Vurubindi, P., Narahari, S. C., Das, A. K., ... & Alenazi, M. J. (2026). Anomaly-based intrusion detection on benchmark datasets for network security: A comprehensive evaluation. Scientific Reports, 16(1), 8507.

Su, D., & Zhang, H. Developing a Data-Driven Computational Framework for Scalable Special Education Practices for Autism and Intellectual Disabilities in the US Public School System.

Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. International Journal of Machine Learning and Cybernetics, 16(9), 6937-6956.

Huang, R. (2026). ConfSched: Confidence-Threshold Routing for Efficient Edge-Cloud Activity Recognition. World Journal of Engineering and Technology, 14(2), 471-486.

Kudin, M., Kurniawan, A., & Insani, C. N. (2026). Analysis of Diabetes Classification Performance Improvement Using Ensemble Bagging and K-Fold. Jurnal Sistem Cerdas, 9(1), 98-113.

Ma, Y. (2026). Industrial Financial Risk Prediction Model Based on Graph Neural Network and Knowledge Graph Inference. Journal of Circuits, Systems and Computers, 35(16), 2650103.

Potharaju, S., Tambe, S. N., Tirandasu, R. K., Kumar, D. A., & Kantipudi, M. P. (2026). Enhancing Cybersecurity in Industrial Internet of Things Systems Using Ensemble Learning Against False Data Injection Attacks. Journal of Current Science and Technology, 16(1), 151-151.

Gao, G., Gao, R., Lu, C., Gao, R., & Kuang, Y. (2026, March). Security Governance Methods and Quantitative Evaluation for Enterprise SMS and Verification Code Systems. In 2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS) (pp. 455-458). IEEE.

Zhang, C., Li, J., Wang, N., & Zhang, D. (2025). Research on intrusion detection method based on transformer and CNN-BiLSTM in internet of things. Sensors, 25(9), 2725.

Chen, F., Liang, H., Yue, L., Xu, P., & Li, S. (2025). Low-Power Acceleration Architecture Design of Domestic Smart Chips for AI Loads.

Liang, R., Fan, F., Liang, Y., & Li, S. (2025). Constructing an Adaptive Optimization Model for Ribbon Recommendation and Interface for User Habits.

Sorostinean, R., Neghina, C., & Gellert, A. (2025). Boosting anomaly detection with unsupervised K-Means and SOM for energy-efficient factory machines. Journal of Intelligent Manufacturing, 1-17.

Wang, Y., Wang, Y., Yin, X., Arias, R., & Chen, J. (2026). Research on Dynamic Assessment of Glucose-Lipid Metabolism and Personalized Drug Response Prediction Based on Wearable Multimodal Sensing.

Eren, K. K., Kucuk, K., Saleh, R. A., Konyar, M. Z., Hardy, O. M., & Khan, S. A. (2026). Distributional drift in IoT intrusion detection systems: implications for cross-dataset generalization. Electronics, 15(11).

Zhu, W., Yang, J., & Yao, Y. (2025). How Cross-Departmental Collaboration Structures Mitigate Cross-Border Compliance Risks: Network Causal Inference Based on ManpowerGroup's Staffing Projects.

Adiputra, I. N. M., Wanchai, P., & Lin, P. C. (2025). Optimized customer churn prediction using tabular generative adversarial network (GAN)-based hybrid sampling method and cost-sensitive learning. PeerJ Computer Science, 11, e2949.

Nian, J., Liang, Z., Liu, H., & Yang, M. (2022). An Efficient Fault-Tolerant Protection Method for L0 BTB. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(3), 1284-1297.

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221.

Gong, M., Deng, Y., Qi, N., Zou, Y., Xue, Z., & Zi, Y. (2025, August). Structure-learnable adapter fine-tuning for parameter-efficient large language models. In International Conference on Artificial Intelligence and Computational Engineering (AICE 2025) (Vol. 2025, pp. 225-230). IET.

Alatawi, M. N. (2025). SAFEL-IoT: secure adaptive federated learning with explainability for anomaly detection in 6G-enabled Smart Industry 5.0. Electronics, 14(11), 2153.

Yin, J., Rao, H., & Huang, Y. (2026, March). Dynamic Modeling and Heterogeneity Analysis of Platform User Behavior Time Series. In 2026 International Conference on AI in Education Technology and Applications (AIETA) (pp. 30-33). IEEE.

Kayit, A. D., & Ismail, M. T. (2025). Leveraging hybrid ensemble models in stock market prediction: A data-driven approach. Data Science in Finance and Economics, 5(3), 355.

Deng, X., & Yang, J. (2026). Design of a Quantitative Futures Trading Model Incorporating Multimodal Feature Fusion and Tail Risk Control. Available at SSRN 6702398.

Abdulganiyu, O. H., Fadi, O., Moukafih, Y., Ait Tchakoucht, T., Saheed, Y. K., Chukwuere, J. E., & Yau, S. (2026). Explainable attention based few shot LSTM for intrusion detection in imbalanced cyber physical system networks. Scientific Reports.

Zhang, Z. (2026). A Study on the Identification of Manipulative Design in Subscription and Payment Interfaces of Digital Consumer Platforms and Its Behavioral Effects. Available at SSRN 6734760.