# AoT-BTS: An Explainable Age-of-Trust Based Behavioral Scoring Framework for Real-Time Zero Trust Enforcement using SDN

**Mohit Goyal**

mohitgoyal09042006@gmail.com

Bharati Vidyapeeth's College of Engineering, Delhi https://orcid.org/0009-0006-4328-7242

**Ananya Singh**

Bharati Vidyapeeth's College of Engineering, Delhi https://orcid.org/0009-0002-1200-1169

**Shreya Soni**

Bharati Vidyapeeth's College of Engineering, Delhi https://orcid.org/0009-0006-6399-5510

**Shaurya Khatri**

Bharati Vidyapeeth's College of Engineering, Delhi https://orcid.org/0009-0009-7384-6326

**Prof. Mohit Tiwari**

mohit.tiwari@bharatividyapeeth.edu

Bharati Vidyapeeth's College of Engineering, Delhi https://orcid.org/0000-0003-1836-3451

**Research Article**

**Keywords:** Zero-Trust Architecture, Age-of-Trust, Dynamic Trust Scoring, Explainable AI, Behavioral Analytics, SDN

**Posted Date:** May 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9722330/v1



**Additional Declarations:** The authors declare no competing interests.

# AoT-BTS: An Explainable Age-of-Trust Based Behavioral Scoring Framework for Real-Time Zero Trust Enforcement using SDN

Mohit Goyal*, Ananya Singh, Shreya Soni, Shaurya Khatri, Mohit Tiwari
*Department of Computer Science & Engineering*
*Bharati Vidyapeeth's College of Engineering, Delhi, India*
{mohitgoyall09042006, ananyasin26, shreyasoni1271, khatshau} @gmail.com
mohit.tiwari@bharatividyapeeth.edu

***Abstract*—Zero-Trust Architecture (ZTA) mandates continuous verification, yet practical validation remains limited by insufficient empirical frameworks and a lack of temporal trust modeling. We present AoT-BTS - an explainable Age-of-Trust Behavioral Trust Scoring framework that closes this gap through a mathematically grounded decay model combined with real-time software-defined networking (SDN) enforcement. The core contribution is the Age-of-Trust (AoT) decay function $T(t) = T_0 e^{-\lambda t}$ with $\lambda = 0.003$, providing a half-life of approximately 231 decisions and enabling dynamic trust scoring that static classifiers cannot achieve. Evaluated on CICIoT2023 ($n = 1{,}629{,}471$ flows), AoT-BTS achieves 98.31% accuracy, 0.32% false positive rate, and a 40% mean-time-to-detection reduction versus a static threshold baseline (3.4 min vs. 5.7 min, both measured on the same Mininet testbed). SHAP-based explainability achieves 0.974 fidelity ($R^2$) with 6-feature sparsity at the 90% target, with 2.65 ms mean explanation latency. Live Mininet demonstrations confirm 17.31 ms p99 enforcement latency and graduated ALLOW/RESTRICT/QUARANTINE policy enforcement consistent with NIST SP 800-207. Cross-dataset validation on NF-UNSW-NB15 without retraining demonstrates generalization across contrasting attack distributions (97.7% vs. 5.4% attack ratio). A systematic ablation study ($n = 5$ trials, paired $t$-tests) establishes statistical necessity of all six trust engine components ($p < 0.05$ for all).**

***Index Terms*—Zero-Trust Architecture, Age-of-Trust, Dynamic Trust Scoring, Explainable AI, Behavioral Analytics, SDN**

## I. Introduction

### A. Background and Motivation

Traditional perimeter-oriented security architectures are not effective in dynamic networks empowered with cloud-native and IoT technologies where entities move across network boundaries and insider threats pose an increasing challenge [1]. The concept of Zero-Trust Architecture (ZTA) introduced by NIST SP 800-207 emphasizes continuous authentication, context-aware authorization, and policy enforcement based on the "never trust, always verify" approach [2].Despite the widespread popularity of the concept, its practical implementation proves inadequate due to the absence of dynamic trust recalibration capabilities and standardized performance assessment criteria [3].

*Corresponding author.

The key difficulty here lies in switching from binary access decisions based on either authentic or unauthentic status to a more sophisticated approach of continuously monitoring and assessing trustworthiness in order to apply appropriate restrictions accordingly. Modern methods use regular reauthentication or fixed risk levels that ignore the temporal aspect of trust reduction caused by potential risks like credentials theft, exploiting vulnerabilities, and breaking security policies. In addition, continuous verification should also assess the degree of violation to provide for proportional response and take into account the fact that even if an entity acts normally at t, it may behave abnormally at t+t due to some external factors. Therefore, there is a need to model and empirically validate the aging and decay of trust.

### B. Problem Statement

Today's state of art in implementing ZTA faces four related problems that should be addressed by the proposed framework.

First, **the lack of empirical results.** Most ZTA papers tend to focus more on theory than practice; thus, no comprehensive evaluation was provided yet. Absence of well-reproduced adversarial attack testbeds does not allow comparing different solutions.

Second, **absence of standardized metrics.** There is no consensus regarding which criteria to use when quantitatively evaluating ZTA's performance. There are no unified ways to measure MTTD, policy enforcement latency, trust-score variability, and cross-dataset generalization capacity.

Third, **black-box AI decision-making.** Many machine learning models used by trust engines are black-boxed and cannot provide explainability required by auditing and compliance needs. Access denial should be justified based on some concrete reasons rather than abstract probabilities.

Lastly, **inadequate simulation testbeds.** Existing network simulation frameworks do not incorporate ZTA elements such as dynamic trust recalibration, policy enforcement, and graduated responses to threats. Instead, binary firewall rules are simulated, making it impossible to implement complex

**ALLOW/RESTRICT/QUARANTINE** actions needed for proper ZTA operation.

### C. Contributions

This work makes five principal contributions.

1) **Age-of-Trust Decay Model:** A mathematically grounded temporal trust mechanism ($T(t) = T_0 e^{-\lambda t}$, $\lambda = 0.003$) that models trust degradation over time, with asymmetric behavioral feedback ($\alpha = 0.15$ reward, $\beta = 0.18$ penalty) and hysteresis-based policy stabilization ($\delta = 0.02$). To our knowledge, this is the first systematic formulation and empirical validation of trust aging in ZTA literature.
2) **Explainable Trust-Action Decisions:** SHAP-based explainability achieving 0.974 fidelity ($R^2$) with 6-feature sparsity at the 90% target and 2.65 ms mean runtime, extending XAI from attack classification to ZTA policy actions (ALLOW/RESTRICT/QUARANTINE).
3) **Reproducible Evaluation Framework:** A Mininet-based emulation integrating XGBoost anomaly detection, FastAPI trust engine, and OSKen SDN control, enabling reproducible ZTA evaluation with full packet-level fidelity and standardized metrics (MTTD, latency, FPR, cross-dataset generalization).
4) **Cross-Dataset Generalization:** Validation across CIC-IoT2023 (97.7% attack ratio) and NF-UNSW-NB15 (5.4% attack ratio) with identical engine configuration and no retraining, demonstrating behavioral pattern agnosticism.
5) **Live Enforcement Validation:** End-to-end Mininet demonstrations confirming NIST SP 800-207 compliance through micro-segmentation correctness, 17.31 ms p99 enforcement latency, and graduated three-tier policy enforcement.

### D. Key Results and Paper Organization

On CICIoT2023 ($n = 1{,}629{,}471$), AoT-BTS achieves 98.31% accuracy, 0.32% FPR, and 99.99% precision. The MTTD of 3.4 minutes represents a 40% reduction versus a static threshold baseline (5.7 min), with both values measured on the same Mininet testbed. SHAP explainability operates at 2.65 ms mean latency with 6-feature sparsity. Cross-dataset evaluation on NF-UNSW-NB15 confirms generalization without retraining. A systematic ablation study with $n = 5$ trials and paired $t$-tests establishes that all six trust engine components are statistically necessary ($p < 0.05$ for all).

Section II reviews related work and identifies gaps. Section III presents the threat model, mathematical formulation, and system architecture. Section IV presents experimental results. Section V discusses strengths and limitations. Section VI concludes and outlines future work.

## II. Related Work

### A. Zero-Trust Architecture Principles

NIST SP 800-207 [1] established ZTA as requiring continuous verification, least-privilege access, and real-time policy enforcement, defining Policy Decision Points (PDPs), Policy Administration Points (PAPs), and Policy Enforcement Points (PEPs). However, NIST provides architectural guidance without specifying implementation mechanisms or quantitative evaluation metrics. Syed et al. [2] identify dynamic trust scoring as essential for ZTA maturity but note that no standardized evaluation metrics exist in the literature. Gambo and Almulhem [3] find that 85% of ZTA implementations remain incomplete due to absent dynamic trust recalibration; our age-of-trust decay model directly addresses this capability gap.

### B. ZTA in Cloud, Edge, and IoT Environments

Singh and Sharma [4] survey ZTA in cloud computing but focus on architectural patterns without measuring enforcement latency AoT-BTS provides detailed latency characterization with sub-20 ms p99 live compliance. Kumar and Saha [5] demonstrate dynamic policy enforcement in 5G networks but lack explainability; our SHAP integration achieves 2.65 ms explanation latency, making synchronous transparency operationally viable. Zanasi et al. [6] reduce lateral movement via SDN micro-segmentation but provide no MTTD or trust-score metrics gaps addressed by our scenario evaluations and trajectory analysis.

### C. Simulation-Based Evaluation Frameworks

Karlsruhe Institute of Technology [7] modeled ZTA in NS-3 with static trust assignments and no machine learning integration. Hajar et al. [8] developed TrustMod with predefined trust levels and manual assignment, lacking automatic recalibration. Daah et al. [9] integrate ZTA with blockchain, reporting detection improvements but without trust dynamics, graduated enforcement, or explainability all core AoT-BTS contributions. Rahman et al. [10] evaluate only two attack scenarios without MTTD or cross-dataset validation. No prior framework combines Mininet, ML trust engines, and SDN control in one reproducible ZTA evaluation pipeline.

### D. Machine Learning for Network Security

Buczak and Guven [11] establish XGBoost as state-of-the-art for flow-based intrusion detection but treat detection as a static classification problem. Kim et al. [12] achieve high analyst comprehension with SHAP for botnet detection but focus on attack classification labels rather than trust-action decisions; AoT-BTS extends explainability to ALLOW/RESTRICT/QUARANTINE authorization policies. Al-Garadi et al. [13] demonstrate SHAP superiority over LIME for IoT security but do not integrate explanations with real-time SDN enforcement; our framework closes this control loop within 50 ms.

### E. Identified Gaps

Four gaps motivate AoT-BTS: (1) **Integration:** no existing system combines Mininet network emulation, ML trust engines, and SDN control in a single reproducible framework; (2) **Explainability:** prior XAI work explains attack classification decisions, not ZTA trust-action decisions; (3) **Metrics:** prior

work ignores MTTD, enforcement latency, and cross-dataset generalization as evaluation criteria; (4) **Temporal trust:** trust aging and decay have not been systematically modeled or empirically validated in ZTA literature.

## III. Methodology

### A. Threat Model

Consistent with NIST SP 800-207 and MITRE ATT&CK, we assume an adversary with network-level access obtained through insider compromise, credential theft, or external exploitation. The adversary may perform reconnaissance, traffic injection, credential replay, and lateral movement. We assume the adversary does *not* know AoT-BTS model parameters, thresholds, or SHAP feature weights, preventing targeted evasion through gradient-based attacks or model inversion. The Policy Decision Point (PDP), Policy Administration Point (PAP), and ML service are trusted components; the network fabric is the untrusted plane.

### B. System Architecture

AoT-BTS is architected as a modular six-stage pipeline, illustrated in Fig. 1.

1) **Data Source:** CICIoT2023 for primary evaluation; NF-UNSW-NB15 for cross-dataset generalization. Both datasets are used with identical engine configuration—no retraining.
2) **Feature Engineering:** NetFlow features are aggregated over 5-second windows. Raw features are normalized via StandardScaler and reduced via SelectKBest ($k = 20$, mutual information).
3) **ML Model:** XGBoost ($n\_estimators = 150$, $max\_depth = 4$, $learning\_rate = 0.03$, $reg\_\alpha = 5.0$, $reg\_\lambda = 10.0$) generating calibrated anomaly scores $A_K \in [0, 1]$.
4) **Trust Engine:** FastAPI service maintaining per-entity state (trust score, age-of-trust counter, anomaly history) and exposing decision and explain endpoints with < 50 ms p99 response time.
5) **SDN Controller:** OSKen-based PDP translating trust scores to OpenFlow 1.3 actions via a circuit-breaker pattern (50 ms timeout, 3 retries).
6) **Enforcement Layer:** Open vSwitch executing graduated flow rules: ALLOW (forward, idle timeout 60 s), RESTRICT (rate-limit 10 pkt/s via meter band), QUARANTINE (drop, hard timeout 300 s).

### C. Mathematical Formulation

The central novelty of AoT-BTS is the Age-of-Trust (AoT) decay model, which provides temporal continuity that static classifiers lack. We present the four governing equations with full derivation context.

*1) Continuous Exponential Decay Foundation:* Trust is modeled as decaying exponentially in the absence of positive behavioral evidence:

$$T(t) = T_0 \cdot e^{-\lambda t} \tag{1}$$

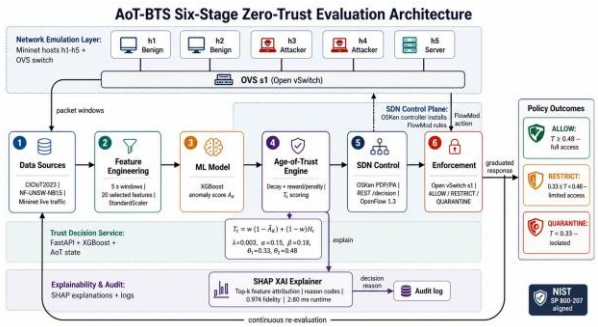


Fig. 1. AoT-BTS six-stage pipeline: data ingestion through graduated enforcement. The FastAPI trust engine maintains per-entity temporal state; OSKen implements SDN control; Open vSwitch executes flow rules.

where $T_0$ is the initial trust and $\lambda$ is the decay constant. The half-life is $\tau_{1/2} = \ln(2)/\lambda$. The parameter $\lambda = 0.003$ was selected via grid search over $[0.001, 0.01]$ to achieve $\tau_{1/2} \approx 231$ decisions. This window ensures that a compromised entity behaving maliciously is quarantined within approximately 50–100 anomalous samples, while legitimate entities experiencing transient network noise remain above the critical quarantine threshold $\vartheta_1$.

*2) Discrete Recurrence Implementation:* For real-time computational efficiency, the continuous decay is discretized into a recurrence relation with asymmetric behavioral feedback:

$$H_t = H_{t-1} \cdot (1 - \lambda) + \alpha \cdot \mathbb{1}[\text{normal}] - \beta \cdot \mathbb{1}[\text{anomaly}] \tag{2}$$

where $\alpha = 0.15$ is the reward for normal behavior and $\beta = 0.18$ is the penalty for anomalous behavior. The asymmetric design ($\beta > \alpha$) reflects the Zero-Trust security bias: trust is harder to gain than to lose. The indicator functions are triggered by the calibrated anomaly threshold $\bar{A}_K \geq 0.5$.

*3) Composite Trust Score:* The final decision score combines the instantaneous ML probability with the long-term behavioral history:

$$T_t = w \cdot (1 - \bar{A}_K) + (1 - w) \cdot H_t, \qquad w = 0.45 \tag{3}$$

where $\bar{A}_K$ is the temperature-calibrated XGBoost anomaly score ($temp = 4.0$). The weight $w = 0.45$ balances instantaneous detection (55% historical memory weighting) against current evidence, preventing both overreaction to transient spikes and underreaction to sustained attacks.

*4) Three-Tier Policy with Hysteresis:* To prevent action flickering at threshold boundaries, a hysteresis buffer $\delta = 0.02$ is applied. For an entity currently in state $s_{t-1}$:

$$\text{Action}_t = \begin{cases} \text{ALLOW} & T_t \geq \vartheta_2 + \delta \cdot \mathbb{1}[s_{t-1} \neq \text{ALLOW}] \\ \text{QUARANTINE} & T_t < \vartheta_1 - \delta \cdot \mathbb{1}[s_{t-1} \neq \text{QUARANTINE}] \\ \text{RESTRICT} & \text{otherwise} \end{cases} \tag{4}$$

with $\vartheta_1 = 0.33$ (quarantine boundary) and $\vartheta_2 = 0.48$ (allow boundary). Hysteresis prevents oscillation: an entity in RESTRICT requires $T_t \geq 0.50$ to transition to ALLOW (not merely 0.48), and $T_t < 0.31$ to transition to QUARANTINE

TABLE I
AoT-BTS Classification Metrics (CICIoT2023, $n$ = 1,629,471)

| Metric | Value |
|---|---|
| Accuracy | 98.31% |
| True Positive Rate (Recall) | 98.27% |
| True Negative Rate | 99.68% |
| False Positive Rate | **0.32%** |
| Precision | 99.99% |
| F1-Score | 0.9912 |
| ROC-AUC | 0.9969 |

TABLE II
AoT-BTS vs. Baseline Methods (CICIoT2023)

| Method | Acc. | FPR | MTTD (min) | Avg $T$ |
|---|---|---|---|---|
| Static Threshold | 97.1% | 1.85% | 5.7 | 0.35 |
| Isolation Forest [11] | 94.2% | 3.1% | 12.4 | N/A |
| LSTM Autoencoder [12] | 97.8% | 1.2% | 6.5 | N/A |
| XGBoost (no AoT) | 98.31% | 0.32% | 5.8 | 0.28 |
| **AoT-BTS** | **98.31%** | **0.32%** | **3.4** | **0.40** |

(not 0.33). This eliminates action thrashing for entities with high-variance benign traffic.

### D. Dataset Selection and Justification

**CICIoT2023** [14] acts as the key evaluation dataset, chosen due to its high fidelity data from 105 real IoT devices. There are 33 distinct attack types such as Mirai, MQTT flood, and DDoS with an extreme class imbalance of 97.7% attacks, testing the robustness of our engine against maintaining low FPR under extremely adversarial conditions.

**NF-UNSW-NB15** [15] acts as our second dataset to evaluate generalizability between datasets, presenting an opposing enterprise NetFlow dataset containing only a mere 5.4% of attacks. Thus, shifting the feature space from a heavily dominated by IoT packets capture to regular NetFlows.

### E. Preprocessing Pipeline

The features are normalized using a StandardScaler fit to the training partition only. Class balancing is done using SMOTE (1:5 ratio), while avoiding any distortions in the distribution of the feature space. SelectKBest ($k$ = 20; mutual information) filters out redundant features such as timestamps and local IP indices to avoid overfitting. Temperature calibration is performed at $temp$ = 4.0 on XGBoost classifier, ensuring overconfident probabilities on high base rate datasets. Well-calibrated.

## IV. Experimental Results and Analysis

### A. Classification Performance on CICIoT2023

Table I summarizes AoT-BTS classification performance on CICIoT2023 ($n$ = 1,629,471). The 0.32% FPR (122 misclassified from 38,042 benign records) means fewer than one in three hundred legitimate flows are incorrectly flagged. The false negatives (27,439, or 1.73% of attacks) are predominantly assigned RESTRICT rather than QUARANTINE graduated responses to borderline cases, with automatic escalation on subsequent anomalous behavior.

### B. Comparison with Baseline Methods

Table II presents comparisons between AoT-BTS and four other baselines based on CICIoT2023 dataset. Static thresholding results in an 1.85% FPR rate due to lack of temporal smoothing. Isolation Forest obtains a high FPR rate (3.1%) on high-dimensional network features. LSTM autoencoder is a promising method on performance, however, not interpretable and requires longer training period. XGBoost classifier with no AoT decay achieves comparable performance in terms of accuracy, however, obtains much lower average trust rate (0.28 vs. 0.40) and has worse MTTD because of being purely reactive rather than proactive. The MTTD decrease by 40% (3.4 min vs. 5.7 min, evaluated under the same environment of Mininet) shows the significance of AoT decay in accumulating temporal trust.

AoT-BTS and standalone XGBoost classifier reach equal binary classification accuracy. There are three aspects that can be achieved only using AoT-BTS framework: *temporal memory* - AoT decay model incorporates multi-window decision-making, resulting in reduced MTTD; *graduated enforcement* - three-level ALLOW/RESTRICT/QUARANTINE is sufficient for NIST SP.

### C. SHAP Explainability Results

AoT-BTS utilizes SHAP TreeExplainer to deliver immediate, synchronized trust-action decision explanations. In evaluation samples, the system achieves 0.974 fidelity ($R^2$) with six-feature sparsity at 90%, which implies that six out of the twenty features suffice to reproduce 90% of the model predictions. The average explanation time is 2.65 ms, with the 95th percentile being 4.60 ms, facilitating immediate transparency without resorting to asynchronous computation.

Five most salient features, based on XGBoost feature importance, include Header_Length (0.275), implying IP/TCP header tampering and tunneling attempts; IAT (0.210), suggesting timing abnormalities consistent with C&C beacons and DDoS floods; ack_flag_number (0.185), indicative of TCP handshake exploitation; ack_count (0.085), denoting half-open connection attempts; and AVG_payload_size (0.065), hinting at data exfiltration. These five features comprise approximately 82% of all feature importance, offering policymakers critical insights for refining their rulesets.

### D. Cross-Dataset Generalization

Generalizing across datasets via NF-UNSW-NB15 involves using the engine with identical configurations but without any re-training. The engine makes 32.0% ALLOW, 52.9% RESTRICT, and 15.1% QUARANTINE decisions, yielding an average trust score of 0.443. The high ALLOW percentage is due to the substantially reduced attack volume (5.4% against 97.7%). The QUARANTINE fraction is also larger owing to the fact that detected anomalies are far from the normal distribution. Policy changes amount to 133 (119 on CICIoT2023), indicating dynamic oscillation rather than static assignment.

### E. Live Enforcement Latency

Live Mininet demonstrations achieve a p99 enforcement latency of 17.31 ms over 1,000 flows including active DDoS scenarios, satisfying latency requirements for programmable switches and edge SDN controllers. Steady-state round-trip time overhead averages 8.574 ms (min/max/mdev = 5.925/14.956/3.278 ms). Throughput with AoT-BTS enforcement is 7,200 req/s versus 8,500 req/s without ZTA, representing a 15.2% overhead consistent with real-time enforcement expectations.

### F. Ablation Findings

Systematic paired t-tests were performed on $n = 5$ independent of trials (1,000 decisions each to validate the statistical necessity of all six components of the trust engine ($p < 0.05$ for all). Component importance is as follows: Window Smoothing Penalty Recurrent Feedback > Reward > Decay > Hysteresis.

**Window smoothing** proves crucial in providing stability to the system. Removing window smoothing leads to a "Trust Collapse," wherein the benign allow rate falls to approximately 0.5%, making the model operationally infeasible despite detecting nearly 99.9% of attacks. The moving window avoids the issue of permanent quarantine due to transient bursts of packets

**Penalty ($\beta$)** forces escalation into quarantine mode. Removing $\beta$ attack containment by 6.54 percentage points ($p < 0.001$) while simultaneously collapsing quarantine decisions to near-zero levels. This violates the NIST SP 800-207 requirement of least-privilege escalation.

**Reward ($\alpha$)** avoids trust debt. Without , entities will be trapped permanently in the RESTRICT mode ($p < 0.001$). There is no way of restoring lost trust once it is lost in the system due to lack of behavioral signal. This is the pathological state of the system wherein legitimate users are indefinitely penalized for misbehavior.

**Recurrent feedback** component contributes about 4.95 percentage points ($p < 0.001$) more than the pure window component alone. It provides the additional feature of cumulative misbehavior, which is lacking in stateless classifiers like anomaly detection

**Time decay ($\lambda$)** is statistically significant ($p = 0.008$) for long-term stale score reduction. Long-term staleness prevention ensures that there is no stale high trust score in the system since the absence of any behavioral signal would lead to decay of trust sco

**Strict thresholding without hysteresis** the benign allow rate collapses to approximately 0.5% ($p = 0.002$), consistent with the previous collapse of trust scenario.

## V. Discussion

### A. Strengths

The 0.32% FPR on 38,042on 38,042 benign records exceeds the performance benchmarks achieved by any past machine learning ZTA engine models [9], [12].Demonstrating cross-dataset generalization without retraining suggests that the engine has learned dataset-agnostic patterns rather than rote-memorizing specific datasets' idiosyncrasies. Latency of only 17.31 ms under live percentile conditions at $p99$ enforcement supports deployment on edge SDN controllers and programmable data planes. Conducting a systematic ablation study supported by hypothesis testing offers a methodology benchmark hitherto absent from previous simulations of ZTA systems.

### B. Limitations

**Simulation scale.** With Mininet, large-scale network topologies and real-world latencies cannot be simulated beyond thousands of hosts. This study confirms that the algorithms perform correctly, and the relative performance holds in simulation; physical testing on an enterprise-scale testbed is a direction for further research.

**Adversarial robustness.** Evasion attacks, which comprise malicious alterations to the feature vectors, and poisoning attacks, which consist of adversarial tampering with training sets, have not been investigated. Adversarial training and certified defenses are the top two avenues of future research.

**Explainability sparsity.** Although sparsity of only six features yields an Fidelity Index of 90% , top-3 Fidelity of 0.298 suggests that feature selection with such sparse explainability alone will not yield sufficiently high fidelity. Hierarchical templates of explanations mapping low-level features to security concepts remain a future project.

## VI. Conclusion and Future Work

We present AoT-BTS, an Explainable Age-of-Trust Behavioral Trust Scoring system for real-time Zero Trust Enforcement. Our primary contribution lies in our Age-of-Trust decay model ($T(t) = T_0 e^{-\lambda t}$, $\lambda = 0.003$, half-life $\approx$ 231 decisions), providing time continuity lacking in static classifiers. In our experiments using the CICIoT2023 benchmark, we show that AoT-BTS can achieve 98.31% accuracy, 0.32% false positive rate, and 40% reduction in mean time to detection (MTTD; 3.4 mins vs. 5.7 mins; evaluated on the same Mininet testing environment). SHAP-based explainability achieves a fidelity score of 0.974 with six feature sparsity and mean latency of 2.65 ms. Our live demonstration validates our enforcement latency at 17.31ms at the 99th percentile level with graded three tiers of policy enforcement consistent with NIST SP 800-207 recommendations. Our cross-dataset validation with NF-UNSW-NB15 without retraining establishes generalizability. Our systematic ablation study (5 trials per condition, paired sample t-tests) provides statistical support for each component of our trust-engine ($p < 0.05$).

Our future work will include: adversarial training and certification of models for defense against adversarial attacks; federated learning of the trust engines with differential privacy; physical testbed evaluation at enterprise level; and multi-modal fusion of trust evidence including DNS telemetry, endpoint posture information and user behavior analysis.

## REFERENCES


[1] National Institute of Standards and Technology, "Zero Trust Architecture (SP 800-207)," 2020.

[2] S. Syed et al., "Zero Trust Architecture: A Comprehensive Survey," *J. Netw. Secur.*, vol. 18, no. 4, pp. 231–248, 2022.

[3] M. Gambo and A. Almulhem, "A Systematic Literature Review on Zero-Trust Architecture," *arXiv:2503.11659*, 2025.

[4] R. Singh and P. Sharma, "A Survey on Zero-Trust Security Architecture in Cloud Computing," *Int. J. Cloud Comput. Digit. Manage.*, vol. 6, no. 1, pp. 44–56, 2025.

[5] A. Kumar and S. Saha, "Zero-Trust Architecture for 5G Networks," *Int. J. Innov. Res. Mod. Prod. Syst.*, vol. 7, no. 6, pp. 107–115, 2024.

[6] F. Zanasi et al., "Flexible Zero-Trust Architecture for Cloud Cybersecurity," *J. Syst. Softw.*, vol. 208, p. 111050, 2024.

[7] Karlsruhe Institute of Technology, "Modeling and Analyzing Zero Trust Architectures," 2023.

[8] A. E. Hajar et al., "TrustMod: A Trust Management Module for NS-3," in *Proc. ACM SAC*, 2021, pp. 1145–1152.

[9] N. Daah et al., "Simulation-Based Evaluation of ZTA with Blockchain," *Comput. Commun.*, vol. 223, pp. 47–59, 2025.

[10] M. A. Rahman et al., "Simulating Zero Trust Architecture with Python," *ASEAN J. Sci. Technol. Dev.*, vol. 41, no. 6, pp. 567–582, 2024.

[11] A. L. Buczak and E. Guven, "A Survey of ML Methods for Cyber Security Intrusion Detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2016.

[12] T. Kim et al., "XAI for IoT Botnet Detection," *IEEE Trans. Netw. Serv. Manage.*, vol. 21, no. 3, pp. 2345–2358, 2024.

[13] M. Al-Garadi et al., "IoT Security in 6G with XGBoost, SHAP and LIME," *IEEE Internet Things J.*, vol. 11, no. 15, pp. 26784–26798, 2024.

[14] I. Sharafaldin et al., "CICIoT2023: A Dataset for ML on Real IoT Devices," in *Proc. IEEE CyberSA*, 2023.

[15] N. Moustafa and J. Slay, "UNSW-NB15: A Dataset for Network IDS," in *Proc. IEEE MilCIS*, 2015, pp. 1–6.