Multimodal Machine Learning for Cybersecurity in Internet of Things Environments: A Literature Review

Abdelaaziz NASSIRI, Azeddine Khyat, Kama El Guemmat, Mohamed Aazi
Computing, Artificial Intelligence and Cyber Security Laboratory (2IACS),
ENSET Mohammedia, Hassan II University of Casablanca, Morocco

***Abstract*—This research provides a comprehensive synthesis of Multimodal Machine Learning (MML) as a transformative paradigm for IoT defense. By integrating heterogeneous data streams, including network flow statistics, device-level telemetry, and behavioral biometrics, MML architectures facilitate a holistic understanding of system states. The algorithmic advancements that were analyzed are classified into hybrid CNN-RNN structures and state-of-the-art cross-modal Transformers, and evaluate their performance across benchmark datasets such as ToN-IoT and CICIoT2023. Quantitative results show that cross-modal Transformers achieve F1-scores between 0.95 and 0.99 across detection tasks, while hybrid CNN-LSTM models range from 0.89 to 0.96. Furthermore, this study addresses the technical "optimization triad" of pruning, quantization, and edge-cloud orchestration required to deploy these models on resource-constrained hardware.**

***Keywords*—Internet of Things (IoT); Multimodal Machine Learning; deep learning, cyber-physical security; data fusion; Intrusion Detection Systems (IDS); Edge AI, Explainable AI (XAI)**

I. Introduction

The digital landscape is currently undergoing a fundamental reconfiguration, driven not by incremental expansion but by the convergence of ubiquitous connectivity, exponential data proliferation, and transformative advancements in computational intelligence. Central to this paradigm shift is the Internet of Things (IoT), a technological framework that has evolved from a disparate collection of connected devices into a central component of contemporary digital infrastructure.

As billions of autonomous nodes from tiny environmental sensors to heavy-duty industrial actuators stream data in real-time, the traditional boundary between the digital and the physical is effectively dissolving into a cyber-physical continuum.

However, security frameworks have struggled to keep pace with this rapid expansion. The massive, fragmented attack surface presents challenges that traditional network defense mechanisms cannot adequately address [1] [7].

A. Context: IoT Convergence and Security Bottlenecks

The rapid integration of IoT into critical infrastructure, including power grids, healthcare systems, and transportation networks, has fundamentally altered the risk profile of modern civilization. In these high-stakes environments, security is not merely a matter of data privacy; it constitutes a non-negotiable requirement for operational safety.

The Industrial Internet of Things (IIoT) illustrates this dynamic. Industrial systems are exposed to cyber threats that can result in physical consequences, as the integration of physical machinery with intelligent electronics has yielded significant efficiency improvements. A potential physical catastrophe can result not only from a software defect but also from a digital exploit within a control loop [4] [5] [6] [8].

The IoT ecosystem introduces several structural challenges that confound traditional IT security approaches:

- Extreme Heterogeneity: The wide variety of hardware architectures and proprietary protocols renders the application of a universal security standard practically infeasible.
- Resource Constraints: Many devices operate under severe power and memory limitations, preventing the execution of standard encryption or heavy anti-malware suites.
- Data Volume: With an estimated 26 billion devices expected by 2030, the resulting data volume makes manual oversight impossible. The industry is now forced to rely on automated, intelligent defense systems [2] [3] [9].

B. Problem Statement: The Limitations of Legacy Defense

Traditional cybersecurity, built for the static "walled garden" networks of a previous decade, faces significant challenges in addressing the dynamic complexity of IoT. Most current systems still rely on unimodal analysis that examines a single stream, such as network logs, to detect anomalies. While this approach keeps computational costs low, it creates a blind spot: it cannot detect multi-vector attacks that span different system layers [1] [9].

Despite being the most widely deployed approach, signature-based IDS is not effective against zero-day exploits. Anomaly-based models often cannot distinguish between benign deviations from baseline behavior and actual attacks, leading to an excessive amount of false positive alerts. These systems have a tendency to fail to distinguish between a true intrusion and harmless network latency or transient sensor malfunction [2] [5] [35] [37].

This limitation of unimodal approaches creates vulnerabilities that sophisticated attackers can exploit. An adversary can execute a sequence of actions that appear benign individually such as a minor CPU spike combined with a small shift in packet timing but collectively indicate a security breach. Without the ability to correlate these scattered signals, traditional systems lack the holistic situational awareness needed to defend complex IoT deployments [17] [18] [12].

C. Motivation: Why Multimodal Approaches are the Future?

The limitations of single-source models have motivated a shift toward Multimodal Machine Learning (MML). MML provides a unified framework to ingest and analyze diverse data types simultaneously. By fusing network traffic, device telemetry, and even behavioral biometrics, MML systems can catch the subtle indicators of an attack that are invisible to a unimodal filter [12] [18] [19].

The fundamental principle is data complementarity. Information that appears as noise in one channel can be clarified by another. For example, a suspected DDoS attack on the network can be validated by a corresponding increase in a device's heat profile (high temperature alarm) or a rapid decrease in accessible RAM. This cross-modal validation enhances detection and significantly reduces false positive rates [2] [11].

1) Contribution and Scope: This review integrates the technological frontier at the intersection of MML and IoT security. The focus is exclusively on the mechanics of data fusion rather than doing a comprehensive study; all models were classified from hybrid deep learning to Transformer-based networks, addressing the challenging issue of executing these resource-intensive models on constrained edge devices [3] [7] [9].

2) Organization of the paper: The remainder of this document is organized as follows: Section II establishes the preliminaries, detailing the IoT threat landscape, the variety of data modalities available, and the core strategies of multimodal fusion. Section III presents the systematic review methodology. Section IV provides taxonomy and classification of MML architecture. Section V discusses applications and use cases. Section VI reviews datasets and evaluation metrics. Section VII provides a critical discussion of trade-offs and limitations. Section VIII identifies open challenges and future directions. Section IX concludes the review.

II. Background And Preliminaries

Developing an effective multimodal security framework requires a comprehensive understanding of the operational environment, the nature of the data involved, and the mathematical mechanisms for integrating that data [1] [3] [7].

A. IoT Security Landscape: Major Threats and Vulnerabilities

The IoT threat landscape is characterized by its breadth and the potential consequences of successful exploitation. Threats are typically categorized based on their intent, origin, and the specific layer of the IoT stack they target [1] [3] [22]. Table I presents the IoT threat taxonomy.

TABLE I. IoT Threat Taxonomy

Threat Category	Targeted Layer	Mechanism	Primary Impact
DDoS Attacks	Network	Overwhelming target resources using a botnet.	System unavailability and network collapse.
Malware/ Botnets	Application/Edge	Mirai, BASHLITE, or ransomware infecting devices.	Information leakage and device takeover.
False Data Injection	Physical/ Sensing	Injecting fraudulent sensor data to mislead control systems.	Grid instability or process failure.
Slow DoS	Network/ App	Stealthy low-volume traffic mimicking legitimate slow nodes.	Resource exhaustion with low visibility.
Adversarial Evasion	Learning	Manipulating inputs to fool ML models.	Compromised detection accuracy.

Sources: [3] [23] [36]

The fundamental vulnerability of IoT systems lies in the "security-by-design" gap. Devices are often deployed with default credentials, unpatched firmware, and a lack of standardized development guidelines. Furthermore, the reliance on lightweight protocols such as IEEE 802.15.4e or MQTT, while efficient for communication, can increase susceptibility to topological manipulation and resource exhaustion attacks [3] [23] [36].

TABLE II. Comparative Benchmarking of Surveyed MML Architectures for IoT Security.

Architecture	Fusion Strategy	Dataset	Acc%	F1	AUC
CNN-LSTM	Hybrid	Custom multi-modal	96,04 [37]	0.960	–
CNN-LSTM / CNN-GRU (IDS)	Hybrid	CICIoT2023 / ToN-IoT	>99.6 [36,37]	0.996	-
RansomFormer	Cross-attention	Custom ransomware	99.5 [43]	0.995	0.998
MT-CMVAD	Transformer fusion	UCF-Crime / CUHK	- [41]	-	0.989
TTGNet-AMD	Hybrid DL + GCN	Drebin / VirusTotal	~98.7 [30]	0.985	0.990
FW-CNN + Metaheuristic	Feature-level	ToN-IoT	98.4 [11]	0.982	-
Soft-Voting Ensemble	Decision-level	ToN-IoT / Bot-IoT	99.1 [49]	0.989	-
AI4FIDS (Federated)	Late fusion (FL)	Multi-domain IIoT	97.8 [40]	0.976	-
Autoencoder (Anomaly	Latent-space	N-BaIoT	97.3 [39]	0.971	0.981
GAN-augmented MML	Feature-level	CICIoT2023	99.3 [39]	0.991	-

Source: [11, 30, 36, 37, 39, 40, 41, 43, 49]

Accuracy and F1 values are reported on the primary dataset cited by each study (see Table II).

B. Overview of Data Modalities in IoT

In a multimodal context, a "modality" refers to a distinct stream of information that captures a specific aspect of a system's state or behavior. IoT ecosystems are rich in these modalities [12] [15] [16] [18] [32] [61]:

- Network Flow and Packets: This includes tabular features like flow duration, protocol type, and connection rates, as well as raw packet-level features such as payload entropy and application-layer semantics (e.g., MQTT message types).
- Device Telemetry: Internal performance metrics such as CPU load, RAM usage, memory mapping, and power consumption patterns.
- Sensor Data: The primary functional output of the device, which may include numerical readings (temperature, pressure) or high-dimensional data (images, audio).
- Static Code Features: Structural information extracted from software binaries, such as opcode sequences, Application Programming Interface (API) call sequences, and function call graphs (FCG).
- Behavioral Biometrics: Patterns derived from human interaction, including keystroke dynamics, gait, and touch gestures.

C. Foundations of Multimodal Machine Learning (MML)

Multimodal Machine Learning is the computational process of integrating information from multiple modalities to perform inference. The effectiveness of an MML system depends largely on its fusion strategy, the stage at which different data types are combined [12] [17] [18] [19] [61].

1) Early Fusion (Data-Level): This strategy involves concatenating raw features from different sources into a single vector before feeding it into the model. While this allows the model to capture deep inter-modal interactions from the beginning, it is highly sensitive to temporal misalignment and differences in data scales.

2) Late Fusion (Decision-Level): Each modality is processed by an independent model, and the individual outputs are merged at the final stage using techniques like majority voting, averaging, or weighted confidence scores. This is more robust to modality-specific noise but often misses the nuanced feature-level dependencies that define complex attacks.

3) Hybrid Fusion (Feature-Level): The most prevalent approach in modern research. It utilizes modality-specific encoders to extract high-level feature representations, which are then integrated in the intermediate layers of the network. This allows for a balance between modality-specific learning and global context integration.

III. Review Methodology

To ensure reproducibility and rigor, this review follows these systematic review steps.

A. Search Strategy

Literature was retrieved from major academic databases covering the period 2018 – 2026.

The digital libraries were especially, but not limited to: Google Scholar, IEEE Xplore, SpringerLink.

The following primary Boolean query was applied consistently across all selected databases:

("IoT" OR "IIoT") AND ("Multimodal Machine Learning" OR "Multi-modal Deep Learning" OR "Data Fusion" OR "Feature Fusion" OR "Cross-modal Attention") AND ("Cybersecurity" OR "Intrusion Detection" OR "Anomaly Detection" OR "Malware" OR "Authentication" OR "IDS" OR "Attack Detection")

B. Inclusion and Exclusion Criteria

Inclusion criteria: 1) Peer-reviewed articles or preprints from 2018-2026; 2) Explicit use of at least two distinct data modalities; 3) Application to IoT security (IDS, malware detection, or authentication); 4) Reporting of quantitative performance metrics.

Exclusion criteria: 1) Theoretical papers without experimental validation; 2) Single-modality studies; 3) non-English publications.

C. Sélection Process

The selection was based on the essential point of multimodality and intrusion detection systems, with papers addressing the most relevant databases and meeting the criterion of multimodality; the second selection criterion was recent references. This allowed us to keep about 60 reference articles.

IV. Taxonomy of MML Architectures for IoT Security

Recent progress in multimodal machine learning for IoT security has focused on creating systems that combine different types of data and make clear how they are all connected. In the contexts of IoT and IIoT, these modalities may include network traffic characteristics, device-level telemetry, system logs, sensor data, and, in some cases, behavioral signals. This section of the study groups the literature by the basic architecture of machine learning, with a focus on deep learning models, attention-based methods, Transformer-centric architectures, and hybrid ensemble frameworks. The goal is to look at how different types of architecture do multimodal fusion, deal with interactions between features from different modes, and solve basic IoT problems like data heterogeneity, temporal dynamics, and limited computing power [1] [7] [9] [34].

A. Deep Learning-Based Multimodal Models

Deep Learning (DL) is the most popular MML paradigm because it can learn hierarchical representations without manual feature engineering [18] [34] [35] [43] (see Table III).

TABLE III. MML ARCHITECTURE TAXONOMY

Architecture Family	Early Fusion	Hybrid Fusion	Late Fusion
CNN-based	Feature concatenation before CNN layers [35]	FW-CNN + metaheuristic hyperparameter tuning [11]	Ensemble: CNN + RF/SVM decision voting [34][48]
RNN/LSTM-based	Multimodal LSTM on concatenated sequences [37]	CNN-LSTM / CNN-GRU — spatial + temporal [6][35][37]	LSTM outputs + federated aggregation [8]
Transformer-based	Token-level cross-modal embedding [33]	RansomFormer cross-attention (bytes+API) [42]; MT-CMVAD [40]	Transformer ensemble with confidence weighting [34]
Hybrid Ensemble	Multi-feature static+dynamic fusion [25][29]	TTGNet-AMD: LSTM + Transformer-GCN [29]; AI4FIDS FL [39]	Soft-voting ensemble over independent models [11][48]
Autoencoder/ GAN	Joint latent-space reconstruction [10]	GAN augmentation + CNN classifier [38]	GAN synthetic data + downstream detector [10][38]

B. Convolutional and Recurrent Neural Network Integration

A common architecture for analyzing network and telemetry data is the combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). CNN layers are typically used to extract spatial or structural patterns from a window of data, while LSTM/GRU layers capture the temporal dependencies essential for identifying multi-stage attacks that unfold over time. The research studied highlights the efficacy of the hybrid CNN-LSTM model, which achieved an average accuracy of 96.04% across various behavioral modalities. In network intrusion detection, CNN-LSTM and CNN-GRU models have demonstrated accuracies exceeding 99.83%, surpassing the CNN model by 0.93% in accuracy, outperforming traditional deep learning techniques by effectively modeling both the spatial structure of packet flows and their sequential order [6] [35] [34] [37][62].

C. Autoencoders and Generative Models

Autoencoders (AE) are frequently employed in multimodal settings to learn a shared latent representation across different data sources. By training an AE to reconstruct "normal" multimodal inputs, any significant reconstruction error in a new input can be flagged as an anomaly. Generative Adversarial Networks (GANs) are increasingly used to enhance MML-based detection. GANs serve a dual purpose: first, they generate synthetic multimodal data to address the common problem of class imbalance, where malicious samples are rare compared to benign traffic, improving the robustness of the classifier. Second, they facilitate adversarial training, helping models recognize and resist "evasion" attacks where adversaries deliberately manipulate one modality to mask anomalies in another [10] [38] [41].

D. Attention and Transformer-Based MML

The most significant architectural shift in recent years is the adoption of the Transformer model and its attention mechanisms. Unlike traditional models that treat all inputs with equal weight, attention mechanisms allow a model to dynamically focus on the most relevant features or modalities for a specific task [10] [33] [40] [42].

1) Cross-modal attention mechanisms: The "cross-attention" mechanism is particularly vital for MML. It allows one modality (the "query") to search for relevant features in another modality (the "key" and "value"). This enables the model to align and bridge the semantic gap between disparate data types. For instance, in malware analysis, the RansomFormer model uses a cross-attention fusion layer to relate static Portable Executable (PE) byte data with dynamic API call sequences. The mathematical foundation of this fusion typically involves a query vector Q derived from one modality (e.g., bytes) and key K and value V vectors derived from another (e.g., APIs). The attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right) \cdot V$$

where, d_k is the dimensionality of the keys. This allows the model to prioritize API calls that are most relevant to specific byte-level signatures, achieving near-perfect accuracy (up to 99.5%) in ransomware detection [42].

Similar transformer-based models like MT-CMVAD have shown significant improvements in video anomaly detection, achieving AUC scores of up to 98.9% while reducing computational complexity [41].

2) Hybrid and ensemble models: Despite the dominance of deep learning, hybrid models that combine traditional machine learning (ML) with DL often provide better results in resource-constrained IoT settings. Traditional models like Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT) are prized for their simplicity, interpretability, and low computational requirements [2] [34] [48].

In a hybrid architecture, a deep learning model might be used for high-level feature extraction, with the final classification performed by a robust ensemble. For example, the FW-CNN model utilizes a one-dimensional CNN for feature learning, but tunes its hyperparameters using a hybrid metaheuristic optimizer to enhance its stability and detection rate. Another approach, the soft-voting ensemble classifier, fuses outputs from different specialized models using a confidence-weighted strategy, which has been shown to significantly reduce false positives in multimodal threat detection [11] [34] [48].

V. APPLICATIONS AND USE CASES

The practical impact of MML in IoT is best illustrated by its application to specific cybersecurity goals, where the fusion of diverse data sources directly addresses the weaknesses of unimodal systems [5] [6] [8] [39].

A. Multimodal Intrusion Detection Systems (M-IDS)

The dominant application of MML is the Intrusion Detection System (IDS). Modern M-IDS frameworks seek to create a "holistic" view of network and device behavior [5] [6] [8] [9] [35] [37].

1) Cyber-physical and industrial fusion: In smart grids and IIoT, M-IDS combines network traffic logs with physical sensor readings. An attacker might attempt a False Data Injection (FDI) to destabilize a power grid. While the individual sensor readings might appear within normal bounds, an MML model can detect the attack by identifying the subtle inconsistency between the sensors' reported states and the corresponding control-layer network commands [4] [5].

2) Federated and privacy-preserving detection: Systems like AI4FIDS leverage Federated Learning (FL) to combine data from multiple domains (e.g., different factories or hospitals), without compromising local data privacy. This allows the system to learn from a global pool of heterogeneous attack signatures while keeping sensitive local data behind its own firewall [8] [39].

3) Packet and metadata analysis: Traditional IDS often ignores the application layer. Newer MML models, such as the one proposed in NAIIDS4IoT, fuse low-level flow statistics with rich application-layer metadata, such as MQTT topic structures and payload entropy. This enables the detection of stealthy "Slow DoS" attacks that mimic legitimate network latency [5] [35].

B. Multimodal Malware Analysis

IoT malware has become increasingly evasive, using packing and encryption to hide its code from static scanners. MML addresses this by fusing static code features with dynamic behavior logs [10] [25] [28] [29] [43] [44] [52]. Table IV presents the multimodal malware analysis.

TABLE IV. MULTIMODAL MALWARE ANALYSIS

Feature Type	Specific Modalities	Core Advantage in MML
Static	Opcode sequences, Function Call Graphs (FCG), Byte-images.	Captures the comprehensive structural intent of the code.
Dynamic	System call traces, Network activity, Memory usage.	Resilient to code obfuscation and packing.
Fused	Cross-attention over structural and behavioral features.	Maximizes detection accuracy (99%+) and attribution to families.

Source: [10] [25] [28] [29] [43] [44] [52]

The TTGNet-AMD framework exemplifies this approach by using LSTM to process temporal opcode sequences while simultaneously using a Transformer-GCN hybrid to analyze the structural FCG. This collaborative interaction allows the model to identify malicious behavior even if the binary has been heavily obfuscated [29].

C. Secure Authentication and Access Control

Authentication in IoT is transitioning from static passwords to Continuous Authentication (CA) based on multimodal identity factors [26] [30] [31] [49] [50].

1) Behavioral and interaction fusion: MML models can verify a user's identity in real-time by fusing motion sensor data (accelerometer, gyroscope) with touch interaction patterns. This creates a "personal motion fingerprint" that is unique to the individual's biomechanics [31].

2) Biometric and liveness detection: Advanced Multi-Factor Authentication (MFA) systems integrate physiological traits (face, fingerprint, iris) with liveness indicators like Electrocardiogram (ECG) signals. Fusing these signals through a Siamese Neural Network architecture ensures that the system is not fooled by high-resolution photos or synthetic fingerprint replicas [30] [49] [50].

3) Continuous risk assessment: CA systems use MML to continuously monitor risk based on current and historical context. If a user's behavioral signature (e.g., typing speed or gait) deviates significantly from their profile, the system can automatically trigger a re-authentication request or lock the device [26] [31].

VI. DATASETS AND EVALUATION

The development of high-performing MML models is dependent on the availability of high-quality, diverse datasets that reflect the unique challenges of the IoT environment [13][24][27][46][47][51].

A. Common Multimodal IoT Security Datasets

Traditional datasets like NSL-KDD or Kyoto 2006+ are largely outdated for IoT research, as they do not capture the specific telemetry and protocols used in modern smart systems. Several benchmark datasets have emerged to fill this gap:

1) ToN-IoT: Sourced from the UNSW Canberra repository, this dataset integrates telemetry data from various sensors (weather stations, smart fridges) with network traffic logs and system event files from heterogeneous IoT devices. It captures a wide range of modern attacks, including ransomware, backdoors, and password cracking [13] [27].

2) Bot-IoT: A realistic dataset generated at the Cyber Range Lab of UNSW-Canberra, containing 72 million records that model IoT botnet activities such as DoS, DDoS, and information theft [51].

3) IoTID20: This dataset models a smart home network with devices like AI speakers, WiFi cameras, and smartphones, capturing 83 features across benign and malicious sessions [47].

4) CICIoT2023: An extensive and realistic dataset that executes 33 different attack types (divided into 7 classes) against a topology of 105 devices, including IoT devices as both attackers and victims [24].

5) N-BaIoT: Addressing the lack of public botnet data, this dataset provides real traffic from nine commercial IoT devices infected by Mirai and BASHLITE botnets [46] [51].

B. Evaluation Metrics and Benchmarking

Evaluating MML security systems requires a move beyond simple accuracy. In cybersecurity, the costs of a False Negative

(missing an attack) and a False Positive (blocking a legitimate user) are often asymmetrical.

1) Precision, Recall, and F1-score: These metrics are essential for evaluating performance on imbalanced datasets. Precision measures the reliability of an alarm, while Recall measures the system's ability to catch all malicious events. The F1-score provides the harmonic mean, which is critical for balancing these trade-offs.

2) Equal Error Rate (EER): Primarily used in authentication, EER identifies the point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal, serving as a primary indicator of a system's robustness.

3) Inference latency and resource overhead: For IoT devices, the time and energy required to make a prediction are as important as the prediction's accuracy. Metrics such as "packets per second", throughput, and model size in kilobytes (KB) are increasingly used to benchmark models for edge deployment [20] [21] [45].

VII. Edge-Cloud Deployment Trade-offs

Deploying Multimodal Machine Learning (MML) models on resource-constrained Internet of Things (IoT) devices necessitates a principled trade-off between detection accuracy and computational efficiency. Within this context, edge computing emerges as a foundational architectural pillar, enabling low-latency inference and reduced bandwidth consumption through localized data processing, while simultaneously reinforcing system resilience via adaptive, on-device responses to environmental stressors or adversarial conditions. Conversely, while cloud-centric platforms offer substantial advantages for large-scale model training, centralized data governance, and long-term policy management, they introduce non-trivial inference delays and inherit systemic vulnerabilities, including heightened privacy risks and exposure to single points of failure. To reconcile these competing imperatives, hybrid edge–cloud architectures distribute computational workloads based on task criticality: real-time anomaly flagging and time-sensitive inference are executed at the edge, whereas computationally intensive cross-modal analysis, periodic model retraining, and high-level coordination are selectively offloaded to the cloud. This collaborative paradigm establishes a complementary, rather than substitutive, relationship between edge and cloud tiers, thereby simultaneously satisfying the demands of real-time responsiveness, systemic scalability, and long-term operational resilience in resource-constrained IoT environments [20][21][45][53] (see Table V).

TABLE V. Quantitative Edge vs. Cloud Inference Trade-Offs

Deployment Mode	Inference Latency	Accuracy Δ -vs. Cloud	Representative Ref.
Full Cloud	50–200 ms	Baseline (0%)	[20][21]
Full Edge (quantised INT8)	5–15 ms	−2 to −4%	[45][58][60]
Selective Edge-Cloud Offload	8–25 ms (local); 80–150 ms (offloaded)	−0.5 to −1.5%	[20][21][53]
Federated (distributed edge)	10–30 ms per node	−1 to −3%	[8][39]

VIII. Open Challenges and Future Directions

The journey from academic research to practical, large-scale deployment of MML in IoT security is hindered by several significant obstacles that define the current research frontier [3] [20] [21] [22] [45] [53].

A. Data Synchronization and Spatiotemporal Alignment

A fundamental challenge in MML is the "asynchronous nature" of IoT data. Different sensors operate at different sampling rates; for instance, a network flow meter might provide updates every second, while a thermal sensor reports every minute, and a camera captures thirty frames per second.

Integrating these discordant streams requires robust alignment techniques. Current research into cross-modal attention mechanisms and transformer-based interpolation offers a path forward, allowing models to learn relationships across time steps even when data is missing or sporadically available. Furthermore, spatiotemporal misalignment where sensors have different spatial resolutions or perspectives remains an open problem, particularly in smart city environments [15] [16] [17] [32] [40].

B. Explainability and Trustworthiness (XAI)

As MML models grow in complexity, they increasingly suffer from the "black-box" issue, where their decision-making logic is opaque to human analysts. In security contexts, this lack of transparency is a critical liability, as security personnel must be able to validate and trust a system's output before taking remediation actions.

The integration of Explainable AI (XAI) tools like SHAP and LIME is becoming mandatory for advanced M-IDS. SHAP (Shapley Additive Explanations), based on game theory, can identify the specific contribution of each feature or modality to a prediction, providing a "visual explanation" that can help an analyst distinguish between a genuine cyberattack and a benign anomaly. Future research must focus on making these XAI methods computationally lightweight enough to run on IoT hardware [14] [41] [54] [55] [56] [57].

C. Resource Constraints and the Optimization Triad

The most persistent challenge is the "resource-performance trade-off". High-accuracy deep learning models, particularly large multimodal transformers, are computationally intensive and require significant memory and energy. IoT devices, however, often have power budgets below 10 watts and only kilobytes (KB) or megabytes of RAM [20] [21] [45] [53] [58] [59] [60].

The "Optimization Triad" addressing this includes:

1) Model compression: This includes pruning (removing redundant weights or neurons), quantization (reducing numerical precision from 32-bit floats to 8-bit integers), and knowledge distillation (training a small "student" model to mimic a large "teacher" model) [58] [59] [60].

2) Hardware acceleration: Development of specialized edge AI chips and hardware-aware pruning strategies that align

the model architecture with the specific processing units of the device [45] [53] [60].

3) Collaborative edge-cloud architectures: Implementing selective cloud offloading, where simple anomalies are processed locally at the edge for low latency, while complex, cross-modal analysis is offloaded to powerful cloud clusters [20] [21] [45] [53].

4) Toward antifragile and autonomous systems: Looking toward 2026 and beyond, the research community is moving toward "Antifragility", the design of systems that not only resist attacks (robustness) or recover from them (resilience) but actually grow stronger and more intelligent through exposure to stress and volatility. This involves integrating continuous streaming learning that can adapt to "concept drift" in real-time without human intervention. Furthermore, the rise of "Agentic AI" suggests a future where autonomous agents will coordinate defense responses across billions of devices, using multimodal insights to predict and mitigate threats before they even occur [3] [22] [61].

IX. Conclusion

The convergence of the Internet of Things and advanced machine learning has fundamentally redefined the parameters of cybersecurity. As this systematic review has demonstrated, the transition from unimodal to multimodal security architectures is not merely a technical upgrade but a necessary paradigm shift for defending complex, interconnected ecosystems. By integrating diverse and heterogeneous data streams—ranging from network flow statistics to behavioral biomechanics—Multimodal Machine Learning (MML) provides the holistic situational awareness required to detect and mitigate contemporary multi-vector threats [1] [3] [7] [22].

The classification of current literature reveals a clear trajectory toward deep learning models augmented by dynamic attention mechanisms. Architectures such as multimodal transformers and hybrid CNN-LSTM networks have established new benchmarks in accuracy for intrusion detection, malware analysis, and continuous authentication. These models effectively leverage the complementarity of different modalities, bridging the semantic gap between low-level telemetry and high-level behavioral patterns [10] [34] [35] [40] [42].

However, the field remains at a critical juncture. The challenges of data synchronization, model explainability, and resource constraints on the IoT edge must be addressed before MML can be deployed universally. The emerging optimization triad of model compression, hardware acceleration, and collaborative edge-cloud architectures provides a promising roadmap for future development. Ultimately, the goal is to move beyond passive defense toward autonomous, antifragile security systems that can self-heal and adapt to an ever-evolving threat landscape, ensuring the long-term safety, privacy, and resilience of our increasingly digitalized world [3] [20] [22] [45] [53] [61].

References

[1] Z. He, D. Davila, S. Bi, T. Wang, and T. Hou, "Machine Learning for Cybersecurity: A Survey of Applications, Adversarial Challenges, and Future Research Directions," *Electronics*, vol. 14, no. 23, p. 4563, Nov. 2025, doi: 10.3390/electronics14234563.

[2] A. Zahoor, W. Abbasi, M. Z. Babar, and A. Aljohani, "Robust IoT security using isolation forest and one class SVM algorithms," *Sci. Rep.*, vol. 15, no. 1, p. 36586, Oct. 2025, doi: 10.1038/s41598-025-20445-4.

[3] B. Alotaibi, "A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions," Dec. 19, 2025. doi: 10.20944/preprints202512.1717.v1./preprints202512.1717.v1 [Online]. Available: https://www.preprints.org/manuscript/202512.1717/v1.

[4] T. Bhuiyan, "AI in Smart Grid Cybersecurity: A Systematic Review of Machine Learning and Deep Learning Approaches against False Data Injection and Other Emerging Attacks," *JCSTS*, vol. 7, no. 8, pp. 1207–1295, Aug. 2025, doi: 10.32996/jcsts.2025.7.8.136.

[5] Y. S, "AI-Augmented Intrusion Detection Systems for Mitigating Advanced Persistent Threats in Cyber-Physical Manufacturing Networks," Oct. 06, 2025. doi: 10.20944/preprints202510.0470.v1./preprints202510.0470.v1 [Online].Available: https://www.preprints.org/manuscript/202510.0470/v1.

[6] S. I. Popoola, Y. Tsado, A. A. Ogunjinmi, E. Sanchez-Velazquez, Y. Peng, and D. B. Rawat, "Multi-Stage Deep Learning for Intrusion Detection in Industrial Internet of Things," *IEEE Access*, vol. 13, pp. 60532–60555, 2025, doi: 10.1109/ACCESS.2025.3557959.

[7] H. A. S. Ali and V. R. J, "Machine Learning for Internet of Things (IoT) Security: A Comprehensive Survey," *IJCNA*, vol. 11, no. 5, pp. 617–659, Oct. 2024, doi: 10.22247/ijcna/2024/40.

[8] M. Devine, S. P. Ardakani, M. Al-Khafajiy, and Y. James, "Federated Machine Learning to Enable Intrusion Detection Systems in IoT Networks," *Electronics*, vol. 14, no. 6, p. 1176, Mar. 2025, doi:10.3390/electronics14061176.

[9] H. Liao *et al.*, "A survey of deep learning technologies for intrusion detection in Internet of Things," *IEEe Access*, vol. 12, pp. 4745–4761, 2024, doi: 10.1109/ACCESS.2023.3349287.

[10] T. Nazmin, "A Dynamic Hierarchical Attention Framework for Multimodal Malware Detection," Aug. 2025, [Online]. Available: https://scholarworks.uark.edu/etd/5933 [Online]. Available: https://www.osti.gov/servlets/purl/2584221.

[11] H. A. Alsalamah and W. N. Ismail, "Evolutionary Computation for Feature Optimization and Image-Based Dimensionality Reduction in IoT Intrusion Detection," *Mathematics*, vol. 13, no. 23, p. 3869, Dec. 2025, doi: 10.3390/math13233869.

[12] V. Singh, "Multimodal deep learning: Integrating text, vision, and sensor data: Developing models that can process and understand multiple data modalities simultaneously," International Journal of Research in Information Technology and Computing. https://romanpub. com/ijaetv4-1-2022. php, vol. 4, no. 1, Jun. 2022, Accessed: Jan. 21, 2026. [Online]. Available: https://romanpub.com/resources/Vol.%204%20No.%201%20(June%2C%202022)%20-%2035.pdf

[13] Z. Cao, Z. Zhao, W. Shang, S. Ai, and S. Shen, "Using the ToN-IoT dataset to develop a new intrusion detection system for industrial IoT devices," *Multimedia Tools and Applications*, vol. 84, no. 16, pp. 16425–16453, 2025.

[14] İ. Kök, F. Y. Okay, Ö. Muyanlı, and S. Özdemir, "Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14764–14779, Aug. 2023, doi: 10.1109/JIOT.2023.3287678.

[15] T. Sadiq and C. W. Omlin, "Sensing in Smart Cities: A Multimodal Machine Learning Perspective," *Smart Cities*, vol. 9, no. 1, p. 3, Dec. 2025, doi: 10.3390/smartcities9010003.

[16] M. J. C. S. Reis, "Internet of Things and Artificial Intelligence for Secure and Sustainable Green Mobility: A Multimodal Data Fusion Approach to Enhance Efficiency and Security," *MTI*, vol. 9, no. 5, p. 39, Apr. 2025, doi: 10.3390/mti9050039.

[17] Q. Ge, C. Wei, and D. Wu, "Abnormal Network User Detection Method based on the Hybrid Neural Network from the Perspective of Multimodal Data Fusion," *J CIRCUIT SYST COMP*, vol. 34, no. 08, p. 2550195, May 2025, doi: 10.1142/S0218126625501956.

[18] A. Saghir, A. Akbar, A. Hasan, and A. Zafar, "Deep learning for multi-modal data fusion in IoT applications," *Mehran University Research*

Journal of Engineering & Technology, vol. 44, no. 1, pp. 75–81, 2025, doi: 10.22581/muet1982.3171.

[19] A. Kate, "Multimodal Machine Learning for Financial Fraud Detection Across Banking and E-Commerce Ecosystems," 2025.

[20] X. Wang and W. Jia, "Optimizing edge ai: a comprehensive survey on data, model, and system strategies.(2025)," *arXiv preprint arXiv:2501.03265*.

[21] S. A. Cajas Ordóñez, J. Samanta, A. L. Suárez-Cetrulo, and R. S. Carbajo, "Intelligent Edge Computing and Machine Learning: A Survey of Optimization and Applications," *Future Internet*, vol. 17, no. 9, p. 417, Sep. 2025, doi: 10.3390/fi17090417.

[22] B. Alotaibi, "A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions," Dec. 19, 2025. doi: 10.20944/preprints202512.1717.v1.

[23] B. Monteiro and J. Granjal, "Trust-Aware Distributed and Hybrid Intrusion Detection for Rank Attacks in RPL IoT Environments," *IoT*, vol. 7, no. 1, p. 4, 2025, doi: 10.3390/iot7010004.

[24] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," *Sensors*, vol. 23, no. 13, p. 5941, Jun. 2023, doi: 10.3390/s23135941.

[25] S. Wang and Y. Wang, "Malware Classification Technology Combining Multimodal Fusion with Deep Learning Algorithms," *Journal of Cyber Security and Mobility*, vol. 14, no. 3, pp. 597–622, 2025, doi: 10.13052/jcsm2245-1439.1434.

[26] B. Alotaibi and M. Alotaibi, "Hybrid Deep Learning Framework for Continuous User Authentication Based on Smartphone Sensors," *Sensors*, vol. 25, no. 9, p. 2817, Apr. 2025, doi: 10.3390/s25092817.

[27] Moustafa, Nour (TON-IoT Network Intrusion Dataset,). [Online]. Available: https://research.unsw.edu.au/projects/toniot-datasets

[28] S. Maganur, Y. Jiang, J. Huang, and F. Zhong, "Feature-Centric Approaches to Android Malware Analysis: A Survey," *Computers*, vol. 14, no. 11, p. 482, Nov. 2025, doi: 10.3390/computers14110482.

[29] J. Feng, M. Wang, J. Song, C. Yang, L. Zhang, and Y. Shen, "TTGNet - AMD: Android malware detection based on multi-modal feature fusion," *PeerJ Computer Science*, vol. 11, p. e3412, Dec. 2025, doi: 10.7717/peerj-cs.3412.

[30] P. P. Jena, K. N. Kattigenahally, S. Nikitha, S. Sarda, and H. Y, "Multimodal Biometric Authentication: Deep Learning Approach," in *2021 International Conference on Circuits, Controls and Communications (CCUBE)*, Bangalore, India: IEEE, Dec. 2021, pp. 1–5. doi: 10.1109/CCUBE53681.2021.9702724.

[31] S. K. Natarajan, A. Abdullah, S. Kaur, and P. Natarajan, "Advancing Multi-Modal Behavioral Biometric Authentication: A Deep Learning Approach with Synthetic Data Generation," *IEEE Access*, 2025.

[32] C. Wang, Y. Xu, and D. Zhang, "Multi-Modal Perception and Fusion for Maritime Autonomy: A Survey," 2025, doi: 0.20944/preprints202510.0977.v1.

[33] R. K. Solanki, A. M. Dhore, and A. R. Gadekar, "AUGMENTING AR/VR EXPERIENCES USING MULTIMODAL TRANSFORMER MODELS," *Int. J. of Electronics Engineering and Applications*, vol. 13, no. 2, 2025, [Online]. Available: https://ijeea.in/wp-content/uploads/15_21.pdf

[34] A. Odeh and A. Abu Taleb, "Ensemble-based deep learning models for enhancing IoT intrusion detection," *Applied Sciences*, vol. 13, no. 21, p. 11985, 2023, doi: 10.3390/app132111985.

[35] K. O. Adefemi, M. B. Mutanga, and O. A. Alimi, "A Hybrid CNN–GRU Deep Learning Model for IoT Network Intrusion Detection," *Journal of Sensor and Actuator Networks*, vol. 14, no. 5, p. 96, 2025, doi: 10.3390/jsan14050096.

[36] S. K. R. Mallidi and R. R. Ramisetty, "Optimizing Intrusion Detection for IoT : A Systematic Review of Machine Learn8ng and Deep Learning Approaches With Feature Selection and Data Balancing," *WIREs Data Min & Knowl*, vol. 15, no. 2, p. e70008, Jun. 2025, doi: 10.1002/widm.70008.

[37] L. Zhen, N. H. Kamarudin, V. J. Kok, and F. Qamar, "Anomaly detection model in network security situational awareness based on machine learning: Limitation, techniques, future trends," *IEEE Access*, 2025. **DOI:** 10.1109/ACCESS.2025.3589620

[38] N. Peppes, E. Daskalakis, T. Alexakis, and E. Adamopoulou, "A Multimodal Framework for Advanced Cybersecurity Threat Detection Using GAN-Driven Data Synthesis," *Applied Sciences*, vol. 15, no. 15, p. 8730, 2025. doi.org/10.3390/app15158730

[39] P. Radoglou-Grammatikis *et al.*, "AI4FIDS: Multimodal Federated Intrusion Detection," *IEEE Trans. Emerg. Topics Comput.*, pp. 1–15, 2025, doi: 10.1109/TETC.2025.3562346.

[40] H. Ding, S. Lou, H. Ye, and Y. Chen, "MT-CMVAD: A Multi-Modal Transformer Framework for Cross-Modal Video Anomaly Detection," *Applied Sciences*, vol. 15, no. 12, p. 6773, 2025, [Online]. Available: https://doi.org/10.3390/app15126773

[41] J. Dong *et al.*, "A Novel Multimodal Data Fusion Framework: Enhancing Prediction and Understanding of Inter-State Cyberattacks," *BDCC*, vol. 9, no. 3, p. 63, Mar. 2025, doi: 10.3390/bdcc9030063.

[42] S. Alzahrani, Y. Xiao, S. Asiri, N. Alasmari, and T. Li, "ansomFormer: A Cross-Modal Transformer Architecture for Ransomware Detection via the Fusion of Byte and API Features," *Electronics*, vol. 14, no. 7, p. 1245, Mar. 2025, doi: 10.3390/electronics14071245.

[43] B. Alotaibi, "Multimodal Deep Learning Fusion for Accurate and Explainable Malware Family Classification," *Applied Sciences*, vol. 15, no. 21, p. 11635, Oct. 2025, doi: 10.3390/app152111635.

[44] M. U. Tanveer, K. Munir, H. J. Alyamani, S. R. Hassan, M. Sheraz, and T. C. Chuah, "Graph-augmented multi-modal learning framework for robust android malware detection," *Scientific Reports*, vol. 15, no. 1, p. 38341, 2025.

[45] R. Kumar and A. Sharma, "Edge AI: a review of machine learning models for resource-constrained devices," *Artificial Intelligence and Machine Learning Review*, vol. 5, no. 3, pp. 1–11, 2024.

[46] F. Taher, M. Abdel-Salam, M. Elhoseny, and I. M. El-Hasnony, "Reliable machine learning model for IIoT botnet detection," *Ieee access*, vol. 11, pp. 49319–49336, 2023. doi: 10.1109/ACCESS.2023.3253432

[47] O. B. J. Rabie, S. Selvarajan, T. Hasanin, A. M. Alshareef, C. K. Yogesh, and M. Uddin, "A novel IoT intrusion detection framework using Decisive Red Fox optimization and descriptive back propagated radial basis function models," *Scientific Reports*, vol. 14, no. 1, p. 386, 2024.doi.org/10.1038/s41598-024-51154-z

[48] P. Zhao, "Distributed SVM-based Multimodal Intrusion Detection Architecture With Incremental Learning And Modality Scalability," *IJCAI*, vol. 49, no. 7, Nov. 2025, doi: 10.31449/inf.v49i7.9350.

[49] A. Ganmati, K. Afdel, and L. Koutti, "Deep Learning-Based Multi-Factor Authentication: A Survey of Biometric and Smart Card Integration Approaches," 2025, *arXiv*. doi: 10.48550/ARXIV.2510.05163.

[50] V. Velmurugan and S. Priyadharshni, "AI-Driven Secure Authentication: A Deep Learning Approach for Multi-Modal Biometric Systems," in *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, IEEE, 2025, pp. 49–56.

[51] Y. Meidan, et al. "detection_of_IoT_botnet_attacks_N_BaIoT," UCI Machine Learning Repository, 2018. [Online]. Available: https://doi.org/10.24432/C5RC8J.

[52] Z. Qian *et al.*, "Evaluating Diverse Feature Extraction Techniques of Multifaceted IoT Malware Analysis: A Survey," *arXiv e-prints*, p. arXiv: 2509.03442, 2025.

[53] S. O. Semerikov, T. A. Vakaliuk, O. B. Kanevska, O. A. Ostroushko, and A. O. Kolhatin, "Edge intelligence unleashed: a survey on deploying large language models in resource-constrained environments," *Journal of Edge Computing*, vol. 4, no. 2, pp. 179–233, 2025.

[54] J. Moss, J. Gordon, W. Duclos, Y. Liu, Q. Wang, and J. Wang, "Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation," *Electronics*, vol. 14, no. 23, p. 4622, 2025.

[55] K. S. Alketbi and A. Mehmood, "A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection," *IEEE Access*, 2025.

[56] G. Rjoub *et al.*, "A survey on explainable artificial intelligence for cybersecurity," *IEEE Transactions on Network and Service Management*, vol. 20, no. 4, pp. 5115–5140, 2023.

[57] S. Nazim, M. M. Alam, S. S. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, "Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM," *Plos one*, vol. 20, no. 5, p. e0318542, 2025, doi: 10.1371/journal.pone.0318542.

[58] R. AGRAWAL and H. KUMAR, "Energy-Efficient AI Model Design for Edge Devices Using Neural Network Pruning and Optimization Techniques," 2023.

[59] J. Greg, 'Quantization and Pruning Strategies for Energy-Efficient TinyML Models', vol. Volume 6 Issue 10, 2025.

[60] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2022, doi: 10.1109/JPROC.2022.3226481.

[61] P. Berrang, "Custom Software Development Blog," 7T.ai, 2025. [Online]. Available: https://7t.ai/digital-transformation-insights-blog/.

[62] M. Latif *et al.*, 'AI-based intelligent sensing detection of cybersecurity threats using multimodal sensor data in smart devices', *Scientific Reports*, vol. 16, no. 1, p. 11091, Feb. 2026, doi: 10.1038/s41598-026-40614-3.