



ADAPTIVE CREDIBILITY-ENHANCED TRUST MANAGEMENT FOR RESILIENT DEVICE INTERACTIONS IN THE SOCIAL INTERNET OF THINGS

¹Etudaiye Abdulmumuni Isah, ¹Fabiya Ademola Odunayo, ¹Lawal Moshood Olatunji

¹Department of Cybersecurity, Federal Polytechnic, Ede, Nigeria

Correspondence Author: etudaiye.abdulmumuni@federalpolyede.edu.ng

Abstract: *The Social Internet of Things (SIoT) introduces social relationships among connected objects to support discovery, cooperation, and service interaction. These relationships create a trust-management challenge because direct evidence may be limited and recommendations can be manipulated. This paper presents an Adaptive Credibility-Enhanced Trust (ACET) model that combines direct trust, credibility-weighted recommendation trust, and social trust with incentive, penalty, and recovery mechanisms. Recommendation credibility is normalized using Softmax weighting, while normalized entropy and weighted disagreement characterize recommendation consistency. The model is evaluated in a controlled simulation of 100 nodes using a Barabási-Albert topology, demand-based recommendations, 500 rounds, four attack conditions, three malicious-node fractions, and 30 seeds. Five configurations are compared: DT, DT+RT, DT+ST, Fixed Hybrid, and ACET, producing 1,800 paired runs. Across all conditions, DT records 93.35% mean accuracy and MAE of 0.2123, whereas ACET records 89.97% accuracy and MAE of 0.3011. ACET nevertheless exceeds recommendation-heavy and fixed-fusion configurations under Bad-Mouthing, Collusion, and the evaluated Hybrid condition, while its relative performance declines under On-Off behaviour. The findings indicate that multi-source trust integration is attack-dependent rather than universally superior and that recommender credibility can reduce the effect of manipulated recommendations when indirect evidence is required.*

Keywords: Social Internet of Things; trust management; recommendation trust; social trust; credibility; Softmax; Bad-Mouthing; Collusion; On-Off.

1. Introduction

The Internet of Things (IoT) connects heterogeneous sensing, computing, communication, and actuation devices that increasingly make autonomous interaction decisions. The Social Internet of Things (SIoT) extends this environment by allowing objects to establish social relationships and exploit those relationships for discovery, navigation, cooperation, and service selection (Atzori et al., 2012). Consequently, an interaction decision may depend not only on a device's own experience but also on information obtained from other objects.

Trust management is important in this setting because a requester may have insufficient direct experience with a provider. Recommendation mechanisms can extend the evidence available to the requester, but they also create an attack surface. A malicious recommender can submit negative reports about honest objects, promote malicious objects, or coordinate with other recommenders. Social relationships provide contextual evidence, but social proximity does not necessarily establish behavioural reliability. Surveys of SIoT trust management identify direct evidence, indirect evidence, social attributes, trust aggregation, updating, and resistance to attacks as recurring design dimensions (Alam et al., 2022; Chahal et al., 2020; Sagar et al., 2024).

The central problem is how to control the influence of heterogeneous trust sources when evidence quality differs. Direct observations are first-hand but may be sparse; recommendations increase coverage but may be manipulated; and social information provides context without necessarily indicating service reliability. The model therefore preserves these distinctions rather than assuming equal source influence.

This paper proposes Adaptive Credibility-Enhanced Trust (ACET), which combines direct, recommendation, and social evidence. The recommendation component assigns weights using recommender credibility through Softmax normalization. Normalized entropy and weighted disagreement describe the concentration and consistency of recommendation evidence. Social trust

uses friendship, community, and interaction-role attributes. The aggregated trust state is updated with explicit incentive, penalty, and recovery mechanisms.

The study makes four contributions: a credibility-conditioned multi-source trust model; controlled comparison against component-level baselines under four adversarial conditions and three malicious-node fractions; joint evaluation of classification and continuous trust-estimation error; and paired seed-level statistical analysis with confidence intervals, effect sizes, and Holm correction.

2. Related Work

SIoT research introduced social networking principles into IoT so that objects can establish relationships and exploit them for improved discovery and interaction (Atzori et al., 2012).

Trustworthiness management subsequently became an important mechanism for deciding whether another object should be relied upon. (Nitti et al.(2014) demonstrated the importance of combining behavioural and social information in SIoT trustworthiness management.

Trust-management studies distinguish direct trust from recommendation-based trust. Direct trust is closely connected to observed behaviour but may provide incomplete coverage, whereas indirect trust extends coverage while making recommender reliability part of the security problem. (Chahal et al. 2020) and (Alam et al. 2022) identify recommendation manipulation and trust aggregation as recurring SIoT challenges.

Social trust adds relationship strength, community membership, and functional roles. (Bangui et al. 2023) note that the meaning of social relationships varies across application domains, while (Sagar et al. 2024) identify multi-dimensional trust, aggregation, and resilience as continuing research issues. Attack resilience is also important because adversaries can manipulate either observed behaviour or the information used to evaluate it; Bad-Mouthing, collusion, and On-Off behaviour represent these two broad attack surfaces (Magdich et al., 2022; Ud Din & Guizani, 2019).

Attack resilience is particularly important because adversaries can manipulate either the behaviour being observed or the information used to evaluate it. Bad-Mouthing attacks generate negative recommendations against honest targets, collusion promotes malicious targets through coordinated recommendations, and On-Off attacks alternate cooperative and uncooperative behaviour. (Magdich et al. 2022) demonstrate the continuing importance of resilience against trust-related attacks, while broader IoT trust surveys emphasize trust updating and attack resistance (Ud Din & Guizani, 2019).

3. Proposed ACET Trust Management Model

3.1 Direct and Recommendation Trust

For requester i and provider j , direct trust is updated from interaction outcomes O at time t using an exponentially weighted history:

$$DT_{ij}^{(t)} = \alpha DT_{ij}^{(t-1)} + (1-\alpha)O_{ij}^{(t)}, \quad \alpha = 0.70$$

When direct history is insufficient, eligible neighbours provide recommendation signals r_k . Each recommender has credibility c_k . ACET transforms credibility into normalized weights using:

$$w_k = \exp(c_k/\tau) / \sum_m \exp(c_m/\tau), \quad \tau = 1.0$$

Recommendation trust is then:

$$RT_{ij} = \sum_k w_k r_k$$

The experimental implementation permits up to eight recommenders. Unlike an unweighted neighbour average, this formulation reduces the influence of recommenders with lower credibility while preserving the relative ordering of credible sources.

3.2 Recommender Credibility

For each requester i and recommender k , credibility combines recommendation agreement, recommender experience with the target, and the requester–recommender social relationship. The implementation uses weights 0.45, 0.30, and 0.25, respectively.

Agreement is updated from observed recommendation/outcome agreement using $\text{credibility_alpha} = 0.70$. Experience is smoothed using the direct-evidence constant $k = 5.0$.

$$C_{ik} = 0.45A_{ik} + 0.30E_{kj} + 0.25S_{ik}$$

where A is the exponentially updated recommendation/outcome agreement, E is recommender experience with target j, and S is the social relationship value between requester i and recommender k.

$$E_{kj} = n_{kj} / (n_{kj} + 5)$$

3.3 Recommendation Consistency

Recommendation concentration is represented by normalized entropy:

$$H = -\sum_k w_k \ln(w_k) / \ln(M)$$

Weighted disagreement around the recommendation trust value is:

$$D = \sum_k w_k (r_k - RT_{ij})^2$$

The implementation combines normalized entropy and disagreement to produce a confidence indicator. With normalized entropy H and weighted disagreement D, disagreement is normalized by the maximum Bernoulli variance scale 0.25 and clipped to [0,1]. Confidence is an evidence-consistency indicator rather than a calibrated probability of correctness.

$$Q = 1 - H(D/0.25)$$

Because confidence measures agreement and concentration rather than truthfulness, unanimous malicious recommendations can still produce high confidence.

3.4 Social Trust and Fixed Hybrid Aggregation

Social trust is represented using friendship relationship strength (FR), community relationship (CR), and interaction role (IR):

$$ST_{ij} = 0.3FR_{ij} + 0.4CR_{ij} + 0.3IR_{ij}$$

The fixed-hybrid configuration uses the base aggregation weights shown below. ACET uses these values as its base weights but subsequently derives effective weights from direct-evidence availability and recommendation confidence.

$$T_{ij} = 0.4DT_{ij} + 0.3RT_{ij} + 0.3ST_{ij} \text{ (Fixed Hybrid)}$$

3.5 Adaptive ACET Fusion and Post-Interaction Update

ACET does not use the fixed 0.4/0.3/0.3 weights as its final fusion weights. Instead, direct evidence progressively increases the effective direct-trust contribution, while recommendation and social evidence are reduced as direct history accumulates. Recommendation influence is additionally modulated by confidence. Here n is the requester–target direct-interaction history and k = 5.0 is the direct-evidence smoothing constant.

$$E = n / (n + k)$$

$$R_DT = 0.4 + 0.6E; \quad R_RT = 0.3(1 - E)Q; \quad R_ST = 0.3(1 - E)$$

$$w_s = R_s / (R_DT + R_RT + R_ST), \quad s \in \{DT, RT, ST\}$$

$$T_{ij} = w_DT DT_{ij} + w_RT RT_{ij} + w_ST ST_{ij}$$

After fusion, the implementation applies the configured incentive or penalty, clips trust to [0,1], and then applies threshold-directed recovery when trust is below $\tau_d = 0.60$. The resulting post-adaptation trust is used for classification.

Following an interaction, successful behaviour receives an incentive of 0.05 and failure or malicious behaviour receives a penalty of 0.10. Values are bounded to [0,1]. When trust is below the decision threshold $\tau_d = 0.6$, recovery moves the value toward that threshold:

$$T_{ij} \leftarrow T_{ij} + \rho(\tau_d - T_{ij}), \quad \text{if } T_{ij} < \tau_d, \quad \rho = 0.10, \quad \tau_d = 0.60$$

The provider is accepted when the resulting trust value is at least 0.6.

3.6 Evaluated Configurations

Configuration	Evidence components	Adaptive mechanisms
DT	Direct	None
DT+RT	Direct + recommendation	No credibility adaptation
DT+ST	Direct + social	No recommendation
Fixed Hybrid	Direct + recommendation + social	Fixed 0.4/0.3/0.3 fusion
ACET	Direct + recommendation + social	Adaptive credibility, entropy/disagreement, evidence-dependent fusion, incentive, penalty, recovery

4. Experimental Methodology

4.1 Design and Environment

The final experiment uses 100 nodes in a Barabási-Albert preferential-attachment network. Demand-based recommendation is used so that recommendations are obtained from eligible neighbouring objects rather than from an unrestricted global pool. Each run contains 500 rounds and up to 20 interaction opportunities per round. Four attack conditions, three malicious-node fractions, and 30 random seeds are evaluated for each of five configurations, giving 1,800 runs.

The five configurations are evaluated using paired interaction traces. For a given seed and attack/fraction condition, requester-provider selections, outcomes, and malicious status are checked for equality across configurations. This isolates differences in trust processing from differences in the underlying interaction sequence.

4.2 Attack Conditions

Bad-Mouthing causes malicious recommenders to issue negative recommendations about honest targets. Collusion causes malicious recommenders to issue positive recommendations about malicious targets. On-Off behaviour alternates cooperative and uncooperative provider states. The simulator uses a 0.05 toggle probability and a 0.90 success probability during cooperative states. The Hybrid condition combines Bad-Mouthing recommendation manipulation with On-Off provider behaviour; it is not a simultaneous implementation of all four attacks.

Malicious-node fractions are 10%, 20%, and 30%. Social attributes are simulated through community and role assignments. The principal network size is fixed at 100 nodes, so the experiment evaluates behaviour at that scale rather than establishing empirical scalability to larger networks.

4.3 Parameters and Metrics

Parameter	Value	Parameter	Value
Nodes	100	Rounds	500
Seeds	30	Interactions/round	20
Topology	Barabási-Albert	Malicious fractions	10/20/30%
DT α	0.70	Threshold	0.60
Social weights	0.3/0.4/0.3	Base aggregation	0.4/0.3/0.3
Incentive	0.05	Penalty	0.10
Recovery	0.10	Max recommenders	8
Credibility initial	0.50	Credibility α	0.70
Recommendation τ	1.0	Direct evidence k	5.0

Classification metrics are accuracy, precision, recall, F1, and false-positive rate (FPR). Trust-estimation quality is measured using mean absolute error (MAE). For each interaction, $y=1$ denotes a malicious provider and $y=0$ an honest provider; the ideal trust target is therefore $1-y$. MAE is computed as the mean of $|\text{trust} - (1-y)|$ across evaluation interactions. Mean trust, recommendation

use, and confidence are also recorded. Lower MAE and FPR indicate smaller error, while higher accuracy, precision, recall, and F1 indicate stronger classification performance.

4.4 Statistical Analysis

Because observations are paired across configurations, seed-level paired comparisons are used. The effects analysis reports mean paired differences, 95% confidence intervals, paired p-values, Cohen's d_z , and win rates. Holm correction is applied across the 288 planned comparisons. This provides a family-wise correction while retaining the paired experimental structure.

5. Results

5.1 Overall Performance

Across all experimental conditions, DT produces the highest mean accuracy and lowest MAE. DT achieves 93.35% accuracy and MAE of 0.2123. DT+ST achieves 93.01% accuracy, ACET 89.97%, DT+RT 87.94%, and Fixed Hybrid 85.51%. The results show that adding information sources does not automatically improve decisions.

Model	Accuracy (%)	Precision	Recall	F1	FPR	MAE
DT	93.35	0.768	0.866	0.806	0.050	0.2123
DT+RT	87.94	0.628	0.890	0.726	0.128	0.2626
DT+ST	93.01	0.756	0.864	0.797	0.054	0.3196
Fixed Hybrid	85.51	0.583	0.907	0.697	0.165	0.3303
ACET	89.97	0.674	0.855	0.743	0.092	0.3011

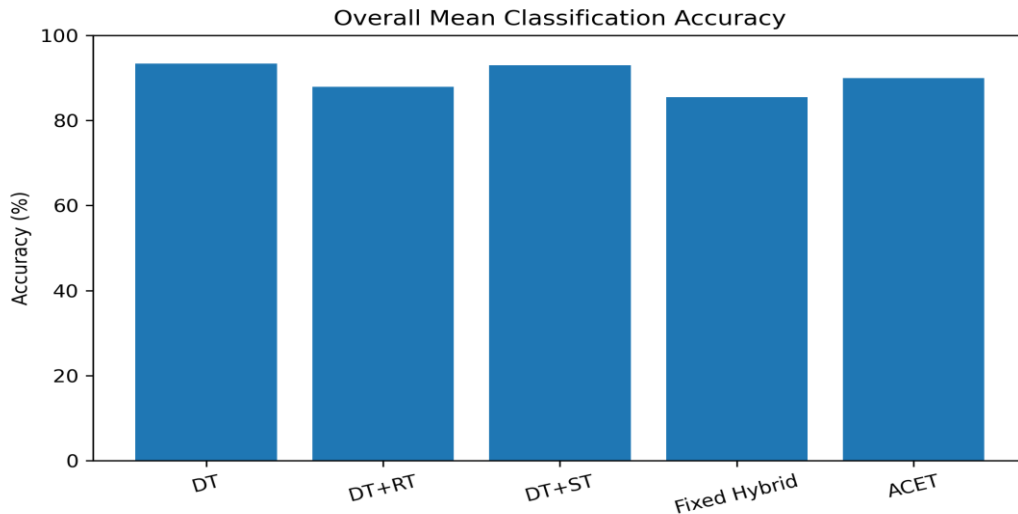


Figure 1. Overall mean classification accuracy (0–100% axis).

DT produces the highest mean accuracy at 93.35%, followed by DT+ST at 93.01%, ACET at 89.97%, DT+RT at 87.94%, and Fixed Hybrid at 85.51%.

The corresponding overall mean internal trust states are 0.7179 for DT, 0.6734 for DT+RT, 0.6415 for ACET, 0.6276 for DT+ST, and 0.6171 for Fixed Hybrid. These values describe the trust states generated by the models and should not be interpreted as independent measures of model quality. Classification metrics and MAE therefore provide the principal basis for performance interpretation.

5.2 Recommendation-Manipulation Attacks

Under Bad-Mouthing, ACET records 93.09%, 91.41%, and 90.53% accuracy at 10%, 20%, and 30% malicious fractions, respectively. DT+RT records 91.13%, 87.46%, and 85.60%, while Fixed Hybrid records 88.93%, 83.54%, and 80.53%. Averaged across the three malicious fractions, these correspond to 91.68%, 88.06%, and 84.33%, respectively, as shown in Table 1. Under Collusion, ACET records 93.09%, 91.44%, and 90.58% accuracy at 10%, 20%, and 30% malicious fractions,

respectively. DT+RT records 91.14%, 87.51%, and 85.71%, while Fixed Hybrid records 88.98%, 83.66%, and 80.73%. Averaged across the three malicious fractions, the corresponding values are 91.71%, 88.12%, and 84.46%.

Attack	ACET	DT	DT+RT	DT+ST	Fixed Hybrid
Bad-Mouthing	91.68	95.91	88.06	95.61	84.33
Collusion	91.71	95.91	88.12	95.61	84.46

Table1: Mean classification accuracy by attack condition, averaged across the 10%, 20%, and 30% malicious fractions.

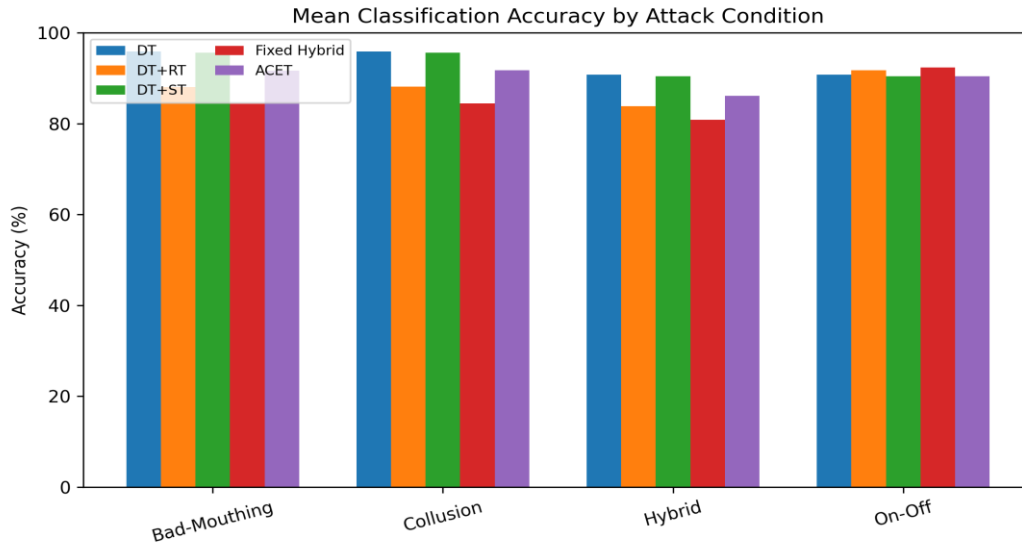


Figure 2. Mean classification accuracy by attack condition (0–100% axis).

At the 20% malicious fraction, the paired comparison between ACET and DT+RT shows a mean accuracy difference of 0.0396 under Bad-Mouthing, with a 95% confidence interval of [0.0320, 0.0471] and Cohen's ($d_z=1.96$). Under Collusion, the corresponding mean difference is 0.0393, with a 95% confidence interval of [0.0318, 0.0467] and Cohen's ($d_z=1.97$). Both paired comparisons remain statistically significant after Holm correction across the 288 planned comparisons ($(p<.001)$).

5.3 Hybrid and On-Off Attacks

Under the Hybrid condition, ACET records 90.43%, 85.79%, and 82.17% accuracy at 10%, 20%, and 30% malicious fractions, compared with 89.16%, 83.17%, and 79.28% for DT+RT; the corresponding differences are 1.26, 2.62, and 2.88 percentage points. Under On-Off, ACET records 92.74%, 90.38%, and 88.03%, while DT+RT records 93.40%, 91.66%, and 90.06% and Fixed Hybrid 93.66%, 92.36%, and 91.11%. At 30%, ACET is approximately 2.03 percentage points below DT+RT.

Attack	ACET	DT	DT+RT	DT+ST	Fixed Hybrid
Hybrid	86.13	90.78	83.87	90.42	80.87
On-Off	90.38	90.78	91.70	90.42	92.38

The attack-specific results show that credibility-aware recommendation processing addresses recommendation manipulation more directly than temporal provider instability. A provider can accumulate positive evidence during cooperative periods and later change state, which can challenge a fixed direct-trust decay parameter.

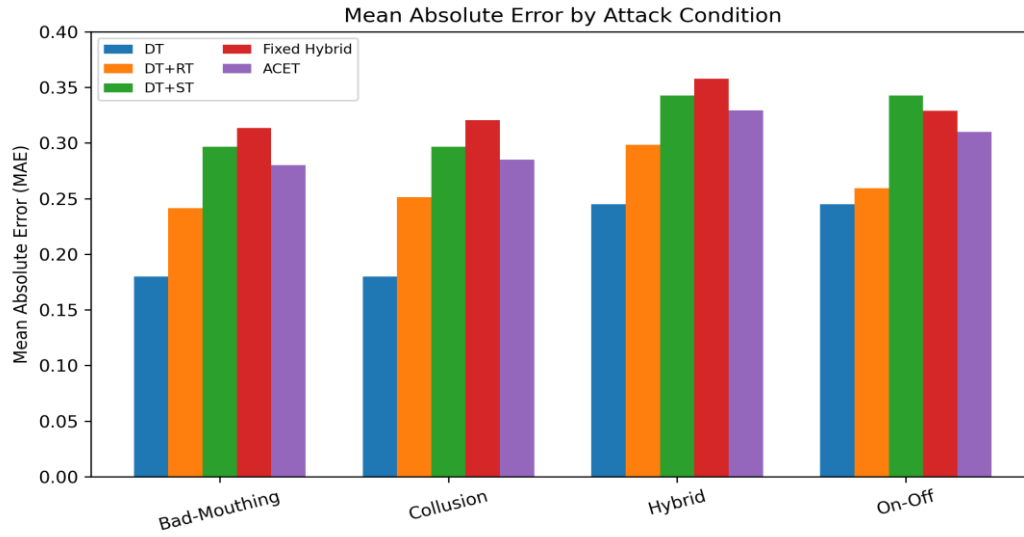


Figure 3. ACET and DT+RT accuracy across attack conditions.

5.4 Trust-Estimation Error

Table 2 presents mean absolute error (MAE) by attack condition, averaged across the 10%, 20%, and 30% malicious-node fractions. DT records the lowest MAE under all four attack conditions: 0.1798 under Bad-Mouthing, 0.1798 under Collusion, 0.2449 under Hybrid, and 0.2449 under On-Off. ACET records MAE values of 0.2801, 0.2849, 0.3294, and 0.3100, respectively. The corresponding values for DT+RT are 0.2416, 0.2513, 0.2983, and 0.2594; for DT+ST, 0.2967, 0.2967, 0.3426, and 0.3426; and for Fixed Hybrid, 0.3134, 0.3207, 0.3578, and 0.3290.

Attack	DT	DT+RT	DT+ST	Fixed Hybrid	ACET
Bad-Mouthing	0.1798	0.2416	0.2967	0.3134	0.2801
Collusion	0.1798	0.2513	0.2967	0.3207	0.2849
Hybrid	0.2449	0.2983	0.3426	0.3578	0.3294
On-Off	0.2449	0.2594	0.3426	0.3290	0.3100

Table 2. Mean MAE by attack condition, averaged across the 10%, 20%, and 30% malicious-node fractions. Lower values indicate smaller deviation between estimated trust and the defined target trust.

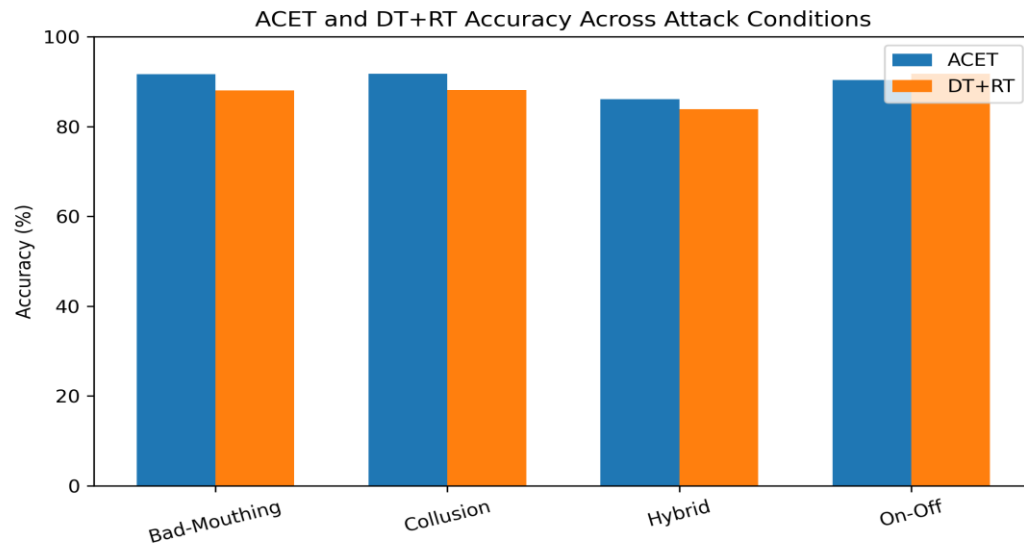


Figure 4. Mean absolute error by attack condition.

The MAE results demonstrate that classification resilience and trust-estimation accuracy are not equivalent: ACET can reduce decision errors relative to recommendation-heavy configurations under

manipulated recommendations while still producing larger continuous estimation error than direct trust.

6. Discussion

The results support a conditional interpretation of multi-source trust. Direct trust performs strongly because first-hand observations are directly connected to provider behaviour and are not manipulated by recommendation attacks in the current experimental design. When sufficient direct evidence exists, adding social or recommendation information can introduce conflicting evidence without improving the final decision.

The value of ACET becomes clearer when recommendations are manipulated. Its credibility mechanism gives different recommendation sources different influence instead of treating them as equally reliable. The increasing ACET advantage over DT+RT and Fixed Hybrid under Bad-Mouthing and Collusion is consistent with this design objective. The result does not imply that credibility weighting defeats all collusion strategies; coordinated attackers that accumulate high credibility or occupy influential network positions may still be effective.

The On-Off findings identify a separate weakness. Source credibility does not detect temporal instability in the provider itself. A future adaptive mechanism could combine source credibility with behavioural change detection, adaptive decay, or temporal consistency measures. Such an extension would make the model more responsive to repeated state transitions. The confidence indicator should also be interpreted carefully. Entropy and disagreement measure the structure of recommendation evidence, not its truthfulness. A group of malicious recommenders can agree and therefore generate high confidence. Collusion detection, source-dependence analysis, or external validation could be added to address this limitation.

7. Limitations and Threats to Validity

The network contains 100 nodes, so the results do not establish scalability to larger networks. Social relationships and attack behaviours are simulated, which improves control but limits direct generalization to operational SIIoT deployments. The baseline configurations are controlled component-level alternatives rather than independent implementations of external state-of-the-art algorithms. The confidence mechanism can assign high confidence to unanimous but incorrect recommendations. The base aggregation weights, thresholds, decay parameters, and recommendation temperature are fixed experimental settings, while ACET derives effective trust-source weights adaptively from direct-evidence availability and recommendation confidence; sensitivity analysis is therefore needed before treating these parameter settings as generally optimal.

8. Conclusion and Future Research

This paper presented ACET, a credibility-enhanced trust-management model for SIIoT interactions. The model integrates direct trust, credibility-weighted recommendation trust, social trust, recommendation consistency indicators, incentive and penalty updates, and recovery. A controlled experiment comprising 1,800 paired runs evaluated five configurations under four adversarial conditions and three malicious-node fractions.

The results show that DT achieves the highest overall accuracy at 93.35% and the lowest overall MAE at 0.2123. ACET achieves 89.97% accuracy and MAE of 0.3011. Nevertheless, ACET provides higher accuracy than DT+RT and Fixed Hybrid under Bad-Mouthing, Collusion, and the evaluated Hybrid condition, with the advantage generally increasing as the malicious fraction rises. Under On-Off behaviour, ACET does not provide the same advantage. The evidence therefore indicates that the benefit of multi-source trust integration is attack-dependent.

Future research should evaluate larger networks and additional SIIoT topologies, adaptive malicious strategies, domain-grounded social attributes, temporal mechanisms for On-Off behaviour, recommender dependence and collusion, confidence calibration, and parameter sensitivity.

References

Alam, S., Zardari, S., Noor, S., Ahmed, S., & Mouratidis, H. (2022). Trust management in Social Internet of Things (SIIoT): A survey. **IEEE Access*, 10*, 108924–108954.
<https://doi.org/10.1109/ACCESS.2022.3213699>

- Atzori, L., Iera, A., Morabito, G., & Nitti, M. (2012). The Social Internet of Things (SIoT): When social networks meet the Internet of Things: Concept, architecture and network characterization. **Computer Networks*, 56*(16), 3594–3608.
<https://doi.org/10.1016/j.comnet.2012.07.010>
- Bangui, H., Búhnová, B., Kušniráková, D., & Halász, D. (2023). Trust management in social Internet of Things across domains. **Internet of Things*, 23*, 100833.
<https://doi.org/10.1016/j.iot.2023.100833>
- Chahal, R. K., Kumar, N., & Batra, S. (2020). Trust management in social Internet of Things: A taxonomy, open issues, and challenges. **Computer Communications*, 150*, 13–46.
<https://doi.org/10.1016/j.comcom.2019.10.034>
- Magdich, R., Jemal, H., & Ben Ayed, M. (2022). A resilient trust management framework towards trust related attacks in the Social Internet of Things. **Computer Communications*, 191*, 92–107. <https://doi.org/10.1016/j.comcom.2022.04.019>
- Nitti, M., Girau, R., & Atzori, L. (2014). Trustworthiness management in the Social Internet of Things. **IEEE Transactions on Knowledge and Data Engineering*, 26*(5), 1253–1266.
<https://doi.org/10.1109/TKDE.2013.105>
- Sagar, S., Mahmood, A., Sheng, Q. Z., Zhang, W. E., Zhang, Y., & Pabani, J. K. (2024). Understanding the trustworthiness management in the social Internet of Things: A survey. **Computer Networks*, 251*, 110611. <https://doi.org/10.1016/j.comnet.2024.110611>
- Ud Din, I., & Guizani, M. (2019). Trust management techniques for the Internet of Things: A survey. **IEEE Access*, 7*, 29763–29789.