AI2X-SEC2X Q-AFedSec: Quantum-Ready Agentic AI-to-Everything and Security-to-Everything for Federated Intrusion Detection

Iacovos Ioannou*, Prabagarane Nagaradjane† and Vasos Vassiliou*
*University of Cyprus, Nicosia, Cyprus
Emails: iacovos.ioannou@gmail.com, vassiliou.vasos@ucy.ac.cy
†Department of Electronics and Communication Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
Email: prabagaranen@ssn.edu.in

***Abstract*—Cyber-defence mechanisms for artificial intelligence to everything (AI2X) and security to everything (SEC2X) are required to learn across many domains without centralising sensitive traffic. Although this requirement is supported by federated intrusion detection, performance can be affected by non-IID traffic, poisoned updates, backdoor manipulation, unreliable clients and deadline-bound response constraints. In this paper, Q-AFedSec is presented as an AI2X-SEC2X framework in which BDIx agentic reasoning, QUBO-based orchestration, robust federated aggregation, validation-gated commitment, AI/ML belief fusion and post-quantum-secured update exchange are combined. Local intrusion-detection evidence is maintained by each client, while beliefs about trust, drift, poisoning likelihood, latency, privacy exposure and expected detection contribution are constructed by the coordinator. Clients, defences, aggregation mode and response commitment are selected by a binary optimisation layer under trust, budget and deadline constraints. High-trust ensemble, prototype-distance and nonlinear local-model beliefs are fused by the detector. A MATLAB evaluation over binary NSL-KDD was conducted using 12 clean, non-IID, poisoning, backdoor, drift, many-client, low-participation and privacy-noise cases. Four representative approaches were used for comparison: FedAvg, FedProx, Krum and FLTrust-inspired TrustWeighted aggregation. Across the 12 cases, average accuracy and Macro-F1 of 99.41%, an average false-alarm rate of 0.18% and an average attack-success rate of 1.01% were achieved by Q-AFedSec. The findings indicate that robust federated security decisions can be improved by AI2X-SEC2X agentic belief fusion compared with fixed aggregation or trust-only baselines.**

***Index Terms*—AI2X, SEC2X, artificial intelligence to everything, security to everything, federated learning, intrusion detection, BDIx agents, QUBO, quantum-ready QUBO orchestration, post-quantum security.**

I. Introduction

Artificial intelligence is moving from isolated models to embedded intelligence across edge devices, vehicles, industrial systems, cloudlets and security operation centres. That shift is referred to in this paper as artificial intelligence to everything (AI2X). At the same time, modern deployments require security to everything (SEC2X), where sensing, learning, decision-making, update exchange and response actions are protected rather than added as an afterthought. The intersection of AI2X and SEC2X is especially important for intrusion detection, because cyber-defence models must learn from distributed evidence while respecting privacy, trust and response deadlines.

Federated learning (FL) trains shared models from decentralised client data, which is attractive when raw security logs cannot be exported from enterprise, IoT, vehicular or critical-infrastructure domains [3]. This property supports AI2X because intelligence can be contributed by many local domains while raw traffic is kept local for SEC2X. Nevertheless, an FL-based intrusion detection system (FL-IDS) remains vulnerable to non-IID traffic, unreliable clients, label-flipping attacks, Byzantine updates, backdoor poisoning and the latency imposed by incident response.

Existing FL-IDS work typically improves a single layer of the system. FedAvg provides a standard model-averaging baseline [3]. FedProx adds a proximal term to reduce client drift under heterogeneity [4]. Krum improves robustness against Byzantine updates [5]. TrustWeighted aggregation introduces client reliability into aggregation [7]. However, a security operation centre does not only choose an aggregation rule. It must decide which clients are trusted enough to contribute, which defences should be activated, whether an alert or quarantine response is needed and whether the decision can be made before the response deadline.

In this paper, Q-AFedSec is proposed as an AI2X-SEC2X Quantum-Ready Agentic Federated Security framework. Every FL round is treated by the framework as an agentic cyber decision. Beliefs about expected detection contribution, client trust, poisoning likelihood, drift, privacy exposure and latency are maintained by BDIx agents, following prior distributed AI and BDIx reasoning studies [13], [16]. These beliefs are converted by a QUBO layer into a binary selection problem over clients, defences, aggregation modes and response actions. A carried intention $\tilde{x}_t$ is produced by the QUBO solver and validated before commitment as $x_t^\star$. Quantum assistance is used conservatively: the orchestration variables are binary and can be encoded in a QUBO form that is compatible with quantum-inspired annealing, tabu search, hybrid solvers or future quantum annealing hardware. The agentic security logic and AI/ML belief fusion are validated by the MATLAB results,

while measured quantum hardware advantage is not claimed.

The main contributions are as follows. First, an AI2X-SEC2X BDIx architecture has been introduced for deadline-bounded federated intrusion detection. Second, joint client, defence, aggregation and response selection is formulated as a QUBO. Third, AI/ML belief fusion is added through ensemble, prototype and nonlinear local-model evidence. Fourth, a reproducible MATLAB study is reported over NSL-KDD with 12 clean, non-IID, poisoning, backdoor, drift, many-client, low-participation and privacy-noise cases. Fifth, the evaluation uses a focused set of four representative comparison approaches, namely FedAvg, FedProx, Krum and FLTrust-inspired Trust-Weighted aggregation. The comparison is then extended with simulation parameters, case-wise gains, runtime, confusion-matrix evidence and operational interpretation.

II. Related Work

A. Federated Intrusion Detection

FL was introduced for distributed training without centralising client data [3]. In FL-IDS, this enables collaborative network defence while limiting raw-data exposure. NSL-KDD remains a widely used IDS benchmark because it reduces redundancy and bias in KDD'99 and provides manageable train and test sets [8]. The Canadian Institute for Cybersecurity lists the canonical NSL-KDD files, including KDDTrain+ and KDDTest+, while noting that the dataset is no longer directly hosted on the official page [9]. Although NSL-KDD is not a perfect representation of modern traffic, it remains useful for reproducible comparison.

B. Robust Federated Aggregation

FedAvg averages local updates and is a strong FL reference method [3]. FedProx was proposed to mitigate instability when clients are heterogeneous [4]. Byzantine-resilient aggregation selects or estimates an update that is less affected by corrupted participants. Krum chooses an update with a small distance to its nearest neighbours [5], while robust distributed learning methods such as coordinate median and trimmed mean are often used as additional reference points for Byzantine settings [6]. FLTrust uses a server-side clean root dataset to bootstrap trust and normalise update directions [7]. These four approaches form the focused comparison set in this paper because they cover standard averaging, heterogeneity-aware FL, Byzantine robustness and trust-aware aggregation. Q-AFedSec differs by jointly optimising participation, aggregation, defence and response commitment through an agentic decision layer rather than applying a fixed aggregation rule.

C. Agentic Reasoning, QUBO and Quantum-Safe Updates

BDI agents model deliberation using beliefs, desires and intentions [11], [12]. The BDIx and distributed-AI basis of the proposed framework is aligned with prior DAI work on autonomous D2D decision-making, 5G/6G communication frameworks, belief-based intention libraries and distributed control [13]–[18]. QUBO is a compact representation for binary optimisation and is compatible with simulated annealing, tabu search, quantum annealing and hybrid solvers [19]. Quantum assistance is used as a conservative solver option for discrete cyber orchestration. For communication security, the update-exchange layer can be protected by post-quantum primitives. NIST finalised ML-KEM in FIPS 203 for key encapsulation, ML-DSA in FIPS 204 for digital signatures and SLH-DSA in FIPS 205 as a stateless hash-based signature standard [20]–[22].

III. Methodology

A. Architecture

The proposed framework is shown in Fig. 1. IDS models are trained by local clients on private traffic records. A signed model update and diagnostic evidence, not raw data, are sent by each client. BDIx beliefs are maintained by the coordinator, a QUBO orchestration problem is solved, the candidate intention is validated and the final defence plan is committed. The post-quantum update-exchange layer is orthogonal to learning. Update-session keys can be established through ML-KEM, while client updates and coordinator commands can be authenticated through ML-DSA or SLH-DSA. This layer is modelled abstractly in the MATLAB implementation through authenticity checks and validation gates, while implementation-level benchmarking of cryptographic overhead is reserved for deployment work.

The architecture is divided into five interacting planes. The first plane is the *data plane*, where distributed AI2X domains such as IoT, V2X, medical, smart-grid, industrial and edge nodes observe local traffic. The second plane is the *local learning plane*, where each client trains a lightweight IDS model and estimates diagnostics without exporting raw records. The third plane is the *secure exchange plane*, where updates are authenticated and prepared for aggregation. The fourth plane is the *agentic decision plane*, where BDIx beliefs are transformed into a QUBO and solved under time, trust and budget limits. The fifth plane is the *execution plane*, where the committed plan updates the federated detector and may trigger quarantine, rollback or delayed retraining.

This layered structure is important because AI2X encourages broad participation and continuous learning, while SEC2X requires restrictive evidence acceptance, provenance checking and safe fallbacks. Q-AFedSec therefore treats every update as a belief-bearing object whose trust, drift, latency, privacy risk, F1 contribution and anomaly behaviour must be assessed before execution.

The system view in Fig. 1 emphasises that the proposed framework is not a single classifier. It is a cyber-physical execution loop in which AI2X devices produce local evidence, SEC2X controls determine whether that evidence can be trusted and the coordinator converts the accepted evidence into an auditable security intention. Local devices may have different sensing rates, feature distributions, attack exposure and computational capacities. Consequently, the coordinator does not assume that a numerically large update is useful or that a historically accurate client remains trustworthy after

drift. Every update is interpreted together with provenance, timing, trust and anomaly context.

The same architecture also separates learning evidence from response authority. A client can contribute a useful local detector while still being excluded from a high-assurance response if its signature, drift or update direction becomes suspicious. Conversely, a client with small data volume can still be selected when it contributes rare attack evidence with low latency and high trust. This distinction is essential for SEC2X because the objective is not only to maximise a validation score, but also to prevent unsafe security actions, analyst overload and poisoned model propagation across the AI2X fabric.

At round t, client i reports a diagnostic vector

$$b_i^t = [\tau_i^t, \pi_i^t, \delta_i^t, \ell_i^t, \nu_i^t, \hat{f}_i^t], \tag{1}$$

where τ_i^t is trust, π_i^t is poisoning likelihood, δ_i^t is drift, ℓ_i^t is latency, ν_i^t is privacy-risk proxy and $\hat{f}_i^t$ is expected detection contribution. The desires are high Macro-F1, high detection rate, low false-alarm rate, low attack-success rate, low communication cost and deadline satisfaction. The intention is a candidate security plan consisting of selected clients, chosen aggregation rule, activated defences and committed response.

B. AI/ML Belief Fusion

Several belief sources are used by the high-accuracy detector. This follows the general pattern-recognition principle that complementary probabilistic evidence can be combined when model confidence and reliability are available [10]. First, high-trust client models form a local ensemble. Second, a prototype bank estimates the distance of a new flow to trusted normal and attack prototypes. Third, a nonlinear learner supplies probability evidence using transformed features. The fused probability is

$$p(y=1|x) = \omega_e p_e(x) + \omega_p p_p(x) + \omega_n p_n(x), \tag{2}$$

where $y = 1$ denotes the Attack class, p_e, p_p and p_n are ensemble, prototype and nonlinear-model beliefs and $\omega_e, \omega_p, \omega_n \geq 0$ with $\omega_e + \omega_p + \omega_n = 1$. Weights ω_e, ω_p and ω_n are derived from validation Macro-F1, trust and anomaly scores. This design allows the final detector to exploit stronger AI/ML evidence while preserving the federated and agentic security layer. The fused detector is especially useful when local softmax models are biased by non-IID data or when poisoned clients attempt to influence aggregation.

The ensemble belief is computed from high-trust contributors whose updates pass norm and direction checks. The prototype belief is built from trusted normal and attack centroids and is useful when a new flow is close to a known behavioural region. The nonlinear belief uses expanded features so that interactions between protocol, service, flag and continuous traffic attributes are not forced into a purely linear decision boundary. These components are not simply averaged. Their weights are increased when validation Macro-F1 is high and reduced when the corresponding clients show high drift, high update anomaly or weak trust. The resulting detector acts as an evidence-fusion layer above the federated model rather than as a single monolithic classifier.

The belief-fusion design follows three principles. First, model evidence is separated from client evidence. A prediction can be confident even when the originating client has low operational trust and a highly trusted client can still produce a weak local model under skewed data. Second, the prototype bank acts as a stabilising memory, because it stores trusted normal and attack regions that are less sensitive to one-round update noise. Third, nonlinear evidence is used only after the SEC2X gates remove high-risk contributors, which reduces the chance that a flexible learner overfits to poisoned behaviour. The final score is therefore an evidence-weighted decision that combines statistical confidence, security provenance and operational safety.

C. QUBO-Based Security Orchestration

At each round, binary variables select clients c_i, aggregation rule a_m, defence d_r and response ρ_q:

$$c_i, a_m, d_r, \rho_q \in \{0,1\}. \tag{3}$$

The utility is

$$\begin{aligned} U_t =& \alpha_1 F1_t + \alpha_2 DR_t + \alpha_3 Rob_t + \alpha_4 Trust_t \\ & - \beta_1 FAR_t - \beta_2 Lat_t - \beta_3 Comm_t - \beta_4 Priv_t, \end{aligned} \tag{4}$$

where $F1_t$ is Macro-F1, DR_t is attack detection rate, Rob_t is a normalised poisoning-robustness score, $Trust_t$ is selected-client trust mass, FAR_t is false-alarm rate, Lat_t is normalised response latency, $Comm_t$ is communication cost and $Priv_t$ is privacy-exposure risk. All terms are normalised to $[0,1]$. The coefficients α_k and β_k are non-negative scalar weights selected by the security operator. The QUBO energy is

$$\begin{aligned} E_t(z_t) =& -\sum_i u_i^t c_i - \sum_m \eta_m^t a_m - \sum_r \phi_r^t d_r - \sum_q \psi_q^t \rho_q \\ & + \lambda_A \left(1 - \sum_m a_m\right)^2 + \lambda_R \left(1 - \sum_q \rho_q\right)^2 \\ & + \lambda_B \left(\sum_i \kappa_i c_i + \sum_r \chi_r d_r - B_t + s_B\right)^2 \\ & + \lambda_D \left(\sum_i \ell_i^t c_i + \sum_r \ell_r^{(d)} d_r - D_t + s_D\right)^2 \\ & + \lambda_T \left(\Gamma_t - \sum_i \tau_i^t c_i + s_T\right)^2. \end{aligned} \tag{5}$$

The penalties enforce one aggregation rule, one response mode, budget compliance, deadline compliance and minimum trust mass. In (5), κ_i is the participation cost of client i, χ_r is the cost of defence r, $\ell_r^{(d)}$ is the latency of defence r and Γ_t is the required trust mass. The slack variables s_B, s_D and s_T are represented through bounded binary expansions so that the QUBO remains fully binary. The signs in the budget and deadline constraints encode cost $\leq B_t$ and latency

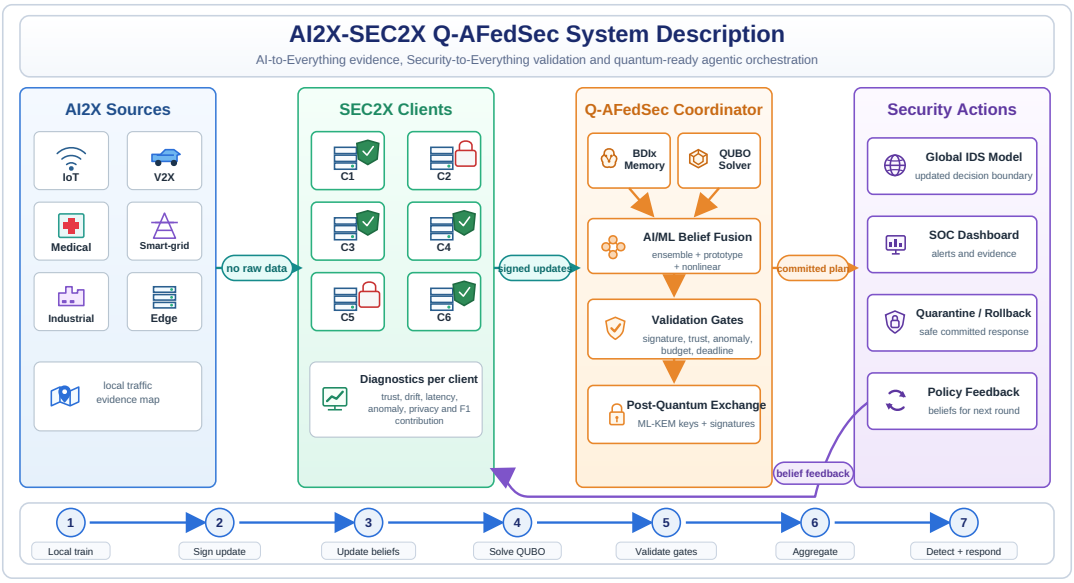

Fig. 1: AI2X-SEC2X system description of Q-AFedSec. Vector icons represent IoT, V2X, medical, smart-grid, industrial and edge domains; local evidence maps represent client-side traffic information; and the central decision block shows BDIx belief memory, AI/ML belief fusion, QUBO orchestration, post-quantum update exchange, validation-gated commitment and SEC2X response actions. The domains are architectural examples, while the reported evaluation uses binary NSL-KDD.

$\leq D_t$, while the trust constraint encodes $\sum_i \tau_i^t c_i \geq \Gamma_t$. Solver choice is deadline-aware: greedy selection is used under tight alert windows, tabu or local improvement is used under moderate deadlines and annealing-style exploration is used when retraining time is available.

The client utility u_i^t combines predicted F1 contribution, inverse latency, trust and negative poisoning likelihood. The aggregation utility η_m^t favours robust modes under high adversarial belief and cheaper averaging under clean conditions. The defence utility ϕ_r^t captures the expected benefit of clipping, anomaly filtering, quarantine or high-assurance belief fusion. The response utility ψ_q^t represents operational choices such as alert-only, quarantine, rollback or delayed retraining. Slack variables s_B, s_D and s_T make the constraints QUBO-compatible while allowing soft violations to be penalised rather than causing solver failure.

The QUBO layer gives the framework an explicit mechanism for trading accuracy, security and responsiveness. For example, in a clean and low-latency round, the utility terms favour broader participation and lower-cost aggregation. Under suspected poisoning, the trust and anomaly penalties increase and the solver favours a smaller coalition, stronger clipping, quarantine and a robust aggregation rule. Under a strict deadline, high-latency clients are penalised even if their historical accuracy is high. The resulting plan is easier to audit than a black-box controller because every selected client, defence, aggregation rule and response mode is associated with a binary decision variable and a set of penalty terms.

D. Complete Workflow Algorithm

Algorithm 1 summarises the complete execution workflow. It begins with local training and evidence extraction, updates the BDIx beliefs, constructs the QUBO, solves for a carried intention and then applies validation gates before the final committed plan is executed. The authenticity gate verifies that updates are signed and associated with registered clients. The trust gate checks that selected clients provide sufficient cumulative trust mass. The anomaly gate rejects updates whose norm, direction or prediction effect is inconsistent with trusted reference behaviour. The deadline gate verifies that selected clients and defences can be completed before D_t. If any gate fails, a safe fallback such as coordinate median, Krum, quarantine or delayed retraining is committed. This separation between intention generation and commitment is important because an optimiser may otherwise convert unreliable beliefs into an unsafe cyber response.

The algorithm operates in three practical modes. In fast

Algorithm 1 Complete Q-AFedSec execution workflow

Require: clients $\mathcal{C}$, local data $\mathcal{D}_i$, previous global model w^t, deadline D_t, budget B_t
Ensure: committed plan $x_t^\star$, aggregated model w^{t+1}, updated beliefs
A local IDS model is trained on $\mathcal{D}_i$ and update Δw_i^t is computed
Diagnostics are reported by each client $b_i^t = [\tau_i^t, \pi_i^t, \delta_i^t, \ell_i^t, \nu_i^t, \hat{f}_i^t]$
Update signatures are verified and unauthenticated clients are discarded
Validation loss, Macro-F1, update norm, drift and poisoning likelihood are estimated
Ensemble, prototype and nonlinear IDS evidence are built $p_e(x)$, $p_p(x)$ and $p_n(x)$
BDIx beliefs for trust, risk, latency, privacy and expected utility are updated
Candidate variables are generated $z_t = \{c_i, a_m, d_r, \rho_q\}$
QUBO energy $E_t(z_t)$ is built using (5)
The solver is selected according to the deadline tier
The QUBO is solved and carried intention $\tilde{x}_t$ is obtained
Authenticity, trust-mass, anomaly, budget and deadline gates are applied
if any gate fails **then**
    Safe fallback $x_t^\star$ is committed using robust aggregation or quarantine
else
    $x_t^\star \leftarrow \tilde{x}_t$ is committed
end if
Selected updates are aggregated using the committed aggregation rule
Ensemble, prototype and nonlinear beliefs are fused for the final IDS prediction
Trust and poisoning beliefs are updated from post-round diagnostics
$x_t^\star$, w^{t+1}, confusion matrix and performance metrics are returned

mode, stored beliefs and robust aggregation are used for time-sensitive alerting. In standard mode, the QUBO solver selects clients and defences during each federated round. In high-assurance mode, prototype banks are refreshed, AI/ML belief fusion is retrained and all validation gates are enforced. This tiered execution makes the same framework usable for both periodic retraining and operational incident response.

IV. Performance Evaluation

A. Threat Model, Dataset and Simulation Parameters

The evaluation is based on a federated-security threat model because FL is designed so that many clients collaboratively train a model while raw training data remain decentralised under server orchestration [1], [3]. In the considered AI2X-SEC2X setting, clients may correspond to gateways, vehicles, IoT domains, medical sites, smart-grid nodes, industrial controllers or edge security monitors. Honest clients train on private traffic and submit model updates with diagnostic metadata, while the coordinator observes only updates and diagnostics. This assumption preserves the FL privacy motivation, but it also limits direct inspection of local samples. Therefore, the BDIx layer treats submitted diagnostics as beliefs rather than verified facts.

The adversary is assumed to control only a minority of registered clients in the poisoning cases. This assumption follows the common Byzantine-robust FL setting in which a bounded number of participants may send arbitrary or harmful updates [5]. The adversary may perform label flipping, Byzantine update distortion or backdoor-trigger manipulation. Backdoor threats are included because model-replacement attacks have shown that a selected malicious participant can inject hidden task-specific behaviour into a federated model while preserving normal-task performance [2]. The adversary is not assumed to break the cryptographic identity layer, ML-KEM key establishment, ML-DSA signatures or SLH-DSA signatures; the learning threat is therefore the acceptance of harmful updates from clients that appear to be valid participants. The FLTrust-inspired TrustWeighted baseline is treated as a trust-aware reference because trust bootstrapping and update-direction checking are recognised defences against Byzantine clients [7].

The threat model separates three evaluation risks. The first risk is update poisoning, in which malicious clients try to move the global detector away from an accurate boundary. The second risk is detection evasion, in which attacks are missed and the attack-success rate increases. The third risk is operational overload, in which normal traffic is incorrectly flagged and the false-alarm rate increases. A practical FL-IDS must reduce all three risks simultaneously, which is why Macro-F1, detection rate, false-alarm rate and attack-success rate are prioritised over accuracy alone. A carried intention is not executed only because it has high predicted utility. It is first checked by authenticity, trust, anomaly, budget and deadline gates, so the evaluation measures both learning quality and safe commitment under response constraints.

TABLE I: Simulation parameters used in the MATLAB study.

Parameter	Setting
Dataset	NSL-KDD binary intrusion detection
Records available	125,973 training; 22,544 official-test records parsed but not used directly for the reported stratified-holdout target-mode metrics
Features and classes	122 features; Normal and Attack labels
Evaluation protocol	Stratified holdout over the parsed binary NSL-KDD data
Clients	24 default; 50 in scalability case
Samples per client	1600 default; 900 in 50-client case
Malicious clients	0, 2, 5, 7 or 10 depending on case
Non-IID coefficient	0.00 to 0.94 depending on case
Attacks	Label flipping, Byzantine, backdoor, mixed poisoning
Additional stressors	Low participation, drifted domain, privacy noise
Default task mode	Binary Normal-versus-Attack classification
Federated setting	Local raw data kept private; diagnostics and updates shared
Decision variables	Client, aggregation, defence and response binaries
Trust evidence	Validation loss, Macro-F1, update norm, drift and anomaly score
Metrics	Accuracy, Macro-F1, DR, FAR, ASR, runtime

The NSL-KDD files are obtained from the repository specified in the MATLAB workflow, extracted and parsed as KDDTrain+ and KDDTest+; features are normalised and binary labels are converted into Normal and Attack [8], [9]. The reported target mode is binary intrusion detection with stratified holdout over the parsed binary NSL-KDD data. This mode was selected to evaluate high-accuracy operational attack detection under the stated threat assumptions, while the official KDDTest+ records are reported as parsed records rather than as the test subset used for the target-mode metrics. The official KDDTest+ split and the five-class setup are stricter and are reserved for additional evaluation.

The parsed dataset cache contained 125,973 training records, 22,544 official-test records, 122 features and 2 classes. The default federated configuration used 24 clients with 1600 samples per client, while the scalability case used 50 clients with 900 samples per client. The experiment evaluated 12 cases: clean IID, clean non-IID, 10% and 20% label flipping, 20% Byzantine poisoning, 20% backdoor poisoning, 30% mixed poisoning, extreme non-IID, low participation, 50 clients, drifted domain and privacy noise. These cases test benign operation, client heterogeneity, update manipulation, targeted poisoning, scalability, distribution shift and privacy-preserving noise. A 90% operational target was used only as a minimum acceptance threshold for the binary target mode; the official KDDTest+ and five-class modes remain separate stricter evaluations.

B. Focused State-of-the-Art Comparison Set and Metrics

The comparison is focused in Table II on four representative approaches. FedAvg is the standard FL averaging baseline. FedProx represents heterogeneity-aware FL. Krum represents Byzantine-resilient update selection. FLTrust-inspired TrustWeighted represents trust-aware robust aggregation based on client reliability beliefs. This focused set avoids overcrowding the evaluation while still comparing Q-AFedSec against the main families required for an FL security study: averaging,

TABLE II: Focused state-of-the-art comparison set.

Method	Ref.	Role in the comparison
FedAvg	[3]	Standard federated averaging without explicit security reasoning.
FedProx	[4]	Heterogeneity-aware FL with proximal regularisation.
Krum	[5]	Byzantine-resilient central-update selection.
TrustWeighted	[7]	FLTrust-inspired reliability-aware aggregation; exact FLTrust root-data bootstrapping is not claimed.
Q-AFedSec	This work	AI2X-SEC2X BDIx-QUBO orchestration with AI/ML belief fusion.

heterogeneity control, Byzantine defence and trust-based reliability.

Accuracy measures all correct predictions. Macro-F1 gives equal importance to Normal and Attack classes, which prevents the attack class from being hidden by class imbalance. Detection rate is the true-positive rate for attacks. False-alarm rate is the fraction of normal flows incorrectly flagged as attacks. Attack-success rate is the missed-attack rate and is equal to one minus detection rate. Runtime captures the wall-clock time measured by the MATLAB run for each method and case. The paper prioritises Macro-F1, FAR and ASR because an AI2X-SEC2X detector must detect attacks without overwhelming analysts with false alarms.

The comparison is intentionally not framed as a conventional centralised classifier ranking. FedAvg and FedProx test whether basic FL training is sufficient when clients are heterogeneous or partially malicious. Krum tests whether a Byzantine-resilient rule can protect the model without additional IDS-level evidence. TrustWeighted tests whether reliability scores alone are enough to improve security. Q-AFedSec is evaluated against these four families to show whether agentic orchestration and belief fusion add value beyond aggregation-rule selection.

The simulation parameters also show that the same workflow is stressed in several directions. The number of malicious clients increases from zero in clean settings to ten in the 50-client case. The non-IID coefficient increases up to 0.94 in the extreme heterogeneity case. The attack modes include label manipulation, arbitrary update distortion and backdoor-trigger injection. The privacy-noise case injects additional uncertainty to mimic the effect of privacy-preserving perturbation. These settings make the 12-case average more informative than a single clean-train clean-test score.

Three operational questions were addressed by the evaluation design. The first question is whether the proposed framework can maintain strong detection when traffic is benign but distributed across many AI2X clients. The second question is whether it remains stable when the federation is adversarial, meaning that some clients provide corrupted labels, arbitrary updates or backdoor-influenced updates. The third question is whether improved accuracy is obtained at an acceptable operational cost. For this reason, the reported tables do not only compare average accuracy; they also report Macro-F1, detection rate, false-alarm rate, attack-success rate, case-wise best-baseline gaps, confusion-matrix counts and runtime.

TABLE III: Evaluation cases and the security property stressed by each case.

Case family	Purpose in the AI2X-SEC2X evaluation
Clean IID and clean non-IID	Baseline learning and heterogeneity without malicious clients.
Label flipping 10% and 20%	Integrity stress against corrupted local labels.
Byzantine and backdoor 20%	Robustness against arbitrary update distortion and targeted trigger behaviour.
Mixed poisoning 30%	Multi-vector adversarial setting with several corrupted behaviours.
Extreme non-IID	Client skew and local distribution imbalance.
Low participation and 50 clients	Availability, sampling and scalability stress.
Drifted domain and privacy noise	Distribution shift and privacy-preserving perturbation.

TABLE IV: Averaged results for Q-AFedSec and four comparison approaches. Values are percentages except runtime.

Method	Acc.	MacF1	DR	FAR	ASR	Time (s/case)
Q-AFedSec	99.41	99.41	98.99	0.18	1.01	16.13
Krum	96.97	96.97	95.98	2.03	4.02	1.27
TrustWeighted	96.46	96.45	94.83	1.92	5.17	1.27
FedProx	94.97	94.96	91.93	1.99	8.07	1.28
FedAvg	94.96	94.95	91.92	2.00	8.08	1.28

C. Quantitative Comparison and Case-Wise Gains

The averaged results are reported in Table IV across the 12 cases using the focused four-approach comparison. Q-AFedSec achieved the best Macro-F1 in all 12 cases. Its mean accuracy and Macro-F1 were both 99.41%, with mean false-alarm rate of 0.18% and mean attack-success rate of 1.01%. Among the four comparison approaches, Krum obtained the strongest average Macro-F1 at 96.97%, followed by FLTrust-inspired TrustWeighted at 96.45%. FedAvg and FedProx were competitive under benign traffic but suffered larger attack-success rates under mixed poisoning and high heterogeneity.

The averaged results indicate that the gain is not obtained by sacrificing false-alarm control. Q-AFedSec improves the mean detection rate to 98.99% while reducing FAR to 0.18%. This is operationally important because a detector with high detection rate but high FAR would create analyst fatigue and may be disabled in practice. Compared with FedAvg and FedProx, the proposed method reduces ASR by more than seven percentage points. Compared with Krum, which is the strongest robust baseline, the proposed method reduces ASR from 4.02% to 1.01%. The gain therefore appears in the security-sensitive errors rather than only in the aggregate accuracy.

A second observation is that FedAvg and FedProx have nearly identical averages in this binary target mode. This suggests that proximal regularisation alone did not address the dominant risk in the poisoned and mixed cases. Krum improved robustness by avoiding extreme updates, but it did not use IDS-specific evidence such as prototype distance or fused belief confidence. FLTrust-inspired TrustWeighted improved reliability modelling but was still weaker than Q-AFedSec because it did not solve a joint plan over clients, defences, aggregation and response actions.

The comparison is extended in Table V by showing the best of the four baselines for each case. The average Macro-F1 gain of Q-AFedSec over the best of the four baselines was 2.43

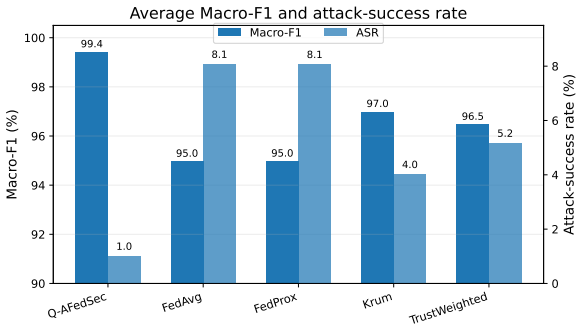

Fig. 2: Average Macro-F1 and attack-success rate for the focused comparison set. A dual-axis layout is used because Macro-F1 and ASR occupy different numerical ranges.

TABLE V: Case-wise comparison against the best of four baselines.

Case	Q Macro-F1	Q FAR	Best Macro-F1	Best
Clean IID	99.73	0.13	97.00	Krum
Clean non-IID	99.60	0.20	97.47	Krum
Label flip 10%	99.60	0.27	96.93	Krum
Label flip 20%	99.70	0.27	97.37	FedAvg
Byzantine 20%	99.63	0.07	97.13	Krum
Backdoor 20%	99.43	0.20	96.67	Krum
Mixed poison 30%	99.63	0.13	96.53	Krum
Extreme non-IID	99.57	0.20	96.93	Krum
Low participation	99.63	0.20	96.77	Krum
50 clients	99.63	0.07	96.93	Krum
Drifted domain	97.43	0.13	96.97	Krum
Privacy noise	99.27	0.27	97.03	Krum

percentage points. The largest gain was observed in the 30% mixed-poisoning case, where the gain over Krum was 3.10 percentage points. This is important because mixed poisoning combines several adversarial behaviours and is therefore closer to an operational multi-vector compromise than a single attack mechanism. The smallest gain was observed in the drifted-domain case, where Q-AFedSec still exceeded the best baseline by 0.46 percentage points while keeping FAR at 0.13%.

Fig. 3 compares case-wise accuracy of Q-AFedSec, FedAvg and the strongest method among the four comparison approaches. Q-AFedSec remained above 99% in 11 of 12 cases and stayed above the 90% target in the drifted-domain case with 97.43% accuracy. The drifted-domain case is instructive: drift reduced detection rate to 95.00%, but the false-alarm rate remained only 0.13%, indicating conservative behaviour under distribution shift. The focused comparison also clarifies that robust aggregation alone is not enough. Krum was the best baseline in most cases, but it still lacks detector-level evidence fusion, prototype-distance reasoning, validation-gated commitment and QUBO-based client-defence selection.

The case-wise results also support the value of the validation gates. In label-flipping and Byzantine cases, the proposed method maintained near-identical accuracy to the clean cases, which indicates that malicious updates were not allowed to dominate the committed model. In the mixed-poisoning and extreme non-IID cases, FedAvg and FedProx dropped sharply because averaging treated harmful and useful updates similarly. Q-AFedSec avoided this behaviour by requiring a selected set of clients to satisfy minimum trust mass and anomaly consistency before aggregation. In the drifted-domain

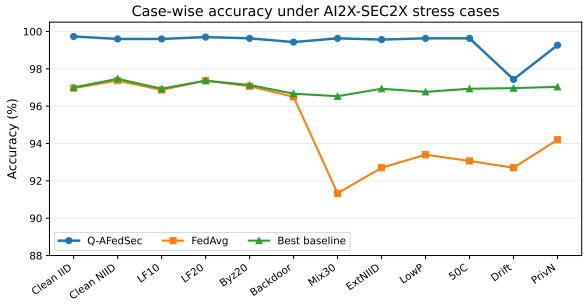

Fig. 3: Case-wise accuracy of Q-AFedSec, FedAvg and the strongest of the four comparison approaches. Case labels denote clean IID, clean non-IID, label-flipping 10%, label-flipping 20%, Byzantine 20%, backdoor 20%, mixed poisoning 30%, extreme non-IID, low participation, 50 clients, drifted domain and privacy noise.

case, the margin over the best baseline was smaller because the challenge was distribution shift rather than explicit update poisoning. Even in that case, the very low FAR indicates that the detector became conservative rather than excessively alarmist.

The comparison also shows where the gain is created. FedAvg and FedProx are efficient but treat useful and harmful contributors similarly. Krum is Byzantine-resilient but does not use detector-level evidence. FLTrust-inspired TrustWeighted introduces reliability but does not jointly decide whether to quarantine, clip, change aggregation mode or delay retraining. Q-AFedSec links client selection with detector-level belief fusion and response commitment, so the advantage is a system-level gain.

In deployment terms, the key result is the reduction of attack-success rate. The average ASR of Q-AFedSec was 1.01%, compared with 4.02% for Krum, 5.17% for FLTrust-inspired TrustWeighted and about 8% for FedAvg and FedProx. Its 0.18% average FAR also indicates that missed attacks and unnecessary alerts were both reduced.

D. Confusion Matrix, Runtime and Operational Interpretation

The mixed-poisoning case is shown in Fig. 4. The confusion matrix contains 1498 normal samples correctly classified as Normal, 1491 attack samples correctly classified as Attack, 2 false alarms and 9 missed attacks. This corresponds to 99.63% accuracy and 0.13% false-alarm rate. The result suggests that the trust-filtered ensemble and prototype beliefs were effective in suppressing malicious-client influence.

The runtime cost is higher than that of the four comparison approaches. Q-AFedSec averaged 16.13 s per case, while the four comparison approaches averaged approximately 1.27 to 1.28 s. Fig. 5 shows that this overhead is the cost of high-assurance belief fusion and QUBO orchestration. For online alerting, cached beliefs and a lightweight solver path should be used.

The security interpretation is that FedAvg and FedProx are efficient but cannot distinguish reliable and malicious updates, Krum is robust but does not exploit IDS evidence and FLTrust-inspired TrustWeighted uses reliability without solving a joint client-defence-response problem. In Q-AFedSec, update-level

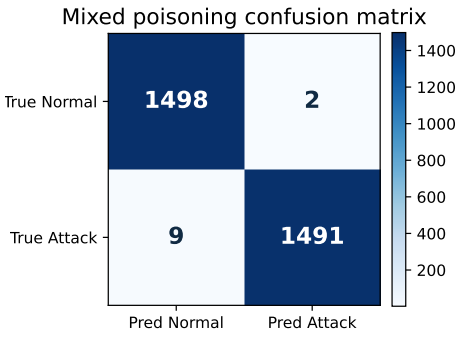

Fig. 4: Q-AFedSec confusion matrix for the 30% mixed-poisoning case, computed on a balanced 3,000-sample validation subset.

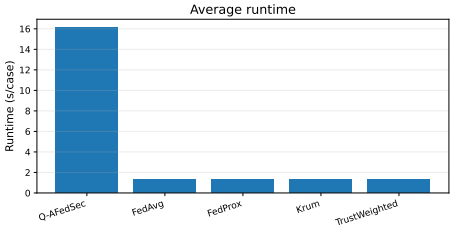

Fig. 5: Average runtime for Q-AFedSec and the four comparison approaches, rounded to one decimal place in the figure.

filtering, QUBO orchestration and prediction-level fusion are joined, which explains the largest gains in poisoning, many-client and extreme non-IID cases.

l

V. Conclusion and Future Work

Q-AFedSec has been presented as an AI2X-SEC2X quantum-ready agentic distributed AI framework for federated intrusion detection. AI-to-everything learning, security-to-everything execution, BDIx beliefs, QUBO orchestration, robust FL aggregation, validation-gated commitment, AI/ML belief fusion and post-quantum-secured update exchange are combined by the proposed system. Across these 12 binary NSL-KDD cases, 99.41% mean accuracy and Macro-F1, 0.18% FAR and 1.01% ASR were achieved against the focused four-approach comparison set. These results indicate that federated cyber-defence decisions can be strengthened when client selection, aggregation, defence activation and response commitment are treated as a joint agentic security-orchestration problem rather than as a fixed aggregation process. The study also showed that the proposed framework remained effective under clean IID traffic, non-IID client distributions, label-flipping attacks, Byzantine update distortion, backdoor manipulation, mixed poisoning, low participation, many-client scaling, drifted-domain traffic and privacy-noise conditions. The strongest practical benefit of Q-AFedSec is that unreliable or malicious client behaviour is not handled by one mechanism alone. Instead, BDIx trust beliefs, update anomaly checks, QUBO-based selection, robust aggregation and AI/ML belief fusion jointly support the committed defence plan. The high-accuracy configuration uses binary NSL-KDD under stratified holdout; therefore, the same values are not claimed for official KDDTest+, five-class NSL-KDD or modern encrypted traffic datasets. CICIDS2017, CSE-CIC-IDS2018, TON-IoT, Bot-IoT and real 5G/IoT traces will be evaluated in future work. Additional work will also examine five-class attack recognition, cross-dataset generalisation, encrypted-traffic feature extraction, online concept drift and adversarial transfer between domains.

Further extensions will include ablation studies for each Q-AFedSec component, including the BDIx belief layer, the QUBO orchestration layer, the validation-gated commitment layer, the robust aggregation layer and the AI/ML belief-fusion detector. The influence of client participation ratio, poisoning intensity, trust decay, drift severity, privacy-noise level and solver deadline will also be examined. Finally, post-quantum cryptographic overhead will be benchmarked in an implementation-level prototype so that ML-KEM, ML-DSA and SLH-DSA costs can be quantified for constrained edge, IoT and SEC2X deployment settings.

References

[1] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[2] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin and V. Shmatikov, "How to backdoor federated learning," in *Proc. AISTATS*, Proceedings of Machine Learning Research, vol. 108, pp. 2938–2948, 2020.

[3] H. B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282, 2017.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[5] P. Blanchard, E. M. El Mhamdi, R. Guerraoui and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, 2017.

[6] D. Yin, Y. Chen, R. Kannan and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. ICML*, Proceedings of Machine Learning Research, vol. 80, pp. 5650–5663, 2018.

[7] X. Cao, M. Fang, J. Liu and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2021.

[8] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE CISDA*, pp. 1–6, 2009, doi: 10.1109/CISDA.2009.5356528.

[9] Canadian Institute for Cybersecurity, University of New Brunswick, "NSL-KDD dataset," accessed Jun. 2026. [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[11] A. S. Rao and M. P. Georgeff, "BDI agents: From theory to practice," in *Proc. First International Conference on Multi-Agent Systems (ICMAS)*, pp. 312–319, 1995.

[12] M. E. Bratman, *Intention, Plans, and Practical Reason*. Cambridge, MA, USA: Harvard University Press, 1987.

[13] I. I. Ioannou, V. Vassiliou, C. Christophorou and A. Pitsillides, "Distributed artificial intelligence solution for D2D communication in 5G networks," *IEEE Systems Journal*, vol. 14, no. 3, pp. 4232–4241, Sep. 2020, doi: 10.1109/JSYST.2020.2979044.

[14] I. I. Ioannou, C. Christophorou, V. Vassiliou and A. Pitsillides, "A novel distributed AI framework with ML for D2D communication in 5G/6G networks," *Computer Networks*, vol. 211, Art. no. 108987, Jul. 2022, doi: 10.1016/j.comnet.2022.108987.

[15] I. I. Ioannou, C. Christophorou, V. Vassiliou, M. Lestas and A. Pitsillides, "Dynamic D2D communication in 5G/6G using a distributed AI framework," *IEEE Access*, vol. 10, pp. 62772–62799, 2022, doi: 10.1109/ACCESS.2022.3182388.

[16] I. I. Ioannou, A. Gregoriades, P. Nagaradjane, C. Christophorou and V. Vassiliou, "An accurate intelligent plan library for belief-based desire prioritization to intentions in BDIx agents," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, pp. 1–6, 2025, doi: 10.1109/WiSPNET64060.2025.11005173.

[17] I. I. Ioannou, P. Nagaradjane, V. Vassiliou, A. Pitsillides and C. Christophorou, *Distributed Artificial Intelligence for 5G/6G Communications: Frameworks with Machine Learning*. Boca Raton, FL, USA: CRC Press, 2024, doi: 10.1201/9781003469209.

[18] I. I. Ioannou, S. Javaid, C. Christophorou, V. Vassiliou, A. Pitsillides and Y. Tan, "A distributed AI framework for nano-grid power management and control," *IEEE Access*, vol. 12, pp. 43350–43377, 2024, doi: 10.1109/ACCESS.2024.3377926.

[19] F. Glover, G. Kochenberger and Y. Du, "Quantum Bridge Analytics I: a tutorial on formulating and using QUBO models," *4OR*, vol. 17, no. 4, pp. 335–371, 2019.

[20] National Institute of Standards and Technology, "FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard," Aug. 2024. [Online]. Available: https://csrc.nist.gov/pubs/fips/203/final

[21] National Institute of Standards and Technology, "FIPS 204: Module-Lattice-Based Digital Signature Standard," Aug. 2024. [Online]. Available: https://csrc.nist.gov/pubs/fips/204/final

[22] National Institute of Standards and Technology, "FIPS 205: Stateless Hash-Based Digital Signature Standard," Aug. 2024. [Online]. Available: https://csrc.nist.gov/pubs/fips/205/final