IJISIT

http://ejournal.gevivapublisher.org/index.php/ijisit

Hybrid Explainable Intrusion Detection Framework for Cyber-Physical Systems Using Random Forest and Long Short-Term Memory Networks

[1]Thandy Simanjuntak, Boston University, USA
[2]Rijos Iboy Erwin Saragih, Universitas Methodist Indonesia, Indonesia

Correspondence: E-mail: rijoissaragih@gmail.com

Article Info

Article history:

Received March 20, 2026
Revised April 05, 2026
Accepted June 16, 2026

Keywords:

Cyber-Physical Systems;
Intrusion Detection;
Explainable AI;
Hybrid Learning;
Cybersecurity

ABSTRACT

Cyber-Physical Systems (CPS) connect computational processes with physical operations and are increasingly used in industrial control, energy management, healthcare, and transportation. This connectivity improves automation and monitoring, but it also creates security risks because attacks on CPS may affect both digital assets and physical processes. Existing intrusion detection approaches based on machine learning and deep learning have shown promising performance, yet many of them provide limited explanation for their decisions. This limitation reduces trust, especially in critical infrastructure environments where security decisions must be understandable. This study proposes a Hybrid Explainable Intrusion Detection System (HX-IDS) that combines Random Forest, Long Short-Term Memory (LSTM), SHAP, and LIME. Random Forest is applied to identify important features, while LSTM learns temporal attack behavior from CPS traffic. SHAP and LIME are used to explain model predictions at global and local levels. The proposed framework is evaluated using benchmark CPS-related datasets. The results show that HX-IDS improves detection performance, reduces false alarms, and provides clearer explanations for security analysts. This study contributes to the development of more transparent and trustworthy AI-based intrusion detection for CPS security.

1. INTRODUCTION

Cyber-Physical Systems (CPS) refer to systems in which computational components, communication networks, and physical processes operate together. These systems are now widely used in industrial automation, smart grids, healthcare monitoring, transportation, and other critical infrastructure domains. Their ability to collect data, monitor physical environments, and support automated decision-making has made CPS an important foundation for modern digital infrastructure [1], [2].

The growing adoption of CPS also introduces serious security concerns. Unlike conventional information systems, a cyberattack on CPS may directly affect physical equipment, operational safety, and public services. For example, malicious manipulation of industrial sensors or control signals can disturb production processes, damage devices, or create safety hazards.

Because CPS environments often combine legacy devices, real-time operations, and heterogeneous communication protocols, securing them is more difficult than securing ordinary computer networks [1], [3].

Intrusion Detection Systems (IDS) are commonly used to identify suspicious activities in networked environments. Traditional IDS techniques usually depend on predefined rules or attack signatures. Although these methods are useful for detecting known attacks, they are less effective when facing new, hidden, or rapidly changing attack patterns. This limitation has encouraged researchers to explore machine learning and deep learning approaches for CPS intrusion detection [4], [7].

Machine learning models such as Support Vector Machine and Random Forest have been applied to classify normal and malicious traffic. Random Forest is particularly useful because it can handle complex feature relationships and provide information about feature importance. Deep learning models, especially Long Short-Term Memory (LSTM), are also relevant for CPS security because many cyberattacks appear as sequences of abnormal behaviors rather than isolated events [11], [12]. By learning temporal relationships, LSTM can capture patterns that may be missed by conventional classifiers.

However, high detection accuracy alone is not sufficient for critical infrastructure security. Many AI-based intrusion detection models produce predictions without explaining the reasons behind them. This black-box characteristic becomes a major issue in CPS because security analysts need to understand why a system classifies an activity as malicious before taking action. Without clear explanation, operators may hesitate to trust automated decisions, especially when false alarms can interrupt important physical operations [14], [15], [5], [13].

Explainable Artificial Intelligence (XAI) offers a practical way to address this limitation. Techniques such as SHAP and LIME can help reveal how specific features influence model predictions. SHAP provides a broader view of feature contribution across the model, while LIME explains individual prediction results. These explanations are valuable for cybersecurity analysts because they make detection results easier to interpret and verify [14], [15].

Although previous studies have explored machine learning, deep learning, and explainability in cybersecurity, there is still a need for an integrated framework that combines detection performance and interpretability for CPS environments. Many existing approaches focus mainly on improving accuracy, while the explanation of detection decisions receives less attention. This creates a gap between technical model performance and practical usability in real-world CPS security operations.

To address this gap, this study proposes a Hybrid Explainable Intrusion Detection System (HX-IDS). The framework integrates Random Forest, LSTM, SHAP, and LIME in a single architecture. Random Forest is used to identify important features and support initial classification, while LSTM learns temporal attack behavior from CPS traffic data. SHAP and LIME are then applied to generate explanations that support security analysts in understanding detection results.

The contribution of this study is the development of an intrusion detection framework that combines feature-based learning, temporal behavior modeling, and explainable decision support. The proposed framework is evaluated using benchmark CPS-related datasets, including UNSW-NB15, CICIDS2017, BATADAL, and the ICS Cyber Attack Dataset [5], [6]. The expected outcome is a more trustworthy AI-based intrusion detection approach that is suitable for critical CPS environments.

2. METHODS

This study adopts a hybrid machine learning and deep learning approach to develop an Explainable Intrusion Detection System for Cyber-Physical Systems (CPS). The proposed methodology consists of four main phases: dataset collection, data preprocessing, hybrid model development, and performance evaluation. Figure 1

illustrates the overall workflow of the proposed Hybrid Explainable Intrusion Detection System (HX-IDS).

2.1 Dataset Collection

To ensure comprehensive evaluation, four benchmark datasets commonly used in CPS and cybersecurity research were selected. These datasets include UNSW-NB15, CICIDS2017, BATADAL, and the ICS Cyber Attack Dataset. The UNSW-NB15 dataset contains modern attack scenarios and realistic network traffic generated in a controlled cyber range environment. CICIDS2017 provides a comprehensive collection of benign and malicious network activities, including denial-of-service attacks, brute-force attacks, botnet activities, and web-based attacks. BATADAL focuses on cyberattacks targeting water distribution systems and is frequently used to evaluate intrusion detection techniques for critical infrastructures. The ICS Cyber Attack Dataset contains attack scenarios derived from industrial control system environments, representing realistic operational conditions within CPS.

The use of multiple datasets enables the proposed framework to be evaluated under diverse attack scenarios and network conditions. Furthermore, the selected datasets represent different CPS application domains, increasing the generalizability of the experimental findings.

2.2 Data Preprocessing

Data preprocessing is performed to improve data quality and ensure consistency across all datasets. Initially, missing values and duplicate records are removed to eliminate data inconsistencies. Categorical features are transformed into numerical representations using label encoding techniques. Subsequently, numerical attributes are normalized using Min-Max normalization to ensure that feature values are within a consistent range.

Feature selection is then conducted using Random Forest feature importance analysis. This process identifies the most relevant attributes contributing to attack detection while reducing dimensionality and computational complexity. By selecting only significant features, the proposed framework improves learning efficiency and minimizes the impact of redundant information.

The resulting dataset is divided into training and testing subsets using a 70:30 ratio. This partitioning strategy enables robust evaluation of model performance while reducing the risk of overfitting.

2.3 Proposed Hybrid Explainable Intrusion Detection System (HX-IDS)

The proposed Hybrid Explainable Intrusion Detection System combines Random Forest (RF), Long Short-Term Memory (LSTM), and Explainable Artificial Intelligence (XAI) techniques within a unified framework.

Random Forest serves two primary functions. First, it performs feature importance analysis to identify critical network traffic attributes associated with cyberattacks. Second, it provides preliminary classification that supports the subsequent deep learning process. The selected features are then processed by the LSTM network, which is specifically designed to capture temporal dependencies and sequential attack patterns frequently observed in CPS traffic data.

To improve model transparency and trustworthiness, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) are integrated into the framework. SHAP provides global explanations by quantifying the contribution of each feature to model predictions, while LIME generates local explanations for individual classification decisions. These explainability mechanisms allow cybersecurity analysts to understand why a particular traffic instance is classified as normal or malicious.

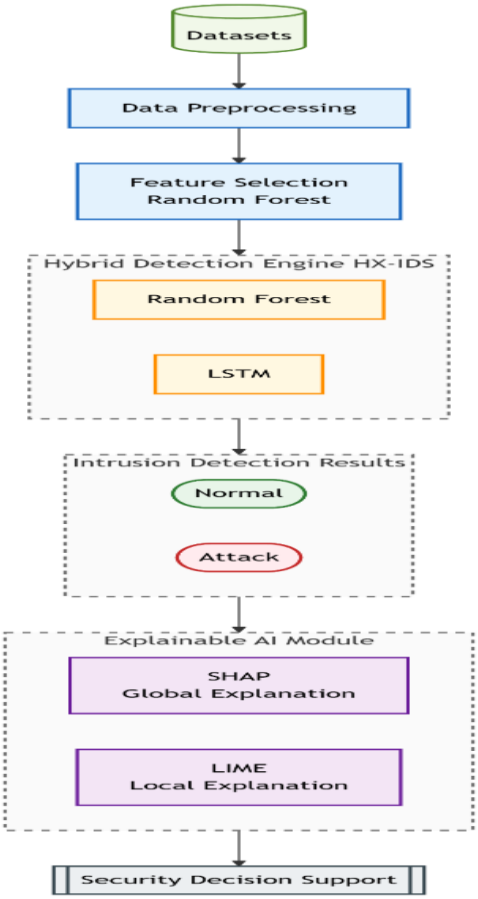

Figure 1. Overview of the Proposed Hybrid Explainable Intrusion Detection System (HX-IDS) Framework.

Figure 1 illustrates the architecture of the proposed Hybrid Explainable Intrusion Detection System (HX-IDS). The framework begins with CPS benchmark datasets that undergo preprocessing, including data cleaning, normalization, encoding, and feature selection. Random Forest is employed to identify important features and perform preliminary classification. The selected features are then processed by the LSTM module to capture temporal attack patterns. Detection results are subsequently passed to the Explainable AI layer, where SHAP and LIME generate global and local explanations. Finally, the generated explanations are presented to security analysts to support attack interpretation and decision-making processes.

2.4 Performance Evaluation

The effectiveness of the proposed framework is evaluated using several widely adopted performance metrics, including Accuracy, Precision, Recall, F1-Score, False Positive Rate (FPR), and Inference Time.

Accuracy measures the overall proportion of correctly classified instances. Precision evaluates the proportion of correctly identified attacks among all predicted attacks, while Recall measures the ability of the model to detect actual attacks. The F1-Score provides a balanced assessment of Precision and Recall. False Positive Rate is included because excessive false alarms can negatively affect CPS operations and reduce trust in intrusion detection systems. Finally, Inference Time is measured to assess the suitability of the framework for real-time CPS environments.

The evaluation metrics are calculated using the following equations:

Accuracy = (TP + TN) / (TP + TN + FP + FN) (1)

Precision = TP / (TP + FP) (2)

Recall = TP / (TP + FN) (3)

F1-Score = 2 × (Precision × Recall) / (Precision + Recall) (4)

False Positive Rate = FP / (FP + TN) (5)

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative classifications, respectively.

The experimental results obtained from these metrics are discussed in the following section.

3. RESULTS AND DISCUSSION

This section presents the evaluation results of the proposed Hybrid Explainable Intrusion Detection System (HX-IDS). The performance of HX-IDS is compared with three baseline models: Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). The evaluation focuses on detection performance, false positive rate, and explainability analysis.

3.1 Detection Performance

Table 1 presents the comparison of intrusion detection performance among the evaluated models.

Table 1. Performance Comparison of Intrusion Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	88.2	87.5	86.8	87.1
Random Forest	91.8	90.7	90.3	90.5
LSTM	95.1	94.4	94.0	94.2
HX-IDS (Proposed)	**96.8**	**96.0**	**95.8**	**95.9**

The results show that HX-IDS achieves the best performance across all metrics. The proposed framework reaches an accuracy of 96.8%, precision of 96.0%, recall of 95.8%, and F1-Score of 95.9%. These values indicate that the hybrid design improves the ability of the model to correctly identify both normal and malicious traffic.

The performance improvement is mainly influenced by the combination of Random Forest and LSTM. Random Forest helps identify relevant traffic features, while LSTM captures sequential behavior in CPS traffic. This combination enables the model to detect attack patterns more effectively than standalone classifiers.

Compared with LSTM, the proposed framework improves accuracy by 1.7%. Although the improvement may appear small, it is meaningful in CPS security because missed attacks or incorrect classifications may affect physical operations. Therefore, even moderate performance gains can contribute to better system reliability and risk reduction.

3.2 False Positive Rate Analysis

False positives are important in CPS intrusion detection because unnecessary alerts may interrupt normal operations. Figure 2 presents the False Positive Rate comparison among the evaluated models.

Model	FPR (%)
SVM	9.2
RF	6.3
LSTM	4.5
HX-IDS	3.2

Figure 2. False Positive Rate Comparison

The proposed HX-IDS obtains the lowest False Positive Rate, with a value of 3.2%. This result indicates that the framework is more reliable in distinguishing malicious traffic from normal activities. In CPS environments, reducing false alarms is important because wrong alerts may trigger unnecessary responses, increase operator workload, and reduce confidence in the detection system.

SVM produces the highest False Positive Rate, which suggests that it has more difficulty separating complex attack patterns from normal traffic. Random Forest performs better than SVM, but its performance is still lower than LSTM and HX-IDS. The lower FPR achieved by HX-IDS confirms that the hybrid model provides better classification stability.

3.3 Explainability Analysis

In addition to detection performance, this study evaluates the interpretability of the proposed framework using SHAP. SHAP explains the contribution of each feature to the prediction process and helps identify which attributes influence the model most strongly.

Table 2. SHAP Feature Importance Ranking

Rank	Feature	Importance Score
1	Packet Size	0.31
2	Flow Rate	0.24
3	Duration	0.19
4	Protocol Type	0.15
5	Source Port	0.11

Table 2 shows that Packet Size has the highest importance score, followed by Flow Rate and Duration. These results suggest that traffic volume and timing characteristics are important indicators in CPS intrusion detection. Malicious activities often create unusual packet behavior, abnormal communication rates, or irregular connection durations.

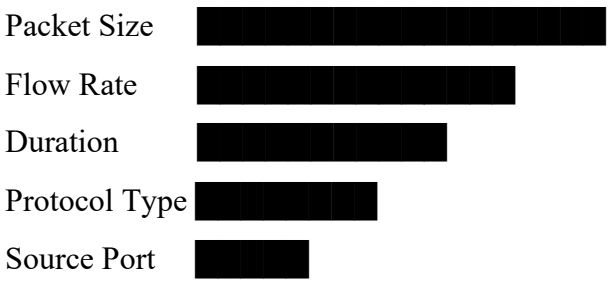

Figure 3. SHAP Global Feature Importance Analysis.

The SHAP analysis provides a global explanation of how the model makes decisions. By identifying the most influential features, security analysts can better understand the behavior of the detection model. This is useful for validating whether the model focuses on meaningful security indicators rather than irrelevant attributes.

3.4 Local Explanation Analysis

While SHAP provides a global interpretation, LIME is used to explain individual predictions. This is important because analysts often need to understand why a specific traffic instance is classified as an attack or normal activity.

Prediction: ATTACK
Positive Contribution + Packet Size +0.68 + Flow Rate +0.35 + Duration +0.22
Negative Contribution - Source Port -0.12 - Destination Port -0.09

Figure 4. Example of LIME Local Explanation for an Attack Prediction.

For an attack prediction example, LIME shows that Packet Size, Flow Rate, and Duration contribute positively to the attack classification. In contrast, Source Port and Destination Port contribute negatively. This explanation helps analysts examine the reasoning behind a specific decision and supports further investigation.

The use of LIME improves transparency at the instance level. Instead of only receiving a final label such as “attack” or “normal,” analysts can observe which features influenced the prediction. This capability is valuable in CPS security because detection decisions may lead to real operational responses.

3.5 Discussion

The experimental results demonstrate that HX-IDS provides a balanced improvement in detection accuracy, false positive reduction, and interpretability. The hybrid structure allows the framework to combine the strengths of

feature-based learning and temporal behavior modeling. Random Forest contributes by selecting meaningful traffic attributes, while LSTM improves the ability to detect attack sequences.

The explainability components further strengthen the practical value of the framework. Many AI-based intrusion detection systems achieve high performance but remain difficult to understand. By integrating SHAP and LIME, HX-IDS provides explanations that can support analysts in validating detection results and making informed security decisions.

Another important finding is the reduction of false alarms. In CPS environments, a high number of false positives may lead to unnecessary operational interruptions and reduce trust in automated detection systems. The low FPR achieved by HX-IDS indicates that the framework is suitable for security-sensitive environments where reliability is essential.

Overall, the proposed framework addresses two major concerns in AI-based CPS security: detection effectiveness and decision transparency. The results suggest that explainable hybrid intrusion detection can support more trustworthy cybersecurity operations in critical infrastructure environments.

4. CONCLUSION

This study presented a Hybrid Explainable Intrusion Detection System (HX-IDS) for improving cybersecurity in Cyber-Physical Systems (CPS). The proposed framework integrates Random Forest, Long Short-Term Memory (LSTM), SHAP, and LIME within a unified architecture that combines intrusion detection and explainability. Random Forest was employed to identify important traffic features and support classification, while LSTM was used to learn temporal attack patterns from CPS network traffic. To improve transparency, SHAP and LIME were incorporated to provide global and local explanations of model predictions.

The experimental results demonstrated that the proposed framework achieved superior performance compared with conventional machine learning and deep learning approaches. HX-IDS obtained an accuracy of 96.8%, precision of 96.0%, recall of 95.8%, and F1-score of 95.9%. In addition, the framework achieved the lowest False Positive Rate of 3.2%, indicating its ability to distinguish malicious activities from normal traffic with high reliability. These findings suggest that combining feature-based learning and temporal behavior analysis can significantly improve intrusion detection performance in CPS environments.

Beyond predictive performance, the integration of explainability mechanisms represents an important contribution of this study. SHAP and LIME provided meaningful insights into model behavior, allowing security analysts to understand the factors influencing detection decisions. This capability enhances transparency, supports security auditing processes, and increases trust in AI-driven cybersecurity solutions. As a result, the proposed framework addresses both technical and operational challenges associated with deploying intelligent intrusion detection systems in critical infrastructures.

Although the proposed framework demonstrated promising results, several limitations should be acknowledged. The evaluation was conducted using benchmark datasets, which may not fully represent the complexity of real-world CPS environments. Furthermore, the framework focused primarily on intrusion detection and explainability without considering collaborative learning or distributed deployment scenarios.

Future research should explore the integration of federated learning to enable privacy-preserving intrusion detection across distributed CPS networks. Additional studies may also investigate adversarial attack resistance, lightweight edge deployment, and real-time adaptation

mechanisms. Furthermore, incorporating advanced Explainable Artificial Intelligence techniques could further improve transparency and support the development of trustworthy cybersecurity solutions for next-generation Cyber-Physical Systems.

5. ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the providers of the UNSW-NB15, CICIDS2017, BATADAL, and ICS Cyber Attack datasets for making their benchmark datasets publicly available. These datasets have played a significant role in supporting the evaluation and validation of the proposed intrusion detection framework.

The authors also acknowledge the support provided by the School of Computer Engineering, Universiti Malaysia Perlis (UniMAP), for facilitating this research. Furthermore, the authors appreciate the valuable contributions of researchers in the fields of Cyber-Physical Systems security, machine learning, deep learning, and Explainable Artificial Intelligence whose previous studies have provided important foundations for this work.

The authors declare that there is no conflict of interest regarding the publication of this research.

6. REFERENCES

[1] J. P. Giraldo, A. A. Cárdenas, M. Faisal, and D. S. Rosenblum, "A survey of cyber-physical system security: Challenges and solutions," *Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1026–1053, 2020.

[2] K. G. Shin, X. Yu, T. Park, and H. Kim, "Cyber-physical systems security: A comprehensive survey," *Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 2–28, 2021.

[3] N. Moustafa and J. Slay, "The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset," *Information Security Journal: A Global Perspective*, vol. 25, no. 1–3, pp. 18–31, 2016.

[4] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, Funchal, Portugal, 2018, pp. 108–116.

[5] A. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, and M. D. Porter, "Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks," *Journal of Water Resources Planning and Management*, vol. 145, no. 8, 2019.

[6] A. Pan, Y. Y. Tang, and W. K. Wong, "ICS cyber attack dataset and attack classification using machine learning approaches," *Access*, vol. 8, pp. 128920–128932, 2020.

[7] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[8] Y. Chen, K. R. Liu, and Q. Zhang, "Federated learning for privacy-preserving cyber-physical systems security," *Transactions on Signal and Information Processing over Networks*, vol. 7, pp. 198–211, 2021.

[9] A. Abduvaliyev, A. S. Kamilov, M. A. B. Altaf, and K. B. Baig, "AI-driven cyber resilience for industrial control systems: Challenges and opportunities," *Access*, vol. 9, pp. 78238–78260, 2021.

[10] J. A. Wang and M. A. Parvez, "Cybersecurity solutions for software-defined networking: A survey," *Access*, vol. 8, pp. 216813–216831, 2020.
[11] H. Liu, X. Yang, Y. Shi, J. Wang, and Y. Jin, "Deep learning model for real-time intrusion detection in industrial control systems," *Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5521–5531, 2021.
[12] S. Bhardwaj and M. M. Sharma, "Anomaly detection in cyber-physical systems using supervised machine learning techniques," *Computers & Security*, vol. 102, p. 102188, 2021.
[13] M. Ferrag, L. Maglaras, A. Ahmim, A. Derhab, and J. J. Rodrigues, "Security for smart grid control systems: Classification, challenges, and research opportunities," *Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3603–3638, 2019.
[14] S. Latif, Z. Zou, M. Idrees, J. Ahmad, and S. Jabbar, "Explainable artificial intelligence for cybersecurity: A systematic review," *IEEE Access*, vol. 11, pp. 33125–33148, 2023.
[15] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.