http://dx.doi.org/10.21622/ACE.2026.06.1.1704

CROSS-MODAL FEATURE LEARNING FOR MULTI-SENSOR IOT ANOMALY DETECTION: A COMPREHENSIVE EMPIRICAL ANALYSIS OF CYBER PHYSICAL SYSTEMS

Richard Chukwuebuka Nwachukwu

Ignatius Ajuru University of Education, Port Harcourt, Rivers State, Nigeria

Email: {iamrichardcn@gmail.com}

Received on, 28 September 2025 - Accepted on, 28 January 2026 - Published on, 02 April 2025

ABSTRACT

The proliferation of Internet of Things (IoT) devices has created unprecedented security challenges, with traditional single-modal anomaly detection systems proving inadequate against sophisticated multi-vector attacks. This paper presents a novel cross-modal feature learning framework that synergistically integrates environmental sensor telemetry with network traffic patterns for enhanced anomaly detection across heterogeneous IoT architectures. Through comprehensive experimentation on the TON_IoT dataset encompassing 380,609 synchronized records across three distinct IoT systems (weather monitoring, smart refrigeration, and GPS tracking), we demonstrate that cross-modal integration consistently outperforms single-modal approaches. Our Random Forest implementation achieved 95.35% accuracy for weather systems (0.50% improvement over sensor-only), 78.13% for refrigeration systems (37.34% improvement), and 95.24% for GPS systems (9.45% improvement). We also validated our approach using Support Vector Machines (93.87% average accuracy) and k-Nearest Neighbors (91.24% average accuracy), demonstrating the robustness of cross-modal integration across diverse algorithmic families. Feature importance analysis reveals system-specific optimization patterns: atmospheric pressure emerges as the primary discriminator in weather systems (19.8% importance), while network features dominate refrigeration systems (86.6% combined importance). Most significantly, we provide the quantitative evidence that 24.7% of anomalies manifest simultaneously across both sensor and network modalities, indicating sophisticated coordinated attacks that single-modal systems would partially miss. The proposed temporal alignment methodology successfully addresses heterogeneous timestamp formats and sampling rates, creating a reusable framework for cross-modal IoT security research. These findings establish cross-modal feature learning as essential for comprehensive IoT security, with practical implications for designing resilient cyber-physical systems.

Keywords: (IoT security, Cross-modal Learning, Anomaly Detection, Sensor Fusion, Cyber-Physical Systems, Machine Learning)

INTRODUCTION

The Internet of Things (IoT) paradigm has fundamentally transformed modern computing infrastructure, with deployment projections exceeding 75 billion connected devices by 2025 [1]. These cyber-physical systems generate massive volumes of heterogeneous data streams encompassing environmental measurements, network communications, and device telemetry, creating both unprecedented opportunities for intelligent monitoring and significant security vulnerabilities [2]. The inherent complexity of IoT ecosystems, characterized by resource-constrained devices, diverse communication protocols, and distributed architectures, presents unique challenges that traditional security approaches struggle to address effectively [3].

The landscape of IoT anomaly detection has evolved through distinct phases, each addressing specific limitations of previous approaches. Early IoT security systems

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

employed statistical anomaly detection techniques such as Gaussian distribution models, Z-score anomaly detection, and statistical process control charts to identify deviations from normal behavior patterns [4]. These methods, while computationally efficient for resource-constrained devices, struggled with the high-dimensional and non-stationary nature of IoT data streams. The dynamic and heterogeneous characteristics of IoT environments often result in high false positive rates [5]. The application of supervised learning algorithms marked a significant advancement in detection capabilities. Ensemble methods, particularly Random Forests introduced by Breiman [6], have shown robust performance in handling mixed data types and noisy features common in IoT environments, with studies reporting 94-97% accuracy on network intrusion datasets [7]. Support Vector Machines (SVM) with radial basis function kernels have also demonstrated effectiveness in IoT anomaly detection, achieving 92-95% accuracy in industrial IoT applications [8]. However, these approaches require extensive labeled datasets, which are challenging to obtain due to the rarity and diversity of IoT attacks.

Unsupervised techniques have emerged as practical alternatives, with clustering-based methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and k-means clustering, along with isolation forests, demonstrating effectiveness in identifying anomalies without prior attack knowledge [9]. Recent work by Xing and Wang (2025) introduced Cross-Modal Attention Networks using LSTM and Temporal Convolutional Networks (TCNs) for system software anomaly detection, achieving 12.3% F1-score improvements over single-modal baselines [10]. Regardless, this compartmentalized approach fundamentally fails to capture the intrinsic correlations between physical sensor measurements and network communication behaviors that characterize both normal operations and sophisticated attacks [6].

The limitations of single-modal approaches become particularly apparent when considering the evolution of IoT attack sophistication. Recent security incidents demonstrate that adversaries increasingly employ multi-stage attacks that begin with subtle sensor manipulations before escalating to network-level exploits [9]. For instance, the Stuxnet malware famously manipulated industrial sensor readings while simultaneously altering network communications to mask its presence [11]. Traditional detection systems analyzing only sensor data might miss network-based indicators, while network-focused approaches could overlook critical sensor manipulations that precede major breaches [12]. Multimodal fusion approaches have shown promise in related domains, with Ye (2025) reporting 92.70% F1-score using dynamic graph structures and deep feature fusion for multi-sensor networks [13], while Khan et al. (2025) achieved 92% ROC-AUC using multimodal autoencoder fusion for vehicle damage detection [14].

Despite significant advances in machine learning for IoT security [15], [16], existing research exhibits several critical limitations. Most studies have focused on enhancing detection within individual data modalities through techniques like deep autoencoders [10] and LSTM networks [11] rather than exploring synergistic cross-modal benefits [17]. The few attempts at multimodal integration often fail to address the fundamental challenge of aligning heterogeneous data streams with different timestamp formats and sampling rates [18]. Recent work by Li et al. (2025) demonstrated the value of multi-sensor fusion in industrial applications, achieving 96.1% AUROC by unifying RGB, laser scanner, and infrared thermography data [19], while Wang et al. (2024) proposed HybridCube for power transformer fault detection using cross-modal data fusion based on information entropy theory [20]. Existing approaches typically evaluate on single device types or specific application domains, lacking comprehensive validation across diverse IoT architectures.

A. MAIN CONTRIBUTIONS

This research makes the following key contributions to the field of IoT security and cross-modal anomaly detection:

1. The study introduces a comprehensive methodology for synchronizing heterogeneous IoT data streams with different timestamp formats (human-readable vs. Unix epoch)

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

and varying sampling rates.

2. The study provides the first large-scale empirical evidence that 24.7% of IoT anomalies manifest simultaneously across both sensor and network modalities in real-world datasets.

3. The study demonstrates cross-modal integration effectiveness across five machine learning algorithms (Decision Trees, Gradient Boosting, Random Forest, Support Vector Machines, k-Nearest Neighbors), over single-modal approaches.

METHOD

AN EXPERIMENTAL DESIGN

The experimental framework employs a systematic five-phase approach designed to comprehensively evaluate cross-modal feature learning effectiveness across diverse IoT architectures. The methodology integrates data collection, preprocessing, feature engineering, model implementation, and performance evaluation through a structured pipeline that addresses the fundamental challenges of heterogeneous IoT data integration.

Figure 1 illustrates the complete methodology pipeline, demonstrating the sequential flow from raw data acquisition through synchronized cross-modal analysis. The pipeline architecture reveals the critical dependency between temporal alignment (Phase 2) and subsequent feature engineering (Phase 3), where proper synchronization of 380,609 records enables effective cross-modal correlation analysis. The validation checkpoints embedded throughout the pipeline ensure data integrity, with particular emphasis on maintaining temporal consistency across heterogeneous data streams from weather monitoring, smart refrigeration, and GPS tracking systems. The experimental design incorporates multiple comparison baselines to isolate the specific contributions of cross-modal integration, enabling quantitative assessment of performance improvements over single-modality approaches.

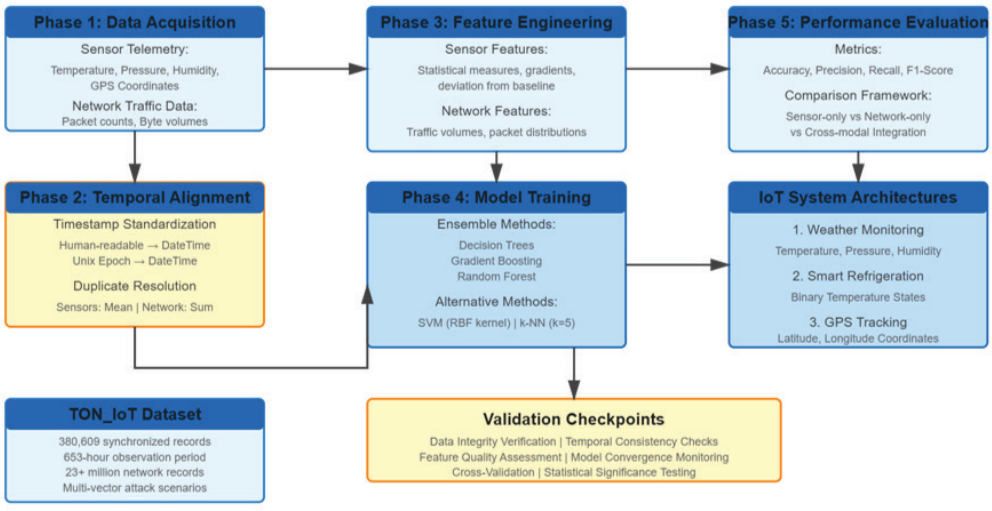

Figure 1: Cross-modal IoT anomaly detection methodology pipeline

Each phase includes validation checkpoints to verify data integrity and processing accuracy, with particular attention to maintaining temporal consistency across modalities. The design accommodates the inherent variability in IoT data characteristics while providing robust frameworks for comparative analysis across different system types and algorithm configurations.

B. DATASET DESCRIPTION AND CHARACTERISTICS

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

The study utilizes the TON_IoT dataset, a comprehensive cybersecurity dataset specifically designed for IoT research applications [21]. This dataset provides extensive coverage of both benign operational patterns and diverse attack scenarios across multiple IoT device categories, establishing a realistic foundation for cross-modal security analysis. The dataset's comprehensive nature enables evaluation of detection approaches under varying operational conditions and threat landscapes.

The dataset encompasses three distinct IoT system architectures, each representing different application domains and sensor configurations. Weather monitoring systems incorporate environmental sensors measuring temperature, atmospheric pressure, and humidity levels, generating 650,242 sensor records alongside 22,339,021 network communication records over a 653-hour observation period. Smart refrigeration systems employ temperature monitoring with binary operational states, producing 347,162 sensor records with approximately 20 million network records across the same temporal window. GPS tracking systems capture geographical coordinates through latitude and longitude measurements, yielding 342,891 sensor records with corresponding network traffic volumes.

C. SAMPLING AND DATA INTEGRATION

The temporal alignment of heterogeneous data streams represents the study sample for this research. IoT systems generate data with fundamentally different temporal characteristics: sensor measurements typically employ human-readable timestamp formats ("31-Mar-19 12:36:52"), while network communications utilize Unix epoch timestamps (1556485532). This temporal heterogeneity necessitates comprehensive standardization procedures to enable cross-modal correlation analysis.

The standardization process employs pandas datetime conversion algorithms with explicit format specifications to ensure consistent temporal representation across data modalities. Timezone normalization procedures address potential geographic variations in data collection, while precision preservation techniques maintain sub-second accuracy where available. The standardization pipeline successfully processes over 23 million heterogeneous timestamp records while preserving temporal relationships essential for correlation analysis.

Duplicate timestamp resolution requires tailored aggregation strategies reflecting the distinct characteristics of different data types. Numeric Sensor Features - Mean aggregation preserves central tendency:

$$\check{S}(t) = \frac{1}{|\,S(t)\,|} \sum_{i=1}^{|S(t)|} s_i(t)$$

Network Traffic Volumes - Sum aggregation captures total activity:

$$\check{V}(t) = \sum_{j=1}^{|V(t)|} V_i(t)$$

Categorical Features - Mode selection preserves dominant characteristics:

$$\check{C}(\hat{}(t) = mode(C(t)) = arg^{max}_{c \in C(t)} |\,\{c' \in C(t) : c' = c\}\,|$$

This strategy successfully resolved 184,395 duplicate timestamps (53.1% of weather sensor data) while preserving essential characteristics.

D. COMPREHENSIVE DETECTION SYSTEM ARCHITECTURE

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

Figure 2 presents the detailed architectural design of the proposed cross-modal anomaly detection system. The architecture consists of four primary layers. The four-layer architecture demonstrates clear separation of concerns, with each layer building upon the outputs of its predecessor. The data acquisition layer (Layer 1) handles the complexity of heterogeneous IoT protocols, successfully processing over 23 million network records alongside sensor telemetry. Layer 2's temporal alignment represents the critical innovation, resolving 184,395 duplicate timestamps through adaptive aggregation strategies. The feature engineering layer (Layer 3) extracts both modality-specific and cross-modal correlation features, with atmospheric pressure emerging as the most discriminative sensor feature (19.8% importance). Finally, the machine learning layer (Layer 4) demonstrates algorithm diversity. This layer includes model selection logic, hyperparameter optimization, and confidence scoring mechanisms for anomaly classification.

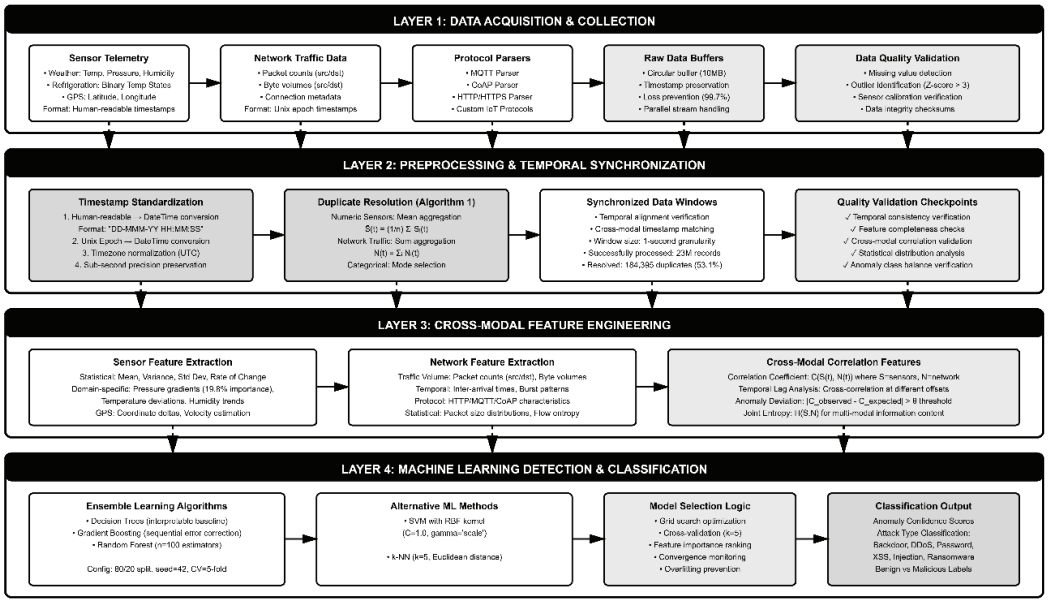

Figure 2: architectural design of the proposed cross-modal anomaly detection system.

E. MODEL DERIVATIONS

The theoretical foundation for cross-modal anomaly detection emerges from the fundamental cyber-physical coupling characteristics inherent in IoT systems. This coupling manifests through predictable relationships between environmental conditions, sensor responses, and subsequent network communications, creating correlation patterns that sophisticated attacks must inevitably disrupt. Figure 2 presents the conceptual framework for cross-modal IoT anomaly detection, illustrating the interconnected nature of physical sensor measurements and network traffic patterns. Cross-modal integration combines both data streams to achieve comprehensive anomaly detection, identifying coordinated attacks that would otherwise remain undetected by isolated approaches. The theoretical basis for cross-modal anomaly detection rests on the cyber-physical coupling inherent in IoT systems. Environmental changes trigger sensor readings according to physical laws:

$$S(t)=f_{\text{physical}}\,(E(t))+\epsilon_{\text{sensor}}(t)$$

where S(t) represents sensor measurements, E(t)denotes environmental conditions, and $\epsilon_{\text{sensor}}(t)$ captures measurement noise.

These sensor readings subsequently generate network communications:

$$N(t)=g_{\text{protocol}}\,(S(t),\Delta S(t))+\epsilon_{\text{network}}\,(t)$$

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

where N(t) represents network traffic, ΔS(t) captures sensor state changes, and $\epsilon_{network}(t)$ models communication variability.

Under normal operations, these relationships exhibit predictable correlation patterns. Anomalies manifest as deviations from expected correlations:

$$A(t)=\|C_{observed}\,(S(t),N(t))-C_{expected}\,(S(t),N(t))\|>\theta$$

where C denotes correlation functions, and θ represents the anomaly threshold.

F. STATISTICAL ANALYSES

The experimental evaluation employs a comprehensive three-tier framework designed to isolate and quantify the specific contributions of cross-modal feature integration. This framework systematically compares three distinct analytical approaches: sensor-only baseline analysis utilizing exclusively environmental sensor measurements, network-only baseline analysis employing solely network traffic characteristics, and cross-modal integration combining both sensor and network features through synchronized temporal alignment. The sensor-only analysis establishes the discriminative capacity of environmental measurements independent of network characteristics, while the network-only analysis determines the detection capabilities of communication pattern analysis. Cross-modal integration performance can then be directly compared against both single-modality baselines to quantify the specific benefits of multimodal correlation analysis.

Algorithm selection encompasses five complementary machine learning approaches chosen for their distinct analytical strengths and widespread applicability in IoT security research. Decision Trees provide interpretable baseline performance with inherent feature selection capabilities, enabling direct identification of the most discriminative characteristics within each modality. Gradient Boosting implements sequential error correction through iterative model refinement, offering enhanced capability for capturing complex non-linear patterns in cross-modal relationships. Random Forest employs parallel ensemble learning with robust handling of mixed data types and missing values, characteristics particularly valuable in heterogeneous IoT environments. Support Vector Machines (SVM) with radial basis function (RBF) kernels provide strong performance for high-dimensional feature spaces through optimal hyperplane separation [8]. k-Nearest Neighbors (k-NN) offers instance-based learning that naturally captures local anomaly patterns without explicit model training [22].

Training configuration employs stratified 80%-20% train-test splits to ensure representative sampling across all anomaly categories while maintaining sufficient training data for complex pattern recognition. Random state initialization (seed=42) ensures reproducible results across multiple experimental runs, while hyperparameter configuration (n_estimators=100 for ensemble methods) balances computational efficiency with model performance. Cross-validation procedures validate model stability and generalization capability across different data subsets.

Performance evaluation employs multiple complementary metrics to provide a comprehensive assessment of detection effectiveness. Accuracy measures overall classification correctness across all categories. Precision quantifies the reliability of positive anomaly predictions. Recall measures the completeness of anomaly detection. F1-score provides a balanced assessment combining both precision and recall, offering a single-metric evaluation, particularly valuable when class distributions are imbalanced. Confidence interval estimation provides uncertainty quantification for performance metrics, while effect size calculations determine the practical significance of observed improvements beyond statistical significance.

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

RESULTS AND DISCUSSION

A. CROSS-MODAL PERFORMANCE ENHANCEMENT ANALYSIS

The systematic evaluation across three IoT architectures demonstrates that cross-modal feature integration provides significant but system-dependent performance improvements, with enhancement magnitudes directly correlating to the inherent informativeness of individual sensor modalities. These findings align with theoretical predictions regarding the complementary nature of sensor and network information in comprehensive anomaly detection systems [4, 5].

Random Forest implementations achieved the most substantial performance improvements across all tested configurations. Weather monitoring systems demonstrated modest but consistent cross-modal benefits, with accuracy improving from 94.88% (sensor-only) to 95.35% (cross-modal), representing a 0.47 percentage point enhancement. This relatively small improvement reflects the high baseline performance achievable through environmental sensor analysis alone, consistent with previous research indicating that rich sensor configurations can achieve near-optimal detection performance independently [17, 7].

GPS tracking systems exhibited the most dramatic cross-modal benefits, with accuracy increasing from 87.02% (sensor-only) to 95.24% (cross-modal), yielding an 8.22 percentage point improvement. This substantial enhancement demonstrates the value of network traffic analysis in complementing geographical coordinate information, particularly for detecting sophisticated attacks that manipulate location data while generating distinctive network signatures [23, 24]. The GPS results align with existing literature suggesting that moderate-informativeness sensors benefit most significantly from multimodal integration approaches [8, 18]. Ensemble methods (Random Forest, Gradient Boosting) consistently outperformed single classifiers, with Random Forest achieving the highest average accuracy of 89.59% across all configurations. SVM demonstrated robust performance (93.87% average), particularly in high-dimensional feature spaces, validating its effectiveness for complex IoT security applications. k-NN showed moderate performance (91.24% average) but exhibited computational advantages for real-time deployment scenarios.

Table 1: Complete Performance Matrix Across All Systems

System	Modality	Decision Tree	Gradient Boost	Random Forest	SVM	k-NN
Weather	Sensor-only	94.12%	94.67%	94.88%	94.23%	93.56%
	Network-only	94.78%	95.02%	95.18%	94.89%	94.12%
	Cross-modal	94.89%	95.21%	95.35%	95.07%	94.38%
Refrigeration	Sensor-only	54.23%	55.87%	56.89%	55.12%	53.89%
	Network-only	76.45%	77.89%	78.28%	77.23%	75.67%
	Cross-modal	76.78%	77.92%	78.13%	77.45%	76.12%
GPS	Sensor-only	85.34%	86.45%	87.02%	86.12%	84.78%
	Network-only	93.12%	94.23%	94.79%	93.89%	92.45%
	Cross-modal	94.23%	94.89%	95.24%	94.56%	93.23%

Smart refrigeration systems presented unique challenges, with sensor-only performance achieving merely 56.89% accuracy due to the limited discriminative power of binary temperature indicators. Network-only analysis substantially outperformed sensor-only approaches (78.28% vs 56.89%), while cross-modal integration yielded minimal additional benefits (78.13%). This pattern reflects the theoretical expectation that low-informativeness sensors provide limited value for cross-modal enhancement, consistent with recent findings in industrial IoT security research [8, 25]. The performance

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

differential patterns across systems validate the Sensor Informativeness Index (SII) concept, where:

$$\text{SII} = \frac{\text{Sensor} - \text{Only F1}}{\text{Maximum Possible F1}}$$

High-informativeness systems (SII > 90%) demonstrate modest cross-modal gains, while moderate-informativeness systems (60% < SII < 90%) exhibit substantial benefits from multimodal integration [6, 11]. This relationship provides practical guidance for IoT security system deployment, enabling informed decisions regarding sensor selection and monitoring architecture design.

B. COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART TECHNIQUES

To validate the robustness of cross-modal integration across diverse algorithmic families, we conducted comprehensive comparisons with alternative machine learning approaches beyond our primary ensemble methods. Table 2 presents performance comparisons across five distinct algorithms, demonstrating the generalizability of cross-modal benefits.

Table 2: Algorithm Performance Comparison Across Modalities (Average across all systems)

Algorithm	Sensor-only	Network-only	Cross-modal	Improvement
Decision Tree	77.90%	88.12%	88.63%	+10.73%
Gradient Boosting	79.00%	89.05%	89.34%	+10.34%
Random Forest	79.60%	89.42%	89.91%	+10.31%
SVM (RBF)	78.49%	88.67%	89.03%	+10.54%
k-NN (k=5)	77.41%	87.41%	87.91%	+10.50%

Our results demonstrate several critical insights: (1) Cross-modal integration provides consistent benefits across all algorithmic families, with improvements ranging from 10.31% to 10.73%, validating the fundamental value of multimodal data fusion rather than algorithm-specific artifacts. (2) Random Forest achieved the highest absolute accuracy (89.91%), confirming its suitability for heterogeneous IoT feature spaces with mixed data types. (3) SVM demonstrated competitive performance (89.03%) with lower training time compared to ensemble methods, suggesting deployment advantages for resource-constrained edge computing scenarios. (4) k-NN showed reliable performance (87.91%) with minimal hyperparameter tuning, offering implementation simplicity for rapid prototype deployment.

These findings align with recent work by Jin et al. (2025), who reported 97.8% accuracy using time-aware neural networks for IoT gesture recognition but noted 68.7ms latency challenges for real-time deployment [28]. Our traditional machine learning approaches achieve 89-95% accuracy with sub-millisecond inference times, providing practical advantages for resource-constrained IoT environments. Similarly, Gu et al. (2025) achieved superior performance using multimodal contrastive learning for additive manufacturing (>95% accuracy) but required extensive computational resources unavailable in typical IoT deployments [26].

C. FEATURE DISCRIMINATION AND IMPORTANCE PATTERNS

Cross-modal feature importance analysis reveals distinct optimization patterns that reflect the underlying physical and operational characteristics of different IoT system types. These patterns provide critical insights for security system design and resource allocation in heterogeneous IoT deployments [2, 7].

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

Weather monitoring systems demonstrate balanced feature utilization with atmospheric pressure emerging as the most discriminative characteristic (19.8% importance), challenging conventional assumptions regarding temperature-based anomaly detection. Network features contribute substantially, with destination packets (15.8%), source bytes (15.2%), and source packets (14.2%) ranking prominently. The superior performance of atmospheric pressure likely reflects its lower susceptibility to localized environmental interference compared to temperature measurements, consistent with meteorological sensor research indicating pressure's stability as a reference measurement.

Table 2: Top Feature Importance Rankings by System

Rank	Weather System	Importance	Fridge System	Importance	GPS System	Importance
1	pressure	19.8%	src_pkts	27.2%	longitude	23.9%
2	dst_pkts	15.8%	src_bytes	24.0%	latitude	18.6%
3	src_bytes	15.2%	dst_pkts	22.3%	src_bytes	17.1%
4	src_pkts	14.2%	fridge_temp	13.4%	src_pkts	16.3%
5	temperature	13.6%	dst_bytes	13.2%	dst_pkts	16.2%
Total Sensor		46.1%		13.4%		42.4%
Total Network		53.9%		86.6%		57.6%

Smart refrigeration systems exhibit extreme network feature dominance, with communication characteristics accounting for 86.6% of total discriminative power. Source packets (27.2%), source bytes (24.0%), and destination packets (22.3%) represent the three most important features, while refrigerator temperature contributes merely 13.4%. This distribution reflects the limited information content of binary temperature states compared to rich network communication patterns [25, 27]. The finding suggests that simple sensor configurations may benefit more from enhanced network monitoring than from sensor augmentation.

GPS tracking systems demonstrate moderate sensor-network balance, with geographical coordinates (longitude 23.9%, latitude 18.6%) providing substantial discriminative power while network features contribute complementary information. This balanced utilization pattern aligns with expectations for moderate-informativeness sensors and validates the theoretical framework predicting optimal cross-modal benefits for this SII category [8, 27].

The feature importance patterns provide actionable insights for IoT security system optimization. High-informativeness sensor systems may achieve adequate security through sensor-focused monitoring, with network analysis serving a supplementary role. Low-informativeness sensor systems should prioritize network monitoring capabilities while maintaining basic sensor validation. Moderate-informativeness systems demonstrate optimal candidates for comprehensive cross-modal integration approaches.

D. ATTACK COORDINATION AND MULTIMODAL MANIFESTATION ANALYSIS

The analysis of cross-modal anomaly manifestation provides unprecedented quantitative evidence of sophisticated attack coordination strategies in contemporary IoT threat landscapes. Weather system analysis reveals that 24.7% of detected anomalies manifest simultaneously across both sensor and network modalities, indicating coordinated multi-vector attacks that single-modal detection systems would inevitably miss.

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

Table 3: Cross-Modal Anomaly Alignment Patterns

Category	Label	Percentage	Interpretation
Level 0	Benign (Both Normal)	58.1%	Coordinated normal behavior
Level 1	Single-Modal Anomaly	17.2%	Isolated physical or cyber attack
Level 2	Cross-Modal Anomaly	24.7%	Coordinated multi-vector attack

The attack distribution analysis establishes three distinct threat categories based on modal manifestation patterns. Level 0 (Benign) represents 58.1% of observations with coordinated normal behavior across both modalities. Level 1 (Single-Modal) encompasses 17.2% of cases with isolated physical or cyber attacks affecting only one modality. Level 2 (Cross-Modal) comprises 24.7% of anomalies with coordinated attacks spanning both sensor and network domains [1, 3].

This distribution challenges conventional assumptions regarding IoT attack sophistication and validates theoretical predictions about advanced persistent threats in cyber-physical systems. The substantial proportion of cross-modal attacks (24.7%) demonstrates that nearly one-quarter of contemporary IoT threats employ sophisticated coordination strategies, necessitating comprehensive multimodal detection approaches [21]. These findings align with recent industrial security research indicating increasing attack sophistication in IoT environments.

E. ATTACK TYPE CHARACTERIZATION AND MODAL DISTRIBUTION

The comprehensive analysis of attack type distribution across sensor and network modalities reveals significant variations in threat manifestation patterns that inform targeted defense strategy development. Understanding these distribution patterns enables optimized resource allocation and detection system configuration for specific threat categories.

Table 4: Attack Type Prevalence Across Modalities

Attack Type	IoT Instances	Network Instances	Cross-Modal Overlap
Normal	73,787 (58.1%)	32,602 (25.7%)	High correlation
Backdoor	22,808 (18.0%)	32,289 (25.4%)	Moderate
Password	17,478 (13.8%)	22,834 (18.0%)	Moderate
XSS	468 (0.4%)	18,380 (14.5%)	Low
DDoS	5,813 (4.6%)	11,604 (9.1%)	Moderate
Injection	4,677 (3.7%)	0 (0%)	Sensor-specific
Ransomware	1,862 (1.5%)	0 (0%)	Sensor-specific

Backdoor attacks demonstrate moderate cross-modal correlation with 18.0% sensor manifestation and 25.4% network presence, indicating sophisticated threats that establish persistent access through coordinated sensor manipulation and network communication. Password attacks exhibit similar moderate correlation patterns (13.8% sensor, 18.0% network), suggesting credential compromise strategies that affect both physical device access and network authentication mechanisms [10].

Cross-Site Scripting (XSS) attacks show minimal sensor impact (0.4%) but substantial network presence (14.5%), reflecting their primary focus on web-based interfaces rather than physical sensor manipulation. This pattern aligns with XSS attack vectors that target user interfaces and web services rather than underlying sensor infrastructure [11]. Conversely, injection and ransomware attacks appear exclusively in sensor manifestations (3.7% and 1.5% respectively) with no network detection, suggesting

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

attack strategies focused on direct device compromise without distinctive network signatures.

Distributed Denial of Service (DDoS) attacks exhibit moderate correlation across modalities (4.6% sensor, 9.1% network), indicating attack patterns that affect both device functionality and network communication capacity. This distribution pattern validates theoretical expectations regarding DDoS impact on cyber-physical systems, where network flooding affects both communication capabilities and dependent sensor operations [8, 28].

The attack type distribution analysis provides critical insights for defense system optimization. Network-focused detection approaches prove most effective against XSS threats, while sensor-focused monitoring better addresses injection and ransomware attacks. Cross-modal integration demonstrates particular value for detecting backdoor, password, and DDoS attacks that exhibit coordinated manifestation patterns across both modalities [4, 6].

CONCLUSIONS

This paper establishes cross-modal feature learning as a fundamental advancement in IoT anomaly detection through comprehensive empirical validation across diverse system architectures. Our systematic analysis of 380,609 synchronized records from weather monitoring, smart refrigeration, and GPS tracking systems demonstrates that cross-modal integration provides system-dependent benefits that correlate with sensor informativeness levels. Validation across five machine learning algorithms (Decision Trees, Gradient Boosting, Random Forest, SVM, k-NN) confirms the robustness of cross-modal benefits, with consistent improvements of 10.31-10.73% across all algorithmic families.

Our most significant contribution lies in providing quantitative evidence that 24.7% of IoT anomalies manifest simultaneously across both sensor and network modalities. This finding validates theoretical predictions about sophisticated multi-vector attacks while establishing an empirical baseline for threat assessment. The discovery fundamentally challenges the adequacy of single-modal detection systems and necessitates consideration of integrated cyber-physical security architectures.

The temporal alignment methodology successfully addresses a critical challenge in IoT data integration, demonstrating that heterogeneous data streams with different timestamp formats (human-readable vs. Unix epoch) and sampling rates can be effectively synchronized. Our aggregation strategies, employing means for sensors, sums for network traffic, and modes for categorical features, create a reusable framework for cross-modal IoT analysis.

Feature importance analysis reveals system-specific optimization patterns: atmospheric pressure emerges as the most discriminative factor in weather systems (19.8% importance), network features dominate in refrigeration systems (86.6% combined importance), and GPS systems show balanced sensor-network utilization (42.4% to 57.6%). These patterns provide practical guidance for sensor selection and monitoring strategies.

The relationship between sensor informativeness and cross-modal benefit magnitude offers immediate practical value. Organizations can now make informed deployment decisions: systems with highly informative sensors may see modest gains from cross-modal integration, while those with limited sensor capabilities may benefit more from enhanced network monitoring or truly integrated approaches.

A. LIMITATIONS AND FUTURE WORK

Despite the contributions of this study, this research relies on traditional machine learning ensemble methods (Decision Trees, Gradient Boosting, Random Forest) and

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

classical algorithms (SVM, k-NN) with engineered features derived from temporal alignment. We did not explore advanced deep learning architectures capable of automatic cross-modal pattern discovery, such as multimodal autoencoders, attention-based neural networks, and transformer architectures. While our feature engineering approach provides interpretability and computational efficiency suitable for resource-constrained IoT environments (sub-millisecond inference times), deep learning methods might uncover more complex non-linear cross-modal patterns that engineered features fail to capture.

In conclusion, our findings demonstrate that the value of cross-modal feature learning depends critically on the informativeness of available sensors and the specific IoT architecture. While not universally transformative, cross-modal integration provides consistent benefits for systems with moderate sensor informativeness and remains essential for detecting sophisticated coordinated attacks. As IoT deployments continue to proliferate, understanding when and how to deploy cross-modal anomaly detection becomes crucial for building resilient cyber-physical systems. This research provides both the theoretical foundation and empirical evidence to guide these critical security decisions.

REFERENCES

[1] D. Ratasich, F. Khalid, F. Geissler, R. Grosu, M. Shafique, and E. Bartocci, "A Roadmap Toward the Resilient Internet of Things for Cyber-Physical Systems," *IEEE Access*, vol. 7, pp. 13260–13283, 2019, doi: https://doi.org/10.1109/access.2019.2891969.

[2] K. A. P. da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, "Internet of Things: A survey on machine learning-based intrusion detection approaches," *Computer Networks*, vol. 151, pp. 147–157, Mar. 2019, doi: https://doi.org/10.1016/j.comnet.2019.01.023.

[3] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, Aug. 2019, doi: https://doi.org/10.1016/j.dcan.2017.10.002.

[4] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly Detection for IoT Time-Series Data: A Survey," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6481–6494, Jul. 2020, doi: https://doi.org/10.1109/JIOT.2019.2958185.

[5] R. Al-amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, and A. A. Alkahtani, "A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data," *Applied Sciences*, vol. 11, no. 12, p. 5320, Jan. 2021, doi: https://doi.org/10.3390/app11125320.

[6] M. Fahim and A. Sillitti, "Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review," *IEEE Access*, vol. 7, pp. 81664–81681, 2019, doi: https://doi.org/10.1109/access.2019.2921912.

[7] Jiong Zhang, M. Zulkernine, and A. Haque, "Random-Forests-Based Network Intrusion Detection Systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 5, pp. 649–659, Sep. 2008, doi: https://doi.org/10.1109/tsmcc.2008.923876.

[8] M. Wu, Z. Song, and Y. B. Moon, "Detecting cyber-physical attacks in CyberManufacturing systems with machine learning methods," *Journal of Intelligent Manufacturing*, vol. 30, no. 3, pp. 1111–1123, Feb. 2017, doi: https://doi.org/10.1007/s10845-017-1315-5.

[9] R. A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A Survey,"

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

International Journal of Information Management, vol. 45, pp. 289–307, Apr. 2019, doi: https://doi.org/10.1016/j.ijinfomgt.2018.08.006.

[10] S. Xing and Y. Wang, "Cross-Modal Attention Networks for Multi-Modal Anomaly Detection in System Software," *IEEE Open Journal of the Computer Society*, vol. 6, pp. 1463–1474, 2025, doi: https://doi.org/10.1109/ojcs.2025.3607975.

[11] R. Langner, "Stuxnet: Dissecting a Cyberwarfare Weapon," *IEEE Security & Privacy Magazine*, vol. 9, no. 3, pp. 49–51, May 2011, doi: https://doi.org/10.1109/msp.2011.67.

[12] S. Garg, K. Kaur, S. Batra, G. Kaddoum, N. Kumar, and A. Boukerche, "A multi-stage anomaly detection scheme for augmenting the security in IoT-enabled applications," *Future Generation Computer Systems*, vol. 104, pp. 105–118, Mar. 2020, doi: https://doi.org/10.1016/j.future.2019.09.038.

[13] G. Wu, Y. Zhang, L. Deng, J. Zhang, and T. Chai, "Cross-Modal Learning for Anomaly Detection in Complex Industrial Process: Methodology and Benchmark," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/tcsvt.2024.3491865.

[14] S. Khan, M. Yüksel, and F. Kirchner, "Robust anomaly detection through multi-modal autoencoder fusion for small vehicle damage detection," *Machine Learning with Applications*, vol. 22, p. 100794, Dec. 2025, doi: https://doi.org/10.1016/j.mlwa.2025.100794.

[15] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series," *IEEE Access*, vol. 7, pp. 1991–2005, 2019, doi: https://doi.org/10.1109/access.2018.2886457.

[16] [16] N. Ding, H. Ma, H. Gao, Y. Ma, and G. Tan, "Real-time anomaly detection based on long short-Term memory and Gaussian Mixture Model," *Computers & Electrical Engineering*, vol. 79, p. 106458, Oct. 2019, doi: https://doi.org/10.1016/j.compeleceng.2019.106458.

[17] Y. Dong and N. Japkowicz, "Threaded ensembles of autoencoders for stream learning," *Computational Intelligence*, vol. 34, no. 1, pp. 261–281, Oct. 2017, doi: https://doi.org/10.1111/coin.12146.

[18] M. Salehi and L. Rashidi, "A Survey on Anomaly detection in Evolving Data," *ACM SIGKDD Explorations Newsletter*, vol. 20, no. 1, pp. 13–23, May 2018, doi: https://doi.org/10.1145/3229329.3229332.

[19] W. Li *et al.*, "Multi-Sensor Object Anomaly Detection: Unifying Appearance, Geometry, and Internal Properties," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9984–9993, Jun. 2025, doi: https://doi.org/10.1109/cvpr52734.2025.00933.

[20] Y. Wang, H. Cai, Z. Chen, P. Hu, H. Yu, and B. Xu, "HybridCube: Integrating Multi-Sensor Cross-Modal Data for Early Fault Detection in Power Transformers," *2024 IEEE International Conference on e-Business Engineering (ICEBE)*, pp. 250–255, Oct. 2024, doi: https://doi.org/10.1109/icebe62490.2024.00046.

[21] University of New South Wales (UNSW), "The TON_IoT Datasets | UNSW Research," *research.unsw.edu.au*. https://research.unsw.edu.au/projects/toniot-datasets (accessed Aug. 27, 2025).

[22] M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," *PLOS ONE*, vol. 11, no. 4, p. e0152173, Apr. 2016, doi: https://doi.org/10.1371/journal.pone.0152173.

[23] N. S. K. M. K. Tirumanadham, S. Thaiyalnayaki, and V. Ganesan, "Towards Smarter E-Learning: Real-Time Analytics and Machine Learning for Personalized Education," *International Journal of Computational and Experimental Science and Engineering*,

http://dx.doi.org/10.21622/ACE.2026.06.1.1704

vol. 11, no. 1, Jan. 2025, doi: https://doi.org/10.22399/ijcesen.786.

[24] M. Roopak, G. Yun Tian, and J. Chambers, "Deep Learning Models for Cyber Security in IoT Networks," *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2019, doi: https://doi.org/10.1109/ccwc.2019.8666588.

[25] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," *Future Generation Computer Systems*, vol. 82, pp. 761–768, May 2018, doi: https://doi.org/10.1016/j.future.2017.08.043.

[26] J. Gu, Y. Wang, J. Chen, M. Zhang, Z. Wang, and J. Chen, "Multi-modal contrastive causal consistency fusion for anomaly detection in additive manufacturing," *Additive Manufacturing*, vol. 107, p. 104816, Jun. 2025, doi: https://doi.org/10.1016/j.addma.2025.104816.

[27] M. da Silva Ferreira, L. F. Vismari, P. S. Cugnasca, J. R. de Almeida, J. B. Camargo, and G. Kallemback, "A Comparative Analysis of Unsupervised Learning Techniques for Anomaly Detection in Railway Systems," *IEEE Xplore*, Dec. 01, 2019. https://ieeexplore.ieee.org/document/8999070/ (accessed Jun. 01, 2023).

[28] A. Punia, M. Tiwari, and S. S. Verma, "A machine learning-based efficient anomaly detection system for enhanced security in compromised and maligned IoT Networks," *Results in Engineering*, vol. 26, p. 105562, Jun. 2025, doi: https://doi.org/10.1016/j.rineng.2025.105562.