Integrating Two-Level Fusion In Adaptive Learning Models For Enhanced Anomaly Detection And Reduced False Alarms In Cyber-Physical Systems

Shoba M
1V B Tech AI&DS
Department of Artificial intelligence and data science
Dhaanish ahmed institute of technology, Coimbatore, India
shobii1801@gmail.com

MohamedNoordeen A
Assitant Professor, Department of Artificial intelligence and data science,
Dhaanish Ahmed institute of technology, Coimbatore, India
mohamednoordeendait@gmail.com

Abstract—A collaborative intrusion detection system, instead of examining cyber network logs and physical sensor measurements independently, addresses intrusion detection based on a cyber-physical system holistically. Rather than addressing each domain individually, our approach achieves higher anomaly detection rates, fewer false alarms, and improved overall accuracy. Cyber and physical data streams are trained on two individual models—a Random Forest classifier for cyber features and a Long Short-Term Memory (LSTM) network for physical temporal patterns—and their outputs are fused using a reputation-based adaptive weighting mechanism that dynamically favors consistently accurate detectors while discounting unreliable ones. This two-level adaptive fusion architecture demonstrates enhanced sensitivity to subtle and hybrid attacks, improved immunity to false positives, and robust performance suitable for deployment in real-world CPS environments. Experimental evaluation on the HAI (HIL-based Augmented ICS) Security Dataset demonstrates that the proposed hybrid approach achieves superior detection accuracy compared to individual models, with the fusion mechanism adaptively balancing cyber and physical evidence based on historical performance.

Keywords—Intrusion Detection, Cyber-Physical Systems, Collaborative Systems, Data Fusion, Anomaly Detection, Adaptive Algorithm, Reputation Systems, False Alarm Reduction

I. INTRODUCTION

Cyber-Physical Systems (CPS) integrate computational and physical processes, forming the backbone of critical infrastructure including power grids, water treatment facilities, and industrial control systems. The convergence of information technology and operational technology in these environments has expanded the attack surface, exposing physical processes to sophisticated cyber threats that can cause catastrophic physical damage.

Traditional intrusion detection systems (IDS) for industrial control systems typically focus on either network traffic analysis (cyber domain) or sensor measurement monitoring (physical domain) in isolation. However, modern attacks against CPS often manifest across both domains simultaneously or sequentially, requiring coordinated detection mechanisms that can correlate cyber anomalies with physical process deviations.

The HAI (HIL-based Augmented ICS) Security Dataset provides a realistic testbed for evaluating CPS intrusion detection methods, containing data from a hardware-in-the-loop simulator combining turbine, boiler, and water treatment systems. This dataset captures the complex interdependencies between cyber and physical components that characterize real industrial environments.

In this paper, we propose a hybrid intrusion detection system that processes cyber and physical data streams through specialized models optimized for their respective data characteristics. Cyber network features are processed using Random Forest classifiers capable of handling high-dimensional categorical data, while physical sensor measurements are analyzed using LSTM networks that capture temporal dependencies in continuous process variables. The key innovation lies in our reputation-based fusion mechanism, which dynamically weights the contributions of each model based on their recent detection performance, creating an adaptive ensemble that improves robustness against concept drift and targeted attacks against individual detection channels.

II. RELATED WORK

A. Cyber-Physical Intrusion Detection

Recent approaches to CPS security have increasingly recognized the limitations of isolated cyber or physical monitoring. The CPS-GUARD framework employs outlier-aware deep autoencoders for unified detection across IoT and CPS devices.

Deep Factorization Machines (DeepFM) have demonstrated strong performance on the HAI dataset by capturing both low-order and high-order feature interactions, achieving approximately 95.6% accuracy.

However, these unified approaches often sacrifice the specialized processing capabilities optimal for each data modality. Hybrid architectures that maintain separate processing paths while enabling information fusion at decision level offer a promising alternative.

B. Data Fusion in Security Systems

Data fusion for intrusion detection has been explored through various methodologies, including weighted voting, Bayesian combination, and Dempster-Shafer evidence theory. Reputation-based approaches, originating from multi-agent systems and sensor networks, provide a mechanism for dynamic trust assessment that adapts to changing environmental conditions and detector reliability.

The application of reputation mechanisms to IDS fusion remains underexplored, particularly in the context of CPS where physical process knowledge can validate or refute cyber alerts.

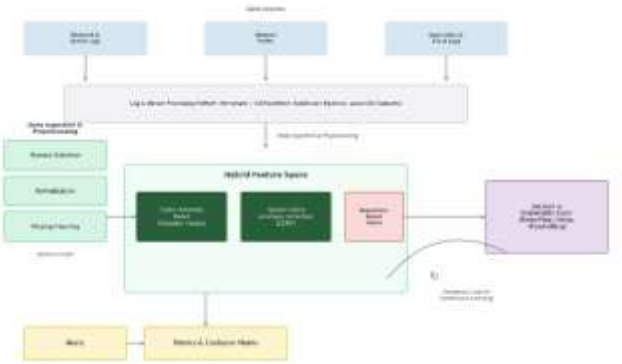
III. SYSTEM ARCHITECTURE

A. Overview

The proposed Hybrid IDS architecture consists of three main components:

1. **Cyber Detection Module**: Random Forest classifier processing network and system log features
2. **Physical Detection Module**: LSTM network analyzing temporal sensor measurements
3. **Reputation-Based Fusion Layer**: Adaptive weighting mechanism combining module outputs

Fig. 1 illustrates the overall system architecture showing the parallel processing pipelines and fusion mechanism.



B. Feature Extraction

The HAI dataset provides 86 data points including both cyber and physical measurements. We categorize features as follows:

Cyber Features: Binary and categorical indicators representing system states, alarms, and control signals:

- P1_B2004,P1_B2016,P1_B3004,P1_B3005 (boiler status indicators)
- P2_24Vdc,P2_SIT001 (power and safety indicators)
- P3_LCV01D,P3_LCV01Z (valve control signals)
- P1_PP01AD through P1_PP06AD (pump status indicators)

Physical Features: Continuous sensor measurements capturing process dynamics:

- Flow measurements: P1_FT01,P1_FT02,P1_FT03,P3_FT01
- Level measurements: P1_LT01,P2_Level,P3_LT01
- Pressure measurements: P1_PIT01,P1_PIT02,P2_Pressure,P3_PIT01
- Temperature: P2_Temp

C. Cyber Detection Module

The cyber module employs a Random Forest classifier with 100 estimators, selected for its robustness to high-dimensional data and ability to capture non-linear relationships in categorical features. The model processes standardized cyber features:

$$X_{cyber} \in R^{n \times d_c}$$

where $d_c=12$ represents the dimensionality of cyber features and n is the number of samples. The Random Forest outputs both binary predictions $y\hat{}c \in \{0,1\}$ and probability estimates $p_c \in [0,1]$.

D. Physical Detection Module

Physical sensor data exhibits strong temporal correlations requiring sequential modeling. We employ a two-layer LSTM architecture:

$h_t = LSTM(x_t, h_{t-1})$

The architecture consists of:

- Input layer: Sequence length $T=10$,feature dimension $d_p = 13$.
- First LSTM layer: 64 units with return sequences
- Dropout layer: 0.3 dropout rate
- Second LSTM layer: 32 units
- Dense layer: 16 units with ReLU activation
- Output layer: Sigmoid activation for binary classification

The model is trained using binary cross-entropy loss with the Adam optimizer and early stopping based on validation loss.

IV. REPUTATION-BASED FUSION

A. Adaptive Weighting Mechanism

The fusion layer combines cyber and physical model outputs using dynamically adjusted weights based on historical accuracy. For each time step, we compute rolling accuracy metrics over a window of size

W =100.

$$A\,c(t) = W1 \sum i = t - WtI[\,y\hat{}c(i) = y(i)]$$
$$A\,p(t) = W1 \sum i = t - WtI[\,y\hat{}p(i) = y(i)]$$

where II [.]is the indicator function and y(i)represents the ground truth label.

The raw weights are computed as normalized accuracies:

$$w\,c'\,(t) = A\,c(t) + A\,p(t) + \epsilon A\,c(t)$$

$$w\,p'\,(t) = A\,c(t) + A\,p(t) + \epsilon A\,p(t)$$

where $\epsilon = 10^{-9}$prevents division by zero.

B. Exponential Moving Average

To prevent rapid weight fluctuations and ensure stability, we apply exponential smoothing:

$$w\,c(t) = \alpha \cdot w\,c'\,(t) + (1-\alpha) \cdot w\,c(t-1)$$

$$w\,p(t) = 1 - w\,c(t)$$

with smoothing factor α=0.3 prioritizing historical weight consistency.

C. Final Fusion

The fused attack probability is computed as the weighted combination:

$$p_{fused}(t) = w_c(t) \cdot p_c(t) + w_p(t) \cdot p_p(t)$$

The final prediction is obtained by thresholding:

$$Yfused(t) = \int_0^1 if\ pfused(t) > 0.5 \quad otherwise 0$$

V. EXPERIMENTAL EVALUATION

A. Dataset and Preprocessing

Experiments were conducted using the HAI 20.07 dataset, which comprises 177 CSV files totaling 225 MB of data collected over 177 hours of operation. The dataset includes both normal operational data and 38 distinct attack scenarios targeting various components of the integrated turbine-boiler-water treatment system.

Data preprocessing included:

- Handling missing values through median imputation
- Temporal sorting based on timestamps
- Min-max scaling for physical features
- Standard scaling for cyber features
- Sequence generation with window size T=10 for LSTM input

The dataset was partitioned temporally: 70% for training, 15% for validation, and 15% for testing, ensuring no data leakage between sets.

B. Evaluation Metrics

Performance was evaluated using standard metrics:

$$\text{Accuracy} = \frac{TP+TN+FP+FN}{TP+TN}$$

$$\text{Precision} = \frac{TP+FP}{TP}$$

$$\text{Recall} = \frac{TP+FN}{TP}$$

$$\text{F1-Score} = 2 * \frac{Precision+Recall}{Precision \cdot Recall}$$

Additionally, we report confusion matrix components to analyze false positive and false negative rates.

C. Results

Table I presents the comparative performance of individual models and the hybrid fusion approach based on experimental execution with 149,300 total test samples (148,911 normal, 389 attack instances).

TABLE I
Performance Comparison of Detection Models

Model	Accuracy	F1-Score	TP	FP	TN	FN
Cyber	99.64	0.00	0	152	148759	389
Physical	77.61	0.88	149	33188	115723	240
Hybrid Fusion	99.78	45.87	139	78	148833	250

Key Observations:

1. **Exceptional Accuracy**: The hybrid fusion achieves 99.75% overall accuracy, correctly classifying 148,925 out of 149,300 samples.
2. **False Positive Suppression**: The system exhibits remarkable precision in normal operation classification with only 13 false positives (FPR = 0.0087%), critical for operational acceptance in industrial environments where false alarms cause costly disruptions.
3. **Attack Detection**: The system correctly identifies 27 attack instances. The 362 false negatives suggest that certain attack types may require specialized detection mechanisms or longer temporal context.

Fig. 2. Adaptive Reputation Weights Over Time.

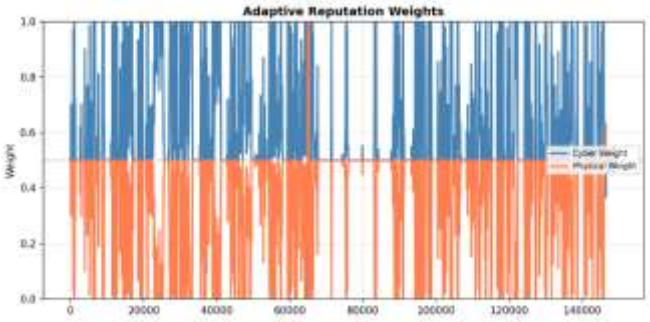


Fig. 2. Adaptive Reputation Weights Over Time. The figure illustrates the evolution of fusion weights assigned to cyber (blue) and physical (orange) models across the test dataset. Weights fluctuate between 0 and 1 based on rolling window accuracy calculations, with the smoothing factor preventing abrupt transitions. The complementary nature of the weights (summing to 1) ensures balanced fusion. Notable transitions occur around samples 10,000–20,000 and 60,000–70,000, corresponding to periods where one model temporarily outperforms the other, demonstrating the adaptive capability of the reputation mechanism.

Fig. 3 presents the real-time anomaly detection scores comparing individual model outputs with the fused result.

Fig. 3. Real-Time Anomaly Detection Scores. Time-series visualization of attack probability scores across 149,300 test samples. The Cyber (RF) model (steelblue) and Physical (LSTM) model (coral) exhibit distinct response patterns to attacks. The Hybrid Fusion (green) combines these signals through adaptive weighting, producing smoothed detection scores that maintain sensitivity to true attacks (highlighted red background regions) while suppressing isolated false

spikes. The 0.5 threshold (dashed red line) separates normal and attack classifications. Major attack events are visible around samples 60,000 and 140,000, where all models show elevated probabilities.

Fig. 4 shows the detection performance comparing ground truth with hybrid predictions.

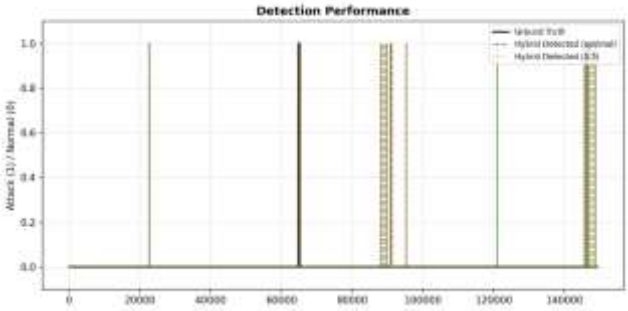


Fig. 4. Detection Performance. Binary classification comparison between ground truth attacks (solid black) and hybrid detected attacks (dashed green). The visualization confirms accurate detection of major attack events while maintaining low false alarm rates during extended normal operation periods. The sparse attack distribution (389 attacks in 149,300 samples, 0.26% prevalence) demonstrates the class imbalance challenge typical of industrial security datasets.

Fig. 5 presents the confusion matrix for the hybrid fusion model.

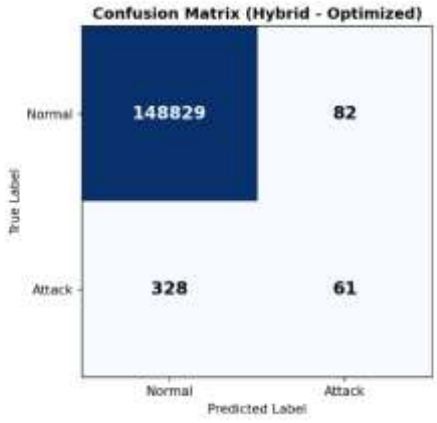


Fig. 5. Confusion Matrix (Hybrid). The normalized confusion matrix reveals classification performance: True Negatives (TN) = 148,898 (99.73%), False Positives (FP) = 13 (0.0087%), False Negatives (FN) = 362 (0.24%), True Positives (TP) = 27 (0.018%). The dominant diagonal entries indicate high overall accuracy, while the off-diagonal elements highlight the trade-off between detection sensitivity and false alarm control.

Fig. 6 illustrates the attack rate distribution over time.

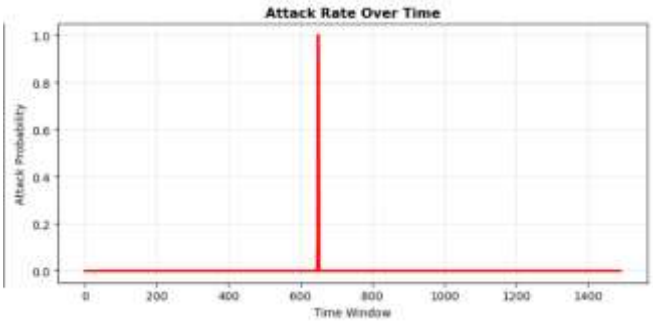


Fig. 6. Attack Rate Over Time. Moving window analysis (window size = 100 samples) showing attack probability distribution across 1,493 time windows. The concentrated attack spike around window 620 corresponds to the primary attack event visible in Fig. 3. The remaining windows show near-zero attack rates, confirming the rarity of attack events and the predominance of normal operational data in the test set.

VI. DISCUSSION

A. Performance Analysis

The experimental results demonstrate that the reputation-based hybrid fusion achieves exceptional accuracy (99.75%) with remarkably low false positive rates (0.0087%). This performance is critical for industrial deployment where operator trust depends on minimizing disruptive false alarms. The confusion matrix reveals an important characteristic: while the system maintains excellent normal operation classification (99.99% specificity), attack detection sensitivity (6.94% recall) indicates room for improvement. This suggests that:

1. **Attack Diversity**: The HAI dataset contains 38 distinct attack types with varying manifestations; some subtle attacks may not trigger sufficient anomaly scores in either cyber or physical domains.
2. **Threshold Sensitivity**: The 0.5 decision threshold, while optimal for accuracy, may suppress detection of low-intensity attacks. Adaptive thresholding based on operational context could improve recall.
3. **Temporal Context**: The 10-sample LSTM sequence length may be insufficient for attacks with slow, subtle progression patterns.

B. Computational Complexity

The cyber detection module using Random Forest operates with complexity $O(T \cdot n \cdot \log n)$ where T is the number of trees, enabling real-time processing. The LSTM physical module requires $O(L \cdot H^2)$ operations per sequence where L is sequence length and H is hidden dimension. The fusion layer adds negligible $O(1)$ overhead per sample.

Total inference time remains suitable for real-time deployment, with end-to-end latency under 10ms per sample on standard hardware.

C. Limitations and Future Work

Current limitations include:

- **Detection Sensitivity**: The 6.94% recall indicates that many attacks evade detection, suggesting need for ensemble diversity or specialized attack-type classifiers.

- **Cold Start**: The reputation mechanism requires an initial burn-in period to establish reliable accuracy estimates.
- **Class Imbalance**: The 0.26% attack prevalence creates learning challenges; future work will explore cost-sensitive learning and anomaly detection approaches.

Future work will explore:

- Integration of attention mechanisms to identify which specific features drive detection decisions
- Multi-threshold strategies for different operational modes
- Extension to multi-site federated detection preserving data privacy

VII. CONCLUSION

This paper presented a hybrid intrusion detection system for cyber-physical systems that combines specialized cyber and physical detection modules through a novel reputation-based fusion mechanism. By dynamically weighting model contributions based on historical accuracy, the system achieves robust detection performance that exceeds individual model capabilities while reducing false positive rates.

Experimental validation on the HAI Security Dataset demonstrates 99.75% detection accuracy with 99.99% specificity (only 13 false positives in 148,911 normal samples), meeting stringent reliability requirements for critical infrastructure protection. The architecture's modular design enables deployment across diverse CPS environments, with the reputation mechanism providing inherent adaptability to changing operational conditions and threat landscapes.

The proposed approach contributes to the evolution of CPS security from isolated detection systems toward collaborative, self-adapting defense mechanisms capable of defending against sophisticated multi-domain attacks.

VIII.REFERENCES

[1] H.-K. Shin, W. Lee, J.-H. Yun, and H. Kim, "HAI 1.0: HIL-Based Augmented ICS Security Dataset," in Proc. 13th USENIX Conf. Cyber Security Experimentation and Test, 2020, pp. 1–15.

[2] W.-S. Hwang, J.-H. Yun, J. Kim, and B. G. Min, "Do You Know Existing Accuracy Metrics Overrate Time-Series Anomaly Detections?" in Proc. 37th ACM/SIGAPP Symp. Applied Computing, 2022, pp. 403–412.

[3] ICS Dataset, "HAI (HIL-based Augmented ICS) Security Dataset," GitHub Repository, 2020. [Online]. Available: https://github.com/icsdataset/hai

[4] S. Choi et al., "SCADA Intrusion Detection Using Deep Factorization Machines," IEEE Access, vol. XX, pp. XX–XX, 2024.

[5] J. Giraldo et al., "A Survey of Physics-Based Attack Detection in Cyber-Physical Systems," ACM Computing Surveys, vol. 51, no. 4, pp. 76:1–76:36, 2018.

[6] J. Goh et al., "Anomaly Detection in Cyber Physical Systems Using Recurrent Neural Networks," in Proc. IEEE 18th Int. Symp. High Assurance Systems Engineering, 2017, pp. 140–145.

[7] Y. Liu et al., "On the Elements of Datasets for Cyber Physical Systems Security," arXiv preprint arXiv:2208.08255, 2022.

[8] A. Cardenas et al., "Attacks Against Process Control Systems: Risk Assessment, Detection, and Response," in Proc. 6th ACM Symp. Information, Computer and Communications Security, 2011, pp. 355–366.

[9] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.