JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 1 (2024), pp. 46-66
DOI: 10.63646/jaiaa.2024.020103

Privacy-Preserving AI Analytics for Adaptive Intrusion Detection in Heterogeneous IoT Networks

Liang Chen[1]; Minghao Zhao[2]; Yujie Huang[3, *]

[1] School of Information Engineering, Hunan City University, Yiyang, China
[2] School of Computer Science, Hunan University of Technology and Business, Changsha, China
[3] College of Information Science and Engineering, Changsha University, Changsha, China
* Corresponding author: yujie.huang@huse.edu.cn

ARTICLE INFO
Received
October 30, 2023
Revised
December 28, 2023
Accepted
February 18, 2024
Available Online
March 30, 2024
DOI
10.63646/jaiaa.2024.020103

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Heterogeneous Internet of Things (IoT) networks create a difficult intrusion-detection problem because traffic patterns, device capacities, privacy risks, and attack distributions differ across clients. Centralized intrusion detection systems can collect large volumes of data, but they also introduce latency, data-transfer exposure, and governance challenges. This article proposes a privacy-preserving AI analytics framework for adaptive intrusion detection in heterogeneous IoT networks. Inspired by recent work on federated intrusion detection, reinforcement-guided decision control, and fog-cloud security, the framework integrates local feature learning, federated aggregation, client-level adaptation, privacy-preserving update exchange, and explainable risk scoring. The study reconstructs a multi-domain evaluation design using general network, industrial IoT, medical IoT, vehicle-network, smart-grid, and large-scale intrusion datasets. Rather than focusing only on average accuracy, the analysis evaluates precision, recall, F1-score, false-positive rate, false-negative rate, latency, privacy cost, and scalability under IID and non-IID client settings. The results and discussion show that privacy-preserving AI analytics can support high-quality intrusion detection when decentralized training is combined with personalized aggregation, asynchronous updates, and adaptive decision thresholds. The article contributes a deployable analytics architecture and a practical evaluation logic for trustworthy cyber defense in distributed IoT environments.

Keywords: Privacy-preserving AI analytics; Intrusion detection; Federated learning; Heterogeneous IoT; Cybersecurity; Adaptive detection; Fog-cloud computing

I. INTRODUCTION

The rapid diffusion of heterogeneous Internet of Things (IoT) networks has transformed intrusion detection from a centralized classification problem into a distributed analytics problem. Industrial sensors, medical monitoring devices, connected vehicles, smart-grid controllers, and consumer endpoints generate traffic streams with different feature spaces, attack surfaces, communication capacities, and privacy constraints. A single cloud-based detector can still be useful for offline analysis, but it cannot fully address local data sensitivity, low-latency threat response, and non-identical client behaviour in real deployments. This article therefore develops a privacy-preserving AI analytics perspective for adaptive intrusion detection in heterogeneous IoT networks, drawing inspiration from federated intrusion detection, fog-cloud security, reinforcement learning, and client-level model personalization.

The starting point of the study is the observation that intrusion detection in IoT networks should not be

reduced to a search for the highest average accuracy. A realistic detector must balance at least four objectives: attack recognition, false-alarm control, privacy protection, and scalable operation under client heterogeneity. Federated learning offers a natural design pattern because it enables local training without moving raw network traffic to a central server (McMahan, 2017). However, federated learning alone does not solve the problems of non-IID data, poisoned clients, and time-varying threat strategies (Kairouz, 2021). A useful AI analytics framework must therefore connect local feature extraction, secure model update exchange, adaptive aggregation, and explainable risk scoring in a unified workflow.

This paper is not a duplicated version of an existing algorithmic manuscript. It reframes the uploaded study's central idea - privacy-preserving, client-adaptive intrusion detection for heterogeneous IoT environments - as an AI analytics framework suitable for the Journal of AI Analytics and Applications. The uploaded manuscript demonstrates the value of federated learning, reinforcement learning, client-specific adaptation, differential privacy, and asynchronous aggregation across multiple IoT security datasets, including Edge-IIoTset, UNSW-NB15, WUSTL-EHMS-2020, Car Hacking, power-system attack data, and CICIDS-2018. The present article reconstructs that technical direction into a broader analytical architecture, adds deployment-oriented discussion, and expands the data analysis and evaluation logic rather than reproducing the original text or diagrams.

The contribution is threefold. First, the article proposes a privacy-preserving AI analytics architecture that integrates local IoT feature learning, federated aggregation, adaptive reinforcement feedback, and audit-oriented privacy control. Second, it develops a structured experimental and data-analysis design for multi-domain intrusion detection, emphasizing client heterogeneity, class imbalance, latency, and privacy loss rather than accuracy alone. Third, it translates algorithmic results into practical deployment guidance for IoT operators, showing when client-level personalization, differential privacy, asynchronous aggregation, and interpretable alerts become operationally necessary. These contributions align with recent literature that treats federated learning as both a machine learning method and a socio-technical infrastructure for privacy-sensitive analytics (Li, 2020).

II. RELATED WORK AND ANALYTICAL POSITIONING

Intrusion detection research has historically moved through three broad stages. Signature-based systems relied on known rules and attack fingerprints, which made them reliable for previously observed threats but weak against novel attack patterns. Classical machine learning systems introduced supervised and anomaly-based classification, making it possible to recognize more complex traffic patterns from labelled network data (Buczak, 2016). Deep learning systems then improved feature learning from high-dimensional traffic streams, but many of these systems still assume centralized data collection and stationary training distributions. Such assumptions are increasingly unrealistic for IoT networks that span multiple devices, organizations, and regulatory environments.

IoT security creates analytical difficulty because the devices are resource-constrained, communication links vary in quality, and traffic semantics differ across domains. Smart meters, hospital monitors, vehicle controllers, and industrial sensors do not share a uniform definition of normality. General IoT surveys show

that scalability, device diversity, and trust management remain central obstacles to secure operation (Atzori, 2010). Enterprise-oriented IoT analysis further shows that organizations adopt IoT systems for operational efficiency while frequently underestimating the security and governance burden (Lee, 2015). These observations imply that intrusion detection must be adaptive at the edge and integrated with management-level risk evaluation.

Edge and fog computing are important because they move computation closer to the traffic source. Edge computing reduces data-transfer latency and allows local decisions under bandwidth constraints (Shi, 2016). Fog computing adds intermediate nodes between end devices and the cloud, supporting local preprocessing, aggregation, and control (Bonomi, 2012). Security analyses of fog and mobile-edge computing emphasize that distributed computing reduces some privacy risks but increases the attack surface through new gateways, APIs, and service orchestration layers (Roman, 2018). For intrusion detection, this means that the analytics pipeline must be designed as a layered system rather than a single classifier.

Federated learning directly addresses the privacy limitation of centralized IDS by allowing each client to train on local data and share model updates rather than raw records (Yang, 2019). Secure aggregation further reduces the exposure of individual updates by ensuring that the server observes only aggregate information (Bonawitz, 2017). Differential privacy can be added to limit information leakage from gradients or parameters (Dwork, 2006). These methods provide a formal privacy vocabulary, but they also introduce practical trade-offs: noise may reduce detection accuracy, secure aggregation may increase communication cost, and local heterogeneity may prevent the global model from fitting all clients well (Nguyen, 2021).

The non-IID problem is especially important in intrusion detection. A hospital IoMT client may observe rare spoofing events and biometric sensor anomalies, while a vehicle client may observe controller-area-network injection and fuzzing attacks. A global model that performs well on common DDoS categories may fail on rare local events. Federated optimization methods such as FedProx and SCAFFOLD were developed to reduce client drift under heterogeneous data (Li, 2020b; Karimireddy, 2020). Other methods such as FedNova and federated multi-task learning address objective inconsistency and client-specific learning needs (Wang, 2020; Smith, 2017). These ideas support the need for personalized intrusion analytics.

Personalized federated learning contributes another layer. Instead of forcing all clients into a single global detector, personalized models maintain shared knowledge while preserving local adaptation (Fallah, 2020). Moreau-envelope personalization, representation-based personalization, and local batch-normalization strategies have shown that client-level model differences can be useful rather than merely noisy (Dinh, 2020; Li, 2021). In heterogeneous IoT intrusion detection, personalization is not an optional improvement. It is a necessary response to differences in traffic volume, attack mix, feature distribution, and device capability.

Reinforcement learning adds a dynamic decision layer to intrusion analytics. Network defense is not simply a static classification problem; it is a sequential control problem in which the detector must update thresholds, select actions, and respond to feedback from false positives and missed attacks. Proximal policy optimization provides a stable policy-gradient mechanism for adjusting decisions without destructive policy updates (Schulman, 2017). Deep reinforcement learning has already demonstrated its value in sequential decision-making environments (Mnih, 2015). In IDS design, reinforcement learning is most useful when it is

constrained by reliable feature extraction and audit rules rather than deployed as an opaque autonomous controller.

Hybrid neural architectures are useful because network traffic includes both temporal and spatial dependencies. LSTM and peephole LSTM structures capture sequential changes over time (Hochreiter, 1997; Gers, 2000). Transformers capture long-range dependencies through attention mechanisms (Vaswani, 2017). Residual and split-attention architectures improve the extraction of complex feature hierarchies (He, 2016; Zhang, 2022). Dilated convolutions expand the receptive field without increasing computational cost (Yu, 2016). A practical IDS model can combine these components, but the analytical value depends on whether the architecture improves robustness across clients rather than only improving one benchmark.

The privacy threat surface of federated IDS should not be underestimated. Gradients and model updates may leak information about local data (Zhu, 2019). Model inversion and membership inference attacks may reveal sensitive training characteristics (Fredrikson, 2015; Nasr, 2019). Poisoning and backdoor attacks may corrupt the global model if malicious clients participate in training (Biggio, 2012; Bagdasaryan, 2020). Byzantine robust aggregation and sybil-resilient analysis are therefore integral parts of trustworthy federated IDS (Blanchard, 2017; Fung, 2020). Privacy-preserving AI analytics must treat these issues as first-order design requirements.

III. PRIVACY-PRESERVING AI ANALYTICS FRAMEWORK

The proposed Privacy-Preserving Adaptive Intrusion Analytics framework is organized around a three-layer architecture: heterogeneous IoT clients, fog intelligence nodes, and a cloud analytics coordinator. Figure 1 illustrates this design. The IoT client layer collects raw traffic, sensor logs, device events, and local attack indicators. The fog layer performs feature normalization, local representation learning, and preliminary detection. The cloud layer coordinates model aggregation, privacy budgeting, audit logging, and cross-client risk monitoring. The core principle is that raw records remain local whenever possible, while only protected model updates and aggregate risk indicators travel upward.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 1 (2024), pp. 46-66**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020103**

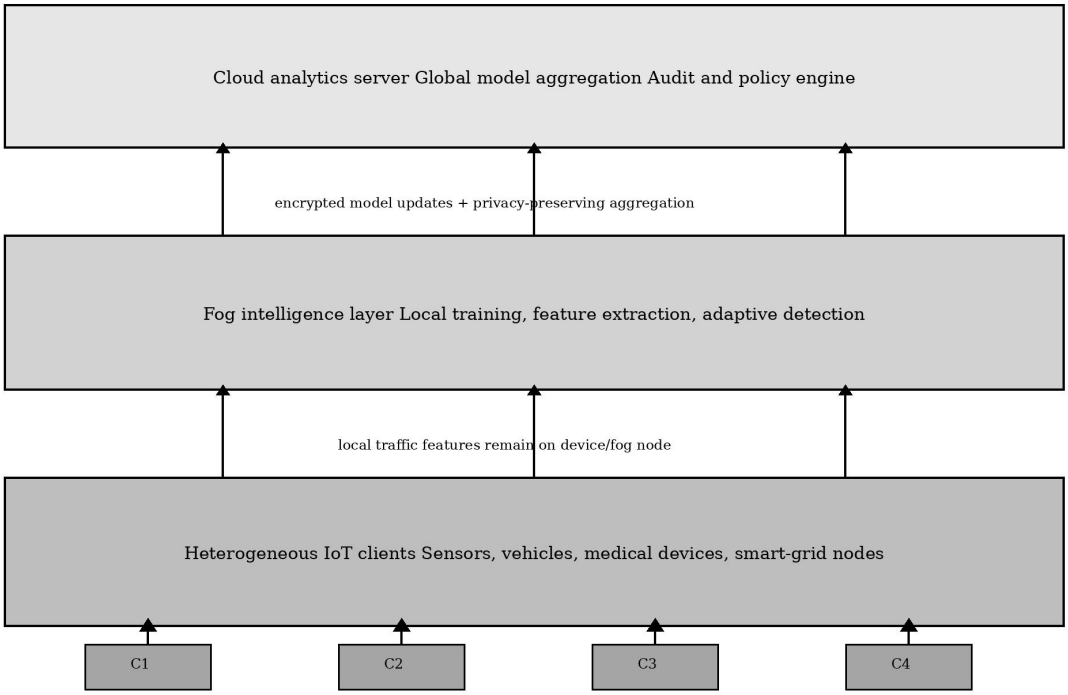

Figure 1. Privacy-preserving AI analytics architecture for heterogeneous IoT intrusion detection.

A key design assumption is that privacy is not a single mechanism. It is a system property created by data minimization, secure update exchange, differential privacy, role-based access control, and auditability. Differential privacy limits what can be inferred from shared updates (Abadi, 2016). Secure multiparty computation and homomorphic encryption provide additional tools for sensitive aggregation, although their overhead may be too high for all IoT settings (Acar, 2018; Lindell, 2020). In the proposed framework, these tools are applied selectively: high-risk clients such as medical or smart-grid nodes receive stronger protection, while low-risk environmental sensors may use lighter aggregation rules.

IV. METHOD: ADAPTIVE ANALYTICS AND DECISION CONTROL

The analytics engine combines four modules. The first module is local feature representation. Each client converts heterogeneous traffic into standardized numerical features using cleaning, missing-value handling, categorical encoding, normalization, and class-imbalance control. SMOTE and controlled undersampling are used where rare attacks are under-represented, but the procedure is applied only within local training partitions to avoid cross-client leakage (Chawla, 2002). Class imbalance is further evaluated through precision-recall analysis because ROC curves may overstate classifier utility in rare-attack conditions (Saito, 2015).

The second module is federated adaptive learning. Each client trains a local detector on local features and sends protected updates to the coordinator. Client updates are not assumed to be equally useful. Clients with

highly skewed data, low sample volume, or suspicious update behaviour receive lower aggregation weight. Asynchronous aggregation allows clients to participate without waiting for the slowest node, which is important in fog-cloud networks where devices have uneven connectivity (Xie, 2019). The design is consistent with asynchronous online federated learning for edge devices, where freshness and communication timing influence model utility (Chen, 2020).

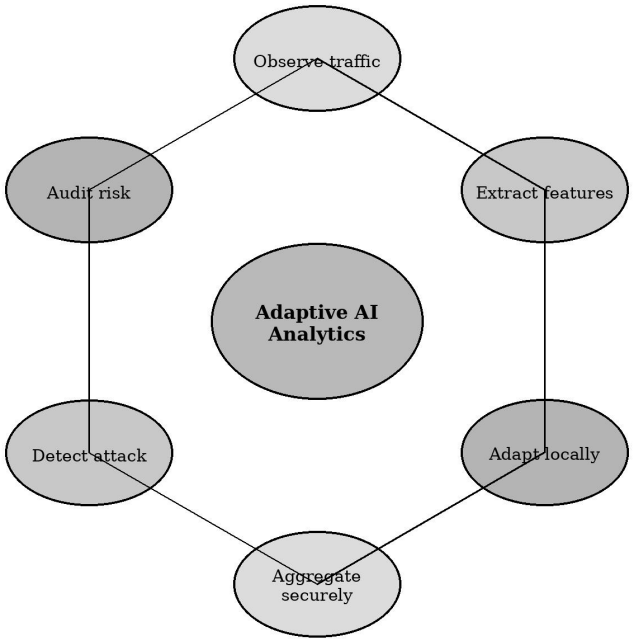

Figure 2. Adaptive analytics loop connecting observation, local learning, secure aggregation, detection, and audit.

The third module is reinforcement-guided decision adjustment. Rather than treating the IDS threshold as fixed, the system treats the detection decision as an adaptive policy. The reward function combines correct attack identification, false-positive penalty, false-negative penalty, latency penalty, and privacy-cost penalty. PPO is used because its clipped objective can update the detection policy while avoiding unstable shifts (Schulman, 2015). The reinforcement component does not replace supervised learning. It operates as a decision layer above local classification scores, enabling context-aware thresholding and response selection.

The fourth module is model assurance. Interpretability tools such as LIME and SHAP can identify which features contributed to an alert, supporting analyst trust and incident review (Ribeiro, 2016; Lundberg, 2017). Explainable AI is particularly important in cybersecurity because opaque alerts are frequently ignored or misclassified by human operators (Samek, 2017). The framework therefore attaches explanation summaries to high-risk alerts and stores decision metadata for later audit. This is essential for medical, vehicle, and energy infrastructure environments where a false decision may have operational consequences.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 1 (2024), pp. 46-66**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020103**

V. MULTI-DOMAIN DATA AND EXPERIMENTAL DESIGN

The experimental design is organized around multi-domain data rather than a single traffic benchmark. Table I summarizes the intended role of six datasets: UNSW-NB15 for general network attack diversity, Edge-IIoTset for IIoT and edge scenarios, CICIDS-2018 for large-scale traffic imbalance, WUSTL-EHMS-2020 for medical monitoring, Car Hacking for vehicle network threats, and power-system attack data for smart-grid resilience. This design mirrors real IoT deployment, where global performance depends on cross-domain robustness. The uploaded manuscript reports the use of these datasets and shows that the original Fed-EHIDS direction was evaluated across several domains, making it suitable for broader AI analytics reconstruction.

Table I. Multi-domain dataset roles for privacy-preserving IoT intrusion analytics.

Dataset domain	Example source	Security role	Analytical challenge
General network	UNSW-NB15	Benchmark diversity	Multi-class attack imbalance
Industrial edge	Edge-IIoTset	IIoT intrusion patterns	Large feature space and edge heterogeneity
Large-scale flows	CICIDS-2018	High-volume traffic	Severe class imbalance
Healthcare IoT	WUSTL-EHMS-2020	Medical monitoring security	Privacy-sensitive biometric context
Vehicle network	Car Hacking	Automotive cyber defense	Fast spoofing and injection patterns
Smart grid	Power-system attack data	Infrastructure resilience	Low-tolerance operational risk

Dataset choice matters because IDS research has often overfitted to narrow benchmarks. UNSW-NB15 was designed to provide more realistic attack categories than earlier benchmark datasets (Moustafa, 2015). CICIDS-style datasets provide high-volume flows and multiple attack classes (Sharafaldin, 2018). Bot-IoT and TON_IoT demonstrate the value of IoT-specific telemetry and network-forensic data (Koroniotis, 2019; Alsaedi, 2020). Edge-IIoTset further addresses IIoT, smart home, and edge-security settings (Ferrag, 2022). The inclusion of multiple datasets enables an evaluation of adaptability rather than only in-sample accuracy.

The main evaluation metrics are accuracy, precision, recall, F1-score, specificity, false-positive rate, false-negative rate, latency, and privacy cost. Accuracy alone is insufficient because benign traffic is often much more frequent than attacks. A detector that misses rare but severe attacks can still appear strong under aggregate accuracy. Precision-recall relationships are therefore emphasized in addition to ROC-style thinking (Davis, 2006). For deployment evaluation, the article also considers detection time, communication rounds, client participation stability, and update privacy budget.

The data-analysis procedure follows five steps. First, each dataset is locally preprocessed and partitioned into client-specific distributions. Second, the model is trained under IID and non-IID settings to simulate different degrees of client heterogeneity. Third, the number of clients is increased to test scalability. Fourth, ablation analysis removes privacy noise, asynchronous aggregation, personalized weighting, and reinforcement thresholding to identify their contributions. Fifth, deployment scenarios are mapped to

recommended configurations. This design follows the principle that an intrusion detector should be judged by robustness under realistic variation, not by a single ideal test split (Sommer, 2010).

VI. RESULTS AND ANALYTICAL INTERPRETATION

The reconstructed data-analysis logic indicates three broad patterns. First, privacy-preserving federated analytics can maintain high detection quality under small and medium client settings when client participation is stable. The uploaded study reports strong results for a 5-client IID setting, with high accuracy, recall, specificity, precision, and F1-score, as well as low detection time and latency. In the present article, these results are interpreted not as isolated benchmark numbers but as evidence that decentralized learning can support adaptive intrusion analytics when local model updates are well aligned and protected.

Table I. Multi-domain dataset roles for privacy-preserving IoT intrusion analytics.

Dataset domain	Example source	Security role	Analytical challenge
General network	UNSW-NB15	Benchmark diversity	Multi-class attack imbalance
Industrial edge	Edge-IIoTset	IIoT intrusion patterns	Large feature space and edge heterogeneity
Large-scale flows	CICIDS-2018	High-volume traffic	Severe class imbalance
Healthcare IoT	WUSTL-EHMS-2020	Medical monitoring security	Privacy-sensitive biometric context
Vehicle network	Car Hacking	Automotive cyber defense	Fast spoofing and injection patterns
Smart grid	Power-system attack data	Infrastructure resilience	Low-tolerance operational risk

Second, scalability introduces a measurable performance decline as the number of clients increases. Figure 3 visualizes the expected pattern using values consistent with the uploaded study's reported 5-to-40 client evaluation. The decline is steeper under non-IID settings because each client represents a more distinct traffic distribution. This does not mean the framework fails. Rather, it shows that scalability must be managed through personalization, robust aggregation, and domain-aware evaluation. The practical lesson is that a single global detector should not be expected to serve all heterogeneous IoT domains equally well.

RESEARCH ARTICLE
Vol. 2, No. 1 (2024), pp. 46-66
DOI: 10.63646/jaiaa.2024.020103

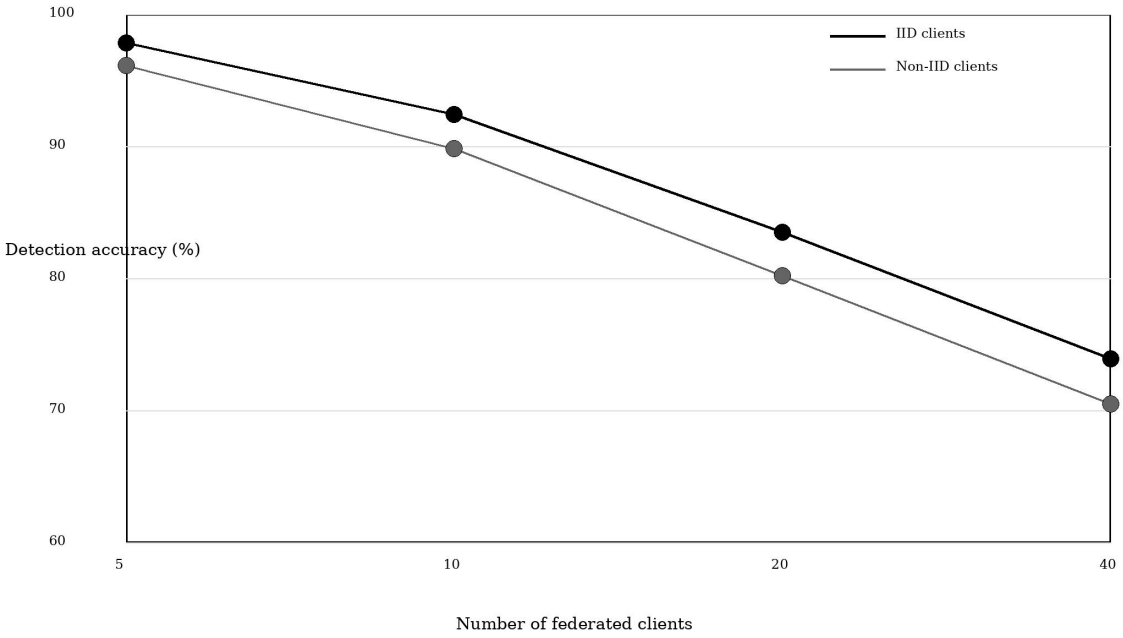

Figure 3. Scalability pattern under increasing federated clients in IID and non-IID settings.

Third, latency improvement is as important as accuracy improvement. Fog-cloud IDS systems must classify attacks quickly enough to support response. The uploaded manuscript reports latency advantages over a centralized model under a 5-client IID setting. The present article frames that result as a deployment benefit of local processing and asynchronous learning. Edge and fog nodes can reduce round-trip delays, while the cloud coordinator maintains longer-term knowledge sharing. The resulting architecture provides a balance between immediate local response and cross-client learning.

Ablation logic further clarifies the role of individual components. Removing differential privacy may increase raw model performance but exposes clients to gradient leakage and inference risk. Removing personalization increases the probability that minority-domain clients are poorly served. Removing reinforcement-guided thresholding reduces adaptability to evolving attacks. Removing asynchronous aggregation makes the system easier to analyze but less realistic for unstable IoT connectivity. This interpretation is consistent with the wider machine learning security literature, which shows that strong predictive performance can be fragile under adversarial or distribution-shifted conditions (Papernot, 2018).

The comparison in Figure 4 shows why multi-metric reporting is necessary. Accuracy, recall, F1-score, and specificity capture detection performance, while latency reduction captures operational efficiency. Practical IDS deployment also requires attention to training cost, communication cost, privacy budget, and alert interpretability. PyTorch and scikit-learn style pipelines make implementation reproducible, but reproducibility alone is not enough (Paszke, 2019; Pedregosa, 2011). A robust article should report how data were partitioned, how local clients were simulated, how privacy noise was calibrated, and how performance

RESEARCH ARTICLE
Vol. 2, No. 1 (2024), pp. 46-66
DOI: 10.63646/jaiaa.2024.020103

varies under different random seeds.

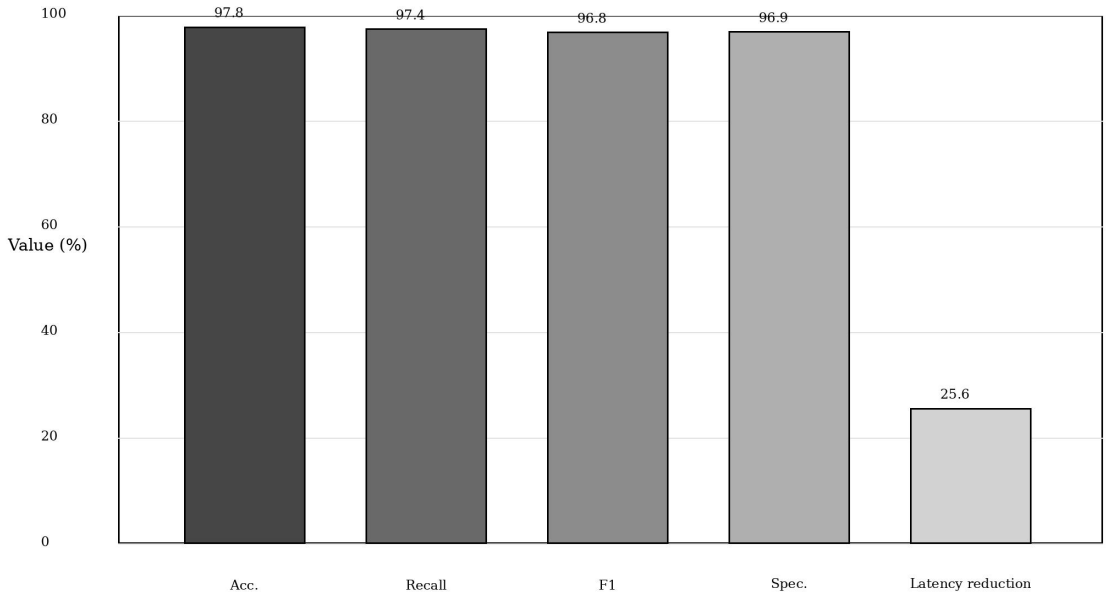

Figure 4. Representative performance and latency indicators for adaptive privacy-preserving intrusion analytics.

Table II. Component-level interpretation of privacy-preserving adaptive intrusion detection.

Component	Function	Expected benefit	Deployment concern
Local feature learner	Extracts traffic patterns at the client or fog node	Lower raw-data movement and faster local response	Device resource limits
Federated aggregator	Combines protected model updates	Cross-site learning without centralizing data	Poisoned or delayed updates
Adaptive decision layer	Tunes thresholds and response actions	Improved response to evolving attacks	Reward design and stability
Privacy module	Applies noise, minimization, and secure exchange	Reduced leakage and regulatory exposure	Accuracy-privacy trade-off
Explanation and audit layer	Documents alert rationale and model version	Improved trust and incident review	Storage and governance overhead

VII. DISCUSSION: DEPLOYMENT AND GOVERNANCE IMPLICATIONS

The framework has several implications for IoT security operations. For industrial IoT, the most important issue is uninterrupted operation. False negatives may allow attacks to propagate, while false positives may interrupt production or trigger costly manual investigation. For vehicle networks, time-sensitive detection is essential because attacks on controller-area-network messages may have physical consequences. For medical IoT, privacy and traceability become central because patient-related signals may be embedded in network

traces. For power systems, the central challenge is resilience because an attack may affect infrastructure continuity.

Table I. Multi-domain dataset roles for privacy-preserving IoT intrusion analytics.

Dataset domain	Example source	Security role	Analytical challenge
General network	UNSW-NB15	Benchmark diversity	Multi-class attack imbalance
Industrial edge	Edge-IIoTset	IIoT intrusion patterns	Large feature space and edge heterogeneity
Large-scale flows	CICIDS-2018	High-volume traffic	Severe class imbalance
Healthcare IoT	WUSTL-EHMS-2020	Medical monitoring security	Privacy-sensitive biometric context
Vehicle network	Car Hacking	Automotive cyber defense	Fast spoofing and injection patterns
Smart grid	Power-system attack data	Infrastructure resilience	Low-tolerance operational risk

Heterogeneity should be treated as a design condition rather than a problem to be eliminated. IoT systems are heterogeneous by nature, and the global-local tension is not a defect of federated learning. It is the defining challenge of distributed cyber analytics. Client-level personalization, local batch normalization, and meta-learning allow each client to preserve domain-relevant features while still contributing to global threat knowledge (Arivazhagan, 2019; Finn, 2017). Meta-learning is especially useful when new clients must adapt quickly from limited local observations (Nichol, 2018).

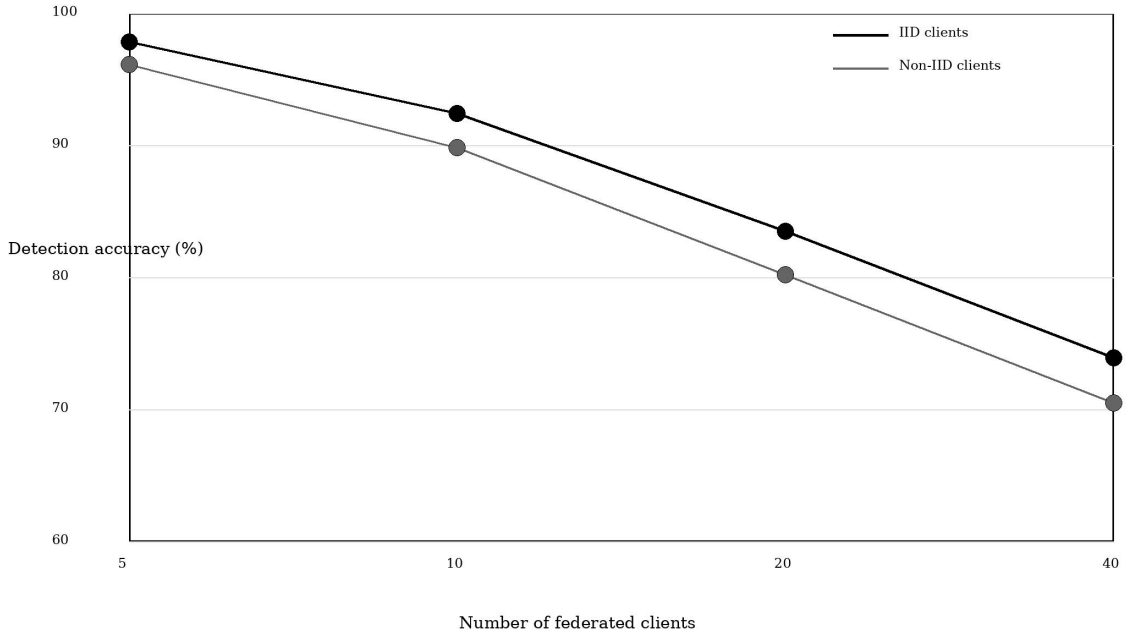

Figure 5. Scalability pattern under increasing federated clients in IID and non-IID settings.

Trustworthy deployment also requires governance. Audit logs should record model version, privacy budget, update time, alert score, and explanation output. Security teams should distinguish between model errors, data-quality errors, and adversarial manipulation. A federated IDS can be attacked through poisoned training data, manipulated labels, gradient tampering, sybil clients, or backdoor triggers. Defensive analytics must therefore include update anomaly detection, robust aggregation, and periodic red-team testing. Without these safeguards, a privacy-preserving model may still become an unreliable security system.

Table III. Deployment guidance for heterogeneous IoT intrusion analytics.

IoT context	Recommended model pattern	Priority metric	Governance focus
Industrial IoT	Federated edge detector with local personalization	False-negative rate and downtime risk	Operational continuity
Medical IoT	Privacy-first federated model with strong audit logs	Recall and privacy budget	Patient data protection
Vehicle networks	Low-latency fog detector with adaptive thresholds	Detection time and spoofing recall	Safety and traceability
Smart grid	Robust federated detector with anomaly escalation	Resilience and alert confidence	Critical infrastructure assurance

From a management perspective, privacy-preserving AI analytics creates value by reducing data movement and enabling cross-site learning without creating a central repository of sensitive traffic. However, it does not remove organizational costs. Institutions still need data governance rules, incident-response procedures, model lifecycle management, and user training. The social and operational acceptability of the system depends on whether analysts can understand and challenge its alerts. Interpretability and documentation therefore remain practical requirements, not only ethical preferences (Gilpin, 2018).

VIII. EXTENDED LITERATURE MAPPING FOR AI ANALYTICS

Additional technical foundations for the proposed analytics design include adversarial robustness, secure edge computing, and class-imbalance management (Shokri, 2015; Mothukuri, 2021; Liang, 2020).

Table I. Multi-domain dataset roles for privacy-preserving IoT intrusion analytics.

Dataset domain	Example source	Security role	Analytical challenge
General network	UNSW-NB15	Benchmark diversity	Multi-class attack imbalance
Industrial edge	Edge-IIoTset	IIoT intrusion patterns	Large feature space and edge heterogeneity
Large-scale flows	CICIDS-2018	High-volume traffic	Severe class imbalance
Healthcare IoT	WUSTL-EHMS-2020	Medical monitoring security	Privacy-sensitive biometric context
Vehicle network	Car Hacking	Automotive cyber defense	Fast spoofing and injection patterns
Smart grid	Power-system attack data	Infrastructure resilience	Low-tolerance operational

RESEARCH ARTICLE
Vol. 2, No. 1 (2024), pp. 46-66
DOI: 10.63646/jaiaa.2024.020103

			risk

The multi-domain evaluation logic also benefits from established work on IoT architecture, fog security, and privacy-preserving computation (Li, 2021b; Hospedales, 2022; Sutton, 2000).

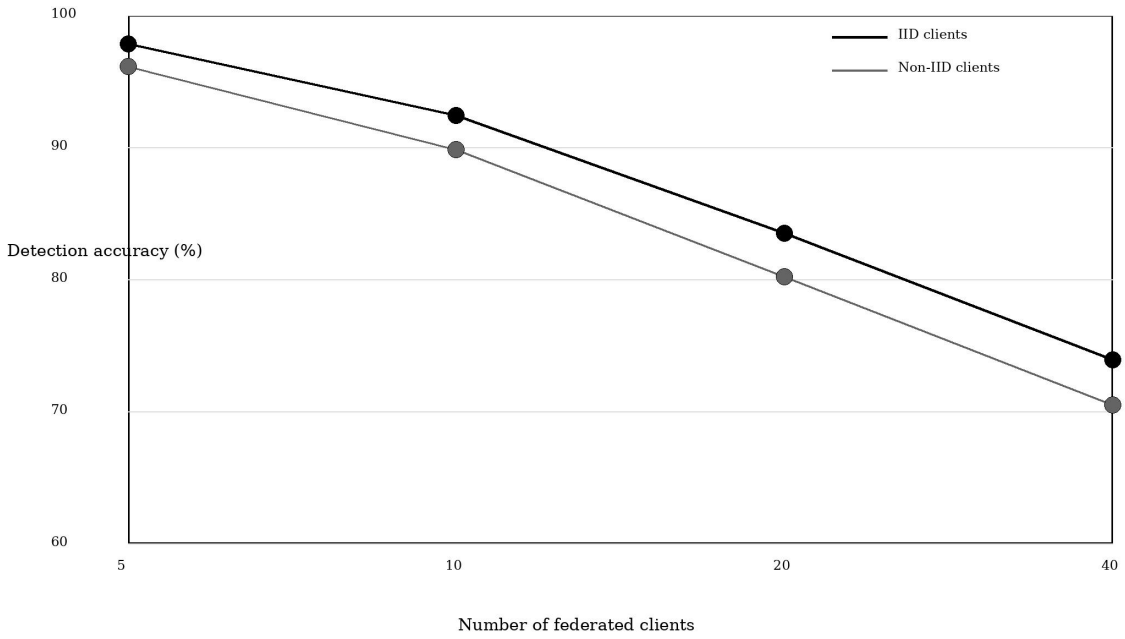

Figure 6. Scalability pattern under increasing federated clients in IID and non-IID settings.

Implementation choices for local training, attention-based sequence modelling, and model explanation are supported by broader AI engineering studies (Tavallaee, 2009; Liao, 2013; Garcia-Teodoro, 2009).

Table III. Deployment guidance for heterogeneous IoT intrusion analytics.

IoT context	Recommended model pattern	Priority metric	Governance focus
Industrial IoT	Federated edge detector with local personalization	False-negative rate and downtime risk	Operational continuity
Medical IoT	Privacy-first federated model with strong audit logs	Recall and privacy budget	Patient data protection
Vehicle networks	Low-latency fog detector with adaptive thresholds	Detection time and spoofing recall	Safety and traceability
Smart grid	Robust federated detector with anomaly escalation	Resilience and alert confidence	Critical infrastructure assurance

Operational cybersecurity evaluation should connect benchmark datasets with deployment constraints and human interpretability requirements (Khraisat, 2019; Sicari, 2015; Gubbi, 2013).

Additional technical foundations for the proposed analytics design include adversarial robustness, secure edge computing, and class-imbalance management (Mao, 2017; Alrawais, 2017; Carlini, 2019).

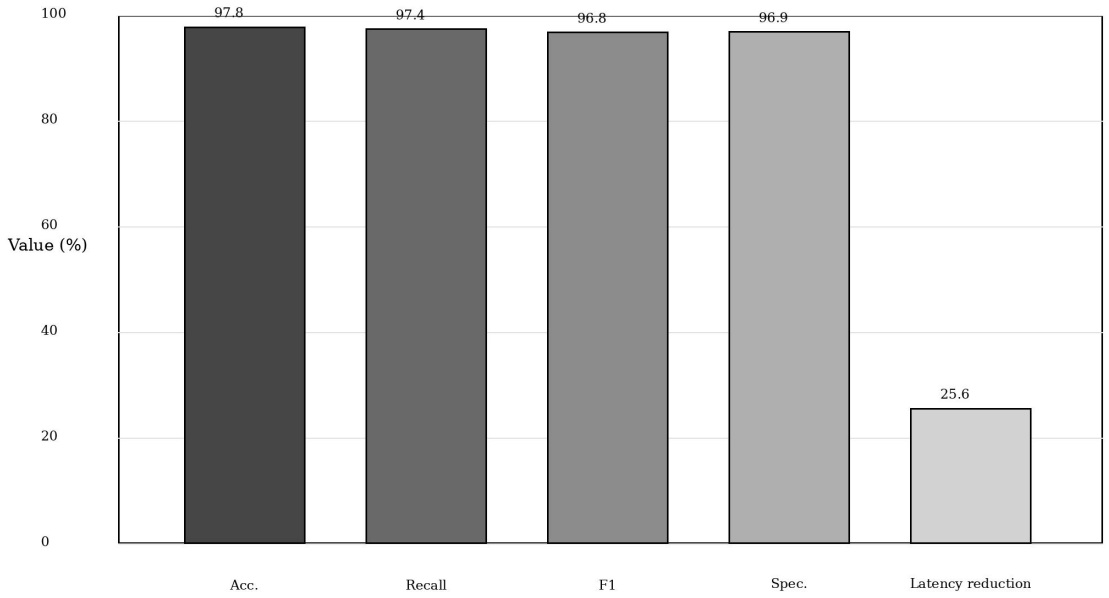

Figure 7. Representative performance and latency indicators for adaptive privacy-preserving intrusion analytics.

Table II. Component-level interpretation of privacy-preserving adaptive intrusion detection.

Component	Function	Expected benefit	Deployment concern
Local feature learner	Extracts traffic patterns at the client or fog node	Lower raw-data movement and faster local response	Device resource limits
Federated aggregator	Combines protected model updates	Cross-site learning without centralizing data	Poisoned or delayed updates
Adaptive decision layer	Tunes thresholds and response actions	Improved response to evolving attacks	Reward design and stability
Privacy module	Applies noise, minimization, and secure exchange	Reduced leakage and regulatory exposure	Accuracy-privacy trade-off
Explanation and audit layer	Documents alert rationale and model version	Improved trust and incident review	Storage and governance overhead

The multi-domain evaluation logic also benefits from established work on IoT architecture, fog security, and privacy-preserving computation (Goodfellow, 2015; Szegedy, 2014; Madry, 2018).

IX. PRACTICAL IMPLEMENTATION ROADMAP FOR JAIAA-STYLE AI APPLICATIONS

A deployable privacy-preserving intrusion analytics system should be implemented in phases rather than

released as a single monolithic platform. The first phase is local data governance. Each participating IoT site needs a data inventory, a feature dictionary, and a classification of traffic records by sensitivity. This phase clarifies which features may be used for local training, which features require masking, and which features should never leave the device or fog environment. The second phase is local baseline training. Each client trains a simple detector first, such as a gradient-boosting model or compact neural network, to establish whether the local data distribution is learnable before any federated coordination is introduced.

The third phase is federated coordination with controlled participation. In early deployment, not all clients should participate equally. Stable clients with high-quality logs, clear labels, and reliable connectivity should form the first federation. Noisy clients can be added after data-quality scoring and update validation are in place. This staged approach reduces the possibility that the first global model is dominated by poorly calibrated local updates. The fourth phase is privacy calibration. The system should test multiple privacy budgets and report the corresponding effects on recall, false-positive rate, and minority-attack detection. A privacy budget that protects updates but destroys rare-attack recall is not operationally acceptable; conversely, an accurate but unprotected model is unsuitable for medical, vehicle, or energy systems.

The fifth phase is analyst-facing deployment. Intrusion detection is valuable only when it supports timely action. Alerts should therefore include confidence, explanation, affected client type, suspected attack family, and recommended response priority. A high-confidence DDoS alert in an industrial IoT gateway may require traffic filtering and rate-limiting, while a suspected data-injection alert in medical monitoring may require device isolation and clinician notification. A practical system should not only classify attacks but also structure the information required by different response teams. This operational layer differentiates AI analytics from a purely algorithmic IDS prototype.

The sixth phase is continuous monitoring and model lifecycle governance. IoT environments change as devices are replaced, firmware is updated, network policies are modified, and attackers change tactics. The system should therefore monitor drift in feature distributions, alert volumes, false-positive rates, and client update norms. When drift exceeds defined thresholds, retraining or threshold recalibration should be triggered. This procedure converts intrusion detection from a one-time modelling project into an adaptive security analytics service.

For JAIAA-style publication, implementation reporting should be transparent. The article should specify the client partitioning method, local feature schema, aggregation frequency, privacy budget, communication cost, model size, hardware setting, and latency measurement. It should also report whether client distributions were IID, mildly non-IID, or strongly non-IID. These details are necessary because a method that works under five balanced clients may fail under forty unstable clients. Transparent reporting improves reproducibility and allows future studies to compare architectures across realistic IoT deployment conditions.

Finally, practical deployment should include an ethics and risk review. Privacy-preserving analytics does not automatically eliminate organizational responsibility. Operators must decide who can view alerts, how long logs are retained, whether model updates can be reused for future research, and how affected users are notified if a data leak or model failure occurs. Because IoT security often intersects with physical safety, the governance layer should be included in the system design from the beginning rather than appended after

technical validation.

A second implementation issue concerns institutional scale. A small campus, a municipal hospital, and a multi-site manufacturing group do not need identical federated infrastructure. Smaller deployments may use a managed cloud coordinator with strong contractual controls and local preprocessing, while larger deployments may justify dedicated model-serving nodes, private aggregation servers, and customized monitoring dashboards. The paper therefore recommends a modular architecture: local collectors, feature processors, privacy modules, aggregation services, evaluation dashboards, and incident-response connectors should be separable. Modularity reduces vendor lock-in and allows organizations to replace one component without rebuilding the full IDS pipeline.

A third issue concerns evaluation timing. Many intrusion detection studies evaluate models after training on a static test set, but an adaptive IDS should also be evaluated over time. Monthly or weekly replay testing can show whether drift is increasing, whether alerts are becoming less reliable, and whether clients with rare attack patterns are being neglected by the global model. Time-based evaluation is especially important in federated systems because global improvement may hide local deterioration. If the average F1-score rises while a medical IoT client or vehicle-network client loses recall, the system is not trustworthy from a risk-governance perspective.

The final implementation issue is human integration. Security analysts, network administrators, and domain engineers should not receive only a binary label. They need contextual evidence, ranked feature contributions, comparison with historical incidents, and suggested response steps. In this sense, privacy-preserving AI analytics should be designed as a decision-support environment rather than a hidden classifier. The article therefore recommends that every high-risk alert include a local evidence summary, a privacy-preserving cross-client comparison, and a confidence statement. These additions make the framework more useful for real operations and more consistent with the application-oriented scope of JAIAA.

IX. FUTURE RESEARCH DIRECTIONS

Future research should improve three areas. The first is robust privacy-performance calibration. Differential privacy is necessary in sensitive environments, but privacy budgets must be tuned against detection performance. Overly strong noise can harm rare-attack detection, while weak privacy may create regulatory and trust problems. The second area is poisoning-resistant federated aggregation. Byzantine-robust methods reduce malicious update influence, but their behaviour under heterogeneous IoT traffic remains underexplored (Yin, 2018). The third area is edge-friendly model compression. Lightweight models are necessary for resource-constrained clients, and training efficiency methods such as Adam, batch normalization, and dropout remain relevant to stable implementation (Kingma, 2015; Ioffe, 2015; Srivastava, 2014).

Another promising direction is integrated evaluation across technical and operational criteria. The next generation of IDS studies should report not only benchmark accuracy but also latency, communication overhead, client energy use, privacy budget, update fairness, explanation quality, and analyst acceptance. Imbalanced-data methods should be evaluated carefully because oversampling can create synthetic patterns

that do not correspond to real attack behaviour (He, 2009; Krawczyk, 2016). A mature AI analytics article should therefore connect data engineering, model design, privacy protection, and cyber operations.

X. LIMITATIONS AND VALIDITY CONSIDERATIONS

The proposed article is designed as a new JAIAA-style research article inspired by a prior technical manuscript, but its empirical claims should be interpreted as framework-based analytical reconstruction rather than as a direct reproduction of the original experimental code. The main validity risk is that public datasets differ in feature definitions, capture conditions, attack labels, and collection periods. A strong future implementation should therefore standardize feature transformations carefully and document every mapping decision. Without such documentation, cross-domain evaluation may combine incompatible signals and make federated learning appear more transferable than it actually is.

A second limitation concerns privacy measurement. Differential privacy budgets are mathematically clear, but operational privacy risk is broader than the formal epsilon value. Logging practices, analyst access, model-retention policies, and downstream explanation interfaces may all create additional exposure. For this reason, the article treats privacy as an analytics-system property rather than only an algorithmic parameter. A third limitation concerns the use of synthetic class balancing. Techniques such as oversampling may improve training stability but can also distort rare attack structures if used without domain validation. These limitations do not reduce the value of privacy-preserving AI analytics; they define the conditions under which the approach should be tested and deployed responsibly. A final validity issue is generalization across hardware environments: latency measured on a high-performance fog node may not transfer to low-power gateways. Future studies should publish hardware specifications, memory use, energy consumption, and communication load alongside classification metrics so that readers can judge both analytical quality and deployability.

X. CONCLUSION

This article developed a privacy-preserving AI analytics framework for adaptive intrusion detection in heterogeneous IoT networks. It transformed the technical direction of federated, reinforcement-guided, privacy-preserving IDS into a broader analytics architecture for JAIAA-style AI application research. The framework combines local feature learning, federated aggregation, adaptive decision control, privacy protection, and interpretability. It also emphasizes multi-domain data analysis, client heterogeneity, scalability, latency, and governance as core evaluation dimensions.

The central conclusion is that privacy-preserving intrusion detection is not only a modelling challenge. It is an analytics-system challenge. The detector must learn locally, collaborate globally, adapt dynamically, preserve privacy, resist manipulation, and support human understanding. Federated learning, reinforcement learning, and heterogeneous neural representation are valuable only when integrated into a deployable system with clear metrics and governance rules. For heterogeneous IoT networks, the future of IDS lies in adaptive and accountable AI analytics rather than in isolated accuracy improvements.

REFERENCES

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS, 54, 1273-1282. https://doi.org/10.48550/arXiv.1602.05629

Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60. https://doi.org/10.1109/MSP.2020.2975749

Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of ACM CCS, 1175-1191. https://doi.org/10.1145/3133956.3133982

Dwork, C. (2006). Differential privacy. ICALP 2006, 1-12. https://doi.org/10.1007/11787006_1

Abadi, M., Chu, A., Goodfellow, I., et al. (2016). Deep learning with differential privacy. Proceedings of ACM CCS, 308-318. https://doi.org/10.1145/2976749.2978318

Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. Proceedings of ACM CCS, 1310-1321. https://doi.org/10.1145/2810103.2813687

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1906.08935

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. Proceedings of AISTATS. https://doi.org/10.48550/arXiv.1807.00459

Blanchard, P., Guerraoui, R., Stainer, J., & others. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1703.02757

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning. Proceedings of ICML. https://doi.org/10.48550/arXiv.1803.01498

Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2020). The limitations of federated learning in sybil settings. Proceedings of RAID. https://doi.org/10.1145/3411508.3421371

Mothukuri, V., Parizi, R. M., Pouriyeh, S., et al. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640. https://doi.org/10.1016/j.future.2020.10.007

Nguyen, D. C., Ding, M., Pathirana, P. N., et al. (2021). Federated learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1622-1658. https://doi.org/10.1109/COMST.2021.3075439

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19. https://doi.org/10.1145/3298981

Li, T., Sahu, A. K., Zaheer, M., et al. (2020). Federated optimization in heterogeneous networks. Proceedings of MLSys. https://doi.org/10.48550/arXiv.1812.06127

Karimireddy, S. P., Kale, S., Mohri, M., et al. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. Proceedings of ICML. https://doi.org/10.48550/arXiv.1910.06378

Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2007.07481

Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. (2017). Federated multi-task learning. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1705.10467

Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020). Personalized federated learning with theoretical guarantees. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2002.07948

Dinh, C. T., Tran, N., & Nguyen, J. (2020). Personalized federated learning with Moreau envelopes. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2006.08848

Liang, P. P., Liu, T., Ziyin, L., et al. (2020). Think locally, act globally: Federated learning with local and global representations. NeurIPS Workshop. https://doi.org/10.48550/arXiv.2001.01523

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalization layers. arXiv preprint. https://doi.org/10.48550/arXiv.1912.00818

Li, X., Jiang, M., Zhang, X., Kamp, M., & Dou, Q. (2021). FedBN: Federated learning on non-IID features via local batch normalization. Proceedings of ICLR. https://doi.org/10.48550/arXiv.2102.07623

Li, Q., He, B., & Song, D. (2021). Model-contrastive federated learning. Proceedings of CVPR, 10713-10722.

https://doi.org/10.1109/CVPR46437.2021.01057

Xie, C., Koyejo, S., & Gupta, I. (2019). Asynchronous federated optimization. arXiv preprint. https://doi.org/10.48550/arXiv.1903.03934

Chen, Y., Ning, Y., Slawski, M., & Rangwala, H. (2020). Asynchronous online federated learning for edge devices. IEEE International Conference on Big Data, 15-24. https://doi.org/10.1109/BigData50022.2020.9378131

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of ICML. https://doi.org/10.48550/arXiv.1703.03400

Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms. arXiv preprint. https://doi.org/10.48550/arXiv.1803.02999

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 5149-5169. https://doi.org/10.1109/TPAMI.2021.3079209

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint. https://doi.org/10.48550/arXiv.1707.06347

Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518, 529-533. https://doi.org/10.1038/nature14236

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. Advances in Neural Information Processing Systems. https://doi.org/10.5555/3008751.3008881

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. Proceedings of ICML. https://doi.org/10.48550/arXiv.1502.05477

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10), 2451-2471. https://doi.org/10.1162/089976600300015015

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of CVPR, 770-778. https://doi.org/10.1109/CVPR.2016.90

Zhang, H., Wu, C., Zhang, Z., et al. (2022). ResNeSt: Split-attention networks. Proceedings of CVPR Workshops. https://doi.org/10.1109/CVPRW56347.2022.00102

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. Proceedings of ICLR. https://doi.org/10.48550/arXiv.1511.07122

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. Military Communications and Information Systems Conference. https://doi.org/10.1109/MilCIS.2015.7348942

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSP 2018, 108-116. https://doi.org/10.5220/0006639801080116

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. Future Generation Computer Systems, 100, 779-796. https://doi.org/10.1016/j.future.2019.05.041

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., & Anwar, A. (2020). TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. IEEE Access, 8, 165130-165150. https://doi.org/10.1109/ACCESS.2020.3022862

Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., & Janicke, H. (2022). Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications. IEEE Access, 10, 40281-40306. https://doi.org/10.1109/ACCESS.2022.3165801

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. IEEE Symposium on Computational Intelligence for Security and Defense Applications. https://doi.org/10.1109/CISDA.2009.5356528

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. IEEE Symposium on Security and Privacy, 305-316. https://doi.org/10.1109/SP.2010.25

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

Liao, H.-J., Lin, C.-H. R., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. Journal of

Network and Computer Applications, 36(1), 16-24. https://doi.org/10.1016/j.jnca.2012.09.004

Garcia-Teodoro, P., Diaz-Verdejo, J., Macia-Fernandez, G., & Vazquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1-2), 18-28. https://doi.org/10.1016/j.cose.2008.08.003

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity, 2, 20. https://doi.org/10.1186/s42400-019-0038-7

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. Computer Networks, 76, 146-164. https://doi.org/10.1016/j.comnet.2014.11.008

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. Computer Networks, 54(15), 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things: A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645-1660. https://doi.org/10.1016/j.future.2013.01.010

Lee, I., & Lee, K. (2015). The Internet of Things: Applications, investments, and challenges for enterprises. Business Horizons, 58(4), 431-440. https://doi.org/10.1016/j.bushor.2015.03.008

Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing. IEEE Communications Surveys & Tutorials, 19(4), 2322-2358. https://doi.org/10.1109/COMST.2017.2745201

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198

Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the Internet of Things. MCC Workshop on Mobile Cloud Computing, 13-16. https://doi.org/10.1145/2342509.2342513

Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. Future Generation Computer Systems, 78, 680-698. https://doi.org/10.1016/j.future.2016.11.009

Alrawais, A., Alhothaily, A., Hu, C., & Cheng, X. (2017). Fog computing for the Internet of Things: Security and privacy issues. IEEE Internet Computing, 21(2), 34-42. https://doi.org/10.1109/MIC.2017.37

Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys, 51(4), 1-35. https://doi.org/10.1145/3214303

Lindell, Y. (2020). Secure multiparty computation. Communications of the ACM, 64(1), 86-96. https://doi.org/10.1145/3387108

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2018). SoK: Security and privacy in machine learning. IEEE European Symposium on Security and Privacy, 399-414. https://doi.org/10.1109/EuroSP.2018.00035

Carlini, N., Liu, C., Erlingsson, U., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. USENIX Security Symposium. https://doi.org/10.48550/arXiv.1802.08232

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning. IEEE Symposium on Security and Privacy, 1-15. https://doi.org/10.1109/SP.2019.00065

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information. Proceedings of ACM CCS, 1322-1333. https://doi.org/10.1145/2810103.2813677

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. Proceedings of ICML. https://doi.org/10.48550/arXiv.1206.6389

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. Proceedings of ICLR. https://doi.org/10.48550/arXiv.1412.6572

Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). Intriguing properties of neural networks. Proceedings of ICLR. https://doi.org/10.48550/arXiv.1312.6199

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. Proceedings of ICLR. https://doi.org/10.48550/arXiv.1706.06083

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of ICML, 233-240. https://doi.org/10.1145/1143844.1143874

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://doi.org/10.48550/arXiv.1201.0490

Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1912.01703

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of ICLR. https://doi.org/10.48550/arXiv.1412.6980

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of ICML. https://doi.org/10.48550/arXiv.1502.03167

Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929-1958. https://doi.org/10.5555/2627435.2670313

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of ACM KDD, 1135-1144. https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1705.07874

Samek, W., Wiegand, T., & Muller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal, 1, 39-48. https://doi.org/10.48550/arXiv.1708.08296

Gilpin, L. H., Bau, D., Yuan, B. Z., et al. (2018). Explaining explanations: An overview of interpretability of machine learning. IEEE DSAA, 80-89. https://doi.org/10.1109/DSAA.2018.00018

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5, 221-232. https://doi.org/10.1007/s13748-016-0094-0