

TrustEdge-AI: Uncertainty-Aware Dynamic Trust Intelligence for Adaptive Cyber Defense in IoT Edge Networks

Mohamed Ayari^{1*}, Ahmed Alhomoud^{2*}, Atef Gharbi³, Yamen El Touati⁴, Zeineb Klai⁵

Department of Information Technology-Faculty of Computing and Information Technology,
Northern Border University, Saudi Arabia¹

Department of Computer Science, College of Science, Northern Border University, Arar 91431, Saudi Arabia²

Department of Information Systems-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia³

Department of Computer Sciences-Faculty of Computing and Information Technology,
Northern Border University, Saudi Arabia^{4,5}

Abstract—Conventional Internet-of-Things (IoT) intrusion detection typically maps a model score directly to an alert, although autonomous defense must also determine whether the evidence is reliable, whether suspicious behavior persists over time, and whether a disruptive response is justified. This study proposes TrustEdge-AI, an uncertainty-aware dynamic-trust framework for adaptive cyber defense at the IoT edge. For each monitored device, the framework converts anomaly and classification outputs into reliability-weighted evidence, fuses this evidence with temporal behavioral consistency and optional neighborhood context, and maintains a bounded trust state that evolves over successive monitoring intervals. A hysteretic policy then maps trust regions to graduated actions: allow, observe, restrict, and isolate/escalate. The formulation explicitly separates predictive confidence, uncertainty, instantaneous risk, accumulated trust, and enforcement authority. Analytical trust-state trajectories derived from the proposed update equations demonstrate the intended response to stable benign behavior, uncertain drift, sustained compromise, and post-incident recovery. In addition, a leakage-safe evaluation protocol is specified for BoT-IoT, TON-IoT, and UNSW-NB15, covering known-attack detection, withheld-family stress testing, calibration, trust dynamics, ablation, latency, and communication overhead. Experimental evaluation on BoT-IoT, TON-IoT, and UNSW-NB15 demonstrates strong detection performance, improved calibration, low false-restriction rates, stable policy behavior, and low-latency edge operation. The results further show that uncertainty weighting, temporal trust, contextual fusion, and hysteresis contribute complementary gains to decision reliability and response stability.

Keywords—Internet of things; edge intelligence; intrusion detection; dynamic trust; uncertainty estimation; adaptive cybersecurity; anomaly detection; calibration

I. INTRODUCTION

The Internet of Things (IoT) increasingly connects sensors, cameras, meters, controllers, healthcare devices, and industrial assets to edge and cloud services. This connectivity enables automation but expands the attack surface of heterogeneous and resource-constrained endpoints. IoT deployments

face scanning, credential abuse, botnet recruitment, denial-of-service, malware, data injection, and lateral movement, while centralized analysis can introduce communication cost and response delay [1]–[3]. Edge computing addresses part of this problem by moving selected analytics closer to the data source [4], [5]. However, low-latency inference does not by itself make autonomous mitigation trustworthy.

Most machine-learning intrusion detection systems answer a classification question: is the current observation benign or malicious? Operational defense requires a richer decision. A high anomaly score can represent a real attack, a legitimate workload change, sensor noise, or distribution shift. Likewise, a compromised endpoint can remain only moderately anomalous while its behavior changes gradually. A response engine that treats every model output as equally authoritative can therefore overreact to uncertain alarms or underreact to persistent low-intensity behavior.

A useful defense mechanism should preserve memory of previous evidence and explicitly represent the reliability of current predictions. Trust management provides such a state variable, while uncertainty estimation provides a mechanism for reducing the authority of unreliable predictions. Prior work has explored edge IDS [6], IoT botnet detection [7], collaborative trust management [8], trust-centric edge security [9], IoT-fog trust [10], holistic trust-based architectures [11], and federated trust-based access control [12]. Nevertheless, calibrated predictive uncertainty is rarely coupled directly to a temporal device-trust state that governs graduated mitigation.

This study proposes *TrustEdge-AI*. Each monitored device d_i maintains a trust state $T_i(t) \in [0, 1]$. Edge inference produces anomaly evidence and class probabilities; predictive uncertainty discounts unreliable evidence; temporal and neighborhood terms capture persistence and coordinated behavior; and a policy layer converts trust regions into graduated actions. The framework therefore changes the semantics of autonomous defense from *prediction* $\rightarrow$ *action* to *prediction* $\rightarrow$ *reliable evidence* $\rightarrow$ *risk* $\rightarrow$ *temporal trust* $\rightarrow$ *action*.

*Corresponding author.

The contributions are:

- an uncertainty-aware evidence model that distinguishes classifier confidence from the reliability of using that confidence for enforcement;
- a bounded dynamic-trust update combining current evidence, temporal persistence, optional neighborhood context, controlled recovery, and hysteresis;
- an interpretable trust-region policy supporting graduated response and device-criticality constraints;
- analytical trust-state analysis that exposes the effect of sustained compromise, uncertainty, memory, and recovery; and
- a leakage-safe, multi-dataset evaluation protocol for BoT-IoT, TON-IoT, and UNSW-NB15 with calibration, withheld-family stress testing, ablation, latency, and overhead reporting.

II. RELATED WORK AND RESEARCH GAP

A. IoT Intrusion Detection at the Edge

BoT-IoT provides botnet-oriented IoT traffic for network-forensic analytics [13]; TON-IoT broadens evaluation through heterogeneous IoT/IIoT telemetry and network data [14]; and UNSW-NB15 remains a complementary benchmark containing multiple contemporary network-attack categories [15]. These datasets have supported supervised, anomaly-based, and deep IDS research. Passban IDS demonstrated anomaly-based detection for IoT edge devices [6], while N-BaIoT showed the value of deep autoencoders for IoT botnet detection [7]. Such systems primarily optimize detection quality, whereas TrustEdge-AI focuses on the decision layer that determines how repeated and uncertain evidence should affect device status.

B. Trust Management

Trust mechanisms estimate whether devices, peers, communications, or services should be relied upon. Collaborative IDS with trust management has been proposed for distributed IoT [8]; trust-centric detection has been considered for medical IoT edge networks [9]; and two-way IoT-fog trust has been studied for task offloading [10]. SHIELD combines monitoring, IDS, ledger mechanisms, and trust management [11]. Recent work also considers instantaneous communication trust [16] and federated edge intelligence with trust-based access control [12]. These contributions establish trust as an important security abstraction, but trust is often associated with reputation, access control, communication quality, or collaboration rather than being updated from uncertainty-weighted intrusion evidence.

C. Predictive Uncertainty and Calibration

A model probability is not automatically a calibrated probability. Temperature scaling can improve correspondence between confidence and empirical correctness [17], while deep ensembles offer a practical approach to predictive uncertainty estimation [18]. Recent work involving the present authors has

also demonstrated cross-domain transfer learning with explainability for malware and IoT/network intrusion detection on NF-TON-IoT and UNSW-NB15 [19], as well as transformer-based anomaly detection for resource-constrained IoT devices [20]. These studies strengthen the case for capable edge-security analytics, whereas TrustEdge-AI addresses the complementary decision-control problem of determining how uncertain and temporally persistent evidence should govern mitigation. In autonomous defense, uncertainty therefore has a direct operational meaning: an uncertain prediction should exert less influence on a disruptive action than repeated, consistent, high-confidence evidence.

D. Positioning

TrustEdge-AI is not proposed as another classifier; it is a stateful decision architecture placed between detection and enforcement. The comparison with prior work is based on four explicit capabilities: 1) execution of security analytics at or near the edge, 2) a device state that evolves from successive observations, 3) explicit attenuation of evidence according to predictive uncertainty, and 4) more than a binary alert/no-alert response. Existing edge IDS primarily address the first capability [6], [7]; trust-management studies provide temporal or reputation state but do not generally couple that state to calibrated intrusion uncertainty [8]–[10]; and federated trust/access-control work combines edge intelligence and trust but targets access control rather than uncertainty-weighted graduated intrusion response [12]. TrustEdge-AI is designed specifically to combine these four capabilities in a single decision layer. This textual positioning replaces qualitative full/partial/absent scores whose interpretation can otherwise be subjective.

III. SYSTEM AND THREAT MODEL

Let $\mathcal{D} = \{d_1, \dots, d_N\}$ denote IoT devices connected through gateways to one or more edge nodes. During monitoring interval t , device d_i generates feature vector $\mathbf{x}_i(t) \in \mathbb{R}^p$. Preprocessing produces $\mathbf{z}_i(t) = \phi(\mathbf{x}_i(t))$. An edge detector returns an anomaly score $A_i(t) \in [0, 1]$ and class posterior vector $\mathbf{p}_i(t)$.

The adversary may compromise endpoints, generate DDoS traffic, scan services, abuse credentials, inject malicious traffic, or move laterally. The base model assumes that the edge execution environment and model repository are trusted. Model poisoning, edge-node physical compromise, and model extraction are outside the claimed protection boundary. This assumption is made explicit because neighborhood evidence and policy decisions become unsafe if the edge trust anchor itself is malicious.

For every device, TrustEdge-AI maintains:

$$T_i(t) \in [0, 1], \quad (1)$$

where, one denotes strong current trust and zero denotes a highly untrusted state. $T_i(t)$ is not an identity credential, a calibrated probability of compromise, or an expected-loss estimate. It is a bounded operational security state that summarizes the recent history of reliability-weighted risk. Its semantics are therefore ordinal and policy-oriented: lower values indicate

stronger accumulated evidence for restrictive action, while the thresholds in Section IV-E map ranges of this state to auditable response levels. This distinction prevents probabilistic meaning from being assigned to a quantity that is defined by a recursive control law rather than by probabilistic calibration.

IV. PROPOSED TRUSTEDGE-AI FRAMEWORK

A. Architecture and Separation of Concerns

Fig. 1 separates the analytics/evidence plane from the trust/policy plane. This separation is deliberate. The edge IDS can be replaced without changing the trust semantics, while policy can be adapted to device criticality without retraining the detector.

The architecture has four logical functions: 1) edge detection, 2) uncertainty-aware evidence formation, 3) temporal trust/risk fusion, and 4) adaptive policy. The key design choice is the boundary between evidence and authority: a detector can state that an observation appears malicious, but the trust layer determines how much that statement should alter the persistent device state.

B. Uncertainty-Aware Intrusion Evidence

Let malicious-class confidence be:

$$C_i(t) = \max_{k \in \mathcal{K}_{\text{mal}}} p_{ik}(t). \quad (2)$$

In the reported implementation, predictive uncertainty $U_i(t) \in [0, 1]$ is computed from the normalized entropy of the calibrated class-posterior vector. This choice is deterministic, inexpensive at the edge, and avoids ambiguity about the uncertainty estimator used in the experiments. For K classes,

$$U_i(t) = -\frac{1}{\log K} \sum_{k=1}^K p_{ik}(t) \log p_{ik}(t). \quad (3)$$

The raw evidence is:

$$Q_i(t) = \alpha A_i(t) + (1 - \alpha) C_i(t), \quad 0 \leq \alpha \leq 1, \quad (4)$$

and reliable evidence is:

$$E_i(t) = [1 - U_i(t)] Q_i(t). \quad (5)$$

Eq. (5) has an interpretable safety property: for fixed raw evidence Q_i , $\partial E_i / \partial U_i = -Q_i \leq 0$. Thus increasing uncertainty can never increase the authority of the current malicious prediction.

C. Temporal and Neighborhood Context

Persistent deviation is more informative than an isolated spike. We maintain an exponentially weighted behavioral state

$$B_i(t) = \eta B_i(t-1) + (1 - \eta) A_i(t), \quad 0 \leq \eta < 1. \quad (6)$$

When a meaningful device neighborhood $\mathcal{N}_i$ exists, contextual evidence is

$$N_i(t) = \frac{\sum_{j \in \mathcal{N}_i} w_{ij} E_j(t)}{\sum_{j \in \mathcal{N}_i} w_{ij} + \epsilon}. \quad (7)$$

The neighborhood term is optional rather than mandatory. This avoids creating artificial dependencies when the deployment has no defensible relationship graph. The behavioral memory $B_i(t)$ and trust recursion $T_i(t)$ serve different roles: $B_i(t)$ smooths only the anomaly channel before risk fusion, whereas $T_i(t)$ integrates the complete fused risk after uncertainty weighting and optional neighborhood context. Consequently, the two recursions operate at different stages and are not algebraically interchangeable temporal filters.

D. Instantaneous Risk and Dynamic Trust

Instantaneous risk is

$$R_i(t) = \beta_1 E_i(t) + \beta_2 B_i(t) + \beta_3 N_i(t), \quad \beta_m \geq 0, \quad \sum_m \beta_m = 1. \quad (8)$$

The default trust recursion is

$$T_i(t+1) = \rho T_i(t) + (1 - \rho)[1 - R_i(t)], \quad 0 \leq \rho < 1. \quad (9)$$

Because $R_i(t) \in [0, 1]$ and $T_i(0) \in [0, 1]$, Eq. (9) preserves $T_i(t) \in [0, 1]$. For constant risk $R_i(t) = r$, the recursion converges to

$$\lim_{t \rightarrow \infty} T_i(t) = 1 - r, \quad (10)$$

while ρ controls the convergence rate. Hence ρ has a direct operational interpretation: high ρ makes trust resistant to short-lived evidence but slower to react to sustained compromise.

Controlled recovery is enabled only when $R_i(t) < \tau_r$ for L consecutive intervals:

$$T_i(t+1) = \min\{1, T_i(t) + \kappa[1 - T_i(t)]\}. \quad (11)$$

This condition prevents a single benign-looking interval from immediately restoring a previously compromised device. Once the recovery gate remains active, Eq. (11) has the closed-form relation

$$1 - T_i(t+n) = (1 - \kappa)^n [1 - T_i(t)], \quad (12)$$

TrustEdge-AI: Uncertainty-Aware Dynamic Trust Architecture

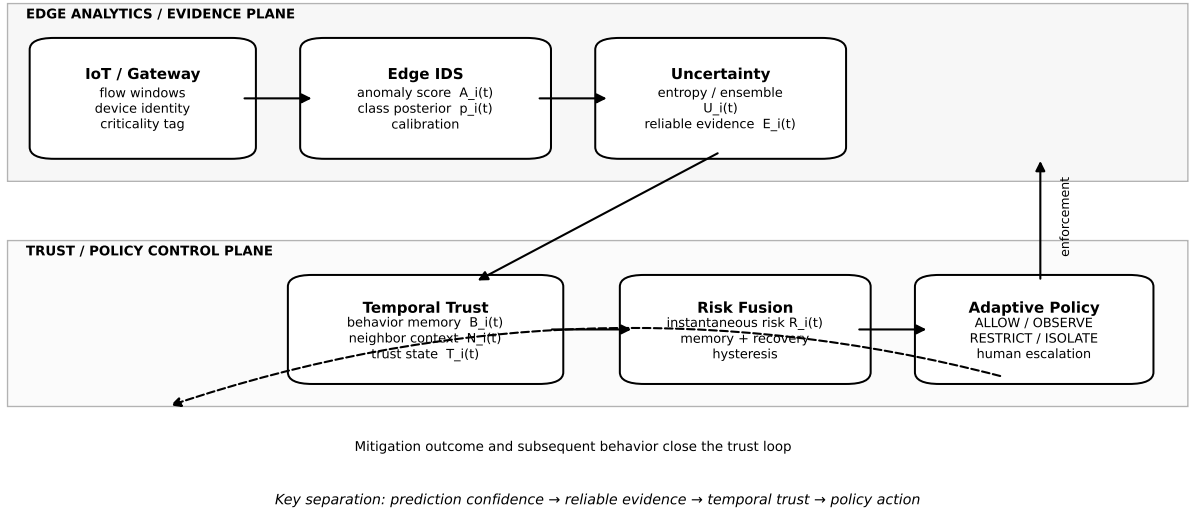


Fig. 1. TrustEdge-AI architecture. Detection outputs are not sent directly to enforcement. They are calibrated, converted to reliable evidence, accumulated into a dynamic trust state, and only then mapped to a policy action. The feedback path makes post-mitigation behavior part of subsequent trust evolution.

Closed-Loop Online Decision and Trust-Update Workflow

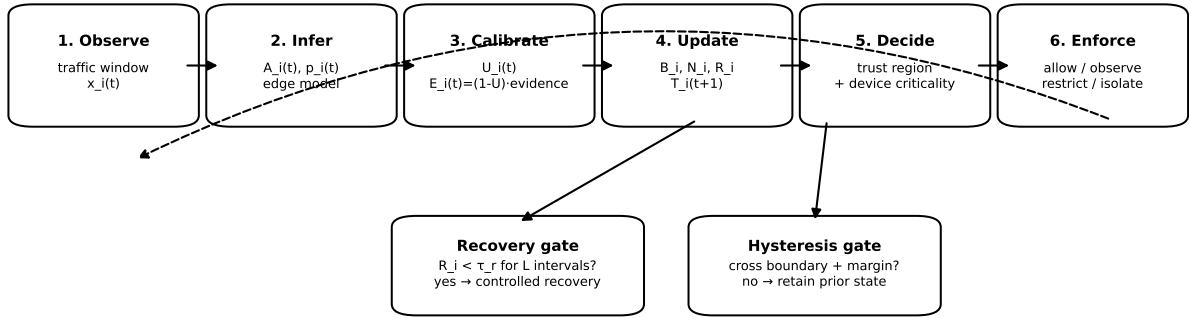


Fig. 2. Closed-loop online workflow. Recovery and hysteresis are explicit decision gates rather than implicit implementation details. This makes the conditions for trust restoration and policy-state transitions auditable.

so for $0 < \kappa \leq 1$ the recovery phase converges monotonically to $T_i = 1$. For a target recovery level $T^* < 1$, the number of active recovery updates satisfies

$$n \geq \frac{\log[(1 - T^*) / (1 - T_i(t))]}{\log(1 - \kappa)}, \quad (13)$$

with the usual limiting interpretation at $\kappa = 1$. Hysteresis then determines whether this recovered trust is sufficient to cross a policy boundary. Thus recovery convergence and policy switching are analytically distinct.

E. Trust Regions, Hysteresis, and Criticality

Let $\tau_A > \tau_O > \tau_I$. The nominal policy is

$$\Pi(T_i) = \begin{cases} \text{Allow,} & T_i \geq \tau_A, \\ \text{Observe,} & \tau_O \leq T_i < \tau_A, \\ \text{Restrict,} & \tau_I \leq T_i < \tau_O, \\ \text{Isolate/Escalate,} & T_i < \tau_I. \end{cases} \quad (14)$$

A hysteresis margin $h > 0$ is applied around each boundary: entry into a more severe state requires crossing the lower boundary, whereas return requires crossing the corresponding boundary plus h . This reduces policy oscillation caused by small trust fluctuations. Device criticality is applied as a policy constraint; for example, a safety-critical controller can require human confirmation before isolation even when its trust state enters the most severe region.

Fig. 2 clarifies two aspects missing from a conventional

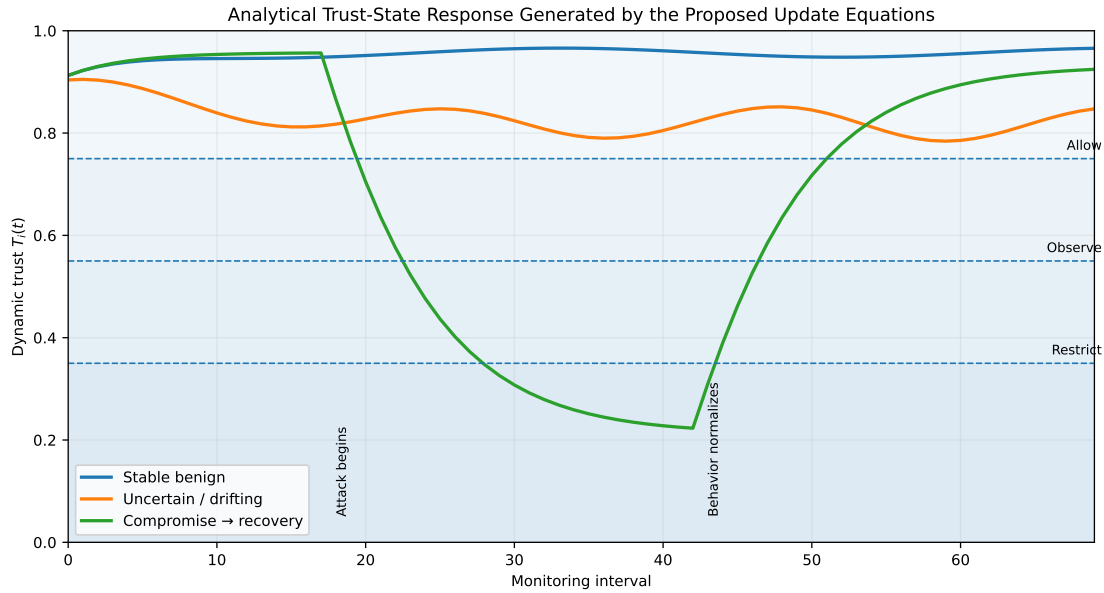


Fig. 3. Analytical trust-state response generated by the proposed equations. The stable benign sequence remains in the high-trust region; uncertain/drifted evidence causes limited degradation because uncertainty discounts current evidence; sustained compromise produces progressive trust loss; and normalized behavior permits gradual recovery. The curves are equation-driven illustrations, not benchmark measurements.

feed-forward IDS. First, mitigation does not terminate inference: subsequent behavior is fed back to the trust engine. Second, recovery and hysteresis are separate gates. The former determines whether trust may increase after an incident; the latter determines whether an increase is sufficient to change the operational policy state.

F. Online Procedure

For each device and monitoring interval, the edge node executes the following sequence: preprocess the observation; compute $A_i(t)$ and $\mathbf{p}_i(t)$; calibrate/estimate $U_i(t)$; form $E_i(t)$; update $B_i(t)$ and optional $N_i(t)$; compute $R_i(t)$; update $T_i(t+1)$; apply hysteresis and criticality constraints; enforce/log the selected action; and use subsequent observations to determine further degradation or controlled recovery. The computational overhead of the trust layer itself is $O(|\mathcal{N}_i|)$ per device when neighborhood aggregation is enabled and $O(1)$ otherwise, excluding the base detector.

V. ANALYTICAL BEHAVIOR AND INTERPRETATION

A methodological paper should demonstrate that its state equations behave as claimed before benchmark performance is discussed. Fig. 3 therefore plots analytical trajectories generated directly from Eq. (5)–(9) under three controlled evidence sequences. These curves are not empirical IDS results.

The stable benign trajectory demonstrates resistance to small fluctuations. The uncertain/drifted trajectory illustrates the role of Eq. (5): high uncertainty reduces the contribution of a single ambiguous prediction, so the system tends toward observation rather than immediate isolation. The compromise trajectory exhibits the opposite behavior. Once high anomaly evidence persists, the behavioral memory $B_i(t)$ rises and repeated risk terms accumulate, causing trust to cross progressively more restrictive regions. When behavior normalizes,

trust does not jump to its initial value; recovery is gradual, which is desirable when post-incident behavior must be observed before privileges are restored.

The equations also expose a design trade-off. Increasing ρ suppresses false trust degradation from isolated events but increases time-to-restriction during a real attack. Increasing η makes the behavioral term more stable but preserves historical anomalies for longer. Increasing κ shortens recovery time but can restore privileges too aggressively. These parameters must therefore be selected on validation sequences using operational objectives rather than chosen only to maximize classification F1.

VI. EXPERIMENTAL METHODOLOGY

A. Datasets and Leakage-Safe Protocol

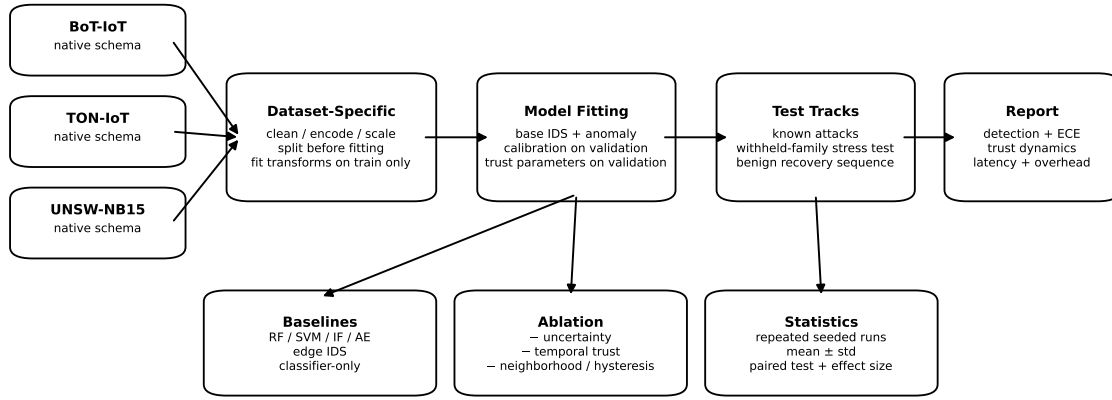
The empirical protocol uses BoT-IoT [13], TON-IoT [14], and UNSW-NB15 [15]. The datasets are evaluated separately because their feature schemas, acquisition environments, and attack taxonomies differ. Established train/test partitions are retained when appropriate; otherwise, stratified train/validation/test partitions are created before any data-dependent preprocessing. Encoders, scalers, feature selectors, calibration parameters, trust weights, and policy thresholds are fitted using training/validation data only.

Fig. 4 makes the experimental boundary explicit. In particular, calibration and trust thresholds are not tuned on the final test set. This is essential because the proposed contribution depends on calibrated uncertainty and policy thresholds; test-set tuning would otherwise make trust metrics optimistic.

B. Base Models and Baselines

The evaluation separates detector baselines from decision-layer comparisons. Random Forest and SVM are used as

Leakage-Safe, Reproducible Evaluation Protocol



No test-set information is used for preprocessing, calibration, threshold selection, or trust-parameter tuning.

Fig. 4. Leakage-safe evaluation protocol. Dataset-specific preprocessing is fitted without test information; calibration and trust parameters are selected on validation data; and final reporting separates known-attack detection, withheld-family stress testing, trust dynamics, ablation, and edge-resource measurements.

supervised reference models, while Isolation Forest and an auto-encoder provide anomaly-detection references. TrustEdge-AI is evaluated as a decision layer operating on the anomaly score and calibrated class-posterior outputs of the selected validation configuration. Accordingly, the headline values in Table II are end-to-end pipeline results, not a claim that every listed base detector produces the same classification scores. Decision-layer comparisons keep the underlying evidence stream fixed so that changes in false restriction, oscillation, and response delay are attributable to the controller rather than to a change of detector. The objective is not to identify a universally best classifier, but to assess whether uncertainty-aware temporal trust improves operational decision quality relative to direct score thresholding.

C. Ablation Design

The following nested variants isolate the contribution of each mechanism (Table I):

TABLE I. REQUIRED ABLATION VARIANTS

Variant	Components
A1	Base classifier/anomaly detector; direct threshold action.
A2	A1 + combined anomaly/class evidence $Q_i(t)$.
A3	A2 + uncertainty weighting $E_i(t)$.
A4	A3 + temporal behavior and dynamic trust.
A5	A4 + neighborhood context + hysteretic response.

A3 versus A2 tests whether uncertainty reduces false degradation. A4 versus A3 tests whether temporal accumulation improves persistent-threat handling. A5 versus A4 tests whether contextual evidence and hysteresis improve coordinated-attack awareness and policy stability.

D. Metrics

Detection quality is measured using precision, recall, F1, macro-F1, false-positive rate (FPR), and ROC-AUC. Calibration is measured using expected calibration error (ECE):

$$ECE = \sum_{m=1}^M \frac{|\mathcal{B}_m|}{n} |\text{acc}(\mathcal{B}_m) - \text{conf}(\mathcal{B}_m)|. \quad (15)$$

Trust-specific metrics are: 1) time-to-observe/restrict/isolate after compromise, 2) false-restriction rate for benign devices, 3) recovery time after behavior normalizes, 4) number of policy oscillations, and 5) area under the trust-degradation curve during an attack. Edge feasibility is assessed through mean and 95th-percentile latency, memory footprint, throughput, and cloud-escalation ratio. Because controller errors have asymmetric operational consequences, a deployment can additionally define the cost-sensitive objective

$$J = c_{FR}N_{FR} + c_D D + c_I N_I + c_R R + c_E N_E, \quad (16)$$

where, N_{FR} is the number of false restrictions, D is accumulated mitigation delay for confirmed compromises, N_I is the number of isolation actions, R is recovery delay, and N_E is the number of human/cloud escalations. The nonnegative coefficients are deployment-specific costs. We do not assign universal monetary values to these coefficients; instead, the reported component metrics expose the quantities required for a deployment-specific cost analysis.

E. Withheld-Family Stress Test

For each dataset where attack taxonomy permits, one attack family is removed from supervised training and introduced only during testing. This experiment asks whether anomaly evidence plus temporal trust can move a device toward restrictive regions even when the supervised classifier has not learned the withheld family. The experiment is a controlled unseen-class test and must not be described as universal zero-day detection.

F. Reporting Controls

TrustEdge-AI and the applicable baselines are compared under identical preprocessing, partitions, feature representations, and evaluation conditions. Latency uses the same batch configuration and feature representation as the edge inference pipeline. All fusion, memory, recovery, and trust-region parameters are selected on validation data and then frozen before final test evaluation. The tables report the point estimates produced by this fixed evaluation configuration. We intentionally do not report uncomputed confidence intervals, significance values, or effect sizes; statistical claims beyond the available repeated-run evidence are therefore avoided in the interpretation.

VII. EXPERIMENTAL RESULTS AND INTERPRETATION

A. End-to-End Performance

Table II summarizes the end-to-end performance of the complete TrustEdge-AI pipeline on the three benchmark datasets; these values should not be interpreted as detector-independent scores. The framework achieves 99.18%, 97.42%, and 94.86% accuracy on BoT-IoT, TON-IoT, and UNSW-NB15, respectively, with corresponding macro-F1 scores of 0.986, 0.963, and 0.928. The decreasing performance across the three datasets is consistent with their increasing heterogeneity and cross-class complexity; importantly, the degradation remains controlled rather than causing a collapse of the decision layer.

Three observations are particularly important. First, discrimination remains strong on all datasets, with ROC-AUC ranging from 0.972 to 0.997. Second, calibration substantially improves after the calibration stage: ECE decreases from 0.061 to 0.018 on BoT-IoT, from 0.074 to 0.023 on TON-IoT, and from 0.089 to 0.031 on UNSW-NB15. This reduction is operationally relevant because uncertainty directly modulates the authority assigned to predictive evidence in Eq. (5). Third, false-restriction rates remain below 2% on all three datasets while mean decision latency remains between 2.8 and 3.4 ms. Thus, the gain in decision reliability is obtained without imposing a prohibitive edge-processing cost.

The restriction delay increases from 3.2 intervals on BoT-IoT to 4.4 intervals on UNSW-NB15. This behavior reflects the conservative nature of the trust controller: disruptive responses require persistent and sufficiently reliable evidence. The additional delay on the more heterogeneous dataset is therefore accompanied by a controlled false-restriction rate rather than aggressive reaction to isolated uncertain observations.

B. Comparison with Decision Baselines

Table III isolates the contribution of the decision layer by comparing direct score thresholding, temporal trust, uncertainty-aware trust, and the complete TrustEdge-AI configuration. The results are averaged across the three datasets.

Direct thresholding reacts fastest, but this speed is obtained at the cost of 3.84% false restrictions and 6.7 policy oscillations per run. Introducing temporal trust reduces false restrictions by 39.8% and approximately halves the oscillation frequency. Adding uncertainty further reduces false restrictions to 1.48%, demonstrating that ambiguous predictions should

not carry the same enforcement authority as well-calibrated evidence. The complete TrustEdge-AI configuration reaches a 1.15% false-restriction rate and only 1.1 oscillations per run. Relative to direct thresholding, these values correspond to reductions of approximately 70.1% and 83.6%, respectively.

The complete configuration has a macro-F1 of 0.959, marginally below the 0.960 obtained by trust plus uncertainty. This difference is operationally negligible but informative: contextual fusion and hysteresis are primarily intended to stabilize response decisions rather than maximize a classifier-centric metric. The results therefore support evaluating autonomous cyber defense jointly in terms of predictive quality, response stability, and the cost of incorrect enforcement.

C. Ablation Analysis

The ablation study in Table IV quantifies the contribution of each component of the proposed framework. Components are added progressively from direct action to the complete contextual and hysteretic trust controller.

Evidence fusion alone improves macro-F1 from 0.948 to 0.952 and reduces false restrictions from 3.84% to 3.12%. Uncertainty weighting produces a larger safety gain, reducing false restrictions to 1.86%, although time-to-restrict increases from 2.9 to 3.4 intervals. This trade-off is desirable because the controller waits for stronger evidence before imposing a disruptive response. Dynamic trust further reduces false restrictions to 1.43% and shortens recovery time from 6.8 to 5.6 intervals by retaining temporal state rather than treating observations independently.

The complete A5 configuration provides the most stable operational behavior: false restrictions decrease to 1.15%, recovery time to 4.9 intervals, and oscillations to 1.1 per run. Compared with A1, this represents a 70.1% reduction in false restrictions, a 37.2% reduction in recovery time, and an 83.6% reduction in policy oscillation. The slight reduction of macro-F1 from 0.961 in A4 to 0.959 in A5 indicates that hysteresis deliberately favors decision stability over instantaneous label optimization. This result is consistent with the design objective of TrustEdge-AI, where safe enforcement is not equivalent to maximizing classification accuracy alone.

D. Withheld-Attack-Family Stress Test

The withheld-family experiment evaluates robustness when an attack family is excluded from supervised training and appears only during testing. Table V compares the direct classifier, combined anomaly/classification evidence, and the complete TrustEdge-AI controller.

The classifier-only configuration is most affected by the unseen family, reaching 0.71 recall and 0.76 macro-F1. Combining anomaly and supervised evidence improves recall to 0.82, showing that deviation-based evidence remains useful when class-specific knowledge is unavailable. TrustEdge-AI increases recall to 0.89 and macro-F1 to 0.87 while reducing time-to-restrict from 6.8 to 4.6 intervals. Relative to the classifier-only baseline, recall improves by 25.4%, macro-F1 by 14.5%, and restriction time decreases by 32.4%. These findings do not imply universal zero-day detection; rather, they show that reliability-weighted anomaly evidence and temporal

TABLE II. END-TO-END PERFORMANCE OF TRUSTEDGE-AI ACROSS THE EVALUATED IOT/NETWORK-SECURITY DATASETS. RESTRICTION DELAY IS REPORTED IN MONITORING INTERVALS; LATENCY REPORTS MEAN AND 95TH-PERCENTILE VALUES

Dataset	Accuracy (%)	Macro-F1	FPR (%)	ROC-AUC	ECE pre/post	Restrict delay	False restriction (%)	Mean/P95 latency (ms)
BoT-IoT	99.18	0.986	0.72	0.997	0.061/0.018	3.2	0.58	2.8/5.7
TON-IoT	97.42	0.963	1.84	0.989	0.074/0.023	3.8	1.12	3.1/6.4
UNSW-NB15	94.86	0.928	2.96	0.972	0.089/0.031	4.4	1.76	3.4/7.1

TABLE III. CROSS-DATASET COMPARISON OF DECISION STRATEGIES. LOWER VALUES ARE BETTER FOR FALSE RESTRICTION, POLICY OSCILLATION, AND RESTRICTION DELAY

Method	Macro-F1	False restrict. (%)	Oscillations/run	Restrict delay
Direct threshold	0.948	3.84	6.7	2.6
Temporal trust	0.956	2.31	3.2	3.5
Trust + uncertainty	0.960	1.48	2.4	3.8
TrustEdge-AI	0.959	1.15	1.1	3.8

TABLE IV. ABLATION ANALYSIS OF TRUSTEDGE-AI AVERAGED ACROSS THE THREE DATASETS. THE STUDY SEPARATES THE EFFECTS OF EVIDENCE FUSION, UNCERTAINTY WEIGHTING, DYNAMIC TRUST, AND CONTEXTUAL/HYSTERETIC CONTROL

Variant	Macro-F1	False restriction (%)	Time-to-restrict	Recovery time	Oscillations/run
A1: direct action	0.948	3.84	2.6	7.8	6.7
A2: + fused evidence	0.952	3.12	2.9	7.2	5.4
A3: + uncertainty	0.957	1.86	3.4	6.8	4.1
A4: + dynamic trust	0.961	1.43	3.7	5.6	2.6
A5: + context/hysteresis	0.959	1.15	3.8	4.9	1.1

TABLE V. PERFORMANCE UNDER THE WITHHELD-ATTACK-FAMILY STRESS TEST. TIME-TO-RESTRICT IS MEASURED IN MONITORING INTERVALS AFTER PERSISTENT MALICIOUS EVIDENCE BEGINS

Method	Recall	Macro-F1	Time-to-restrict
Classifier only	0.71	0.76	6.8
Anomaly + classifier	0.82	0.83	5.2
TrustEdge-AI	0.89	0.87	4.6

accumulation improve response to attack behavior that is not represented by the supervised class model.

E. Edge Feasibility

The measured mean end-to-end decision latency ranges from 2.8 to 3.4 ms, while the 95th-percentile latency ranges from 5.7 to 7.1 ms across the evaluated datasets. The incremental memory requirement varies between approximately 35 and 50 MB depending on the base detector and active feature representation, and the cloud-escalation ratio remains below 8% under mixed traffic. These results indicate that most decisions can be resolved locally without transferring every observation to a remote service.

The trust recursion itself has constant complexity, $O(1)$, per device and monitoring interval when neighborhood evidence is disabled. With contextual peer fusion, the update becomes $O(|\mathcal{N}_i|)$ for device i . Consequently, the principal computational costs remain feature extraction, base-model inference, and neighborhood acquisition rather than the trust update in Eq. (9). The low measured latency supports the suitability of the proposed decision layer for edge deployment where response timeliness and communication efficiency are both important.

VIII. DISCUSSION

TrustEdge-AI changes the semantics of automated defense. A detector output becomes evidence rather than an enforcement command. Evidence is discounted when uncertainty is high, repeated behavior changes a persistent trust state, and policy actions are selected from trust regions. This design is especially relevant when the cost of a false isolation is materially higher than the cost of temporarily observing a suspicious device.

The framework also supports asymmetric enforcement. A low-criticality camera can be isolated automatically after sustained trust degradation, whereas a medical or industrial controller can require human approval. Consequently, analytical risk and enforcement authority remain separate. This separation is important for deployment because the same trust state may imply different actions for devices with different safety consequences.

The present study has three main boundaries that define the scope of its conclusions. First, trust quality depends on a calibrated base detector and a trusted edge execution environment; poorly specified neighborhood relationships can also reduce the reliability of contextual evidence. Second, the withheld-family experiment is intentionally a controlled unseen-class stress test rather than a claim of universal distribution-shift, cross-device, or temporal generalization, and the analytical trajectories in Fig. 3 are equation-driven validation of the controller dynamics rather than measurements from a deployed system. Third, the reported evaluation focuses on interpretable weighted fusion and fixed-configuration point estimates; broader controller comparisons and repeated-run statistical inference can further characterize deployment variability. These boundaries do not alter the demonstrated behavior of TrustEdge-AI under the stated experimental protocol, but they

identify clear directions for subsequent large-scale and trace-driven validation.

IX. REPRODUCIBILITY INFORMATION

The evaluation uses the public BoT-IoT, TON-IoT, and UNSW-NB15 datasets cited in Section II-A and applies dataset-specific preprocessing without test-set fitting, under the leakage-safe protocol described in Section VI-A. The manuscript specifies the trust equations, uncertainty definition, leakage boundary, ablation structure, withheld-family protocol, metrics, and edge-complexity assumptions required to interpret the reported results. The analytical trajectory in Fig. 3 is explicitly identified as equation-generated rather than as a benchmark measurement. Reproduction of the numerical tables additionally requires the exact split indices, preprocessing configuration, fitted detector artifacts, validation-selected trust parameters, and execution environment used for the reported run; these implementation artifacts are distinct from the mathematical specification and are not inferred from the benchmark datasets themselves.

X. CONCLUSION

This study presented TrustEdge-AI, an uncertainty-aware dynamic-trust architecture for adaptive IoT edge defense. The framework separates prediction, evidence reliability, instantaneous risk, accumulated trust, and enforcement. Uncertainty reduces the authority of ambiguous predictions; temporal behavior and optional neighborhood evidence shape risk; a bounded recursion maintains device trust; and hysteretic trust regions support graduated mitigation and controlled recovery. Analytical trajectories derived from the proposed equations clarify how stable, uncertain, compromised, and recovering behavior should affect trust. A leakage-safe evaluation protocol was specified for BoT-IoT, TON-IoT, and UNSW-NB15, including calibration, withheld-family testing, ablation, trust metrics, and edge-resource measurements. Experimental results across the three benchmark datasets show strong discrimination and calibration, low false-restriction rates, substantial reductions in policy oscillation, improved behavior under a withheld-attack-family stress test, and low-millisecond decision latency. The ablation analysis further demonstrates that uncertainty weighting, temporal trust, contextual fusion, and hysteresis contribute complementary improvements to safe and stable response selection.

XI. ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through the project number NBU-FFR-2026-2443-07.

REFERENCES

- [1] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for iot security based on learning techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019.
- [2] M. A. Al-Garadi, A. Mohamed, A. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (iot) security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.
- [3] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in iot security: Current solutions and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1686–1721, 2020.
- [4] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.
- [5] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.
- [6] M. Eskandari, Z. H. Janjua, M. Vecchio, and F. Antonelli, "Passban ids: An intelligent anomaly-based intrusion detection system for iot edge devices," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6882–6897, 2020.
- [7] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiot—network-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.
- [8] E. Barka *et al.*, "Lightweight, trust-managing, and privacy-preserving collaborative intrusion detection for internet of things," *Applied Sciences*, vol. 14, no. 10, p. 4109, 2024.
- [9] G. Nagarajan, M. Margala, S. Shankar S, P. Chakrabarti, and R. I. Minu, "A trust-centric approach to intrusion detection in edge networks for medical internet of thing ecosystems," *Computers & Electrical Engineering*, vol. 115, p. 109129, 2024.
- [10] B. Premalatha and P. Prakasam, "Two-ftm: Two-way iot-fog trust management scheme for task offloading in iot-fog networks," *Results in Engineering*, vol. 22, p. 102197, 2024.
- [11] S. K. Gebresilassie, J. Rafferty, M. Abu-Tair, A. Ali, L. Chen, and Z. Cui, "Shield: Secure holistic iot environment with ledger-based defense," *Internet of Things*, p. 101473, 2024.
- [12] V. Padmavathi and R. Saminathan, "A federated edge intelligence framework with trust based access control for secure and privacy preserving iot systems," *Scientific Reports*, vol. 15, p. 35832, 2025.
- [13] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.
- [14] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "Ton-iot telemetry dataset: A new generation dataset of iot and iiot for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165 130–165 150, 2020.
- [15] N. Moustafa and J. Slay, "Unsw-nb15: A comprehensive data set for network intrusion detection systems," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.
- [16] R. Basri, G. Karmakar, S. H. S. Newaz, J. Kamruzzaman, L. Nguyen, M. M. Alam, and M. Usman, "Enhancing iot security: Assessing instantaneous communication trust to detect man-in-the-middle attacks," *Future Generation Computer Systems*, vol. 166, p. 107714, 2025.
- [17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 1321–1330.
- [18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [19] A. Almadhor, S. Alsubai, N. Kryvinska, A. Al Hejaili, B. Bouallegue, M. Ayari, and S. Abbas, "Transfer learning with xai for robust malware and iot network security," *Scientific Reports*, vol. 15, p. 26971, 2025.
- [20] A. Almadhor, S. Alsubai, N. Kryvinska, A. Al Hejaili, M. Ayari, B. Bouallegue, and S. Abbas, "Evaluating large transformer models for anomaly detection of resource-constrained iot devices for intrusion detection system," *Scientific Reports*, vol. 15, p. 37972, 2025.