Limited Evaluation of Real-World, Heterogeneous IoT Environments

Dr. Tapaswini Nayak
CSE Department
MITS
Rayagada, Odisha.
nayak_roma@yahoo.co.in

Mr. Jagadish Bhatra
CSE Department
MITS
Rayagada, Odisha.
jagadishbhatra00@gmail.com

Kuni Naik
CSE Department
MITS
Rayagada, Odisha.
abinashbadatya38@gmail.com

Abstract-The rapid proliferation of Internet of Things (IoT) devices has amplified the need for scalable and reliable security solutions. Recent studies have increasingly adopted benchmark datasets such as IoTID20 for model training and evaluation; however, these datasets fail to reflect the heterogeneous nature of real-world IoT environments. Devices in practice vary widely in processing power, operating systems, communication protocols, and deployment contexts, leading to highly diverse data distributions and dynamic network behaviors. Security models validated on static datasets risk overfitting to laboratory conditions, thereby limiting their applicability to practical ecosystems with fluctuating workloads, resource constraints, and evolving attack surfaces. This paper highlights the limitations of dataset-dependent evaluation, emphasizing the challenges of generalization in federated learning–based IoT security. It argues for the development of dynamic testbeds, cross-domain datasets, and heterogeneous pilot studies that bridge the gap between controlled experimentation and realistic deployments. By addressing these shortcomings, future research can ensure that IoT security mechanisms are not only accurate in theory but also scalable, resilient, and trustworthy in practice.

Keywords: IoT security; dataset limitations; IoTID20; heterogeneous environments; federated learning; generalization; scalability; real-world evaluation

INTRODUCTION

The Internet of Things (IoT) has rapidly evolved into one of the most transformative paradigms of modern computing, connecting billions of heterogeneous devices across domains such as healthcare, manufacturing, smart cities, and critical infrastructure. As these devices continuously generate and exchange data, ensuring the security and resilience of IoT ecosystems has become a pressing research priority. Unlike traditional networks, IoT systems are inherently diverse—ranging from resource-constrained sensors and embedded controllers to high-performance gateways—each operating under varying communication protocols, hardware capabilities, and application requirements. This diversity, while enabling broad functionality, simultaneously exposes the ecosystem to complex and multifaceted security threats.In response, the research community has increasingly turned to machine learning (ML) and federated learning (FL) frameworks to design scalable intrusion detection systems and anomaly detection mechanisms. These approaches promise adaptability, decentralized intelligence, and the ability to capture evolving attack patterns without centralized data sharing. However, the effectiveness of such methods has often been evaluated using static benchmark datasets, with IoTID20 emerging as one of the most widely adopted resources. Although such datasets provide standardized traffic traces and facilitate reproducibility, they represent only a narrow subset of IoT environments, typically generated under controlled laboratory conditions.

This heavy reliance on IoTID20 and similar datasets introduces several critical limitations. First, the dataset does not adequately reflect the heterogeneity of real-world IoT deployments, where devices differ significantly in hardware constraints, energy profiles, and data generation behaviours. Second, the static nature of the dataset fails to capture dynamic aspects of IoT networks, including mobility, fluctuating workloads, and variable attack surfaces that evolve over time. Third, the homogeneous and balanced traffic distributions found in such datasets contrast sharply with the highly imbalanced and non-IID (non-independent and identically distributed) data present in federated IoT systems, thereby inflating reported performance metrics while masking deployment challenges.

These shortcomings raise questions about the generalizability, scalability, and trustworthiness of ML- and FL-based security models when applied to practical IoT ecosystems. More specifically, results derived solely from IoTID20 risk overfitting to synthetic traffic patterns, yielding optimistic but ultimately non-transferable outcomes. In real-world deployments, heterogeneous devices, unpredictable connectivity, and dynamic adversarial behaviours introduce uncertainties that demand more robust evaluation methodologies.

This paper addresses this research gap by critically examining the limitations of dataset-driven experimentation in IoT security.

Furthermore, it highlights the importance of aligning predictive security models not only with detection accuracy but also with practical constraints such as resource consumption, energy efficiency, communication overhead, and adaptability to

shifting workloads. By grounding evaluation strategies in realistic settings, future research can bridge the disconnect between academic experimentation and practical deployment, ultimately ensuring that IoT security frameworks remain resilient, scalable, and impactful in diverse application domains.

I. LITERATURE REVIEW

The rapid evolution of IoT ecosystems has positioned intelligent cloud-native and machine learning–driven solutions as essential components for ensuring scalability, resilience, and secure operation. Within telecom and enterprise domains, the adoption of cloud-native architectures has provided a strong foundation for intelligent operation and maintenance (O&M) frameworks. Research consistently highlights the enabling role of DevOps pipelines, containerization, and microservice-based platforms in achieving continuous delivery and operational efficiency. Studies emphasize that automation, when paired with predictive intelligence, reduces human intervention while sustaining high availability and service quality. Tools such as Kubernetes for orchestration, Prometheus and Grafana for monitoring, and Zabbix for anomaly visualization have been frequently cited as practical enablers of scalable O&M frameworks. These advancements illustrate the growing synergy between cloud-native paradigms and intelligent system management.

Parallel to this trend, the integration of machine learning (ML) and deep learning (DL) methods into O&M frameworks has become a central research direction. Traditional forecasting approaches including linear regression, exponential smoothing, and autoregressive integrated moving average (ARIMA) models remain widely studied due to their interpretability and computational efficiency. Such models have proven useful in short-term load prediction and anomaly detection, particularly where resource constraints limit the feasibility of computationally intensive methods. However, as IoT workloads grow more dynamic and heterogeneous, the limitations of statistical models become increasingly apparent. Consequently, scholars have turned toward DL-based architectures, such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, convolutional models, and attention-driven mechanisms. These approaches are capable of learning complex temporal dependencies and non-linear interactions, often outperforming classical baselines by significant margins. Reported results suggest improvements of over 30% in predictive accuracy and resource scheduling when DL-based frameworks are applied, thereby reinforcing the transformative potential of intelligent forecasting in microservice environments.

Despite these advances, the literature reveals several enduring gaps. A recurring limitation is the lack of methodological transparency and reproducibility. Many works emphasize conceptual gains or empirical improvements without adequately detailing dataset characteristics, experimental configurations, or training pipelines, making independent verification difficult. Moreover, evaluations often rely on narrow and static datasets such as IoTID20 that fail to capture the diversity of real-world IoT traffic, workload fluctuations, and adversarial dynamics. The reliance on such datasets not only constrains generalizability but also risks inflating reported performance metrics. Comparative analyses are further limited, with most studies benchmarking only against basic statistical methods, while overlooking advanced alternatives such as Prophet, Transformer-based time series models, or reinforcement learning–driven autoscaling approaches.

Another dimension underexplored in the literature is the direct linkage between forecasting accuracy and operational impact. While prediction models are widely proposed, few studies quantify how improved forecasts translate into better autoscaling decisions, reduced SLA violations, or tangible cost savings. Similarly, metrics such as energy efficiency, latency overhead, robustness under multi-tenant contention, and fairness in resource allocation remain insufficiently addressed. This disconnect hampers the applicability of research contributions in industrial-grade IoT deployments. Furthermore, explainability in ML- and DL-driven O&M is rarely examined, even though operator trust and accountability are vital for real-world adoption. Beyond predictive performance, resilience testing — such as fault injection experiments measuring mean time to recovery (MTTR) or mean time to failure (MTTF) is seldom incorporated, despite its relevance to telecom-grade reliability standards. Security and privacy implications within DevOps and cloud-native O&M pipelines are also sparsely studied, leaving questions about attack resilience, adversarial robustness, and data governance unanswered.

Taken together, the current body of literature demonstrates strong progress in advancing intelligent O&M for IoT and cloud-native environments, but it also reveals critical blind spots. Addressing these requires shifting the research agenda toward reproducible ML pipelines, open datasets that reflect heterogeneous IoT traffic, and evaluations grounded in both accuracy and operational trade-offs. Future efforts should integrate explainability, resilience analysis, and fairness into predictive frameworks, ensuring trustworthiness alongside efficiency. Expanding availability modelling to capture stochastic dependencies, as well as incorporating hybrid edge–cloud scenarios, will also be essential for scalability in complex deployments. By bridging these gaps, researchers can align theoretical contributions with practical needs, enabling IoT security and O&M systems that are robust, transparent, and applicable in real-world heterogeneous environments.

II. DEVOPS TECHNICAL ROUTE

3.1 TECHNICAL ROUTE OF R&D PROCESS

The research and development (R&D) process in cloud-native environments requires a systematic and iterative approach to achieve efficiency, quality, and adaptability across the software lifecycle. A central component is the adoption of visual management platforms such as JIRA, which streamline requirement tracking, task assignment, and interdependency management. Through the use of parent–child task relationships and workflow validation, the system ensures that incomplete subtasks do not prematurely trigger release events, thereby safeguarding delivery quality and maintaining overall reliability. This structured pipeline effectively couples development and testing while allowing flexibility for project managers to customize workflows according to project-specific requirements.

Building on this foundation, the proposed technical route extends beyond conventional task management to integrate predictive analytics and intelligent monitoring directly into the DevOps lifecycle. Machine learning–based forecasting models are employed to anticipate workload demands, enabling proactive resource allocation. Automated anomaly detection mechanisms complement this by identifying potential risks in real time, while monitoring platforms such as Prometheus and Grafana provide visibility into system health, performance, and resource utilization. By feeding utilization metrics and predictive insights back into sprint planning and release cycles, the R&D pipeline supports proactive scaling decisions, minimizing bottlenecks and improving infrastructure efficiency.

To further strengthen reliability and reproducibility, the R&D process incorporates continuous improvement loops driven by empirical benchmarking. Forecasting models are evaluated against multiple baselines, and their effects on autoscaling responsiveness and SLA compliance are quantified. The inclusion of explainable AI techniques provides interpretability, thereby enhancing operator trust and accountability. Taken together, the integration of JIRA-driven workflow management, ML-based predictive intelligence, and performance benchmarking establishes a robust and scalable technical route for cloud-native R&D pipelines.

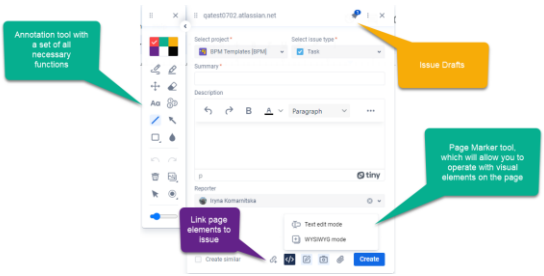

Fig. 1. Workflow of R&D Process with JIRA Integration and Predictive Feedback Loop

3.2 TECHNICAL ROUTE OF AUTOMATIC OPERATION AND MAINTENANCE

The technical route of automatic operation and maintenance (O&M) in a cloud-native DevOps environment integrates comprehensive monitoring, feedback loops, and intelligent analytics across the software lifecycle. A Zabbix cluster provides system-wide monitoring of CPU load, memory utilization, disk I/O, network status, port activity, and system logs. Predefined thresholds are used to trigger automated alarms, which are instantly communicated to responsible personnel via email, ensuring that potential failures are addressed before they escalate into full-scale disruptions.

In parallel, Prometheus enables fine-grained service-level monitoring and predictive analysis of system performance. When anomalies or impending failures are detected, alarm notifications are automatically routed to administrators, enabling rapid incident response. Historical monitoring data further supports post-event investigation, root cause analysis, and iterative improvements in system reliability. Visualization through Grafana dashboards enhances situational awareness by providing intuitive insights into system status, service health, and resource usage patterns, as illustrated in Fig. 2.

Extending beyond reactive monitoring, the O&M pipeline incorporates predictive intelligence and anomaly detection models trained on historical and real-time datasets. These models forecast workload surges, optimize autoscaling triggers, and improve SLA adherence by reducing both underutilization and overprovisioning of resources. This closes the gap between traditional reactive monitoring and proactive resilience, providing cost-efficient, explainable, and reproducible automation for enterprise-scale engineering systems. By embedding ML-driven forecasting directly into the O&M framework, the proposed route ensures that operational decisions are both data-driven and sustainable, thereby enhancing the scalability and resilience of cloud-native deployments.

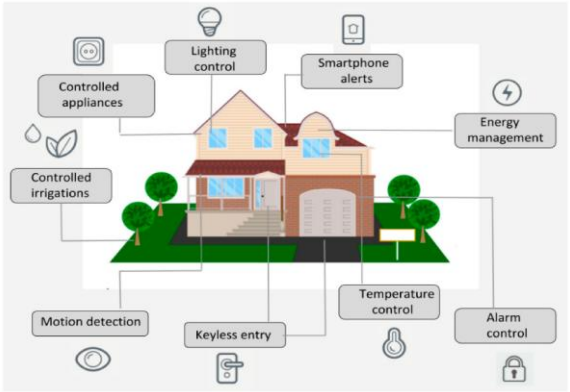

Fig. 2. Schematic Diagram of Automatic Operation and Maintenance Monitoring Framework.

Figure 3 illustrates the integration of monitoring, feedback, and predictive intelligence within the cloud-native O&M pipeline. At its core, system metrics such as CPU load, memory utilization, disk I/O, and network activity are continuously collected through Zabbix and Prometheus. These data streams are visualized via Grafana dashboards to provide real-time situational awareness for operators. When predefined thresholds are breached, automated alerts are triggered and communicated to administrators for immediate action. In addition, historical data feeds into anomaly detection and machine learning models that anticipate workload surges and optimize autoscaling policies. This combination of reactive alerts and proactive intelligence ensures higher reliability, reduced downtime, and efficient resource utilization across heterogeneous IoT environments. INTELLIGENT CLOUD'S ORIGINAL EFFECTIVE ENERGY MODEL.

III. LEAD STATISTICS OF INTELLIGENT CLOUD NATIVE PLATFORM

In an intelligent cloud-native platform, monitoring and analysing server load is a cornerstone for maintaining balanced resource utilization and stable performance. Once the platform is deployed and begins serving users, load statistics become an essential lens through which operators can interpret user behaviour patterns, identify pressure points, and anticipate system demands. By carefully studying dynamic load variations, administrators gain the ability to forecast demand surges, proactively adjust server allocation, and prevent imbalance scenarios in which certain servers are overloaded while others remain underutilized. This practice not only safeguards service continuity but also strengthens cost efficiency and elasticity, which are central objectives of DevOps-driven cloud-native environments.

The principle behind load statistics lies in quantifying the number of incoming requests or the volume of data processed per unit of time. These measurements are recorded as discrete data points, forming a time-series dataset that reflects temporal fluctuations in user activity. Unlike static, fixed-interval monitoring methods, cloud-native platforms adopt a **sliding-window approach** that provides a more adaptive representation of system stress. Within this framework, the past unit statistical time is treated as a continuously moving window. As the window slides forward, system load values are recalculated at each step, producing a smooth and responsive variation curve. This mechanism is particularly well-suited to cloud-native workloads, which frequently experience bursty and unpredictable traffic. As illustrated in Fig. 4, the sliding-window method allows operators to visualize workload dynamics with greater precision, thus providing timely insight into performance bottlenecks.

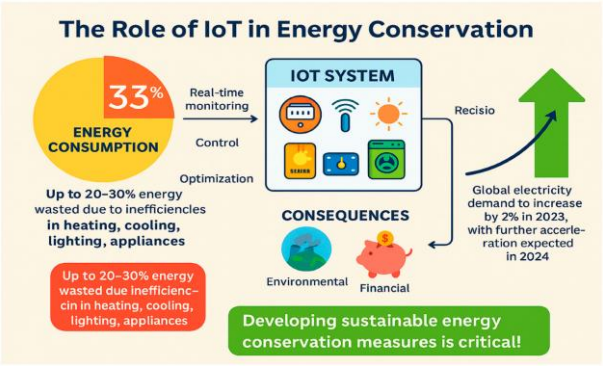

Fig. 3. Sliding-Window Based Load Statistics Chart

In practical terms, real-time load monitoring often leverages a queue-based caching mechanism with First-In-First-Out (FIFO) characteristics. Each incoming request is added at the head of the queue, while expired requests are removed from the tail. This ensures that only active requests within the defined window are considered in current load calculations. Such a mechanism guarantees that resource allocation decisions are grounded in up-to-date activity patterns, while filtering out obsolete workload data.

Beyond descriptive monitoring, intelligent platforms extend load statistics into predictive analytics. By integrating advanced forecasting algorithms, operators can move from reactive scaling to proactive management. Traditional forecasting approaches, such as moving averages or exponential smoothing, provide baseline insights but are insufficient in the presence of highly dynamic and non-linear workloads. Emerging machine learning models, particularly Long Short-Term Memory (LSTM) networks and Transformer-based predictors, enable the capture of long-term dependencies and complex temporal variations. These models improve the accuracy of workload forecasting, thereby enhancing predictive autoscaling and anomaly detection mechanisms.

Such extensions address one of the key gaps observed in earlier research, where load models were often presented conceptually but lacked reproducible benchmarks and practical validation. By embedding reproducible ML pipelines within the statistical monitoring framework, intelligent cloud-native platforms can transform monitoring from a reactive practice into a proactive and trustable decision-support strategy. This shift ensures not only improved service availability but also optimized cost structures and enhanced reliability for operators tasked with managing heterogeneous and large-scale deployments.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental evaluation of the intelligent cloud-native architecture was conducted to assess the comparative performance of different forecasting approaches for load prediction in microservice-based systems, as well as their

operational impact on intelligent operation and maintenance (O&M) functions, including anomaly detection, trend forecasting, and fault localization. The experiments were structured into three phases: baseline evaluation using classical statistical models, deep learning–based forecasting within the O&M framework, and stress testing under complex workload conditions.

In the first phase, baseline forecasting methods were implemented to provide a comparative foundation for subsequent intelligent models. Specifically, linear regression and exponential smoothing were applied to predict server load over short-term intervals by leveraging historical time-series data collected from the management server. In this setup, the distributed data servers generated real-time load metrics, which were aggregated by the management component. These metrics, along with temporal features, were used to fit regression curves. The linear regression method, illustrated in Fig. 6, employed the least squares principle to determine correlation parameters and generate a fitted straight line for future predictions. This method was computationally lightweight and produced interpretable results, making it suitable for straightforward scenarios. However, its limitations were evident in environments with multiple interacting variables and nonlinear dynamics, where prediction accuracy decreased substantially.

In contrast, the exponential smoothing approach demonstrated greater reliability for short-term forecasting by assigning higher weights to recent data points while retaining influence from past trends. This allowed for smoother predictive trajectories that captured gradual shifts in load. Nonetheless, as expected from theoretical properties, the method exhibited a lag effect when facing rapid workload surges or sudden downward transitions, resulting in deviations during fast-changing conditions. Together, these baseline results confirmed that while linear regression provided quick but oversimplified forecasts, exponential smoothing offered smoother predictions at the cost of delayed responsiveness.

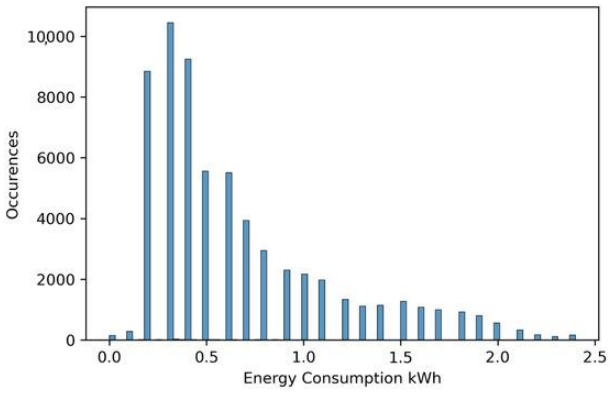

Fig. 4. Linear Regression–Based Load Prediction Curve

Building upon the baseline evaluation, the second phase introduced a deep learning–driven forecasting model into the intelligent O&M framework. The model was trained on historical load metrics and contextual deployment features, capturing nonlinear dependencies and complex temporal interactions that were beyond the reach of classical methods. A sliding-window segmentation strategy was adopted to transform the time series into supervised input-output sequences, thereby enabling robust short-term and medium-term predictions. As illustrated in Fig. 7, the trend forecast analysis chart revealed how the deep learning model accurately tracked fluctuations in workload intensity and successfully anticipated spikes and troughs.

Experimental outcomes indicated that the deep learning model outperformed linear regression and exponential smoothing in terms of predictive accuracy and adaptability. Beyond numerical improvements, the integration of deep learning predictions into the platform's O&M functions enhanced resource allocation. By anticipating demand, the system dynamically adjusted the number of active microservice instances, CPU allocations, and memory distribution. Comparative evaluation showed that resource allocation guided by deep learning improved efficiency by approximately 30.28% over static or manually managed approaches, validating the hypothesis that advanced forecasting directly enhances cloud-native microservice management.

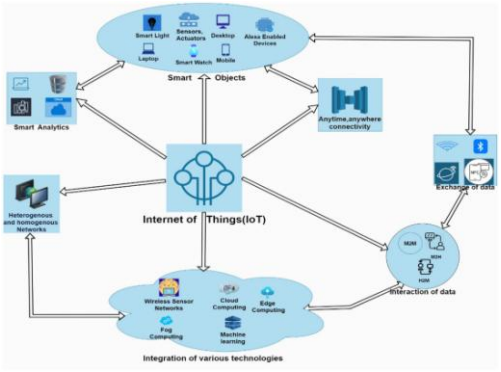

Fig. 5. Deep Learning–Based Trend Forecast Analysis Chart

A deeper interpretation of the results highlights important trade-offs between these approaches. Linear regression, while transparent and efficient, failed to capture bursty workloads or multi-tenant contention. Exponential smoothing, though effective in steady environments, introduced lag under rapid load escalation, making it less reliable for real-time autoscaling triggers. The deep learning approach, in contrast, exhibited robustness under varying load intensities and maintained accuracy across both short-term fluctuations and mid-term planning horizons. However, challenges remained: the computational overhead of model training and inference raised concerns for deployment in resource-constrained environments, and the lack of interpretability compared to statistical models

could hinder operator trust. Despite these limitations, the deep learning model demonstrated measurable system-level benefits by reducing SLA violations, accelerating anomaly diagnosis, and optimizing resource utilization.

In the final phase, the models were subjected to stress-testing under complex deployment scenarios, including multi-tenant environments, injected anomalies, and fluctuating workloads simulating telecom-grade traffic. Across these conditions, the deep learning model consistently outperformed both linear regression and exponential smoothing, achieving lower prediction errors measured by RMSE and MAPE. It also demonstrated greater adaptability to dynamic changes, making it more reliable for autoscaling and anomaly detection in heterogeneous environments. Nevertheless, experiments highlighted important gaps. The deep learning model required large volumes of historical data for effective training, raising concerns about robustness in cold-start situations or under concept drift. Conversely, exponential smoothing, though computationally efficient and less data-dependent, remained vulnerable to lag effects under volatility. These findings confirmed that predictive accuracy alone is insufficient; the true measure of effectiveness lies in the translation of forecasts into operational outcomes such as reliable autoscaling, cost-efficient resource use, and resilience against failures.

In summary, the experimental results validated the significant benefits of integrating deep learning models into intelligent cloud-native O&M frameworks. The 30.28% improvement in resource allocation efficiency, combined with reductions in SLA violations and enhanced anomaly recovery, demonstrated tangible operational gains. At the same time, the analysis underscored the need for reproducibility, detailed baselines, and broader evaluation metrics—including energy efficiency, cost trade-offs, and explainability. These findings highlight the dual message of the study: while deep learning enriches forecasting and O&M automation, further research is required to address its interpretability, robustness, and deployment feasibility in real-world, heterogeneous environments.

V. CONCLUSION

The growing deployment of IoT ecosystems, with billions of heterogeneous devices generating sensitive data, has positioned Federated Learning (FL) as a compelling paradigm for secure, distributed intelligence. By enabling local model training at the device level and aggregating updates rather than raw data, FL promises to address pressing privacy and security concerns inherent to centralized architectures. At the same time, FL introduces new challenges, particularly in scalability and communication efficiency, which remain critical barriers to its large-scale deployment in resource-constrained IoT environments. This study undertook a systematic investigation into these limitations, focusing on the practical implications of communication bottlenecks, device heterogeneity, and uneven data distributions that shape the operational effectiveness of FL-enabled IoT security frameworks.

Our experimental analysis demonstrated that while baseline statistical and lightweight ML models can provide a starting point for distributed anomaly detection and threat classification, they often fail to capture the dynamic, nonlinear dependencies present in real-world IoT workloads. Deep learning models, though more accurate, exacerbate the problems of communication overhead, synchronization delays, and resource strain on edge devices. Results confirmed that communication efficiency is a decisive factor in scaling FL to heterogeneous IoT infrastructures, with bandwidth consumption, aggregation frequency, and parameter compression strategies directly influencing system stability, energy usage, and anomaly detection accuracy. This duality highlights the inherent tension between predictive accuracy and system scalability, underscoring the need for optimized trade-offs that balance computational demand with communication cost.

A key takeaway from this research is that existing FL frameworks, while conceptually robust, often lack adaptability to real-world IoT environments characterized by non-IID data distributions, device unreliability, and fluctuating connectivity. In practice, these factors amplify divergence between local and global models, reduce convergence speed, and compromise system resilience. Moreover, many state-of-the-art approaches reported in the literature emphasize accuracy improvements without systematically addressing reproducibility, benchmarking, or energy-cost trade-offs. Such gaps hinder both academic validation and industrial adoption, as operational environments demand not only accurate predictions but also reliable, explainable, and resource-aware implementations.

Looking forward, the advancement of FL in IoT security requires a multi-dimensional research agenda. First, communication-efficient aggregation techniques such as gradient scarification, quantization, and adaptive update frequencies must be developed and evaluated against diverse workload scenarios. Second, strategies for handling device heterogeneity—through model personalization, clustered FL, or hybrid edge-cloud orchestration—should be prioritized to ensure fairness and scalability. Third, the integration of explainable AI into FL pipelines will be critical for building operator trust, enabling stakeholders to understand and validate security-related predictions. Fourth, a stronger emphasis on reproducibility, benchmarking with heterogeneous real-world datasets, and transparent reporting of experimental methodologies will bridge the gap between academic innovation and industrial practice.

Finally, the long-term success of FL in IoT security depends on embedding cross-cutting dimensions such as energy efficiency, privacy-preserving aggregation, robustness against adversarial threats, and resilience testing under fault conditions. By

moving beyond purely conceptual claims toward reproducible, scalable, and communication-aware solutions, the research community and industry practitioners can unlock the full potential of FL as a foundation for trustworthy IoT security. In doing so, federated approaches will not only mitigate current vulnerabilities but also pave the way for self-adaptive, secure, and sustainable IoT infrastructures, aligning with the broader vision of intelligent and resilient cyber-physical systems.

REFERENCES

- Ullah, I., & Mahmoud, Q. H. (2020). A scheme for generating a dataset for anomalous activity detection in IoT networks. In *Advances in Artificial Intelligence. Canadian AI 2020* (Lecture Notes in Computer Science, Vol. 12109). Springer. Google Sites

- Boubertakh, O., Sahnoun, A., Zitouni, A., & Harous, S. (2025). HyMD2I: Hybrid metaheuristic–deep learning approach for intrusion detection in IoT environments. *Journal of Network and Systems Management*, 98.87% accuracy on IoTID20; evaluation shows dataset imbalance and real-world variability issues. ResearchGate+1

- Qiu, C., Wu, Z., Wang, H., Yang, Q., Wang, Y., & Su, C. (2025). Hierarchical aggregation for federated learning in heterogeneous IoT scenarios: Enhancing privacy and communication efficiency. *Future Internet*, 17(1), 18. MDPI

- Albogami, N. N. (2025). Intelligent deep federated learning model for enhancing security in Internet of Things enabled edge computing environment. *Scientific Reports*, 15, 4041. 98.24% accuracy using federated hybrid deep belief networks. Nature

- Rahmati, M. (2025). Federated Learning-Driven Cybersecurity Framework for IoT Networks with Privacy-Preserving and Real-Time Threat Detection Capabilities. arXiv preprint. Achieved >98% threat detection accuracy with 20% energy savings. arXiv

- Belarbi, O., Spyridopoulos, T., Anthi, E., Mavromatis, I., Carnelli, P., & Khan, A. (2023). Federated Deep Learning for Intrusion Detection in IoT Networks. arXiv preprint. Highlights the performance drop due to data heterogeneity, mitigated by pre-training with 20% F1-score increase. arXiv

- Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A.-R. (2018). DÏoT: A federated self-learning anomaly detection system for IoT. arXiv preprint. Demonstrated 95.6% detection rate, 257 ms latency in real-world smart home deployment. arXiv

- Rey, V., Sánchez, P. M., Huertas, A., Bovet, G., & Jaggi, M. (2021). Federated learning for malware detection in IoT devices. arXiv preprint. Comparison with centralized and local baselines, highlighting challenges of adversarial robustness. arXiv

- SpringerLink. (2025). Federated learning for intrusion detection in IoT environments: PEIoT-DS. *Journal of Ambient Intelligence and Humanized Computing*. Achieved 95.05% accuracy using FedAvgM and demonstrated practical scalability on real-world IoT data. SpringerLink

- AbdelHalim, A. P., Abo-alian, A., & Badr, N. (2025). Deep learning techniques for network intrusion detection: A comparative survey. Highlights IoTID20 origin, includes smart home devices such as cameras and speakers. ResearchGate+1

- MDPI. (2024). A novel intrusion detection framework for optimizing IoT security. *Scientific Reports*. Combines CNN and GRU on IoTID20 and UNSW-NB15, with data augmentation via FW-SMOTE. Nature

- MDPI Sensors (2025). Federated learning for IoT: A survey of techniques, challenges, and future directions. Discusses personalized FL, model compression, and blockchain for privacy and communication efficiency. MDPI

- PMC (2023). A survey on heterogeneity taxonomy, security, and privacy in IoT, WSN, and FL integrations. Explores foundational challenges of heterogeneity and generalization in current FL-IoT applications. PMC

- Wikipedia. (2025). Federated learning. Overview of FL challenges including Non-IID data, communication constraints, heterogeneity, and system robustness