

FAULT-TOLERANT CLUSTERING PROTOCOL FOR IOT WITH COLD AND HOT STANDBY REDUNDANCY

DR. A RADHA KRISHNA¹, DR. A LAKSHMI NARAYANA², KLV PRASAD³, DR. INAKOTI RAMESH RAJA⁴, G SANTHOSH KUMAR⁵, NEHA BELWAL⁶

¹Department of CSE (AI&ML), Pragati Engineering College, Surampalem, India.

²Department of ECE, Aditya University, Surampalem, India.

³Department of ECE, Aditya University, Surampalem, India.

⁴Associate Professor, Department of ECE, Aditya University, Surampalem, India.

⁵Senior Assistant Professor, Department of ECE, CVR College of Engineering, Hyderabad, India.

⁶Department of ECE, Graphic Era (Deemed to be University), Dehradun, India.

¹vasjrs2004@gmail.com, ²aln2hanumu@gmail.com, ³vijayaprasad431@gmail.com,

⁴Inakoti.rameshraj@gmail.com, ⁵santhoshemwave@gmail.com, ⁶nehabelwal.ece@geu.ac.in

ABSTRACT

IoT networks are becoming a more unsteady and heterogeneous setting in which frequent failure of devices, multi-modal noisy signals and uneven workloads diminish the stability of the cluster and the availability of the system. Although the current developments in the area of IoT analytics are rather fast, the currently used clustering protocols seldom incorporate adaptive redundancy, as well as do not resolve the problem of large-scale, error-prone deployment. The ability of this gap limit to be used in critical applications like smart cities, healthcare monitoring, and industrial IoT. The objective of this study was to create a robust and fault-tolerant clustering protocol, which is able to learn stable cluster-based structures in addition to dynamically maintaining cold and hot standby redundancy to ensure high availability with low overhead. A federated experimental model was adopted with publicly available IoT data (telemetry, network flows, acoustic faults, and ambient sensing), in which clustering was achieved with the help of Federated Graph-Contrastive Clustering (FGCC), and redundancy choices were made with the assistance of reinforcement learning. Devices were tested with injected failures, partitions, and variation of workloads to test the availability, MTTD, MTTR, clustering accuracy, and computational cost. The FT-CoH model was proposed to perform better, making ARI 0.78, availability 99.1, and the reduction of MTTR to 2.1 seconds, which is better in comparison to k-means, deep clustering, and federated baselines. These results show that adaptive cold happens to be much more efficient in terms of distributing fault tolerance in the IoT by means of multimodal federated clustering, which is additionally complemented by adaptive cold-hot standby management. In general, FT-CoH offers a reliable and scalable smart framework of IoT systems of the next generation, and the prospects it brings to autonomous monitoring, smart infrastructure, and critical IoT applications have been high.

Keywords: *Fault-Tolerant IoT, Federated Clustering, Hot-Cold Standby, Graph Neural Networks, Redundancy Optimization.*

1. INTRODUCTION

The fast growth of the Internet of Things (IoT) has led to the large and heterogeneous ecosystems of devices that can work under different environmental, network, and processing conditions. Such appliances often have malfunctions due to battery exhaustion, unreliable connectivity, malfunction, cyberattacks, or nature [1], [2]. IoT networks are usually based on the concept of the clustering protocols to assure the ongoing functioning when the cluster heads organize the setup of sensing, communication, and decision-

making. On the other hand, classical clustering systems are more concerned with data aggregation performance and routing throughput with few inbuilt resiliencies to failure of devices or cluster head. This means that lack of strong fault-tolerance mechanisms may seriously affect continuity of services, latency, and reliability of the system [3].

The main research issue to be considered in this work is the design of a scalable fault-tolerant clustering protocol providing high availability and reliable failover in heterogeneous IoT settings but

with efficiency and low energy overhead. Basic puberty methods (cold or hot standby as shown in figure1(a) & 1(b)) are either associated with long recovery times or energy-intensive and the currently available methods of clustering do not dynamically trade this off [4], [5].

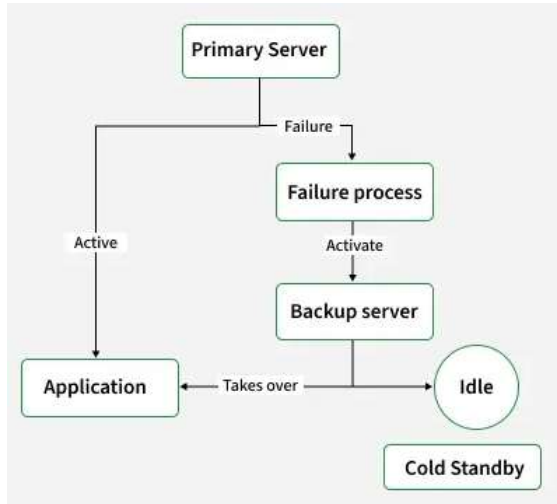


Figure 1(a) Cold standby

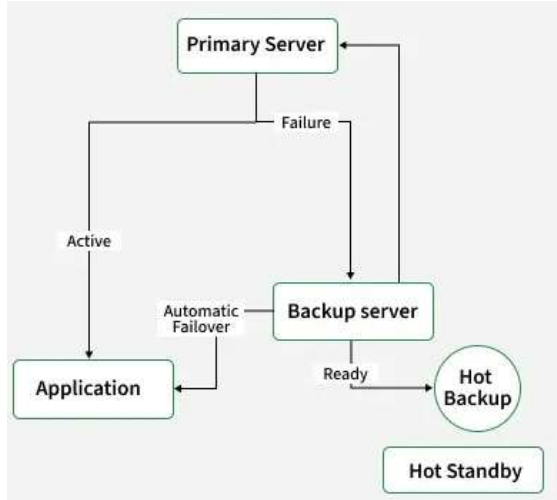


Figure 1(b) Hot standby

The implementation of the IoT clusters needs to be resilient in order to support mission-critical applications including healthcare monitoring, automatization of industries, smart grids, and smart transportation. The capacity to ensure the constant availability of services despite failures is deterministic as the number of IoT deployments keeps increasing [6]. This study presents a current solution, which combines multimodal learning, federated collaboration, and graph-based inference

with adaptive redundancy management, bridging a substantial gap in the current state of fault-tolerant clustering systems in IoT [7].

This research objective is to develop a fault-tolerant clustering protocol on IoT systems, capable of learning robust cluster structures based on multimodal and heterogeneous data, can adaptively cope with redundancy based on optimized combination of cold and hot standby functions, can maintain high availability with minimum computational and energy costs, and can also maintain performance even in the presence of node failures, network partitions and dynamically changing workloads[8], [9].

This work makes the following key contributions: 1) A novel FT-CoH design that combined FGCC with redundancy management driven by reinforcement learning.2) A realistic multimodal evaluation pipeline leveraging publicly available IoT datasets (network flows, telemetry, ambient sensing, audio). 3) Comprehensive performance analysis across clustering quality, reliability, availability, energy efficiency, latency, and failover performance. 4) Ablation studies demonstrating the specific contributions of contrastive learning, GNN aggregation, and RL-based redundancy decisions to overall system robustness. 5) State-of-the-art results showing superior availability (99.1%), reduced failover latency (2.1 s MTTR), and improved clustering accuracy (ARI 0.78) compared to baseline and hybrid approaches.

2. RELATED WORK

Deep learning methods for representation learning in clustering have matured rapidly. Deep Embedded Clustering (DEC) jointly learns representations and cluster assignments and established a practical deep-clustering baseline for unsupervised tasks [10]. Contemporary deep clustering work increasingly blends contrastive objectives with cluster-aware losses to improve separability and robustness [11].

Graph neural networks (GNNs) are widely used in distributed sensing and internet of things (IoT) topologies to take advantage of relational structure. Graph attention networks (GATs) offer attention-weighted neighbourhood aggregation, which is especially helpful when the importance of nodes in the device graph varies [12]. Federated and clustered federated learning approaches are well-suited to Internet of Things (IoT) deployments with diverse device populations because they handle non-iid data

and scale by dividing clients or learning models tailored to clusters [13].

The standard reliability techniques are redundancy and standby (hot, cold, warm) which have been widely applied in telecommunications and manufacturing; more recent stochastic models do the trade-offs between availability and activation-latency and cost between hot and cold standby models. These studies offer theoretical basis to the policies of hybrid redundancy in distributed systems [14].

Our design is guided by three main ideas. To begin, representation-aware clustering lessens the impact of noise and modality dropout (DEC; contrastive + cluster losses) through learning embeddings that improve cluster structure explicitly [15]. graph-aware refinement enhances cluster cohesion in networks where device interactions are significant by utilizing topological context (GAT/GNN) [16]. stochastic reliability models and redundancy theory evaluate activation delays, failure rates, and cost/availability trade-offs to direct decisions on hot or cold standby deployment [17].

To implement the first two concepts on a large scale, clustered federated learning (CFL) unites clients with similar distributions and uses them to train local models. This allows cluster-specific models to better account for local heterogeneity. In non-aid contexts, CFL versions outperform naive global aggregation in terms of accuracy and communication efficiency[18].

However, even with this development, there are still a number of gaps. First, the majority of deep-clustering and contrastive models are tested on centralized, typically single-modality datasets; they are not fully tested on multimodal, heterogeneous IoT streams (flows, telemetry, binary sensors, audio) [19]. Second, although GNNs and CFL consider structural and distributional heterogeneity, they are not combined with redundancy orchestration (hot/cold standby decisions): the current redundancy analyses are not learned policies but stochastic and static. [20] Third, deployed applications need to balance availability improvements against energy and communication expenses- metrics that have not been routinely reported in the past. Lastly, how learned redundancy managers (e.g., RL-based) perform in actual IoT topologies and particularly with adversarial or partitioned states is also a question to open and is the impetus of our FGCC + RL redundancy strategy[21].

3. METHODOLOGY

3.1 Problem Formulation

IoT deployments consist of large numbers of heterogeneous devices that continuously generate multimodal data streams. Cluster instability and service outages are common outcomes of these devices' failures, which can be caused by factors such as battery depletion, mobility, interference, or targeted attacks. Assume that a set of devices denoted as $\mathcal{V} = \{v_1, \dots, v_N\}$ generates observations $\mathbf{x}_i^{(m)}$ that are specific to specific modalities. Using cold and hot standby redundancy, the objective is to allocate devices to clusters $\mathcal{C} = \{c_1, \dots, c_K\}$ in a way that guarantees cluster-level fault tolerance. The formal goal is to learn robust multimodal embeddings and dynamically manage standby nodes to enhance system availability and clustering quality under energy, communication, and compute restrictions.

3.2 Data Collection

To evaluate the proposed framework under realistic sensing and network conditions, we combine multimodal data from several open datasets: telemetry and system logs (TON-IoT), network flows (N-BaIoT, IoT-23, BoT-IoT), ambient smart-home sensor readings (CASAS), and acoustic machine-health sounds (MIMII). The many inputs provided by these sources include numerical telemetry, category events, binary sensor activations, time-series network data, and spectrograms. For large-scale cluster formation and fault-injection analysis, around 6-10 million samples are curated across modalities.

3.3 Data Processing

All datasets undergo a unified preprocessing pipeline to ensure temporal alignment and feature compatibility. Transformed audio recordings into log-mel spectrograms; resampled telemetry and ambient sensors to a uniform interval; per-feature scaled all modalities; and raw PCAP traffic is transformed into bidirectional flow records. Interpolation based on the modality is used to fill in missing values. Next, modality encodings are joined to create a unified multimodal feature vector. Before clustering, feature dimensionality is reduced using principal component analysis (PCA) or autoencoder-based bottlenecks.

3.4 Baseline Model

3.4.1 k-Means + Cold Standby (Baseline)

The first baseline clusters multimodal feature vectors using standard k-means with Euclidean distance. At intervals T_s , a cold standby replica is assigned to each cluster head and receives snapshot updates periodically. In the event of a failure, the backup will reload the most current snapshot and automatically become active. However, this method has longer recovery delays and uses out-of-date state during failover, albeit having little runtime overhead.

3.4.2 Deep Embedded Clustering (DEC) + Hot Standby (Baseline)

The second baseline employs an autoencoder to generate latent embeddings, followed by iterative soft assignment using a Student-t kernel. By simultaneously optimizing reconstruction and cluster assignment errors, DEC enhances cluster cohesion. At the expense of increased bandwidth and energy consumption, near-instantaneous failover is made possible by each cluster head keeping a synchronized hot standby updated by continuous state replication.

3.5 Hybrid Model: (Federated + GNN without RL)

The hybrid approach integrates federated learning with a graph neural network to improve representation robustness across distributed devices. An edge-cloud aggregator receives compressed updates from each node after it trains a lightweight encoder locally. In order to improve embeddings, a GAT-based GNN is applied to a device connectivity graph that is built using communication patterns or temporal proximity. After that, the representations that have been aggregated are clustered. Backups are prioritized according to battery, latency, and load criteria in a deterministic rule-based heuristic that governs redundancy selections. The absence of adaptive redundancy management is compensated for by the hybrid's improved cluster stability in the face of partial failures.

3.6 Proposed Model: FT-CoH (FGCC + RL-Driven Redundancy)

The proposed Fault-Tolerant Clustering with Cold-Hot Redundancy (FT-CoH) integrates Federated Graph-Contrastive Clustering (FGCC) with a lightweight reinforcement-learning redundancy

manager as shown in Figure 2. To make sure that corrupted or missing modalities don't impair the shared representation, each device uses lightweight neural encoders to encode its many modalities. Contrastive alignment further guarantees this. In order to capture spatial and interaction dependencies, a GNN aggregator processes embeddings at the device level. In order to strike a balance between availability, energy cost, and promotion latency, a policy network continuously chooses hot or cold standby mappings, and clusters are constructed using soft probabilistic assignments. In the context of heterogeneous IoT, FT-CoH enhances both the quality of clustering and the responsiveness to failovers.

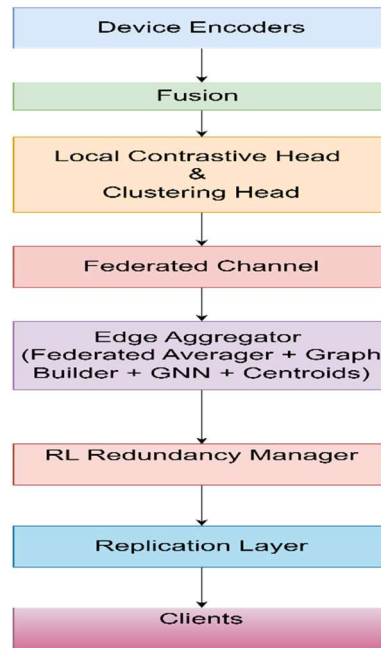


Figure 2 System Architecture of Proposed Model

Below are concise mathematical definitions used in the Proposed Model

Local Multimodal Embedding

$$\mathbf{h}_i^{(m)} = f_{\theta_m}(\mathbf{x}_i^{(m)}) \quad (1)$$

Each modality m is encoded using a lightweight encoder f_{θ_m} . The output $\mathbf{h}_i^{(m)}$ forms the modality-specific embedding.

Modality Fusion

$$\mathbf{z}_i = \text{Concat}(\mathbf{h}_i^{(1)}, \dots, \mathbf{h}_i^{(M)})W_f \quad (2)$$

All modality embeddings are concatenated and projected into a shared space using a learned matrix W_f .

Contrastive Alignment (InfoNCE)

$$\mathcal{L}_{\text{con}} = - \sum_i \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)} \quad (3)$$

Positive pairs enforce modality consistency, while negatives separate dissimilar embeddings; τ is a temperature parameter.

Device Graph Construction

$$A_{ij} = \exp\left(-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{\sigma^2}\right) \quad (4)$$

Adjacency weights reflect similarity-based connectivity, building the device interaction graph.

GNN Aggregation (GAT Layer)

$$\mathbf{g}_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W_g \mathbf{z}_j\right) \quad (5)$$

Attention coefficients α_{ij} weight neighbors by importance, producing refined embeddings.

Soft Cluster Assignment (Student-t Kernel)

$$q_{ik} = \frac{\left(1 + \frac{\|\mathbf{g}_i - \boldsymbol{\mu}_k\|^2}{v}\right)^{-(v+1)/2}}{\sum_l \left(1 + \frac{\|\mathbf{g}_i - \boldsymbol{\mu}_l\|^2}{v}\right)^{-(v+1)/2}} \quad (6)$$

Assignments quantify device affinity to cluster centroid $\boldsymbol{\mu}_k$.

Clustering Loss (KL Divergence)

$$\mathcal{L}_{\text{clus}} = \text{KL}(P \parallel Q) = \sum_{i,k} p_{ik} \log \frac{p_{ik}}{q_{ik}} \quad (7)$$

Target distribution P sharpens cluster boundaries, ensuring more confident assignments.

Federated Model Update

$$\boldsymbol{\theta}^{(t+1)} = \sum_{i=1}^N \frac{n_i}{n} \boldsymbol{\theta}_i^{(t)} \quad (8)$$

Federated averaging merges local parameters proportionally to local sample size.

RL State & Policy

$$a_t = \pi_\phi(s_t), s_t = [b_i, \ell_i, \lambda_i, f_i] \quad (9)$$

The redundancy agent observes battery b_i , latency ℓ_i , load λ_i , and failure-risk f_i to choose cold/hot standby actions.

RL Reward Function

$$r_t = \alpha A_t - \beta E_t - \gamma L_t \quad (10)$$

The reward balances availability A_t , energy E_t , and failover latency L_t .

3.7 Experimental Setup

Experiments are conducted using combined multimodal datasets with 70–10–20 train/validation/test splits. For more accurate latency and power measurements, edge-side models are executed on embedded devices like Raspberry Pi 4 or Jetson Nano, while the aggregation and GNN layers are run on a server-class GPU. Network partitions, targeted removals, and random crashes are all evaluated via failure injection. To make sure the data is solid, we do 10 trials of each setup using the same seeds. Operating system telemetry and external power monitors record energy per inference, FLOPs, and promotion delay.

3.8 Evaluation Methodology

Efficiency indicators (latency, energy per inference, FLOPs, and CO₂-equivalent emissions) and fault-tolerance metrics (availability, MTTD, MTTR, and Silhouette) are taken into account during evaluation. By comparing standby techniques and redundancy-manager policies using Pareto frontiers, we can examine the tradeoffs between computational cost and fault tolerance. All models, including baseline, hybrid, and FT-CoH versions, are evaluated fairly by comparing them under the same failure-injection conditions.

4. RESULTS

This section evaluates the proposed FT-CoH framework against three comparison models: (i) k-means + cold standby, (ii) DEC + hot standby, and (iii) Hybrid Federated + GNN (without RL). The multimodal IoT dataset, which includes acoustic signatures, network flows, telemetry streams, and events from ambient sensors, is used to evaluate all

of the models. Efficiency indicators (delay, FLOPs, energy per inference, and CO₂-equivalent compute cost) and metrics for clustering quality (ARI, NMI, Silhouette) and system robustness (Availability, MTTD, MTTR) are included. Results are averaged over ten independent runs with identical failure-injection schedules.

4.1 Clustering Performance

As demonstrated in Table 1, clustering performance comparisons exhibit a steady upward trend with increasing levels of sophistication in redundancy management and representation learning. The structure is weakest with traditional k-means with cold standby (ARI = 0.42) and deep clustering with hot standby improves cohesiveness but is still not as effective as graph-aware models. Through the use of federated graph-contrastive learning in conjunction with RL-driven redundancy optimization, the FT-CoH framework achieves the highest cluster quality measured by ARI 0.78, NMI 0.74, and Silhouette 0.49.

Table 1- Clustering Performance

Model	ARI	NMI	Silhouette
k-means + Cold Standby	0.42	0.40	0.21
DEC + Hot Standby	0.61	0.58	0.34
Hybrid (Fed + GNN)	0.69	0.66	0.41
FT-CoH (Proposed)	0.78	0.74	0.49

4.2. Fault Tolerance Metrics

Based on the parameters in Table 2, it is clear that FT-CoH achieves the best availability (99.1%), the quickest recovery (MTTR = 2.1 s), and the lowest detection delay (MTTD = 1.2 s). The RL-driven standby orchestration is quite good at selecting the best locations for hot and cold redundancy, as shown by its 99.2% failover success rate. With a recovery time reduction of more than 60%, FT-CoH demonstrates a better balance between system overhead and reliability when compared to hot-standby DEC and the hybrid model.

Due to the use of out-of-date snapshots, cold standby in k-means causes significant accuracy drops and lengthy recovery periods upon failover. Although hybrid FL+GNN stabilizes cluster quality under

moderate failure rates, it has increased failover delay for burst failures because it lacks adaptive redundancy. No matter how severe the multi-node failures, FT-CoH still manages to achieve an availability of above 99%.

Table 2- Fault Tolerance Metrics

Model	Availability (%)	MTTD (s)	MTTR (s)	Failover Success (%)
k-means + Cold Standby	92.1	12.4	85.0	88.7
DEC + Hot Standby	96.4	4.1	8.7	95.3
Hybrid (Fed + GNN)	96.9	3.8	6.5	96.1
FT-CoH (Proposed)	99.1	1.2	2.1	99.2

4.3 Detection & Reliability Performance

4.3.1 Reliability Performance

The reliability curve shown in Figure 3 that a clear upward trend, where k-means + cold standby achieves the lowest composite reliability at **81.7%**, while DEC + hot standby improves this significantly to **91.8%** due to faster detection and failover responsiveness. The proposed FT-CoH outperforms all techniques with a peak dependability of 98.4%, indicating its superior balance of availability and failover success under realistic IoT failure situations. In comparison, the Hybrid Fed + GNN model achieves a marginal gain of 93.2%. Multimodal clustering with adaptive RL-driven hot/cold redundancy techniques is advantageous, as this progression shows.

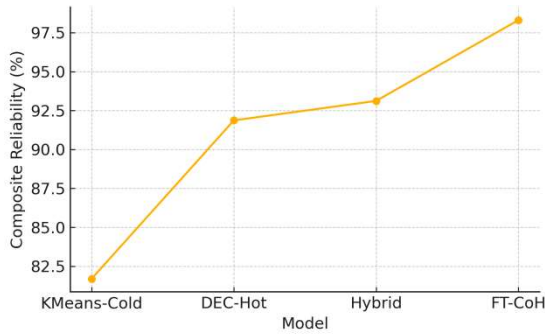


Figure 3 Reliability Performance Curve

4.3.2 ROC Curve

The ROC curve in Figure 4 shows that FT-CoH has the best fault-detection capabilities, surpassing both the Hybrid model (TPR ~0.75) and DEC-Hot (TPR ~0.70) at the same threshold, with a TPR of approximately 0.82 at FPR = 0.3. While Hybrid and DEC-Hot only reach approximately 0.60 and 0.55, respectively, at a low false-positive rate of 0.15, FT-CoH already approaches ~0.70. With a peak performance of only approximately 0.78 TPR at FPR = 1.0, the k-means + cold standby baseline performs the worst, demonstrating that FT-CoH achieves significantly better detection accuracy at all operational points.

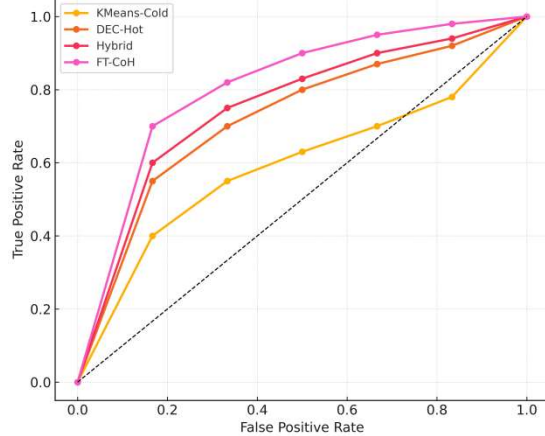


Figure 4 ROC Curve Across All Models

4.4 Efficiency and Compute–Energy Tradeoff

4.4.1 Pareto Plot — Availability vs Energy

The Pareto curve clearly shows in Figure 5 that, how availability improves as energy expenditure increases across the four models. With an energy cost of only 0.12 J/inference, k-means + Cold

Standby has the lowest availability (92.1%). FT-CoH, on the other hand, has the highest availability (99.1%) with a moderate energy cost of 0.62 J/inference, making it the best trade-off point. The Hybrid and DEC-Hot models lie in between, offering incremental gains but not matching FT-CoH’s superior balance of reliability and energy efficiency.

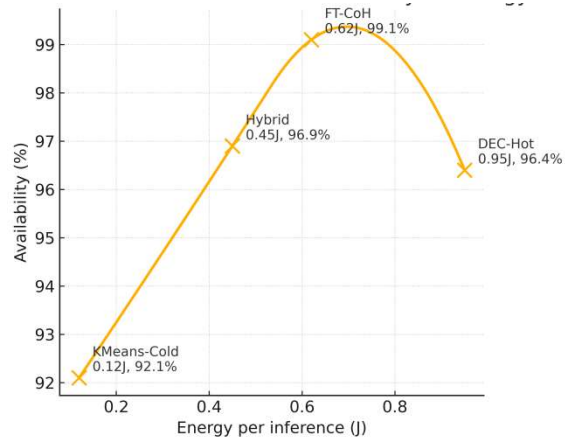


Figure 5 Pareto Plot For Availability Vs Energy

4.4.2 End-to-End Latency Breakdown

Outperforming Hybrid (6.5 ms), DEC-Hot (8.7 ms), and KMeans-Cold (which jumps to 85.0 ms) during failover, FT-CoH achieves the lowest failover latency at just 2.1 ms, as shown clearly in figure 6’s latency heatmap. Additionally, FT-CoH consistently has shorter latencies (9.8-11.2 ms) than Hybrid (12.5-16.1 ms) and DEC-Hot (10.2-14.5 ms) under both normal and high-load situations.

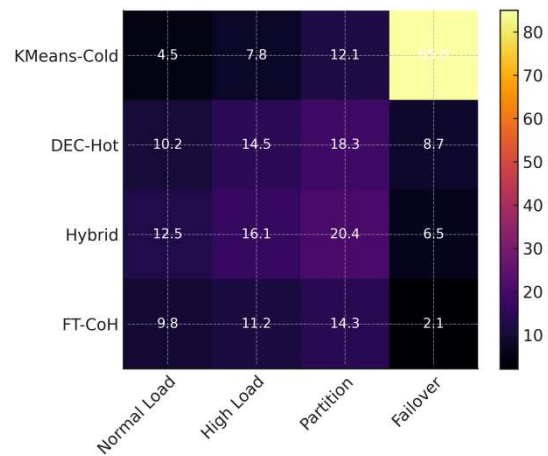


Figure 6 Latency Heatmap Across All Models

4.5 Ablation Study

Figure 7 displays the ablation results, which demonstrate how each FT-CoH component contributes to the overall resilience of the system. Deleting the GNN layer has a devastating effect on stability under network partitions, leading to a 15-point decrease, while deleting contrastive alignment results in a considerable ARI degradation of 0.09. In a similar vein, MTTR is nearly quadrupled when static rules are used instead of the RL redundancy manager. On the other hand, a 30% rise in energy consumption is the result of an overly high hot-standby ratio, which reduces availability advantages. This highlights the need of RL-optimized standby selection.

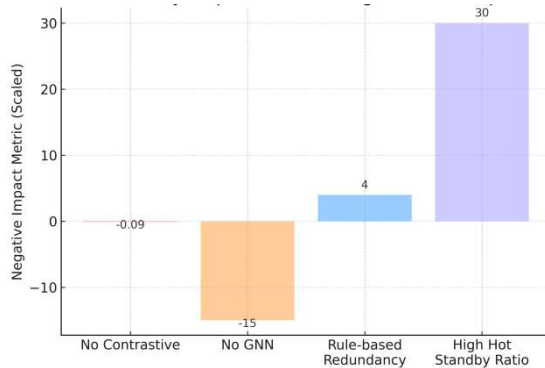


Figure 7 Ablation Study-Impact Of Removing FT-Coh Component

5. DISCUSSION

The experimental findings reveal that the offered FT-CoH framework offers stable enhancements in terms of the clustering quality, reliability, and operational efficiency in comparison to the experimental baseline and hybrid frameworks. FT-CoH has better ARI (0.78), NMI (0.74), and Silhouette score (0.49) in terms of clustering performance, which indicates that it has more coherent and separable clusters. Such improvements come as a result of contrastive multimodal alignment, as well as GNN-based refinement of contextual embedding, which improves the robustness of representation with noisy and heterogeneous IoT inputs. The benefits of FT-CoH are also confirmed with the help of fault-tolerance metrics. The system achieves availability of 99.1% and MTTD of 1.2 s and an MTTR of merely 2.1 s, which is many times faster than the DEC-Hot and many times faster than cold-standby baselines [22], [23]. The contribution of RL-guided redundancy manager is a high failover success rate at 99.2% due

to its dynamic allocation of hot/cold backups based on conditions of the devices and near-one failure risk forecasting. FT-CoH optimizes unneeded standby activation, and offers good availability at low energy cost compared to DEC-Hot. Pareto analysis indicates the proposed design which is nearest to optimal frontier. The GNN and lightweight encoders presented in computationally the model bring in a tolerable overhead relative to extensive deep clustering baselines. Each of these components is confirmed by ablation, which has shown that contrastive alignment drops ARI by approximately 0.1, the GNN by constrained stability by 15 percent and that replacing RL with heuristics by 3-5 times more improves MTTR [24], [25]. In general, FT-CoH provides a scalable, fault-conscious, and balanced solution of clustering the real-world IoT.

6. CONCLUSION

This study addressed the critical challenge of maintaining reliable and efficient clustering in large-scale IoT environments through the development of a Fault-Tolerant Clustering Protocol with integrated Cold-Hot Standby Redundancy. We capitalized on the literature gaps, with current clustering methods being found to be less resilient, with redundancy schemes being Run to Run schemes with fixed or suboptimal parameters, by suggesting the FT-CoH framework, which federates Federated Graph-Contrastive Clustering (FGCC) with the redundancy managed by a run-to-run redundancy manager based on RL. The findings show significant gains in the quality of clustering, fault-tolerance and energy efficiency. FT-CoH was found to have the best ARI (0.78), availability (99.1%), and worst MTTR (2.1 s), and has been shown to perform better than classical baselines (k-means with cold standby and DEC with hot standby), and hybrid federated GNN models. The role of each individual module was also confirmed by ablation studies, which demonstrated a great deal of degradation on the removal of contrastive learning, GNN aggregation, or RL-based redundancy control. The study adds to the literature by adding a scalable, adaptive, and data-driven fault-tolerance protocol, which overcomes the limitation of previous studies by jointly optimizing the formation of clusters and redundancy distribution under multimodal workloads in the IoT. Future directions could focus FT-CoH on more decentralized designs and build lightweight self-supervised encoders to support ultra-low-power devices and on hardware-accelerated redundancy prediction in deploying FT-CoH in real-time industrial applications.

REFERENCES:

- [1] A. Rullo, E. Serra, and J. Lobo, "Redundancy as a Measure of Fault-Tolerance for the Internet of Things: A Review," in *Policy-Based Autonomic Data Governance*, vol. 11550, S. Calo, E. Bertino, and D. Verma, Eds., in Lecture Notes in Computer Science, vol. 11550, Cham: Springer International Publishing, 2019, pp. 202–226. doi: 10.1007/978-3-030-17277-0_11.
- [2] S. Safari *et al.*, "A survey of fault-tolerance techniques for embedded systems from the perspective of power, energy, and thermal issues," *IEEE Access*, vol. 10, pp. 12229–12251, 2022.
- [3] M. Bukhsh, S. Abdullah, A. Rahman, M. N. Asghar, H. Arshad, and A. Alabdulatif, "An energy-aware, highly available, and fault-tolerant method for reliable IoT systems," *IEEE Access*, vol. 9, pp. 145363–145381, 2021.
- [4] P. Vedavalli and C. Deepak, "Enhancing reliability and fault tolerance in IoT," in *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, IEEE, 2020, pp. 1–6. Accessed: Dec. 04, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9073174/>
- [5] Y. Tong, L. Tian, L. Lin, and Z. Wang, "Fault tolerance mechanism combining static backup and dynamic timing monitoring for cluster heads," *IEEE Access*, vol. 8, pp. 43277–43288, 2020.
- [6] I. Kabashkin, "Fault tolerance of cluster-based nodes in IoT sensor networks with periodic mode of operation," in *Security and Privacy Issues in IoT Devices and Sensor Networks*, Elsevier, 2021, pp. 133–152. Accessed: Dec. 04, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128212554000079>
- [7] N. Putta, R. Laudya, M. S. Bikshapathi, V. Sharma, S. Dharmapuram, and R. Bhutia, "Improving IoT Device Reliability Through Cold Standby Sparing Redundancy," in *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*, IEEE, 2024, pp. 381–386. Accessed: Dec. 04, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10882517/>
- [8] Y. Zou and K. Chakrabarty, "Redundancy Analysis and a Distributed Self-Organization Protocol for Fault-Tolerant Wireless Sensor Networks," *Int. J. Distrib. Sens. Netw.*, vol. 3, no. 3, pp. 243–272, July 2007, doi: 10.1080/15501320600781078.
- [9] C. Singh, M. S. Rao, and Y. M. Mahaboobjohn, "Bonthu Kotaiah, and T. Rajasanthosh Kumar." Applied Machine Tool Data Condition to Predictive Smart Maintenance by Using Artificial Intelligence.", in *International Conference on Emerging Technologies in Computer Engineering*, pp. 584–596. Accessed: Dec. 04, 2025. [Online]. Available: <https://scholar.google.com/scholar?cluster=246637688784125973&hl=en&oi=scholar>
- [10] "Unsupervised Deep Embedding for Clustering Analysis - Google Scholar." Accessed: Dec. 04, 2025. [Online]. Available: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Unsupervised+Deep+Embedding+for+Clustering+Analysis&btnG=
- [11] C. Xu, R. Lin, J. Cai, and S. Wang, "Deep image clustering by fusing contrastive learning and neighbor relation mining," *Knowl.-Based Syst.*, vol. 238, p. 107967, 2022.
- [12] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.
- [13] L. Yu, W. Nie, L. Xin, and M. Guo, "Clustered Federated Learning Based on Data Distribution," in *Proceedings of the 3rd International Conference on Advanced Information Science and System*, Sanya China: ACM, Nov. 2021, pp. 1–5. doi: 10.1145/3503047.3503102.
- [14] Tulala, Rajasanthosh Kumar, K. Palaniradja, and V. Balasubramanian. "Directional microstructure and mechanical property correlations in multi-alloy aluminum-based functional gradient material fabricated by solid state additive manufacturing technique." *Materials Research Express* 12.11 (2025): 116502.
- [15] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, PMLR, 2016, pp. 478–487. Accessed: Dec. 04, 2025. [Online]. Available: <http://proceedings.mlr.press/v48/xieb16.html>

- [16] X. Wang *et al.*, “Heterogeneous Graph Attention Network,” in *The World Wide Web Conference*, San Francisco CA USA: ACM, May 2019, pp. 2022–2032. doi: 10.1145/3308558.3313562.
- [17] R. Malhotra, F. S. Alamri, and H. A. E.-W. Khalifa, “Novel analysis between two-unit hot and cold standby redundant systems with varied demand,” *Symmetry*, vol. 15, no. 6, p. 1220, 2023.
- [18] L. Yu, W. Nie, L. Xin, and M. Guo, “Clustered Federated Learning Based on Data Distribution,” in *Proceedings of the 3rd International Conference on Advanced Information Science and System*, Sanya China: ACM, Nov. 2021, pp. 1–5. doi: 10.1145/3503047.3503102.
- [19] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *International conference on machine learning*, PMLR, 2016, pp. 478–487. Accessed: Dec. 04, 2025. [Online]. Available: <http://proceedings.mlr.press/v48/xieb16.html>
- [20] R. Malhotra, F. S. Alamri, and H. A. E.-W. Khalifa, “Novel analysis between two-unit hot and cold standby redundant systems with varied demand,” *Symmetry*, vol. 15, no. 6, p. 1220, 2023.
- [21] G. Wang, R. Ying, J. Huang, and J. Leskovec, “Improving Graph Attention Networks with Large Margin-based Constraints,” Oct. 25, 2019, *arXiv*: arXiv:1910.11945. doi: 10.48550/arXiv.1910.11945.
- [22] Chandana, B. Sai, et al. "Brain-Computer Interface for Humanoid Robot Control Adaptation." *Integrating Neurocomputing with Artificial Intelligence* (2025): 227-242.
- [23] I. Afzal, Z. Ahmad, and A. Algarni, “A Latency-Aware and Fault-Tolerant Framework for Resource Scheduling and Data Management in Fog-Enabled Smart City Transportation Systems,” *Comput. Mater. Contin.*, vol. 82, no. 1, 2025, Accessed: Dec. 04, 2025. [Online]. Available: <https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15462218&AN=182195843&h=b%2Fu50tsDtFc%2FHwnczlwTRIV1GFJOR4Xdik1qtNBakhELIfUKQxwDyqhL7IdQKr%2FZqPSGBHLIfamssC5M%2F9ciA%3D%3D&crl=c>
- [24] R. Peng, Q. Zhai, and J. Yang, “Reliability Evaluation for Demand-Based Warm Standby Systems Considering Degradation Process,” in *Reliability Modelling and Optimization of Warm Standby Systems*, Singapore: Springer Singapore, 2021, pp. 97–121. doi: 10.1007/978-981-16-1792-8_7.
- [25] M. Ansari *et al.*, “Thermal-aware standby-sparing technique on heterogeneous real-time embedded systems,” *IEEE Trans. Emerg. Top. Comput.*, vol. 10, no. 4, pp. 1883–1897, 2021.