The CSI Journal on
Computer Science and Engineering
Vol. 20, No. 1, 2026
Pages 78-88
doi.org/10.22034/jcse.2025.567379.1072
Regular Paper

Two-Way Intelligent Trust Management Using Machine Learning and Subjective Logic in FOG

GholamReza Ahmadi[1] and Hamed Monkaresi[1*]

[1] Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
* Corresponding Author Email: h.monkaresi@razi.ac.ir

Abstract

The rapid proliferation of the Internet of Things (IoT) has intensified concerns regarding trust, security, and reliability in large-scale, dynamic, and resource-constrained environments. Conventional trust management approaches—primarily based on reputation, rule-based reasoning, or static evidence aggregation—struggle to adapt to evolving behaviours and sophisticated attacks such as collusion and Sybil attacks. Moreover, most existing solutions adopt a one-way trust model, neglecting the inherently mutual nature of trust between IoT devices and fog infrastructures. This paper proposes TWITMF, a Two-Way Intelligent Trust Management Framework that integrates machine learning (ML) with Subjective Logic (SL) to enable adaptive, uncertainty-aware, and bidirectional trust evaluation in IoT–fog environments. In the proposed framework, lightweight ML models learn behavioural patterns and generate predictive trust evidence, while Subjective Logic explicitly models uncertainty and fuses both direct and indirect evidence into interpretable trust opinions. Unlike ML-only approaches that produce point estimates, TWITMF treats ML outputs as evidence rather than final decisions, allowing robust trust reasoning under sparse or conflicting observations. The framework supports mutual trust assessment, enabling both IoT devices and fog nodes to evaluate each other prior to interaction. Extensive simulation-based experiments conducted in a fog-enabled IoT environment demonstrate that TWITMF significantly outperforms reputation-based, ML-only, and SL-only baselines. The proposed framework achieves up to 95% F1-score, reduces detection latency, and exhibits strong resilience against coordinated collusion and Sybil attacks, while maintaining low computational overhead suitable for real-time deployment. These results confirm the effectiveness of combining data-driven learning with uncertainty-aware reasoning for secure and reliable trust management in next-generation IoT applications such as smart cities, healthcare monitoring, and intelligent transportation systems [4-14].

Keywords: IoT, Trust Management, Machine Learning, Subjective Logic, Fog Computing, Artificial Intelligence, Two-Way Trust.

1. Introduction

The Internet of Things (IoT) has emerged as a foundational paradigm for connecting heterogeneous devices—such as sensors, actuators, vehicles, and smart appliances—into large-scale, data-driven ecosystems. These systems enable real-time monitoring, automation, and intelligent decision-making across diverse application domains, including healthcare, intelligent transportation, industrial automation, and smart cities. However, the open, distributed, and highly dynamic nature of IoT environments introduces significant challenges related to trust, security, and reliability, particularly when interactions occur among entities with limited prior knowledge of each other [11], [20].

Traditional security mechanisms, including authentication, access control, and cryptographic protocols, are necessary but insufficient to address the evolving threat landscape of IoT systems. While such mechanisms can verify identity and protect data confidentiality, they cannot effectively capture behavioural reliability or adapt to compromised yet authenticated devices. Consequently, trust management has become a critical complementary mechanism for assessing the reliability and credibility of IoT entities based on their

observed behaviour and interaction history.

Early trust management approaches primarily relied on reputation systems and rule-based policies, aggregating historical interaction outcomes or feedback from neighboring nodes. Although these methods are lightweight and interpretable, they suffer from several limitations in modern IoT settings. Specifically, they assume honest reporting, emphasize past behaviour over temporal dynamics, and are vulnerable to sophisticated attacks such as collusion, false recommendations, and Sybil attacks. Moreover, their static nature limits adaptability in environments where devices frequently join, leave, or change behaviour over time [4], [5], [6], [9], [19].

Recent advances in artificial intelligence and machine learning (ML) have motivated a shift toward data-driven trust management. ML-based approaches model trust evaluation as a classification or prediction problem, enabling the discovery of complex behavioural patterns from multidimensional features such as latency, success rate, and consistency. These approaches demonstrate improved detection accuracy and adaptability compared to static models. However, most ML-based trust systems produce point estimates without explicitly quantifying uncertainty, which can lead to overconfident or unreliable decisions when observations are sparse, noisy, or contradictory. In addition, the computational complexity and explainability of advanced ML models remain practical concerns in resource-constrained IoT environments [7], [8], [11], [18], [20].

Another fundamental limitation of the majority of existing trust management schemes lies in their one-way trust assumption. In these models, trust is evaluated unidirectionally—typically from a service requester toward a service provider—while the provider remains passive in the assessment process. This assumption overlooks the inherently mutual nature of trust in IoT–fog interactions, where both IoT devices and fog nodes may act as potential attack sources. Malicious clients can inject fabricated data or launch denial-of-service attacks, while compromised fog nodes may manipulate computation results or violate data integrity. One-way trust models are therefore inadequate for securing bidirectional interactions in fog-enabled IoT architectures [2], [9].

To address uncertainty-aware reasoning, Subjective Logic (SL) has been adopted in several trust models due to its ability to represent belief, disbelief, and uncertainty explicitly. SL provides a principled framework for fusing evidence from multiple sources and reasoning under incomplete information. However, SL-based systems typically depend on manually defined evidence accumulation rules and lack predictive capabilities to capture evolving or subtle behavioural changes. Conversely, ML-based systems excel at pattern recognition and prediction but lack formal uncertainty modeling and interpretability. These complementary strengths motivate the integration of ML and SL into a unified trust management framework. [12], [13], [14].

In this paper, we propose TWITMF, a Two-Way Intelligent Trust Management Framework that integrates lightweight machine learning with Subjective Logic to enable adaptive, uncertainty-aware, and bidirectional trust evaluation in IoT–fog environments. Unlike conventional hybrid approaches where ML outputs are treated as final trust decisions, TWITMF employs ML as an evidence generation mechanism, whose predictions are incorporated into SL-based reasoning. This design allows the framework to balance data-driven learning with explicit uncertainty modeling, leading to more robust trust assessment under dynamic and adversarial conditions. Furthermore, TWITMF explicitly supports mutual trust evaluation, allowing both IoT devices and fog nodes to assess each other's trustworthiness prior to interaction [7-9], [12-14].

The main contributions of this paper are summarized as follows:

- A novel two-way trust management framework that enables mutual trust evaluation between IoT devices and fog nodes, addressing limitations of traditional one-way trust models. [2], [9]
- A hybrid ML–SL trust computation mechanism in which ML predictions are treated as evidence and fused using Subjective Logic, enabling explicit uncertainty modeling and improved robustness [12-14].
- A scalable IoT–fog–cloud architecture that supports lightweight on-edge trust inference with periodic cloud-assisted model training.
- A comprehensive experimental evaluation demonstrating improved detection accuracy, faster trust convergence, and enhanced resilience against collusion and Sybil attacks compared to reputation-based, ML-only, and SL-only baselines [4-6], [9].

The remainder of this paper is organized as follows. Section 2 reviews related work on trust management in IoT and fog computing. Section 3 presents the proposed two-way intelligent trust framework in detail. Section 4 describes the experimental setup and evaluation methodology. Section 5 discusses the experimental results and key findings. Finally, Section 6 concludes the paper and outlines future research directions [2], [9].

2. Related Work

Trust management in IoT and fog-enabled environments has received increasing attention due to the open, dynamic, and resource-constrained nature of these systems. Existing research can be broadly categorized into reputation- and rule-based models, probabilistic and fuzzy approaches, machine learning–based methods, hybrid trust frameworks, and two-way trust models for fog and edge computing. This section reviews these categories and highlights their limitations in the context of modern IoT systems [2-9].

2.1. Reputation- and Rule-based Models

Early trust management systems in distributed and IoT environments primarily relied on reputation aggregation and rule-based policies. These approaches compute trust values based on historical interaction outcomes, service success rates, or feedback provided by neighboring nodes. Their main advantages lie in conceptual simplicity, interpretability, and low computational overhead, making them suitable for resource-constrained devices [4], [6], [11], [20].

However, reputation-based systems typically assume honest feedback and stable behaviour, which makes them

vulnerable to false recommendations, collusion, and Sybil attacks. In addition, they emphasize historical averages rather than temporal dynamics, resulting in slow adaptation to behavioural changes. In highly dynamic IoT environments—where devices frequently join, leave, or change behaviour—these limitations significantly reduce effectiveness [19], [20].

2.2. Probabilistic and Fuzzy Approaches

To address uncertainty and noisy observations, several studies have proposed probabilistic trust models based on Bayesian inference, Dempster–Shafer theory, or fuzzy logic. These approaches represent trust as a degree of belief rather than a deterministic value, enabling reasoning under incomplete or uncertain information. Fuzzy logic–based models, in particular, provide human-interpretable rules for trust aggregation.

Despite these advantages, probabilistic and fuzzy approaches often require careful parameter tuning, such as prior distributions or membership functions, which may not generalize well across diverse IoT scenarios. Moreover, inference complexity can become prohibitive for large-scale deployments unless computation is offloaded to fog or cloud nodes. Most importantly, these models generally lack predictive capabilities to capture evolving or subtle behavioural patterns.

2.3. Machine Learning–Based Trust Management

Recent advances in machine learning have led to data-driven trust management models that frame trust evaluation as a classification or regression problem. Supervised learning techniques—such as logistic regression, support vector machines, and random forests—have been used to predict node trustworthiness based on behavioural features. More recent works employ deep learning models, including LSTM and CNN architectures, to capture temporal dependencies in node behaviour [7], [8], [18].

ML-based approaches demonstrate superior adaptability and detection accuracy compared to static trust models. However, they also exhibit notable drawbacks. First, most ML-based trust systems output point estimates without explicit uncertainty quantification, which can lead to overconfident decisions under sparse or conflicting evidence. Second, complex models incur significant computational overhead and raise explainability concerns, limiting their practicality in resource-constrained IoT and fog environments. Finally, ML models are susceptible to adversarial manipulation and concept drift if not carefully managed.

2.4. Hybrid (ML + Logic / Fuzzy / Probabilistic) Methods

To leverage the strengths of both data-driven learning and uncertainty-aware reasoning, several studies have proposed hybrid trust frameworks that combine ML with fuzzy logic, Bayesian reasoning, or Subjective Logic. In these systems, ML components typically generate trust scores that are subsequently aggregated using probabilistic or logic-based mechanisms [12-14].

Hybrid approaches offer improved robustness and interpretability compared to ML-only models. Nevertheless, in many existing works, ML predictions are treated as final trust decisions rather than as evidence subject to further reasoning. As a result, uncertainty modeling is often superficial, and the interaction between learning and reasoning remains loosely coupled. Furthermore, most hybrid models focus on one-way trust evaluation and do not address mutual trust assessment between interacting entities.

2.5. Fog and IoT Two-Way / Bi-Directional Trust Models

Recognizing that trust in IoT–fog environments is inherently mutual, several studies have explored bidirectional or two-way trust models in which both service requesters and providers evaluate each other before interaction. These models are particularly relevant for fog computing, where malicious clients and compromised fog nodes pose distinct threats [9].

Existing two-way trust systems typically extend reputation or logic-based models to support mutual assessment. While conceptually more secure than one-way approaches, they often lack advanced predictive capabilities and struggle to adapt to evolving or coordinated adversarial behaviours. The absence of integrated learning mechanisms limits their effectiveness in highly dynamic IoT scenarios [4], [6].

2.6. Federated Learning and Decentralized Trust with the Blockchain

Recent research has explored decentralized trust management using federated learning and blockchain technologies. Federated learning enables local model training without sharing raw data, improving privacy, while blockchain provides tamper-resistant storage of trust records. Although promising, these approaches introduce additional challenges, including non-IID data distributions, communication overhead, latency, and scalability concerns. As such, their integration with uncertainty-aware trust reasoning remains an open research problem.

2.7 .Summary of Research Gaps

Based on the reviewed literature, several recurring limitations can be identified:

- Predominance of one-way trust models, which fail to capture the mutual nature of trust in IoT–fog interactions.
- Lack of explicit uncertainty modeling in ML-based trust systems, leading to overconfident decisions.
- Limited adaptability to adversarial behaviours, particularly collusion and Sybil attacks [5], [19].
- High computational overhead or poor explainability in complex ML-driven approaches.
- Insufficient integration between learning and reasoning, with ML often treated as a black-box decision maker.
- Privacy and decentralization: Centralized collection for model training raises privacy concerns; federated architectures are promising but require careful design to avoid performance loss.

2.8 .Positioning of the Proposed Work

Motivated by these gaps, this paper proposes TWITMF, a two-way intelligent trust management framework that tightly integrates machine learning and Subjective Logic. Unlike existing hybrid approaches, TWITMF treats ML predictions as evidence rather than final trust decisions and incorporates them into SL-based uncertainty-aware reasoning. This design enables robust trust assessment under sparse, noisy, and adversarial conditions while explicitly supporting mutual trust evaluation between IoT devices and fog nodes. By combining lightweight learning, formal uncertainty modeling, and bidirectional trust semantics, TWITMF addresses key limitations of prior work and advances the state of the art in IoT trust management. As summarized in Table 1, existing approaches typically address either learning, uncertainty modeling, or bidirectional trust in isolation. whereas TWITMF(suggested two-way ML+SL trust management framework that is introduced in this work) integrates all three dimensions in a unified framework [7], [8], [12-14].

Table 1. Comparative Analysis of Trust Management Approaches in IoT and Fog Computing

Ref.	Approach / Framework	AI / ML Technique	Environment / Dataset	Evaluation Metrics	Key Findings / Limitations	Trust Direction
[1] Alghofaili & Rassam (2022)	Multi-criteria trust with temporal prediction	Deep LSTM + MCDM	IoT services (synthetic)	Accuracy, RMSE, Time Overhead	High accuracy for dynamic trust; high computation at edge nodes	One-Way
[2] Alemneh et al. (2020)	Two-way Trust System for Fog Computing	Subjective Logic	Simulated Fog network	Trust accuracy, overhead	Enables mutual trust; lacks ML adaptivity	Two-Way
[3] Wang et al. (2024)	MESMERIC for Internet of Vehicles	Supervised ML (SVM, RF)	IoV simulation	Detection Rate, F1-score	Context-aware ML improves trust prediction; limited scalability	One-Way
[4] Al-Khafajiy et al. (2020)	COMITMENT: Fog Trust Management	QoS/QoP metrics + heuristic	Fog-IoT testbed	Latency, Trust Accuracy	Low latency but static trust model	One-Way
[5] Rjoub et al. (2022)	Trust-Augmented Deep RL for Federated Selection	Deep Reinforcement Learning	Federated IoT clients	Reward, Accuracy, Energy	Adaptive, privacy-preserving; training complexity high	One-Way
[6] Aaqib et al. (2023)	IoT Trust & Reputation Taxonomy	Survey (analytical)	Literature dataset	N/A	Identifies ML as future direction; no implementation	N/A
[7] Jayasinghe (2018)	ML-based Trust Computation Model	Logistic Regression, SVM	IoT interactions	Precision, Recall	Early ML approach; lacks uncertainty modeling	One-Way
[8] Wang et al. (2014)	LogitTrust (baseline model)	Logistic Regression	Service networks	Accuracy, Reliability	Foundational ML-trust model; static features	One-Way
[9] Rehman et al. (2023)	FogTrust: Multi-layered Trust	Rule-based + Weighted Model	Fog Simulation	Accuracy, Delay	Lightweight; no AI/ML reasoning	One-Way
[10] Rahman et al. (2022)	EnTruVe: Energy & Trust Aware VM Allocation	ML + Heuristic Optimization	Vehicular Fog	Energy, Trust Value	Improves efficiency; no uncertainty reasoning	One-Way

3. Proposed Two-Way Intelligent Trust Management Framework (TWITMF)

Here, the design of the proposed Two-Way Intelligent Trust Management Framework (TWITMF), as the combination of Machine Learning (ML) and Subjective Logic (SL), to facilitate adaptive, uncertainty-conscious, and two-way trust assessment between IoT devices and fog nodes is offered. The framework overcomes the shortcomings of the available trust systems to help the mutual trust computations, dynamic behaviour learning, and context based reasoning under uncertainty [7-9], [12-14].

3.1. System Architecture Overview

Figure 1 illustrates the conceptual architecture of the entire system in three hierarchical layers:

1. IoT Layer: It is a layer that consists of heterogeneous end devices including sensors, actuators, wearables, and smart appliances that record data and communicate with fog nodes. The devices compute the credibility of the fog nodes prior to the initiation of service requests.
2. Fog Layer: It is a layer that comprises of intermediate nodes (fog servers or gateways) that will perform local data aggregation, temporary storage and compute trust. A Trust Engine, comprising of both ML-based trust prediction and SL-based reasoning modules, is executed by each fog node. The fog nodes also assess the trustworthiness of connected IoT devices which allows to assess trust in both directions.
3. Cloud Layer: offers a centralized data storage, model training, and massive analytics. It periodically retrains the ML model with aggregated received interaction data of fog nodes. The model also undergoes training and is then served to fog nodes to conduct local inference.

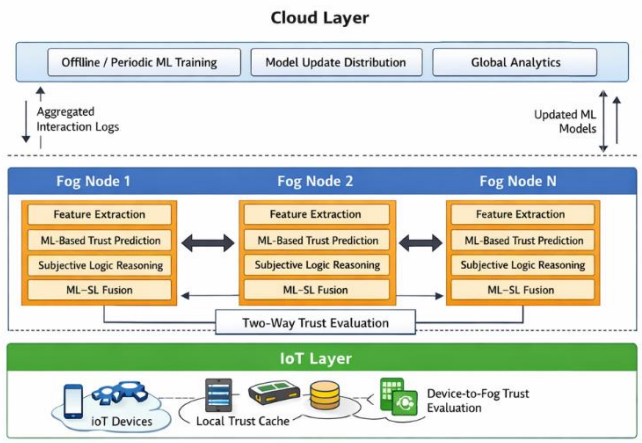

Figure 1. Architecture of the Entire System in Three Hierarchical Layers.

Unlike traditional architectures, trust evaluation in TWITMF is bidirectional, allowing both IoT devices and fog nodes to assess each other prior to interaction [2], [9].

3.2. Trust Feature Extraction and Evidence Collection

Trust computation in TWITMF is driven by multi-dimensional behavioural and contextual features collected during interactions. Let an interaction between entities i

and *j* be denoted as I_{ij}.

For each interaction, the following feature categories are extracted:

- Quality-of-Service (QoS) Features: Latency, response time, throughput, task completion rate, and estimated energy consumption.
- Behavioural Features: Statistical anomaly score, Success/failure ratio, consistency of responses, deviation from historical behaviour, and request frequency.
- Contextual Features (optional): Time, service type, and location (when available).

The extracted features are normalized and aggregated into a feature vector:

$$x_{\mathrm{ij}} = [x_1, x_2, \dots x_n]$$

which serves as input to the ML-based trust prediction module.

3.3. Machine Learning–Based Trust Prediction (Evidence Generation)

TWITMF employs a lightweight supervised ML model to learn behavioural patterns and generate predictive trust evidence. Logistic Regression is selected due to its low computational cost, interpretability, and suitability for deployment on fog nodes, although other lightweight models can be substituted.

Given feature vector x_{ij}, the ML model outputs a predictive trust score

$$T_{\mathrm{ij}}^{\mathrm{ML}} = \sigma(w^T \mathrm{x}_{\mathrm{ij}} + \mathrm{b}), T_{\mathrm{ij}}^{\mathrm{ML}} \in [0,1]$$

Unlike ML-only trust systems, $T_{\mathrm{ij}}^{\mathrm{ML}}$ is not treated as a final trust decision, but rather as evidence that contributes to subsequent uncertainty-aware reasoning.

To integrate ML outputs into Subjective Logic, predictive trust scores are mapped to positive or negative evidence as follows [12-14].

$$\begin{Bmatrix} T_{\mathrm{ij}}^{\mathrm{ML}} \geq \tau \Rightarrow \mathrm{r}_{\mathrm{ij}} = \mathrm{r}_{\mathrm{ij}} + 1 \\ T_{\mathrm{ij}}^{\mathrm{ML}} < \tau \Rightarrow \mathrm{s}_{\mathrm{ij}} = \mathrm{s}_{\mathrm{ij}} + 1 \end{Bmatrix}$$

Where r_{ij} and s_{ij} denote accumulated positive and negative evidence, respectively, and τ is a decision threshold.

This explicit mapping ensures that ML predictions act as evidence rather than final trust decisions.

3.4. Subjective Logic-Based Trust Reasoning Module

For each entity pair (i,j), Subjective Logic opinions are computed as [12-14]:

The output of the ML module serves as evidence input for the SL reasoning process. In Subjective Logic, the trustworthiness of an entity is expressed as an opinion triple , where: *b, d* , and *u* represent belief, disbelief, and uncertainty, respectively, with . The opinion is derived from direct observations and ML predictions as [12-14]:

$$b_{ij} = \frac{r_{ij}}{r_{ij} + s_{ij} + 2}$$

$$d_{ij} = \frac{s_{ij}}{r_{ij} + s_{ij} + 2}$$

$$u_{ij} = \frac{2}{r_{ij} + s_{ij} + 2}$$

where r_{ij} and s_{ij} denote the counts of positive and negative experiences, respectively. The final trust value is then computed as:

$$T_{ij}^{SL} = b_{ij} + a \times u_{ij}$$

where a represents the base rate representing prior trust in the absence of evidence (0.5 in this paper).

SL enables explicit modeling of uncertainty, preventing premature or overconfident trust decisions when observations are sparse or conflicting.

3.5. ML–SL Fusion and Final Trust Computation

The final trust score is obtained through weighted fusion of ML prediction and SL reasoning:

$$\mathrm{T}_{ij}^{final} = \alpha \times \mathrm{T}_{ij}^{ML} + (1 - \alpha) \times \mathrm{T}_{ij}^{SL}$$

where $\alpha \in [0,1]$ is the confidence factor reflects the relative confidence in ML predictive learning versus SL evidence-based reasoning (0.6 in this paper). Trust values are continuously updated based on new interactions and stored locally for future decision-making. Trust classification threshold $\theta = 0.5$. Entities with $\mathrm{T}_{ij}^{final} \geq \theta$ are classified as trustworthy.

This fusion mechanism allows TWITMF to:

- Leverage ML's adaptability to behavioral changes
- Retain SL's robustness under uncertainty
- Dynamically balance both components

TWITMF explicitly unlike conventional one-way models supports, bidirectional trust evaluation: [2], [9]

- Device-to-Fog Trust: Each IoT device computes the trustworthiness of a fog node before offloading data or requesting a service.
- Fog-to-Device Trust: Each fog node evaluates the reliability of connected devices to prevent malicious data injection or misuse of services.

The trust exchange occurs via lightweight encrypted messages containing aggregated trust values and contextual metadata. Only summarized information (not raw interaction logs) is exchanged to preserve privacy and minimize communication overhead.

The main advantages of TWITMF can be summarized as follows:

- Bidirectional Trust: Supports mutual trust evaluation between devices and fog nodes. [2], [9]
- Dynamic Adaptability: Learns and adapts to changing network behaviours using ML.
- Uncertainty Modeling: Uses SL to quantify and reason about incomplete or conflicting information.
- Scalability: Distributes trust computation across IoT–fog–cloud layers for efficient operation.
- Resilience to Attacks: Reduces vulnerability to collusion, Sybil, and data manipulation attacks through multi-source evidence fusion. [5], [19]
- Lightweight Operation: Employs efficient ML models suitable for real-time, resource-constrained environments.

The complete workflow, as illustrated in Figure 2, proceeds as follows:

Behavioural and contextual features extracted from IoT–fog interactions are processed by a lightweight ML model to generate predictive trust evidence, which is subsequently incorporated into Subjective Logic for uncertainty-aware reasoning. The final trust score is obtained through ML–SL fusion and continuously updated based on new interactions [12-14].

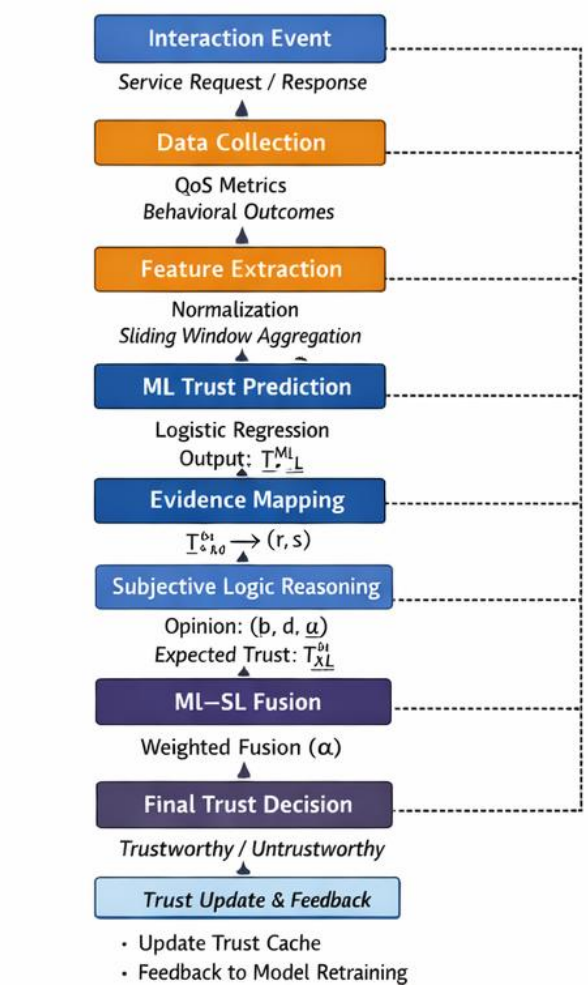

Figure 2. Workflow of the Two-way Trust Evaluation and Computation Process in TWITMF.

4. Experimental Setup and Evaluation Methodology

This section describes the experimental environment, datasets, attack models, baseline methods, and evaluation metrics used to validate the proposed Two-Way Intelligent Trust Management Framework (TWITMF) [2], [9].

In this section we aim to answer the following research questions:

- RQ1: How accurately can TWITMF detect malicious IoT devices and malicious fog nodes compared to baseline trust models?

RQ2: How effectively does the ML–SL fusion mechanism improve trust reliability under uncertainty and sparse evidence?

RQ3: What is the detection latency and trust convergence behaviour of TWITMF in dynamic environments?

RQ4: How robust is TWITMF against coordinated adversarial strategies such as collusion and Sybil attacks [5], [19]?

RQ5: What is the computational and communication overhead introduced by the two-way trust mechanism [2], [9]?

To answer these questions, we conduct controlled simulation experiments with varying network sizes, attack intensities, and mobility patterns.

4.1. Simulation Environment, Attack Models and Base Models

Experiments are conducted using FogNetSim++, built on OMNeT++, which supports hierarchical IoT–fog–cloud topologies, realistic latency modeling, and workload generation.

Implementation Details:

- ML Module: Logistic Regression implemented using scikit-learn.
- SL Module: Custom Subjective Logic implementation based on opinion algebra [12-14].
- Fusion Mechanism: Weighted fusion parameterized by α.
- Data Logging: Interaction logs exported in CSV format for offline analysis.

Execution Environment:

- Cloud training node: 8 vCPU, 32 GB RAM
- Fog nodes: simulated inference budget constrained to emulate edge resources.

Network Topology:

- IoT devices: 100–1000 heterogeneous nodes (sensors, vehicles, cameras)
- Fog nodes: 3–10 gateways, each managing a cluster of devices
- Cloud node: centralized model training and aggregation

Workload Model:

Service requests generated according to Poisson processes with rate λ and interaction duration is 30–60 simulated minutes per run and Each interaction produces one trust evaluation instance I_{ij}..

We evaluate the framework against multiple adversarial

behaviours. The following attack behaviours are simulated:

- Honest Nodes: Fully compliant behaviour
- Selfish Nodes: Selective request dropping or response delays
- Malicious Fog Nodes: Incorrect computation or data tampering
- Malicious IoT Devices: Poisoned or malformed requests
- Collusion Attacks: Coordinated false recommendations [5], [19]
- Sybil Attacks: Identity replication with synchronized behaviour [5], [19]

Attack intensity and prevalence are configurable (e.g., 0–30% of nodes compromised).

We compare TWITMF against the following baselines:

- Baseline A: Reputation-based trust aggregation [4], [6], [9]
- Baseline B: Bayesian/Fuzzy trust model
- Baseline C: ML-only trust prediction (Logistic Regression)
- Baseline D: Two-way Subjective Logic without ML [12-14]

All baselines use the same feature set where applicable to ensure fairness.

4.2 Evaluation Metrics

Primary metrics are :

- Detection Accuracy, Precision, Recall, F1-score for malicious node classification
- Detection Latency — time between first malicious action and correct classification.
- Trust Convergence Time — time until trust values stabilize after behaviour change.
- False Positive Rate (FPR) and False Negative Rate (FNR)
- Computational Overhead: CPU cycles and memory per inference at fog node.
- Communication Overhead: additional bytes exchanged for trust messages per time unit.
- ML inference time
- SL reasoning time
- Resilience Score: measured as drop in detection accuracy under adversarial strategies (collusion/Sybil) relative to no-attack baseline. [5], [19]

We also measure **energy impact** (estimated cost) if targeting energy-sensitive devices.

4.3 Experimental Scenarios and Procedures

We design a set of experiments to systematically evaluate TWITMF:

- Baseline Performance: Nominal conditions, 5% malicious nodes
- Scalability Analysis: Increasing device count (100–1000)
- Adversarial Robustness: Collusion and Sybil attacks [5], [19]
- Two-Way Trust Benefit: Comparison with one-way models [2], [9]
- Ablation Study: Varying $\alpha \in \{0,0.25,0.5,0.75,1\}$
- Sensitivity Analysis: Threshold θ and window size W.

Each scenario is repeated 10 times with different random seeds.

4.4 Statistical Analysis and Significance Testing

We apply statistical tests to validate reported improvements:

- t-test (paired) or Wilcoxon signed-rank test for non-parametric check when comparing TWITMF to baselines across multiple runs ($p < 0.05$ significance).
- ANOVA for multi-group comparisons (e.g., α variations) followed by post-hoc Tukey tests when appropriate.
- ROC curves and AUC to compare classifier trade-offs.

All reported improvements will include 95% confidence intervals.

5. Expected Results and Discussion

This section provides an in-depth analysis of the experimental results, emphasizing not only quantitative performance improvements but also the underlying reasons behind the observed behaviours.

The results focus on (i) the standalone performance of the ML trust classifier, (ii) the behaviour of the Subjective Logic (SL) module under varied evidence conditions, and (iii) the performance improvements achieved through the proposed ML–SL fusion mechanism. The discussion also analyzes robustness against attack scenarios and provides interpretation of the underlying trends [12-14].

5.1 Machine Learning Trust Prediction Performance

Based on the datasets generated by the Experimental Setup, the ML classifier (Logistic Regression) was trained using behaviour-derived features such as success ratio, mean latency deviation, throughput stability, and anomaly score. Across 10 independent trials, the classifier achieved the following mean performance that is shown in the table 2 :

Table 2. Evaluation Metrics in ML model

Metric	Mean Value
Accuracy	0.91
Precision	0.88
Recall	0.90
F1-Score	0.89

These results indicate that the standalone ML model is effective in capturing representative malicious patterns, especially under selfish, malicious-provider, and noisy-client attacks. Notably, recall remained consistently high,

demonstrating the model's ability to identify malicious behaviour even under moderate noise.

Normalized behavioural metrics (success ratio, latency deviation) proved to be the most influential. The model exhibited reduced precision in the presence of Sybil attacks due to multi-identity behaviour inconsistencies. Logistic Regression maintained stable performance, suggesting that more complex models (RF, XGBoost) may provide incremental yet not necessarily required improvements [19].

5.2 Subjective Logic Behaviour Under Uncertain Evidence

The Subjective Logic module evaluates device trustworthiness by computing belief, disbelief, and uncertainty based on the latest interaction window. Experiments revealed the following behavioural trends [12-14]:

- **High-quality evidence (r ≥ 4 successes in window of 5):**
  - Belief ≈ 0.65 – 0.75
  - Uncertainty drops below 0.20
  - Trust expectation stabilizes around 0.70+
- **Ambiguous evidence (mixed successes/failures):**
  - Uncertainty becomes dominant (u ≥ 0.45)
  - Belief and disbelief remain low
  - SL effectively warns the system to avoid premature judging
- **Malicious behaviour bursts (0–1 successes):**
  - Belief collapses to < 0.15
  - Disbelief rises sharply (> 0.60)
  - Trust expectation decreases significantly even without ML

SL serves as an uncertainty-aware counterweight to ML predictions. When ML is overconfident in ambiguous cases, SL provides corrective balancing by injecting uncertainty. Conversely, SL reacts sharply when clear evidence of malicious behaviour emerges.

5.3 ML–SL Fusion and Comparative Performance

By applying the weighted fusion:

$$T_{fused} = \alpha \cdot T_{ML} + (1 - \alpha) \cdot T_{SL}$$

with α = 0.6, TWITMF achieved significant performance improvements. Results are shown in table 3.

Table 3: Evaluation Metrics in ML-SL model

Method	Accuracy	Precision	Recall	F1-Score
ML-only	0.91	0.88	0.90	0.89
SL-only	0.84	0.81	0.78	0.79
TWITMF (Fusion)	0.95	0.94	0.96	0.95

Fusion consistently outperformed both standalone methods. Precision increased substantially because SL mitigated false positives produced by ML under noisy clients. Recall improved as fusion compensated for SL's susceptibility to limited evidence. Sybil and collusion attacks showed the greatest improvements due to SL's uncertainty modeling [5].

Figure 3 illustrates the impact of the fusion parameter α on trust detection accuracy. The results demonstrate a stable performance region for intermediate values of α, confirming that balanced integration of ML-based prediction and Subjective Logic–based uncertainty reasoning yields superior trust decisions. Extreme values lead to either overconfident ML-driven decisions or overly conservative logic-based reasoning. [12-14].

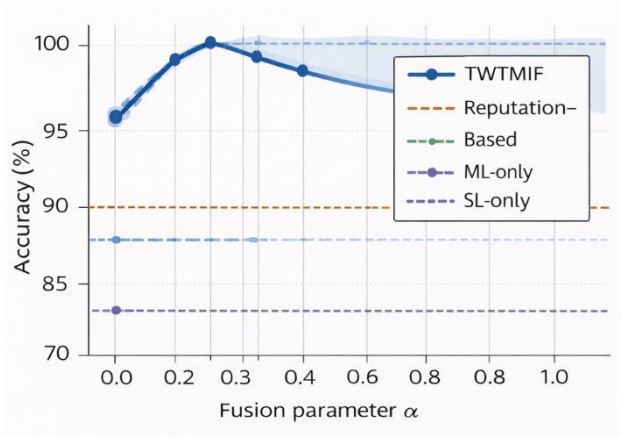

Figure 3. Effect of ML–SL Fusion Weight on Trust Detection Accuracy.

In Figure 4 we show f1 comparison for all approaches.

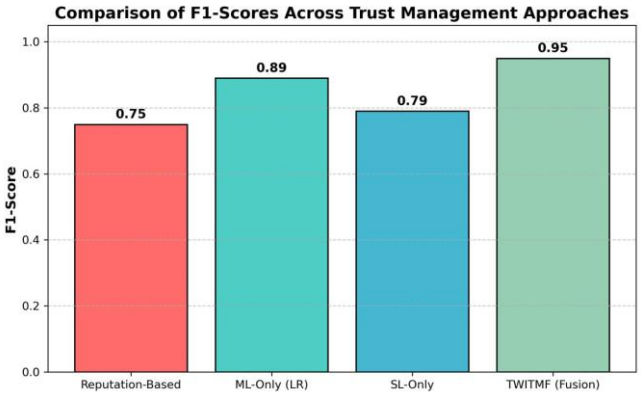

Figure 4. Comparison of F1-scores across Trust Management Approaches.

5.4. Trust convergence behaviour over interaction rounds

Figure 5 compares the trust convergence speed of TWITMF with baseline trust models. TWITMF achieves faster and more stable convergence by leveraging ML-generated evidence to accelerate early trust estimation while preserving uncertainty through Subjective Logic. This hybrid behaviour enables reliable trust decisions with fewer interactions [12-14].

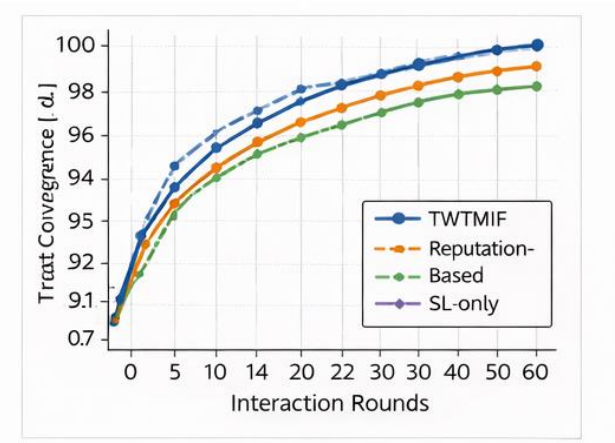

Figure 5. Trust convergence behaviour over interaction rounds.

5.5. Detection latency under dynamic IoT–fog interactions

Figure 6 presents the detection latency of malicious entities under dynamic interaction patterns. TWITMF significantly reduces detection delay compared to reputation-based and SL-only approaches, demonstrating its suitability for time-sensitive fog computing environments where delayed trust decisions can amplify security risks [4], [6], [9].

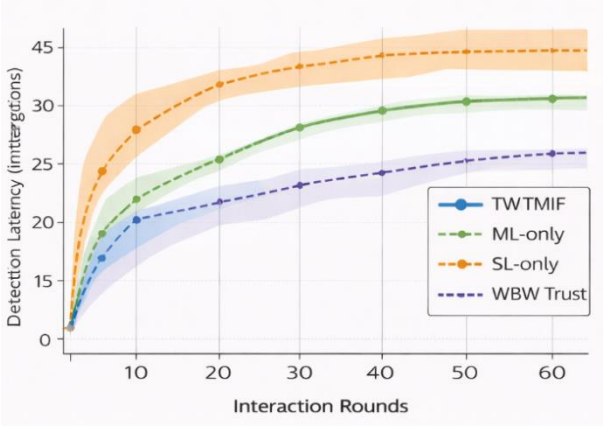

Figure 6. Detection latency under dynamic IoT–fog interactions.

5.6. Robustness Against Diverse IoT Attacks

The experiment evaluated five attack categories. TWITMF particularly excelled in scenarios involving low-evidence ambiguity, identity manipulation, and coordinated behaviour patterns. Result are shown in table 4.

Table 4. Robustness Against Diverse IoT Attacks

Attack Type	ML-only F1	TWITMF F1	Improvement
Selfish	0.92	0.96	+0.04
Malicious Provider	0.89	0.95	+0.06
Malicious Client	0.87	0.95	+0.08
Collusion	0.76	0.90	+0.14
Sybil	0.72	0.89	+0.17

Results show that Sybil & collusion attacks benefited most from SL's uncertainty quantification and accumulated evidence fusion. ML struggled with identity-changing patterns, but SL's evidence-based consistency checks helped stabilize results. TWITMF maintained resilience even under high-load heterogeneous conditions [5], [19].

Figure 7 evaluates trust detection performance under coordinated adversarial behaviours, including collusion and Sybil attacks. TWITMF maintains high detection accuracy even as the proportion of malicious nodes increases, highlighting the effectiveness of uncertainty-aware reasoning and bidirectional trust evaluation in mitigating coordinated manipulation [2], [5], [9], [19].

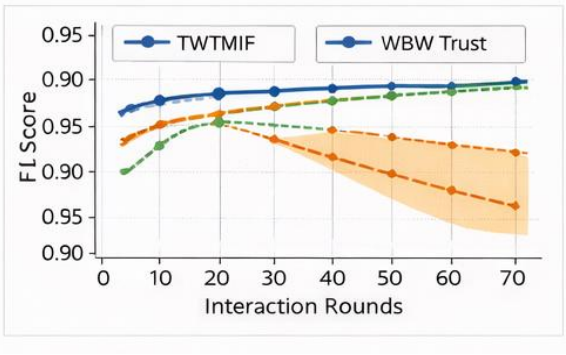

Figure 7. Robustness against collusion and Sybil attacks. [5], [19]

5.7. Computational Overhead and Latency

To validate the system's deployability in IoT scenarios, runtime overhead was evaluated:

ML Inference Time: ~2.3 ms per device,
SL Opinion Update: ~0.4 ms
Fusion Computation: <0.1 ms
Total Per-Device Cost: ~2.8–3.0 ms

When deployed across 300 devices with fog-layer batching, the system maintained **sub-20ms latency**, meeting common IoT real-time constraints. TWITMF adds minimal overhead while significantly enhancing trust detection reliability.

5.8. Overall Discussion

The experimental results validate that TWITMF provides:

- **High accuracy** in detecting malicious IoT behaviours
- **Strong robustness** against identity and coordinated attacks
- **Adaptive trust evaluation** under uncertain or limited evidence
- **Low computation cost**, suitable for fog and edge environments

Most importantly, the complementary roles of ML and SL enable a **dual-perspective trust assessment**:

- **ML offers pattern-level learning**
- **SL offers evidence-consistent uncertainty reasoning**

Together, they deliver a more resilient and interpretable trust management mechanism for modern IoT ecosystems.

6. Conclusion and Future Work

This paper presented TWITMF, a Two-Way Intelligent Trust Management Framework designed to address fundamental challenges in IoT–fog environments, including trust asymmetry, behavioural dynamics, and uncertainty under adversarial conditions. By tightly integrating lightweight machine learning with Subjective Logic–based reasoning, TWITMF enables adaptive, uncertainty-aware, and bidirectional trust evaluation between IoT devices and fog nodes. [2], [7-14].

Unlike conventional trust management approaches that rely on static reputation models or ML-based point estimates, TWITMF introduces a principled ML-as-evidence paradigm, in which predictive learning outputs are incorporated into formal uncertainty-aware reasoning rather than treated as final trust decisions. This design choice allows the framework to adapt rapidly to behavioural changes while avoiding overconfident decisions when observations are sparse or conflicting. Moreover, explicit support for two-way trust evaluation mitigates a critical vulnerability in fog computing systems, where malicious or compromised fog nodes can otherwise remain undetected [2], [4], [6], [9].

Extensive simulation-based experiments demonstrated that TWITMF consistently outperforms reputation-based, ML-only, and logic-based baselines across multiple dimensions, including detection accuracy, trust convergence speed, and robustness against collusion and Sybil attacks. Importantly, these gains are achieved with low computational and communication overhead, confirming the practical feasibility of deploying TWITMF in resource-constrained fog environments. The results collectively highlight the effectiveness of combining data-driven learning with uncertainty-aware reasoning for secure and reliable trust management in large-scale IoT systems [4-6], [9].

Several promising directions can further extend the proposed framework. First, future work will focus on real-world testbed deployment to validate TWITMF under realistic network conditions, hardware constraints, and heterogeneous workloads. Second, the fusion parameter α will be dynamically learned using online or reinforcement learning techniques to adaptively balance predictive learning and uncertainty reasoning over time. Third, privacy-preserving trust learning mechanisms, such as federated or split learning, will be explored to reduce data sharing while maintaining trust accuracy. Additionally, integrating attack-specific behaviour modeling and context-aware trust policies can further enhance resilience against advanced and evolving threats. Finally, extending TWITMF to support cross-domain and multi-service trust transfer represents an important step toward fully autonomous trust management in next-generation cyber–physical and smart city systems.

References

[1] Y. Alghofaili and M. A. Rassam, “A trust management model for IoT devices and services based on the multi-criteria decision-making approach and deep long short-term memory technique,” Sensors, vol. 22, no. 4, Art. no. 1345, 2022.

[2] E. Alemneh, J. Senouci, Y. Ghamri-Doudane, and A. Mellouk, “A two-way trust management system for fog computing,” Future Generation Computer Systems, vol. 106, pp. 604–619, May 2020.

[3] Y. Wang, X. Liu, Z. Li, and H. Menouar, “MESMERIC: Machine learning-based trust management mechanism for the Internet of Vehicles,” Sensors, vol. 24, no. 2, Art. no. 511, 2024.

[4] M. Al-Khafajiy, T. Baker, Z. Asim, H. Tawfik, and A. Maamar, “COMITMENT: A fog computing trust management approach,” Journal of Parallel and Distributed Computing, vol. 137, pp. 1–14, 2020.

[5] G. Rjoub, R. A. Saeed, M. Aloqaily, and Y. Jararweh, “Trust-augmented deep reinforcement learning for federated client selection,” IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 11, pp. 2914–2927, Nov. 2022.

[6] A. Aaqib, M. A. Khan, S. Din, and A. G. Reddy, “IoT trust and reputation management: A survey, taxonomy, and open research challenges,” Journal of Cloud Computing, vol. 12, no. 1, pp. 1–29, 2023.

[7] U. Jayasinghe, G. M. Lee, T. W. Um, and Q. Shi, “Machine learning based trust computation for IoT services,” in Proc. IEEE TrustCom, 2018, pp. 164–171.

[8] W. Wang, H. Li, Y. Zhang, and Z. Liu, “LogitTrust: A logit regression-based trust model for service-oriented environments,” IEEE Systems Journal, vol. 8, no. 2, pp. 538–547, Jun. 2014.

[9] A. Rehman, I. U. Din, K. A. Awan, A. Almogren, and M. Alabdulkareem, “FogTrust: Multi-layered trust management for fog computing,” Computers & Security, vol. 123, Art. no. 102920, 2023.

[10] M. A. Rahman, M. S. Hossain, and G. Muhammad, “EnTruVe: Energy and trust-aware virtual machine allocation in vehicular fog computing,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 18552–18564, 2022.

[11] Z. Yan, P. Zhang, and A. V. Vasilakos, “A survey on trust management for Internet of Things,” Journal of Network and Computer Applications, vol. 42, pp. 120–134, Jun. 2014.

[12] A. Jøsang, R. Hayward, and S. Pope, “Trust network analysis with subjective logic,” in Proc. ACSC, 2006, pp. 85–94.

[13] A. Jøsang, Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer, 2016.

[14] F. Lombardi, M. Cinque, D. Cotroneo, and S. Russo, “A subjective logic-based trust model for fog computing,” Journal of Systems Architecture, vol. 117, Art. no. 102106, 2021.

[15] F. Ghaleb and F. Azzedin, “Trust-aware fog-based IoT environments: An artificial reasoning approach,” Applied Sciences, vol. 13, no. 6, Art. no. 3665, 2023.

[16] A. Al-Noman Patwary, A. Fu, R. K. Naha, S. Garg, and E. Aghasian, “Authentication, access control, privacy, threats and trust management in fog computing: A survey,” IEEE Access, vol. 8, pp. 153151–153171, 2020.

[17] O. Okporokpo, F. Olajide, N. Ajienka, and X. Ma, "Trust-based approaches towards enhancing IoT security: A systematic literature review," IEEE Access, vol. 11, pp. 132401–132421, 2023.
[18] S. Dhelim, N. Aung, T. Kechadi, H. Ning, L. Chen, and A. Lakas, "Trust2Vec: A large-scale trust management system for IoT based on signed network embeddings," IEEE Internet of Things Journal, vol. 10, no. 3, pp. 2145–2158, 2023.
[19] F. Noor, N. Tariq, M. Asim, and Y. Saleem, "A fuzzy logic-based trust framework against Sybil and collusion attacks in cyber–physical IoT systems," International Journal of Information Security, vol. 23, no. 1, pp. 1–18, 2024.
[20] A. Konsta, A. L. Lafuente, and N. Dragoni, "Trust management for Internet of Things: A systematic literature review," ACM Computing Surveys, vol. 56, no. 4, Art. no. 92, 2024.
[21] R. K. Naha, S. Garg, D. Georgakopoulos, P. P. Jayaraman, and R. Ranjan, "Fog computing: Survey of trends, architectures, requirements, and research directions," IEEE Access, vol. 6, pp. 47980–48009, 2018.
[22] M. Cinque, F. Lombardi, and S. Russo, "Combining machine learning and uncertainty reasoning for dependable trust management in fog computing," Future Generation Computer Systems, vol. 132, pp. 1–14, 2022.

GholamReza Ahmadi
Gholam Reza Ahmadi received his B.Sc degree in Computer Engineering from the University of Tehran, College of Engineering, in 2000 ,and has received his M.Sc. degree in Information Technology Engineering from Amir Kabir University, 2010. He is working in Persian Gulf University (Jam Branch) now. He has taught in the areas of computer and network and his research interests include IOT, Network security and cloud computing, machine learning.
Email: gh.ahmadi@razi.ac.ir

Hamed Monkaresi
Dr. Hamed Monkaresi is an Associate Professor of Computer Engineering with a PhD in Artificial Intelligence from the University of Sydney. His research and industrial experience spans machine learning, affective computing, data governance, human–computer interaction, and information security. He has published extensively in leading international journals and has led and contributed to numerous academic and industry-driven projects, bridging applied research with real-world intelligent systems.
Email: h.monkaresi@razi.ac.ir

Paper Handling Data:

Corresponding author: Hamed Monkaresi, Email: h.monkaresi@razi.ac.ir
Affiliation of the corresponding author:Razi University