Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 03, Issue 08, August 2026,
Publish Date: 27-08-2026
PageNo.07-13

A Hybrid Explainable AI Framework for Multi-Class Anomaly and Intrusion Detection in UAV and Industrial IoT Networks

Dr. Haruto Nakamura
Department of Artificial Intelligence, Tokyo Institute of Advanced Computing, Tokyo, Japan

Dr. Aiko Tanaka
Department of Intelligent Information Systems, Osaka Institute of Technology and AI, Osaka, Japan

ABSTRACT

The convergence of unmanned aerial vehicles (UAVs), industrial Internet of Things (IIoT) devices, wireless sensor networks, and networked control infrastructures has created highly heterogeneous communication environments in which anomalous and malicious activities may propagate across multiple layers. Conventional intrusion detection approaches that rely on a single classifier frequently face difficulties in distinguishing multiple attack classes under heterogeneous traffic, particularly when communication characteristics, payload structures, and network relationships vary considerably. This paper proposes a conceptual hybrid explainable artificial intelligence (XAI) framework for multi-class anomaly and intrusion detection in UAV and IIoT networks. The framework combines heterogeneous traffic preprocessing, feature-level representation, hybrid ensemble classification, anomaly scoring, and post-hoc explanation into a unified detection pipeline. Its theoretical foundation draws upon the security properties of network coding, secure wireless communication, image-quality assessment, and representation-oriented information processing discussed in the supplied literature. In particular, secure network coding studies establish the importance of protecting information flows and communication structures, while research on opportunistic relaying and physical-layer secrecy highlights the security implications of wireless communication conditions (Cai and Yeung, 2002; Khan and Chatzigeorgiou, 2017; Tajbakhsh et al., 2018). The proposed architecture is designed to distinguish multiple attack categories while simultaneously providing human-interpretable explanations concerning influential features, confidence, and predicted attack class. The resulting framework offers a theoretically grounded direction for security monitoring in heterogeneous UAV-IIoT environments, although empirical validation using representative multi-class datasets remains necessary before operational deployment.

KEYWORDS: Explainable AI, Intrusion Detection, UAV Networks, Industrial IoT, Multi-Class Classification, Anomaly Detection, Ensemble Learning, Network Security, Wireless Security.

INTRODUCTION

The increasing integration of UAVs and industrial IoT systems has transformed traditionally isolated infrastructures into interconnected cyber-physical communication environments. UAVs may operate as mobile sensing, surveillance, or communication platforms, whereas IIoT devices continuously exchange operational information with gateways, controllers, edge nodes, and cloud services. Such connectivity improves flexibility and responsiveness but also increases the number of communication paths through which abnormal or malicious activity may emerge.The security problem is particularly challenging because UAV and IIoT networks are heterogeneous by design. A UAV may communicate through dynamically changing wireless links, while industrial sensors may generate highly periodic traffic. Gateways and control nodes may simultaneously process information from devices with substantially different communication patterns. Consequently, an intrusion detector trained around a single traffic representation may not adequately represent the diversity of legitimate and malicious behavior.

The security of communication infrastructure is also inseparable from the protection of information flows. Research on secure network coding demonstrates that communication security must account not only for

individual transmitted messages but also for how information propagates through a network (Cai and Yeung, 2002). Subsequent work on restricted wiretap sets further illustrates that security requirements can depend on the particular structure of information exposure within a network (Cui et al., 2012). These principles are relevant to UAV-IIoT environments because a security event may affect multiple communication paths rather than a single endpoint.

Another challenge is the interpretability of machine-learning-based intrusion detection. A detector that simply reports attack or normal provides limited assistance to security operators. In a multi-class environment, the distinction between reconnaissance, flooding, manipulation, or other abnormal behaviors may influence the appropriate response. Explainability is therefore not merely a presentation feature; it can become an operational component that allows analysts to examine why a particular traffic instance was assigned to a specific class.

The principal objective of this research is to develop a theoretically grounded hybrid XAI architecture capable of detecting multiple classes of anomalies and intrusions across heterogeneous UAV and IIoT networks while providing interpretable evidence for its decisions. The framework integrates heterogeneous feature processing, complementary learning models, ensemble decision-making, anomaly scoring, and explanation generation.

The contribution is conceptual rather than an empirical performance claim. Since no experimental dataset or measured results were supplied, the Results section presents framework-level findings and analytically derived expectations rather than fabricated accuracy, precision, recall, or F1-score values.

LITERATURE REVIEW

2.1 Communication security and network structure

The foundations of secure communication in networked environments are established by research on network coding. Cai and Yeung (2002) demonstrated the importance of considering information flow when designing secure network communication. This perspective is particularly relevant to distributed UAV-IIoT systems, in which information may traverse multiple intermediate nodes before reaching a destination.

Cui et al. (2012) extended the security discussion to nonuniform or restricted wiretap configurations. Their work indicates that communication security cannot always be treated as a uniform property across all network links. For heterogeneous industrial and UAV infrastructures, this observation supports the need for security models capable of incorporating differences among communication channels, node roles, and exposure conditions.

Wireless communication introduces additional complexity. Khan and Chatzigeorgiou (2017) examined opportunistic relaying and random linear network coding in secure and reliable communication, demonstrating the interaction between reliability, relaying, coding, and security. Similarly, Tajbakhsh et al. (2018) examined network coding for physical-layer secrecy. Together, these studies provide a theoretical basis for considering communication context as part of intrusion detection rather than treating network traffic as isolated observations.

Vien et al. (2014) investigated energy-delay tradeoffs and relay positioning in wireless butterfly networks. Although their primary focus differs from intrusion detection, the work is useful for understanding why network topology and communication constraints matter in wireless systems. A UAV network operating under mobility and resource limitations cannot necessarily be analyzed using assumptions appropriate for static enterprise networks.

2.2 Information representation and quality assessment

Several supplied references focus on image representation, steganography, and image-quality assessment. Hashad et al. (2005) examined a robust steganography technique using discrete cosine transform insertion, while Po-Yueh and Lin et al. (2006) presented a discrete-wavelet-transform-based approach for image steganography. These studies illustrate the broader principle that meaningful detection may depend strongly on how raw information is transformed into discriminative representations.

Qin et al. (2019) surveyed coverless image steganography and thereby highlighted the difficulty of identifying concealed information when conventional assumptions about embedded content do not hold. Although these works address image security rather than UAV-IIoT intrusion detection directly, their methodological relevance lies in representation diversity and the challenge of detecting hidden or indirect information patterns.

Hore and Ziou (2010) compared PSNR and SSIM as image-quality metrics. Their work demonstrates that different evaluation measures capture different aspects of information quality. This principle translates conceptually into intrusion detection: relying on a single evaluation criterion may provide an incomplete representation of model behavior.

Wang et al. (2004) proposed an image-quality assessment perspective based on error visibility and structural similarity. Again, the direct domain is image processing, but the methodological implication for this framework is important: detection systems should evaluate both

individual deviations and structural patterns rather than treating every deviation as equally informative.

2.3 Data composition and heterogeneous information

Grubinger et al. (2006) introduced a benchmark resource for visual information systems. The significance of this work for the present framework is methodological: standardized benchmark resources are necessary for comparing systems operating on complex information spaces. A comparable principle applies to UAV-IIoT intrusion detection, where datasets should contain diverse devices, traffic patterns, and attack classes.

Porter and Duff (1984) addressed compositing digital images, establishing a representation-oriented perspective in which multiple information components can be combined into a coherent output. For the proposed security architecture, this provides a conceptual analogy for feature fusion: heterogeneous information streams can be combined before a final decision is produced.

Shankar and Elhoseny (2019) specifically addressed secure image transmission in wireless sensor network applications. The work provides a closer connection between information processing and wireless sensor security, reinforcing the importance of securing data while it moves through resource-constrained networks.

Overall, the supplied literature reveals a research gap. The references collectively address secure network communication, wireless security, information representation, and evaluation, but they do not provide an integrated explainable multi-class intrusion-detection architecture specifically designed for heterogeneous UAV-IIoT environments. The proposed framework addresses this gap by combining these conceptual foundations into a unified security-monitoring pipeline.

METHODOLOGY

3.1 Framework architecture

The proposed architecture consists of six interconnected stages: heterogeneous traffic acquisition, preprocessing and normalization, multi-view feature representation, hybrid classification, anomaly and confidence analysis, and explainability.

At the first stage, traffic is collected from UAV communication links, IIoT sensors, gateways, edge nodes, and relevant control interfaces. Instead of assuming that all devices generate statistically equivalent traffic, the architecture maintains contextual information such as source category, destination role, communication direction, temporal behavior, and protocol-level characteristics.

The second stage transforms raw observations into a consistent analytical representation. Missing values, duplicated observations, inconsistent scales, and unsuitable categorical encodings are processed before model training. Normalization is particularly important because features describing packet characteristics, temporal intervals, byte counts, and connection frequencies can have substantially different numerical ranges.

The third stage employs multi-view feature representation. Rather than relying on one feature family, the framework separates features into complementary groups. Statistical traffic features capture volume and frequency; temporal features characterize changes over time; communication-context features represent relationships among nodes; and payload or content-oriented features represent observable information characteristics. This design reflects the broader lesson from representation-oriented research that the quality of a decision depends on the representation supplied to the analytical model (Po-Yueh, Lin et al., 2006; Wang et al., 2004).

3.2 Hybrid classification mechanism

The central detection component is a hybrid ensemble. Instead of depending on a single learner, multiple complementary classifiers generate intermediate predictions. One component is optimized for nonlinear feature relationships, another emphasizes probabilistic or statistical discrimination, and a third captures interaction patterns among features.

Let the base learners be represented by f1(x),f2(x),…,fm(x)f_1(x), f_2(x), \ldots, f_m(x), where xx denotes an observed network instance. Each learner generates a class-probability vector:

Pi(y|x)=[pi1,pi2,…,piK]P_i(y|x)=[p_{i1},p_{i2},\ldots,p_{iK}]

where KK represents the number of attack or anomaly classes.

The ensemble combines these probability vectors through weighted aggregation:

P(y|x)=∑i=1mwiPi(y|x)P(y|x)=\sum_{i=1}^{m}w_iP_i(y|x)

subject to:

∑i=1mwi=1\sum_{i=1}^{m}w_i=1

The final class is determined from the highest aggregated probability. The use of multiple learners reduces dependence on the inductive assumptions of any single model. This is particularly important in heterogeneous environments where different traffic families may exhibit different statistical characteristics.

The framework also introduces an anomaly score alongside the class prediction. A sample can therefore be classified into a specific category while simultaneously receiving an indication of how strongly it differs from the expected normal distribution. This distinction is important because a low-confidence prediction may warrant human investigation even when the classifier assigns it to an existing class.

3.3 Multi-class detection strategy

The system is designed to distinguish normal traffic from multiple anomaly and intrusion categories rather than treating every malicious event as a binary attack. A generic class structure may include normal behavior, reconnaissance-like behavior, denial-of-service or flooding behavior, communication manipulation, unauthorized access behavior, and previously unseen anomalous behavior.

The distinction between known and potentially unknown behavior is achieved through a dual-decision mechanism. First, the ensemble determines the most probable known class. Second, confidence and anomaly measurements are evaluated against predefined thresholds. If the predicted class probability is low while the anomaly score is high, the observation is assigned to an unknown or uncertain category.

This mechanism is preferable to forcing every observation into a known class. In dynamic UAV-IIoT environments, new attack patterns may not perfectly resemble previously observed categories. The proposed architecture therefore treats uncertainty as information rather than as classification failure.

3.4 Explainability layer

The XAI layer translates the ensemble decision into an interpretable explanation. For each detected event, the system produces three forms of information: the predicted class, the confidence associated with that prediction, and the features that contributed most strongly to the decision.

For example, suppose an industrial gateway receives an unusual burst of connections from a UAV-associated endpoint. The classifier might assign the observation to a flooding-related class because connection frequency, packet-rate deviation, temporal burstiness, and source-destination behavior collectively differ from the learned baseline. The explanation layer would expose these influential characteristics rather than presenting only a numerical label.

Explainability is particularly valuable for security analysts because two alerts with identical labels may require different responses. A decision dominated by temporal irregularity may suggest a different investigation path from one dominated by unusual communication relationships.

The proposed interpretation strategy also supports model auditing. If an ensemble repeatedly bases decisions on an irrelevant or unstable feature, analysts can identify a potential data-quality or model-design problem. Consequently, XAI becomes part of model validation rather than merely a visualization mechanism.

3.5 Security and communication context

The framework incorporates network context because security cannot be separated from communication structure. The secure network-coding literature emphasizes that the confidentiality and reliability of information depend on how information traverses network structures (Cai and Yeung, 2002; Cui et al., 2012). Accordingly, the proposed detector considers node relationships and communication paths in addition to isolated packet-level characteristics.

Wireless reliability and security are similarly treated as contextual properties. Opportunistic relaying and physical-layer secrecy research demonstrates that communication conditions can influence security outcomes (Khan and Chatzigeorgiou, 2017; Tajbakhsh et al., 2018). Therefore, the detector should distinguish a change caused by normal wireless variability from a change caused by malicious manipulation.

3.6 Evaluation design

Because no experimental dataset is supplied, this study does not invent numerical results. A complete empirical evaluation should measure accuracy, macro-precision, macro-recall, macro-F1, false-positive rate, false-negative rate, detection latency, computational overhead, and explanation consistency.

Macro-averaged metrics are particularly important in multi-class detection because high-frequency classes can otherwise dominate aggregate performance. Confusion matrices should additionally be used to identify systematic confusion between attack categories.

The evaluation should compare the hybrid framework against individual base learners and non-explainable baselines. Ablation experiments should remove one ensemble component or one feature family at a time to determine whether performance gains result from genuine complementarity. Explainability should also be evaluated by testing whether the features identified as influential remain stable under small perturbations of the input.

RESULTS

The principal finding derived from the proposed architecture is that heterogeneous UAV-IIoT intrusion detection should be treated as a multi-view classification problem rather than a single-feature anomaly-recognition task. The literature on secure communication indicates that network structure, communication paths, and exposure conditions affect security behavior (Cai and Yeung, 2002; Cui et al., 2012). Consequently, a detector that ignores network context may fail to distinguish structurally different events that appear similar at the packet level.

A second finding is the theoretical suitability of ensemble decision-making for heterogeneous traffic. Different traffic sources may produce distinct distributions, and therefore a single learner can become disproportionately sensitive to particular feature relationships. Combining complementary models provides a mechanism for reducing this dependency. The approach is consistent with the broader methodological lesson from the supplied representation and information-processing literature that different representations can expose different characteristics of the same underlying information (Po-Yueh, Lin et al., 2006; Qin et al., 2019).

A third finding concerns the separation of classification confidence from anomaly magnitude. A classifier may assign an observation to the nearest known attack category even when that observation is substantially different from previously observed examples. The proposed dual mechanism avoids this limitation by maintaining an explicit uncertainty pathway. This is particularly relevant to UAV-IIoT systems, where operational conditions can evolve over time.

The fourth finding concerns explainability. A multi-class detector becomes more operationally useful when its output contains both the predicted class and evidence supporting that prediction. The proposed explanation layer can expose influential temporal, statistical, and communication-context features. This allows security operators to distinguish between high-confidence routine deviations and potentially significant attacks.

Finally, the supplied literature indicates that evaluation must be multidimensional. Image-quality studies demonstrate that alternative metrics can capture different properties of system output (Hore and Ziou, 2010; Wang et al., 2004). By analogy, intrusion-detection evaluation should not depend solely on accuracy. Class imbalance, minority-class recall, false alarms, latency, and explanation stability should be considered together.

These findings remain architectural and theoretical because the supplied material contains no UAV-IIoT benchmark dataset or experimental measurements. Accordingly, numerical performance superiority cannot be claimed without empirical validation.

DISCUSSION

The proposed framework contributes to intrusion-detection research by connecting three traditionally separated concerns: communication security, heterogeneous machine learning, and interpretability. The communication-security literature establishes that network structure and information exposure influence security, while the proposed architecture translates this insight into contextual features for machine learning (Cai and Yeung, 2002; Cui et al., 2012).

The principal theoretical implication is that intrusion detection should not be understood solely as a pattern-recognition problem. In UAV-IIoT environments, the meaning of an observed pattern depends on the node generating it, the communication relationship involved, the temporal context, and the expected operational behavior. A burst of packets from one device may represent legitimate operational activity under one condition and malicious flooding under another.

The hybrid ensemble introduces a corresponding practical advantage. Security monitoring systems operate under heterogeneous conditions, and complementary models can capture different aspects of the traffic distribution. However, ensemble complexity creates a trade-off. Increasing the number of models may improve representational diversity but also increases computational requirements and can complicate explanation generation. This is particularly significant for edge-based UAV and IIoT nodes with limited computational resources.

Explainability creates another trade-off. Highly detailed explanations may assist expert analysts but overwhelm operators if every feature is reported. The framework therefore favors ranked feature contributions and confidence information rather than exposing the complete internal structure of every model. Such a design is consistent with the need to make security decisions operationally understandable.

The communication-security references also reveal an important limitation. Network coding and physical-layer security mechanisms address communication protection, whereas intrusion detection focuses on identifying suspicious behavior. These functions should not be treated as interchangeable. A secure communication mechanism can reduce exposure without necessarily identifying compromised endpoints, while an intrusion detector can identify suspicious behavior without guaranteeing confidentiality or communication integrity. The proposed

framework therefore positions detection as one component within a broader security architecture.

Another limitation concerns data representation. The image-processing references demonstrate the importance of representation and quality assessment, but their direct findings cannot be transferred quantitatively to UAV-IIoT intrusion detection. Their contribution to this paper is methodological rather than domain-specific. Similarly, the benchmark-oriented work of Grubinger et al. (2006) supports the importance of systematic evaluation, but a dedicated UAV-IIoT benchmark remains necessary.

Future empirical work should therefore construct a heterogeneous dataset containing UAV traffic, industrial sensor traffic, gateway interactions, normal operational variations, and multiple attack categories. Cross-environment testing would be especially important because a model that performs well within one network may fail when transferred to another. Explainability should also be tested independently from predictive performance, including stability, faithfulness, and usefulness to human analysts.

CONCLUSION

This paper proposed a hybrid explainable AI framework for multi-class anomaly and intrusion detection in heterogeneous UAV and Industrial IoT networks. The framework integrates heterogeneous traffic preprocessing, multi-view feature representation, ensemble classification, anomaly scoring, uncertainty handling, and explainable decision generation.

The conceptual foundation combines insights from secure network coding, wireless communication security, information representation, and evaluation research. Secure network-coding studies emphasize the significance of information-flow structure, while wireless-security research demonstrates the interaction among reliability, communication conditions, and confidentiality (Cai and Yeung, 2002; Khan and Chatzigeorgiou, 2017; Tajbakhsh et al., 2018). These principles motivate the incorporation of network and communication context into the proposed detector.

The principal contribution is the integration of multi-class prediction with an explicit uncertainty mechanism and human-interpretable evidence. Rather than producing only an attack label, the framework is designed to indicate what type of event has been detected, how confidently it has been classified, and which characteristics contributed to the decision.

The framework nevertheless remains a research architecture rather than a validated empirical system. No experimental dataset or measured performance was provided, and therefore numerical claims have deliberately not been fabricated. Future research should validate the architecture using representative UAV-IIoT datasets, perform comparative and ablation experiments, assess computational overhead, and evaluate explanation reliability. Such validation would determine whether the proposed hybrid XAI approach can provide measurable improvements over conventional intrusion-detection models while maintaining sufficient transparency for operational cybersecurity environments.

REFERENCES

1. N. Cai and R. W. Yeung, “Secure network coding,” in Proceedings IEEE International Symposium on Information Theory,. IEEE, 2002, p. 323.
2. T. Cui, T. Ho, and J. Kliewer, “On secure network coding with nonuniform or restricted wiretap sets,” IEEE Transactions on Information Theory, vol. 59, no. 1, pp. 166–176, 2012.
3. M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The iapr tc-12 benchmark: A new evaluation resource for visual information systems,” in International workshop ontoImage’2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC’06, Genoa, Italy,, vol. 2, pp 13-23, 22-May 2006.
4. A. I. Hashad, A. S. Madani, and A. E. M. A. Wahdan, “A robust steganography technique using discrete cosine transform insertion,” in 2005 International Conference on Information and Communication Technology. IEEE, 2005, pp. 255–264.
5. A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in 2010 20th international conference on pattern recognition. IEEE, 2010, pp. 2366–2369.
6. A. S. Khan and I. Chatzigeorgiou, “Opportunistic relaying and random linear network coding for secure and reliable communication,” IEEE Transactions on Wireless Communications, vol. 17, no. 1, pp. 223–234, 2017.
7. Y. Po-Yueh, H.-J. Lin et al., “A dwt based approach for image steganography,” International Journal of Applied Science and Engineering, vol. 4, no. 3, pp. 275–290, 2006.
8. T. Porter and T. Duff, “Compositing digital images,” in Proceedings of the 11th annual conference on Computer graphics and interactive techniques, 1984, pp. 253–259.
9. J. Qin, Y. Luo, X. Xiang, Y. Tan, and H. Huang, “Coverless image steganography: a survey,” IEEE access, vol. 7, pp. 171 372–171 394, 2019.
10. K. Shankar and M. Elhoseny, Secure image transmission in wireless sensor network (WSN) applications. Springer, 2019.
11. S. E. Tajbakhsh, J. P. Coon, and G. Chen, “Network coding for physical layer secrecy,” IEEE Wireless

Communications Letters, vol. 7, no. 4, pp. 642–645, 2018.

12. Q.-T. Vien, H. X. Nguyen, B. G. Stewart, J. Choi, and W. Tu, “On the energy–delay tradeoff and relay positioning of wireless butterfly networks,” IEEE Transactions on Vehicular Technology, vol. 64, no. 1, pp. 159–172, 2014.

13. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

14. S. Zhang, S. C. Liew, and P. P. Lam, “Hot topic: Physical-layer network coding,” in Proceedings of the 12th annual international conference on Mobile computing and networking, 2006, pp. 358–365.

15. Islam, M.S., Ahmed, F., Ishtiaq, W. et al. Advanced explainable ensemble models for multi-class intrusion detection in heterogeneous drone and industrial networks. J. Inf. Secur. 2026, 14 (2026).

16. K. Ramamurthy, R. K. Konduru and N. Amanmadov, "EvoGraphCoder: An Evolutionary Graph-Reasoning Framework for Self-Adaptive Software Engineering," in IEEE Access, vol. 14, pp. 63063-63076, 2026, doi: 10.1109/ACCESS.2026.3686019.

17. Geo Philip, Paulson, Artificial Intelligence and Machine Learning Applications in Project Schedule Forecasting: A Predictive Framework for Time-Control in Complex Building Projects (April 16, 2026). Available at SSRN: https://ssrn.com/abstract=6588119 or http://dx.doi.org/10.2139/ssrn.6588119