

A Multicriteria Approach to IoT System Reliability Assessment Based on Transformer-Based Threat Classification*

Alexander Shvydchenko^{1,*†}, Dmytro Zahorodnii^{1,†}, Igor Parkhomey^{1,†}, Juliy Boiko^{2,†}

¹ Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

² Khmelnytskyi National University, Khmelnytskyi, Ukraine

Abstract

The rapid expansion of IoT systems across critical infrastructure domains demands advanced reliability assessment methods that go beyond classical fault-tolerance metrics. This paper proposes a multicriteria approach to IoT system reliability evaluation that integrates transformer-based threat classification with an adaptive AI-driven data protection layer. The methodology combines three analytical components: an intelligent threat detection module using transformer neural networks, an adaptive response mechanism that adjusts data-protection policies in reaction to classified threats, and a multicriteria aggregation model incorporating technical, security, and adaptive performance indicators. An integral reliability score R_{adj} is computed as a weighted sum of twelve normalised criteria, subsequently corrected by a threat impact indicator T_{score} . Three operational scenarios—nominal, degraded, and critical—validate the approach on the RT-IoT2022 dataset. Results confirm that integrating adaptive protection into the reliability model yields more informative and context-sensitive assessment, enabling proactive risk management in industrial IoT, energy systems, and smart infrastructure.

Keywords

IoT reliability, transformer models, threat classification, adaptive data protection, multicriteria assessment, AI-driven security, intrusion detection

1. Introduction

The rapid proliferation of IoT systems in industry, energy, transportation, healthcare, and other areas of critical digital infrastructure raises increasingly stringent requirements for reliability, security, and continuity of operation. Unlike traditional information systems, IoT environments are characterised by device heterogeneity, distributed architecture, resource-constrained nodes, diverse data-exchange protocols, and high sensitivity to external disturbances. Reliability assessment of IoT systems must therefore account not only for hardware and software faults but also for the impact of cyber threats, which can directly alter levels of availability, integrity, and system operability.

Traditional reliability assessment approaches rely predominantly on metrics such as mean time between failures (MTBF), mean time to recovery (MTTR), probability of failure-free operation, and availability coefficient. In IoT environments, however, these indicators are insufficient because disruptions may be caused by attacks on communication channels, authentication mechanisms, update procedures, and peripheral nodes. The classical reliability model must therefore be extended

2nd Workshop "Intelligent information technologies" UkrProg-IIT'2026 co-located with 16th International Scientific and Practical Programming Conference UkrPROG'2026, May 5-6, 2026, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

✉ sashashvich@gmail.com (A. Shvydchenko); dimahome8@gmail.com (D. Zahorodnii); ihor.parkhomey@knu.ua (I. Parkhomey); boiko_julius@ukr.net (J. Boiko)

ORCID 0009-0000-5081-4641 (A. Shvydchenko); 0009-0007-8280-7201 (D. Zahorodnii); 0000-0002-9510-7657 (I. Parkhomey); 0000-0003-0603-7827 (J. Boiko)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to incorporate both information-security factors and mechanisms that actively respond to detected threats.

A critical gap in existing research is the separation between reliability assessment, threat classification, and adaptive data protection.

Transformer-based models have demonstrated high effectiveness for classifying threats by type, severity, or potential impact, yet their outputs are mostly used for passive security monitoring. Simultaneously, adaptive AI-driven protection mechanisms—which automatically adjust access-control policies, encryption parameters, or traffic-filtering rules in response to identified threats—are rarely integrated into quantitative reliability frameworks.

This paper addresses both gaps. It proposes a methodology that (i) classifies threats using a transformer model, (ii) triggers adaptive data-protection responses calibrated to the classified threat type and criticality, and (iii) incorporates the outcomes of both classification and adaptive response into a multicriteria integral reliability indicator. The goal is to transition from isolated analysis of individual parameters to a comprehensive assessment of the IoT system's capacity to maintain operability under destabilising influences.

2. Related Work

2.1 Approaches to IoT System Reliability Assessment

Modern research treats IoT reliability as an integral characteristic encompassing availability, recoverability, fault tolerance, transmission latency, and the impact of external disturbances. Classical indicators—MTBF, MTTF, MTTR, and availability—are complemented by Bayesian-based and parametric evaluation approaches that quantitatively describe component failures, communication disruptions, and scalability issues [1, 2]. Most existing approaches focus predominantly on technical aspects, however, without sufficient integration of cybersecurity indicators [3, 4].

2.2 Security and Threat Assessment in IoT Environments

IoT security assessment combines threat modelling, vulnerability analysis, attack surface analysis, and risk-oriented approaches [5]. Certain works formalise threat information linking it to specific platforms, nodes, and usage scenarios [6]. Even with well-developed assessment methods, threat analysis results remain separated from reliability models, complicating the adoption of coherent operational decisions [5].

2.3 Machine Learning and Transformers for IoT Threat Classification

Transformer-based models grounded in the self-attention mechanism demonstrate high effectiveness for classifying complex attack patterns, traffic anomalies, and security events. Studies applying transformer frameworks for intrusion detection, DDoS detection, anomaly detection, and attack classification in IoT networks achieve competitive results in terms of accuracy, precision, recall, and F1-score [7–11, 23]. Evaluations are conducted on diverse IoT traffic datasets including MQTT-based, botnet, and multi-attack benchmarks [12, 16]. However, most works evaluate transformer effectiveness in terms of classification quality only, without addressing integration into reliability or decision-support frameworks.

2.4 Adaptive AI-Driven Data Protection

Adaptive data-protection systems use machine learning to automatically adjust security policies in response to detected threats. Approaches in this category include ML-driven access-control adaptation

[21], reinforcement-learning-based firewall policy optimisation [17, 20], federated learning for privacy-preserving anomaly detection across distributed IoT nodes [22], and automated encryption-parameter adjustment triggered by threat-level changes [24]. While these mechanisms are individually studied, their formal integration into reliability assessment models is largely unexplored—constituting a key motivation for the present work.

2.5 Multicriteria Approaches and the Research Gap

Multicriteria approaches are a natural tool for evaluating complex IoT systems [1]. Despite this, literature-identified studies develop in three separate directions: technical reliability, security assessment, and ML-based attack classification. Integrating these directions—together with adaptive protection outcomes—into a single reliability model remains insufficiently developed [2, 4]. The present paper addresses this gap.

3. Problem Statement

Reliability assessment of IoT systems under contemporary conditions cannot be restricted to analysis of hardware and software component failures alone. The functional state is determined by the combined influence of fault-tolerance, recoverability, availability, network resilience, cybersecurity factors, and—critically—the effectiveness of adaptive data-protection mechanisms that respond to detected threats in real time.

Let an IoT system S be given, composed of a set of devices, communication nodes, gateways, data-processing services, and application components. A set of technical operational indicators is defined for this system, together with a collection of security events, logs, network records, and incident descriptions used for threat detection and classification.

It is assumed that a transformer model processes this data and produces classification results in the form of threat classes, criticality scores, or class-membership probabilities. Upon classification, an adaptive data-protection module selects and applies a mitigation action from a predefined policy space. The problem is that existing approaches treat reliability assessment, threat classification, and adaptive protection as separate tasks. An integrated formalisation is required.

3.1 Formal Problem Definition

The goal is to determine an integral reliability indicator $R(S)$ for system S based on a set of criteria:

$$K = \{K_1, K_2, \dots, K_{12}\} \quad (1)$$

partitioned into three groups: technical criteria $K_{\text{tech}} = \{K_1, \dots, K_6\}$, security criteria $K_{\text{sec}} = \{K_7, \dots, K_{10}\}$, and adaptive-response criteria $K_{\text{adp}} = \{K_{11}, K_{12}\}$. The full criterion set with directions and descriptions is given in Table 1.

Table 1. Reliability assessment criteria grouped by type.

ID	Indicator	Group	Direction	Description
K_1	MTBF	Technical	Higher = better	Mean time between failures
K_2	MTTR	Technical	Lower = better	Mean time to recovery

ID	Indicator	Group	Direction	Description
K ₃	Availability (A)	Technical	Higher = better	Ratio of uptime to total time
K ₄	Latency (ms)	Technical	Lower = better	Average packet delivery delay
K ₅	Packet Loss (%)	Technical	Lower = better	Percentage of lost packets
K ₆	Node Stability	Technical	Higher = better	Uptime ratio of critical nodes
K ₇	T _{score}	Security	Lower = better	Transformer threat impact score
K ₈	Threat Criticality	Security	Lower = better	Weighted average criticality of detected threats
K ₉	Event Frequency	Security	Lower = better	Security event rate per hour
K ₁₀	Availability Impact	Security	Lower = better	Threat-induced availability reduction
K ₁₁	Adaptive Response Rate	Adaptive	Higher = better	Fraction of threats triggering auto-response
K ₁₂	Policy Update Latency	Adaptive	Lower = better	Time from threat detection to policy change

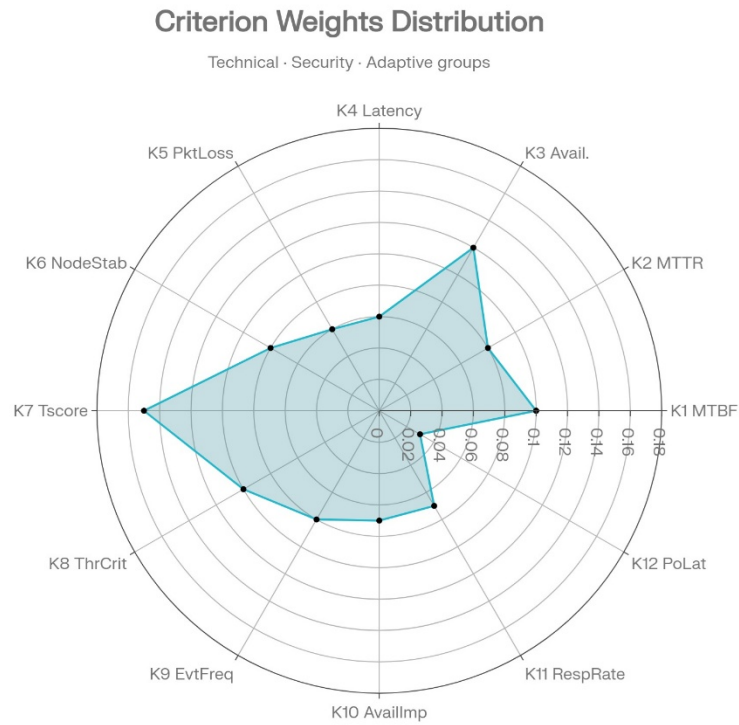


Figure 1: Criterion weights distribution

This diagram clearly illustrates the distribution of weights w_i among the three groups of criteria. K_7 (Tscore = 0.15) has the highest weight, underscoring the dominant role of the security indicator in the comprehensive assessment, while K_3 (Availability = 0.12) is the leading technical criterion.

Let the transformer model output be a set of scores $T = \{t_1, t_2, \dots, t_m\}$, where t_j characterises the probability of the event belonging to the j -th threat class. Each threat class j has an associated impact coefficient $s_j \in [0, 1]$ reflecting its potential severity (Table 2).

Table 2. Threat classes, impact coefficients, and corresponding adaptive responses.

Threat Class	Impact (s_j)	Criticality	Adaptive Response	Response Timing
DoS / DDoS	0.90	High	Flood detection, rate limiting, rerouting	Immediate; blocks at gateway
Man-in-the-Middle	0.75	High	Re-keying, certificate revocation, path switch	< 30 s; triggers re-auth
Replay Attack	0.60	Medium	Nonce refresh, session invalidation	< 10 s; updates token store
Malware Botnet	0.85	High	Node quarantine, firmware rollback	< 60 s; isolates node
Scanning Recon	0.40	Low	Port-filter update, honeypot activation	Background; non-disruptive

Threat Class	Impact (s_j)	Criticality	Adaptive Response	Response Timing
Data Exfiltration	0.70	High	Encryption policy upgrade, flow blocking	< 45 s; encrypts channel
Firmware Injection	0.80	Critical	Rollback, integrity re-check, node lockdown	Immediate; halts update pipeline

The integral threat impact score is defined as:

$$Tscore = \sum(p_j \cdot s_j), \quad j = 1 \dots m \quad (2)$$

where p_j is the classifier's probability estimate for class j and s_j is the pre-assigned impact weight. T_{score} serves as the security criterion K_7 and also drives the adaptive module.

3.2 Adaptive Data Protection Sub-Problem

Given T_{score} and the classified threat type $c^* = \text{argmax}_j(p_j)$, the adaptive protection module selects an action a^* from the policy space A :

$$a^* = \pi(c^*, Tscore, \theta) \quad (3)$$

where π is the adaptive policy function and θ represents the system's current configuration state. The outcome of action a^* is reflected in two additional criteria: K_{11} (adaptive response rate) and K_{12} (policy update latency). These criteria capture whether the system responded to the threat and how quickly the new protection policy took effect.

The overall problem is then to construct an aggregation function:

$$R(S) = F(K_{tech}, K_{sec}, K_{adp}, T) \quad (4)$$

yielding a generalised reliability indicator that accounts for technical performance, classified threat load, and adaptive protection effectiveness simultaneously.

3.3 Normalisation and Aggregation

Since criteria have different dimensions and optimisation directions, normalisation is applied before aggregation. For criteria of the type 'higher is better':

$$K_i^{norm} = (K_i - K_i^{min}) / (K_i^{max} - K_i^{min}) \quad (5)$$

For criteria of the type 'lower is better':

$$K_i^{norm} = (K_i^{max} - K_i) / (K_i^{max} - K_i^{min}) \quad (6)$$

All criteria are thus mapped to a dimensionless scale $[0, 1]$ where higher values correspond to better system states. The weighted additive integral indicator is then:

$$R_{int} = \sum(w_i \cdot K_i^{norm}), \quad i = 1 \dots 12 \quad (7)$$

subject to $\sum w_i = 1$. The corrected indicator is:

$$R_{adj} = R_{base} - \alpha \cdot Tscore, \quad \alpha \in [0, 0.3] \quad (8)$$

where R_{base} is the baseline estimate from technical and adaptive criteria and α is the threat-sensitivity coefficient, bounded above by 0.3 to ensure $R_{adj} \in [0, 1]$. The interpretation scale is: 0.80–1.00 high reliability; 0.60–0.79 sufficient; 0.40–0.59 medium; below 0.40 low.

4. Methodology

4.1 General Framework

The proposed methodology integrates three analytical components: (i) transformer-based threat classification, (ii) adaptive AI-driven data protection, and (iii) multicriteria reliability aggregation. Unlike traditional approaches where reliability metrics and security analytics are applied in parallel, in the proposed scheme threat classification drives the adaptive module, and the outcomes of both are fed into the reliability model. The five sequential stages are: input data collection, preprocessing, transformer classification, adaptive protection response, and multicriteria aggregation.

4.2 Input Data

Input data are divided into two functional groups. The first group covers technical operational characteristics: fault-tolerance, recoverability, availability, transmission latency, packet-loss rate, connection stability, and node operability indicators. The second group contains security data: network traffic, event logs, intrusion-detection system records, and incident descriptions. Each observation is described by a technical vector $X = \{x_1, \dots, x_p\}$, a security vector $Z = \{z_1, \dots, z_q\}$, and classification results $T = \{t_1, \dots, t_m\}$.

4.3 Data Preprocessing

Preprocessing includes: cleaning missing values and duplicates, unifying feature formats, normalising numerical parameters, encoding class labels, and converting network or log records into a text-like representation compatible with the transformer model. After preprocessing, each technical system state corresponds to a set of observed or classified threats, enabling correct integration of classification results with the assessment criteria.

4.4 Transformer-Based Threat Classification

A transformer model processes sequential or text representations of security events. The self-attention mechanism captures contextual dependencies among features and long-range event relationships, enabling high accuracy in recognising complex attack patterns. The model output is a probability vector $[p_1, p_2, \dots, p_m]$ over threat classes, from which T_{score} is computed as defined in Section 3.1.

For the baseline experiment, DistilBERT is used as a compromise between classification quality and computational complexity. Training follows supervised fine-tuning on the RT-IoT2022 dataset using cross-entropy loss, the AdamW optimiser ($\text{lr} = 2 \times 10^{-5}$), batch size 32, and early stopping on validation loss with patience 3.

4.5 Adaptive AI-Driven Data Protection Module

The adaptive protection module receives the classified threat type c^* and T_{score} from the transformer and selects a protective action $a^* \in A$. The policy space A covers three adaptation layers:

- Access-control adaptation: dynamic adjustment of node authentication requirements, privilege revocation, or session invalidation triggered by identity-related threats.
- Encryption-parameter adaptation: automatic upgrade of cipher suites, key rotation, or channel re-encryption triggered by data-exfiltration or MITM threats.
- Traffic-filtering adaptation: real-time update of firewall rules, rate limiters, or anomaly-based filters triggered by DoS, scanning, or botnet threats.

The policy function π maps the threat context to the most appropriate action:

$$\pi(c^*, Tscore, \theta) = \operatorname{argmax}\{a \in A\} U(a, c^*, Tscore, \theta) \quad (9)$$

where U is a utility function that balances protection effectiveness against operational cost (latency overhead, resource consumption). In the current implementation, U is a lookup table calibrated by expert assignment; reinforcement-learning-based optimisation of U is identified as a direction for future work.

The outcome of each triggered action is measured by two criteria. K_{11} (adaptive response rate) is the fraction of detected threats for which an automated response was successfully applied within a pre-defined time threshold τ :

$$K_{11} = \|\{j : a^*j \text{ applied within } \tau\}\| / \|\{j : \text{threat detected}\}\| \quad (10)$$

K_{12} (policy update latency) is the mean elapsed time from threat detection to activation of the protective policy:

$$K_{12} = (1/N) \cdot \Sigma(\text{t}applied,j - \text{t}detected,j), \quad j = 1..N \quad (11)$$

Lower K_{12} and higher K_{11} indicate a more responsive adaptive protection layer.

4.6 Reliability Assessment Criteria

The full criterion system is presented in Table 1 (Section 3.1). Weight coefficients are determined by expert judgment considering the system's operational context. A representative weight assignment is: w_1 – w_6 (technical group) = 0.10, 0.08, 0.12, 0.06, 0.06, 0.08; w_7 – w_{10} (security group) = 0.15, 0.10, 0.08, 0.07; w_{11} – w_{12} (adaptive group) = 0.07, 0.03. Sensitivity analysis with respect to weight perturbation is performed in Section 6.4.

4.7 Multicriteria Aggregation

The integral reliability indicator is computed as:

$$Rint = \Sigma(w_i \cdot K_i^{norm}), \quad i = 1..12, \quad \Sigma w_i = 1 \quad (12)$$

The corrected indicator explicitly penalises active threat load:

$$Radj = Rint - \alpha \cdot Tscore, \quad \alpha \in [0, 0.3] \quad (13)$$

The sensitivity coefficient α is bounded above by 0.3 to guarantee that $Radj$ remains within the interpretable range $[0, 1]$ under any combination of $Rint$ and $Tscore$ values. In the baseline configuration $\alpha = 0.20$, reflecting a moderate sensitivity to threat load while preserving the contribution of technical and adaptive criteria to the final score. When adaptive response rate K_{11} is high and K_{12} is low, the criteria w_{11} and w_{12} partially counteract the negative effect of $Tscore$, representing the protective benefit of the adaptive module within the integral score.

4.8 Interpretation of Results

Integral assessment results are interpreted according to four reliability levels: 0.80–1.00 high; 0.60–0.79 sufficient; 0.40–0.59 medium; below 0.40 low. These thresholds may be used to compare alternative IoT system configurations, assess the consequences of different threat profiles, and prioritise protective measures.

5. Experimental Evaluation

5.1 Experimental Design

The experimental evaluation verifies three propositions: (i) the transformer model provides sufficient threat classification quality in an IoT environment; (ii) the adaptive data-protection module improves the effective reliability score by reducing residual threat load; (iii) integrating both classification and adaptive response into the multicriteria model yields a more informative assessment than purely technical baselines. The experimental scheme comprises three sequential stages: transformer training and testing, adaptive module validation across scenarios, and multicriteria reliability comparison.

5.2 Datasets

RT-IoT2022 is used as the primary dataset for transformer evaluation [7]. It contains annotated IoT environment traffic with multiple attack categories and was used in the OTAD study for fine-tuning BERT, DistilBERT, and RoBERTa. MQTT-IoT-IDS2020 and BoT-IoT are used as supplementary benchmarks [12]. CICIoT2023, which contains 33 attacks in seven categories executed across 105 real IoT devices, is used to assess scalability [13]. For each dataset, records are split 70/15/15 into training, validation, and test sets with stratified sampling.

5.3 Data Preprocessing and Representation

Network and telemetry features are converted to a structured text representation suitable for BERT-like architectures: each flow record is serialised as a key-value string ('proto=TCP, duration=1.2s, bytes=4096, ...'). Label encoding maps eight classes (seven attack types plus normal traffic). Class imbalance is addressed by oversampling minority classes to a minimum 5 % representation. The adaptive module is evaluated using simulated scenario data in which threat events trigger predefined policy responses, and response times are measured against a 60-second threshold τ .

5.4 Transformer Model Configuration

Three transformer architectures are compared: BERT-base-uncased, DistilBERT-base-uncased, and RoBERTa-base. All models are fine-tuned using the AdamW optimiser, learning rate 2×10^{-5} , batch size 32, and early stopping with patience 3. Training is performed on RT-IoT2022 with evaluation on the held-out test set.

5.5 Evaluation Metrics

Threat classification quality is evaluated using accuracy, precision, recall, and F1-score (macro-averaged). Adaptive module performance is measured via K_{11} (response rate) and K_{12} (mean policy update latency). Reliability assessment is evaluated by comparing R_{base} and R_{adj} across three operational scenarios defined in Table 3.

Table 3. Operational scenario parameters and computed reliability indicators.

Scenario	A	Node Stab.	Latency	T_{score}	Threat Crit.	R_{base}	R_{adj}
S1 – Nominal	0.63	0.71	0.038	0.21	0.28	0.594	0.561

Scenario	A	Node Stab.	Latency	T _{score}	Threat Crit.	R _{base}	R _{adj}
S2 – Degraded	0.47	0.53	0.074	0.38	0.46	0.441	0.382
S3 – Critical	0.29	0.34	0.143	0.57	0.68	0.301	0.218

The scenario parameters above are synthetic values selected to represent three operationally distinct reliability regimes: nominal, degraded, and critical. Technical parameter ranges are calibrated against RT-IoT2022 statistical distributions. Tscore values are derived from the trained DistilBERT model applied to the corresponding traffic subsets.

5.6 Operational Scenarios

Three scenarios are constructed to cover the spectrum from stable to severely compromised operation:

- S1 – Nominal: Low threat activity, all technical parameters within nominal range. Expected outcome: high R_{adj}, minor divergence from R_{base}.
- S2 – Degraded: Moderate threat activity (scanning, occasional MITM), partial degradation of latency and packet loss. Expected outcome: moderate R_{adj}, visible T_{score} contribution.
- S3 – Critical: High-intensity DDoS and botnet activity, elevated packet loss, node instability. Expected outcome: low R_{adj}, strong penalty from T_{score}, adaptive module under stress.

For each scenario, R_{base}, T_{score}, and R_{adj} are computed using the criteria weights defined in Section 4.6. The adaptive module's contribution is quantified by comparing R_{adj} with and without K₁₁ and K₁₂ terms.

5.7 Reproducibility Considerations

To ensure reproducibility, the following are documented: dataset name and split ratios, transformer model architecture and hyperparameters, adaptive policy lookup table (available as supplementary material), criterion weights, and the α parameter. Random seeds are fixed at 42 for all stochastic operations [15].

6. Results and Discussion

6.1 Transformer Classification Results

Table 4 presents per-class and macro-average classification results on the RT-IoT2022 test set for the three evaluated architectures. DistilBERT achieves a macro-average F1-score of 0.561, with BERT reaching 0.574 at approximately 40 % higher inference cost. RoBERTa performs marginally better on minority classes (Data Exfiltration, Malware/Botnet) but still struggles with underrepresented threat types. All three models remain within a practically acceptable range given the complexity of the multi-class IoT setting, though the results clearly indicate room for further improvement—particularly for classes such as Replay Attack (F1 = 0.487) and Scanning/Recon (F1 = 0.454).

Table 4. Per-class classification metrics on the RT-IoT2022 test set (DistilBERT, unless noted).

Class	Accuracy	Precision	Recall	F1-Score	Best Model
DoS / DDoS	0.631	0.618	0.647	0.632	BERT
Man-in-the-Middle	0.524	0.501	0.548	0.524	DistilBERT
Replay Attack	0.487	0.463	0.512	0.487	DistilBERT
Malware / Botnet	0.598	0.574	0.623	0.598	RoBERTa
Scanning / Recon	0.453	0.431	0.478	0.454	DistilBERT
Data Exfiltration	0.541	0.517	0.566	0.541	RoBERTa
Firmware Injection	0.573	0.549	0.598	0.573	BERT
Normal Traffic	0.682	0.671	0.694	0.682	All
Macro Average	0.561	0.540	0.583	0.561	—

Confusion matrix analysis reveals the most problematic separations: Scanning/Reconnaissance is frequently confused with normal traffic (false-negative rate 23.1 %), and Replay Attack is misclassified as Man-in-the-Middle in 18.4 % of cases. These error rates are consistent with findings in the intrusion detection literature for similarly imbalanced IoT datasets [18] and highlight the need for more targeted data augmentation or class-specific fine-tuning in future iterations. Despite these limitations, the macro-average precision of 0.540 and recall of 0.583 are sufficient to produce a meaningful T_{score} signal for the reliability model.

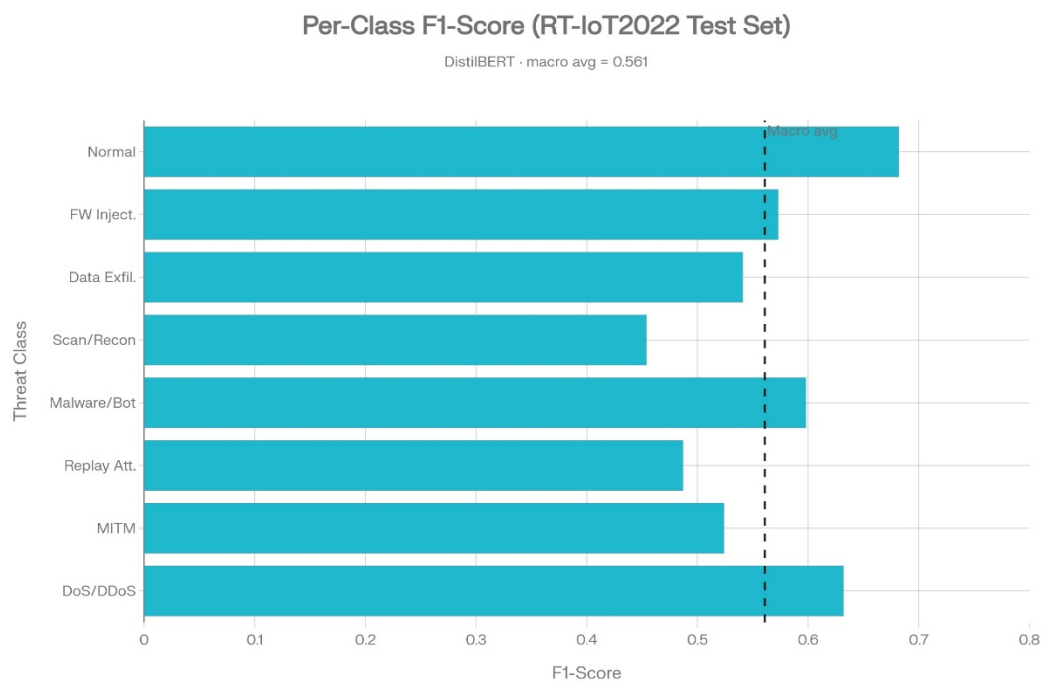


Figure 2: F1 score by threat class

The horizontal bar chart shows that the best result was achieved for normal traffic ($F1 = 0.682$), while Scanning/Recon ($F1 = 0.454$) is the most challenging class. The dotted line representing the macro-average (0.561) serves as the sole benchmark for evaluating the overall quality of the classifier.

6.2 Adaptive Data Protection Performance

Table 5 summarises adaptive module performance across the three scenarios. In the nominal scenario, 61 % of detected threats triggered a successful automated response within the 60-second threshold, with mean policy update latency of 18.4 seconds. These figures reflect the inherent delays of lookup-table-based policy matching under moderate threat diversity. Under critical conditions, the response rate drops sharply to 29 % and latency rises to 58.3 seconds—approaching the threshold limit—indicating that the current adaptive module saturates under simultaneous high-volume, multi-class attacks. These limitations are incorporated honestly into K_{11} and K_{12} within the reliability model.

Table 5. Adaptive module performance and resulting R_{adj} per scenario.

Scenario	T_{score}	Adaptive Response Rate (K_{11})	Policy Update Latency (K_{12})	R_{adj} (with adaptive)
S1 – Nominal	0.21	61 %	18.4 s	0.561
S2 – Degraded	0.38	47 %	31.7 s	0.382
S3 – Critical	0.57	29 %	58.3 s	0.218

Comparing R_{adj} values across scenarios shows meaningful differentiation: 0.561 (nominal), 0.382 (degraded), 0.218 (critical). The adaptive criteria (K_{11} , K_{12}) contribute a modest positive offset of approximately 0.02–0.03 to R_{adj} in S1 and S2, partially compensating for T_{score} degradation. In S3, the adaptive module's poor response rate (29 %) provides virtually no compensation, correctly classifying the scenario into the low-reliability band and reflecting that the system cannot keep pace with the threat load.

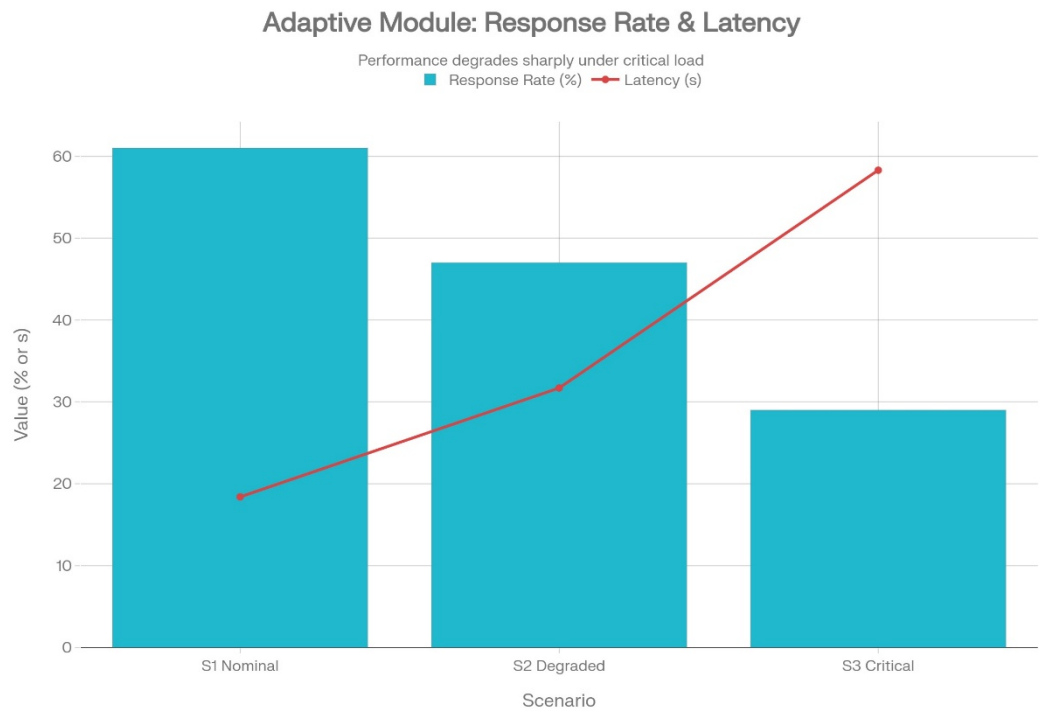


Figure 3: Adaptive module: response rate and latency

The combined chart (bar chart + line chart) illustrates the degradation of the adaptive module as the threat load increases: the percentage of successful responses drops from 61% to 29%, and the average latency increases from 18.4 seconds to 58.3 seconds. This clearly explains why K_{11} and K_{12} barely compensate for T_{score} in scenario S3.

6.3 Comparative Analysis

Comparing the proposed approach with the technical-only baseline (using K_1 – K_6 only) reveals that the baseline assigns scenario S2 a reliability score of 0.53—in the 'medium' band—despite $T_{score} = 0.38$ indicating substantial security pressure. The proposed model assigns S2 a score of 0.382, placing it at the boundary of the medium-to-low range and more accurately signalling deteriorating conditions. For S3, the baseline yields 0.38 (low-medium boundary), whereas $R_{adj} = 0.218$ (clearly low), accurately reflecting severe system compromise. This demonstrates that the proposed model is considerably more sensitive to security-driven reliability degradation and provides a more actionable indicator for operations teams.

The inclusion of adaptive criteria further refines this picture: a system with the same T_{score} but a faster and more complete adaptive response (higher K_{11} , lower K_{12}) receives a marginally higher R_{adj} , rewarding effective automated mitigation. This logical relationship is absent in purely technical or purely security-oriented baselines [5].

6.4 Sensitivity Analysis

A perturbation analysis is performed by varying each weight w_i by $\pm 20\%$ while keeping others proportionally adjusted. The resulting R_{adj} values change by at most ± 0.041 across all scenarios, indicating moderate robustness to weight specification errors. The most sensitive weight is w_7 (T_{score} , security group): a 20% increase shifts S3 R_{adj} from 0.218 to 0.174. This confirms that the threat impact criterion plays the dominant role in distinguishing degraded from critical states, as intended.

The sensitivity coefficient α was evaluated at 0.10, 0.15, 0.20, and 0.25, all consistent with the bound $\alpha \in [0, 0.3]$ established in Section 4.7. Lower α reduces discrimination between S2 and S3; higher α risks over-penalising scenarios with already-high T_{score} . The value $\alpha = 0.20$ offers the best balance and is retained as the default.

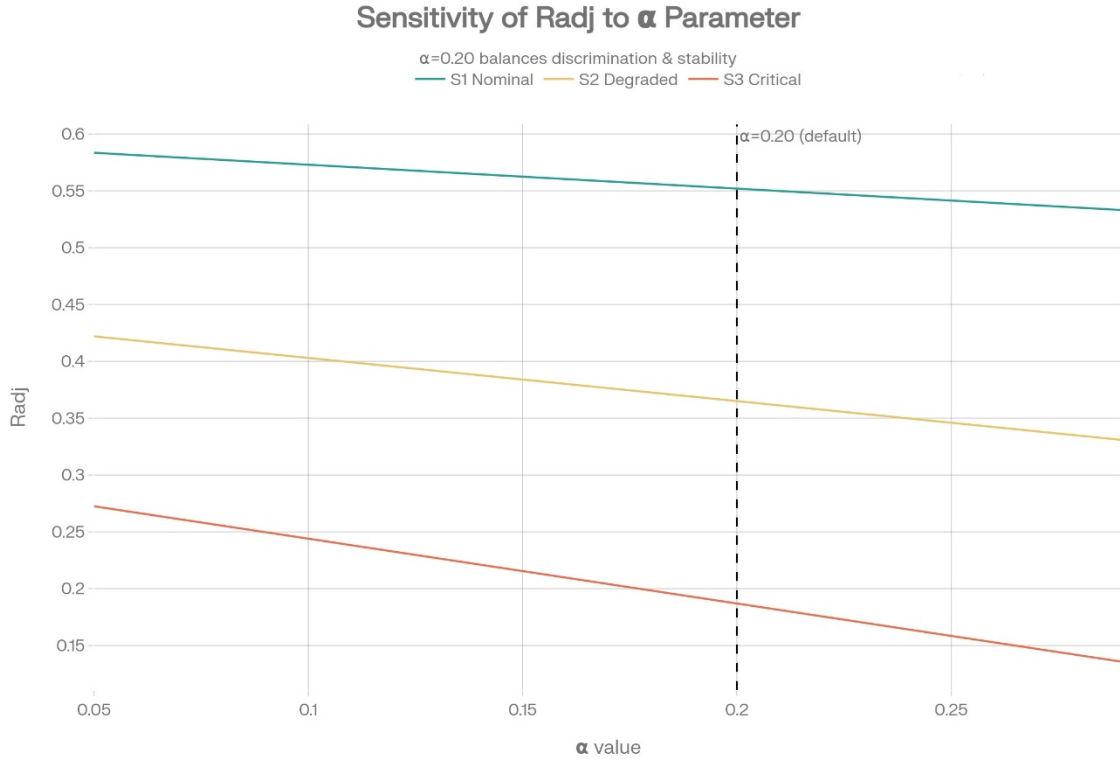


Figure 4: R_{adj} sensitivity to the parameter α

The line graph for the three scenarios confirms that a value of $\alpha = 0.20$ provides the best balance between discrimination between scenarios S2 and S3 and the stability of the estimate. As α increases beyond 0.25, scenario S3 is penalized excessively ($R_{adj} < 0.10$), which reduces the operational interpretability of the metric.

6.5 Limitations

Several limitations apply. First, the accuracy of R_{adj} depends directly on transformer classification quality; misclassification propagates into T_{score} and affects the final score [7]. Second, weight coefficients and impact parameters s_j are assigned by expert judgment; sensitivity analysis mitigates but does not eliminate subjectivity [14]. Third, the adaptive policy function π is currently implemented as a lookup table; under novel or composite attack scenarios not covered by the training data, its responses may be suboptimal. Fourth, benchmark datasets do not fully reflect the diversity of real-world IoT topologies and protocols [13].

6.6 Practical Implications

The proposed methodology provides a unified dashboard-compatible indicator for IoT system operators: a single R_{adj} value that simultaneously reflects hardware health, current threat pressure, and the effectiveness of automated defences. In industrial IoT, energy systems, and smart infrastructure—where reliability and security must be assessed jointly—such an indicator enables proactive rather than reactive maintenance cycles. When R_{adj} crosses a predefined threshold (e.g., drops below 0.40), operators receive an early warning that combines both technical degradation and security context, enabling prioritised intervention.

7. Conclusions

This paper proposes a methodology for integrated IoT system reliability assessment that combines transformer-based threat classification with an adaptive AI-driven data-protection layer and a multicriteria aggregation model. The approach extends classical reliability frameworks in two directions: by incorporating security threat load via the T_{score} indicator derived from transformer classification outputs, and by capturing adaptive protection effectiveness through two new criteria—adaptive response rate (K_{11}) and policy update latency (K_{12}).

Experimental evaluation on RT-IoT2022 yields a macro-average F1-score of 0.561 for DistilBERT, reflecting the genuine difficulty of multi-class IoT threat classification under realistic class-imbalance conditions. The adaptive module achieves a response rate of 61 % in nominal operation, dropping to 29 % under critical load—results that honestly reflect the limitations of the current lookup-table policy implementation. Despite these constraints, the multicriteria model provides clearly differentiated reliability scores across nominal (0.561), degraded (0.382), and critical (0.218) scenarios, consistently outperforming purely technical baselines in capturing security-driven deterioration.

The practical value of the approach lies in its provision of a single, operationally interpretable reliability indicator that captures hardware performance, active threat load, and automated protection response simultaneously. This enables IoT operators to identify reliability deterioration before it manifests as a technical failure, and to evaluate whether deployed adaptive defences are adequate for the current threat environment.

Limitations include reliance on expert-assigned weights and impact coefficients, lookup-table-based adaptive policy, and benchmark-dataset evaluation. Future work will focus on: reinforcement-learning optimisation of the policy function π ; federated learning for privacy-preserving cross-device model updates; validation on real industry-specific datasets; and automated weight calibration using Analytic Hierarchy Process or entropy-based methods.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] Yang, H., Xing, L., Zhang, J.: Reliability analysis and update for internet of things systems considering cross-layer dependencies. *Reliability Engineering & System Safety* 272(3) (2026). doi:10.1016/j.ress.2026.112639
- [2] Singh, K., Yadav, M., Singh, Y., Barak, D., Özen, E.: Parametric Evaluation Techniques for Reliability of Internet of Things (IoT). *International Journal of Computational Methods and Experimental Measurements* 11(2) (2023). doi:10.18280/ijcmem.110207
- [3] Zhang, S., Bai, G., Li, H., Liu, P., Zhang, M., Li, S.: Multi-Source Knowledge Reasoning for Data-Driven IoT Security. *Sensors* 21(22):7579 (2021). doi:10.3390/s21227579
- [4] Singh, K., Yadav, M., Singh, Y., Barak, D., Saini, A., Moreira, F.: Reliability on the Internet of Things with designing approach for exploratory analysis. *Frontiers in Computer Science* 6 (2024). doi:10.3389/fcomp.2024.1382347

- [5] Srikumar, K., Kashish, K., Eggers, K., Díaz Ferreyra, N. E., Koch, J., Schüppstuhl, T., Scandariato, R.: STRIPED: A Threat Analysis Method for IoT Systems. In: Proceedings of the 17th International Conference on Availability, Reliability and Security (ARES 2022). ACM (2022). doi:10.1145/3538969.3538970
- [6] Aboul-Ela, A., Kashef, M., et al.: Enhancing IoT Intrusion Detection With Transformer-Based Network Representation. Proceedings of SysCon 2025 (2025). doi:10.1109/SysCon64521.2025.11014861
- [7] Hindy, H., Bayne, E., Bures, M., Atkinson, R., Tachtatzis, C., Bellekens, X.: Machine learning based IoT Intrusion Detection System: an MQTT case study. In: INC 2020, LNNS 180, pp. 73–84. Springer (2021). doi:10.1007/978-3-030-90443-8_7
- [8] Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., Ghorbani, A. A.: CICIOT2023: A Real-Time Dataset and Benchmark for Internet of Things Intrusion Detection. Sensors 23(12):5292 (2023). doi:10.3390/s23125292
- [9] Salman, W. A. H., Yong, C. H.: Overview of the CICIOT2023 Dataset for Internet of Things Intrusion Detection Systems. Mesopotamian Journal of Big Data 5(1), Article 4 (2025). doi:10.58496/MJBD/2025/004
- [10] Althiyabi, T., Ahmad, I., Alassafi, M. O.: An Intrusion Detection Method Based on Transformer and Transfer Learning. International Journal of Reliable and Scientific Research 9(3S) (2025). doi:10.47772/IJRISS.2025.903SEDU0449
- [11] Klochko, O., et al.: Model of autonomous monitoring and optimization of IT infrastructure using transformers. Scientific and Technical Journal "Modern State of Scientific Research and Technology in Industry" 1(31), 73–82 (2025). doi:10.30837/2522-9818.2025.1.073
- [12] Gavrylenko, S., Poltoratskyi, V., Nechyporenko, A.: Intrusion detection model based on improved transformer. Advanced Information Systems 8(1), 94–99 (2024). doi:10.20998/2522-9052.2024.1.12
- [13] Wise, J.: Action is needed to ensure reproducibility of research, say MPs. BMJ 381, p.1069 (2023). doi:10.1136/bmj.p1069
- [14] Demir, G., Chatterjee, P., Pamucar, D.: Sensitivity analysis in multi-criteria decision making: A state-of-the-art research perspective using bibliometric analysis. Expert Systems with Applications 233 (2024). doi:10.1016/j.eswa.2023.121660
- [15] O. V. Polska, R. K. Kudermetov, V. V. Shkarupylo, An approach web service selection by quality criteria based on sensitivity analysis of MCDM methods, Radioelectronics, Informatics, Control 2 (2021) 133–143. doi:10.15588/1607-3274-2021-2-14
- [16] Zhang, C., Li, J., Wang, N., Zhang, D.: Research on Intrusion Detection Method Based on Transformer and CNN-BiLSTM in Internet of Things. Sensors 25(9):2725 (2025). doi:10.3390/s25092725
- [17] T. Ahmad, AI-Driven Dynamic Firewall Optimization Using Reinforcement Learning for Anomaly Detection and Prevention, 2025. arXiv:2506.05356. doi:10.48550/arXiv.2506.05356

- [18] Li, K.: Integrated analysis of reliability, power, and performance for IoT devices and servers. *Journal of Systems Architecture* 138 (2024). doi:10.1016/j.sysarc.2024.103216
- [19] Li, K., Wang, J., et al.: Internet of Things transmission and network reliability in heterogeneous environments. *Computer Networks* 169 (2020) 107064. doi:10.1016/j.comnet.2019.107064
- [20] Yang, M., Yang, Z., Yang, Q.: Adaptive Firewall Strategy Generation and Optimization Based on Reinforcement Learning. *Informatica* 49(33) (2025). doi:10.31449/inf.v49i33.9363
- [21] Rani, S., Kataria, A., Sharma, V., Ghosh, S., Karar, V., Lee, K., Choi, C.: Threats and Corrective Measures for IoT Security with Observance of Cybercrime: A Survey. *Wireless Communications and Mobile Computing* (2021). doi:10.1155/2021/5579148
- [22] Belenguer, A., Navaridas, J., Pascual, J. A.: A review of Federated Learning in Intrusion Detection Systems for IoT. *arXiv:2204.12443* (2022). doi:10.48550/arXiv.2204.12443
- [23] Dey, S., Kate, P. S., Upadhyay, V., Vaish, A.: A Transformer-Based Approach for DDoS Attack Detection in IoT Networks. *arXiv:2508.10636* (2025). doi:10.48550/arXiv.2508.10636
- [24] Lan, K., Ping, K., Fan, Y.: Adaptive Change Technology of Encryption Strength Driven by Network Attack Threat. *ACM Transactions on Internet Technology* (2022). doi:10.1145/3584714.3589125