



Original Research Paper

Counterfactual Causal Graph Memory with Dual-Drift Attribution and Trust-Gated Continual Learning for Zero-Day Intrusion Detection

S Vimal¹, M. Kaliappan², Gaurav Dhiman^{3,4,5}, R. Renuga Devi⁶

¹Department of Computer Science and Engineering, KIT- Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India

²Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India

³Department of Computer Science and Engineering, Yuan Ze University, Tao Yuan, Taiwan

⁴University Centre for Research and Development, Chandigarh University, Mohali, India

⁵Division of Research and Development, Lovely Professional University, Phagwara, India

⁶Department of Computer Science and Engineering, KIT- Kalaignar Karunanidhi Institute of Technology, Coimbatore, Tamil Nadu 641402 India

Email: kaliappan@ritrjpm.ac.in, svimal@kitcbe.ac.in, gdhiman0001@gmail.com, renugadevime@gmail.com

Abstract—Network intrusion detection systems are increasingly asked to operate in an open, non-stationary world in which genuinely novel zero-day attacks appear and the benign traffic environment continually drifts. Signature and closed-set classifiers fail on unseen attacks because they confidently map every flow to a known class, and naive online learners that adapt to streaming traffic risk silently absorbing the very novel attacks they should flag. This paper presents a framework that unifies robust zero-day detection with safe continual adaptation for the Internet of Things (CausalTwin-IDS). Each network flow is encoded through two behavioral streams, a flow-statistics stream and an Internet-Protocol interaction-graph stream, into a class-relevant causal representation, while contextual attributes are routed to a separate nuisance representation. A counterfactual intervention that permutes nuisance factors across device-group environments and enforces invariance of the causal code and its prediction makes the attack representation robust to benign environment change rather than merely correlated with it. Known attack families are summarized by a multi-prototype memory, and zero-day flows are detected by a source-only split-conformal score that fuses prototype distance, free energy, tri-view graph disagreement, and predictive entropy, with a finite-sample false-alarm guarantee obtained without ever observing a zero-day sample. A dual-drift attribution module separates benign environment drift from genuine attack-causal drift using two kernel maximum-mean-discrepancy signals, and a trust gate admits an online prototype update only when a flow is confidently known, tri-view consistent, and accompanied by attack-rather-than-environment drift, structurally preventing zero-day contamination of the memory. On the real ToN-IoT network benchmark with four attack families held out as zero-day, CausalTwin-IDS attains an H-score of 87.9%, a zero-day detection area under the receiver operating characteristic curve of 0.923, and an open-set classification-rate area of 0.798, exceeding softmax, OpenMax, energy, and autoencoder baselines, while its trust gate retains 90.8% known-attack accuracy under 50% zero-day contamination where a naive updater collapses to 64.1%. Every reported number is produced by an executed run in which zero-day labels are used only for final evaluation, never for training, calibration, or adaptation.

Keywords: Zero-day attack detection, intrusion detection system, open-set recognition, causal representation learning, counterfactual intervention, prototype learning, conformal prediction, concept drift, continual learning, Internet of Things, ToN-IoT.

***Corresponding Author:**

Received: 25th July, 2026 | Revised: 26th August, 2026 | Accepted: 18th September, 2026 | Available Online: 24th September, 2026

(DOI): 10.70102/AEJ.2026.18.7s.44

1. Introduction

Intrusion detection systems (IDS) are a cornerstone of network defense, yet the environment in which they operate has become fundamentally open and non-stationary. New attack techniques, so-called zero-day attacks whose signatures were never observed during training, appear continually, and the benign traffic of a monitored network drifts as devices, services, and usage patterns change. Signature-based detectors cannot recognize attacks they have never seen, and supervised closed-set classifiers, however accurate on known attack families, are structurally unable to flag a novel one: they assign every flow to one of the known classes, typically with high confidence, so a zero-day attack is silently misclassified as benign traffic or as the wrong known attack. The problem addressed in this paper is open-world intrusion detection for the Internet of Things: accurately classifying known attack families, detecting and rejecting zero-day attacks, and safely adapting the detector as the traffic environment drifts, all without access to labels for the novel attacks.

Three coupled difficulties make this problem hard. First, the discriminative signal for an attack is entangled with the benign network context in which it occurs; a detector that keys on contextual artifacts such as particular source addresses or ports will generalize poorly when the environment shifts, and will confuse an environment change for an attack or vice versa. Second, the events of greatest importance, zero-day attacks, are by definition absent from training, so any rejection threshold must be calibrated without them, yet remain valid when a real novel attack finally arrives. Third, an operational IDS must adapt online to keep pace with drift, but adapting on an unlabeled traffic stream that contains undetected zero-day attacks risks absorbing those attacks into the model's notion of normal behavior, a failure mode we term memory contamination. An effective open-world IDS must therefore address representation robustness, calibrated novelty rejection, and contamination-safe adaptation jointly.

A large and growing literature attacks parts of this problem. Zero-day detection has been pursued with unsupervised anomaly detectors, autoencoders, and generative models; graph neural networks have been applied to flow-based detection; and concept-drift and continual-learning methods adapt models over time. However, these threads remain largely separate, and several gaps persist for the specific task of calibrated, drift-aware, contamination-safe zero-day detection. Most anomaly detectors provide a novelty score but no finite-sample guarantee on the false-alarm rate, which is critical when analyst time is scarce. Few detectors disentangle the causal attack signal from the benign environment, so they cannot distinguish a novel attack

from an ordinary environment change. And the drift-adaptation and novelty-detection literatures rarely meet: online IDS learners seldom guard against contamination by undetected attacks, while novelty detectors seldom adapt. No prior IDS, to our knowledge, couples counterfactual causal disentanglement, a source-only conformal novelty guarantee, dual attribution of drift into benign versus attack components, and a trust-gated update rule that is provably conservative against zero-day contamination.

This paper closes those gaps with CausalTwin-IDS. The framework encodes each flow through two behavioral streams, flow statistics and an Internet-Protocol interaction graph, into a class-relevant causal representation, and routes contextual attributes to a separate nuisance representation. It enforces genuine causal robustness through a counterfactual intervention that swaps nuisance factors across device-group environments and requires the causal code and its attack prediction to be unchanged, directly implementing the do-intervention that defines a causal factor. Known attack families are summarized by a multi-prototype memory, and zero-day flows are detected by a source-only conformal score fusing prototype distance, free energy, a tri-view graph-disagreement signal, and predictive entropy, whose evidence weights are selected by a pseudo-open simulation that never observes a real zero-day. A dual-drift attribution module separates benign environment drift from genuine attack-causal drift, and a trust gate refreshes the prototype memory only for flows that are confidently known, tri-view consistent, and accompanied by attack rather than environment drift, so undetected attacks and ambiguous flows are structurally prevented from contaminating the memory. Rejected flows accumulate in a zero-day memory that is clustered for later analysis, without using any labels.

The contributions of this paper are as follows. (1) We propose a dual-stream causal-nuisance disentanglement for network flows that fuses flow-statistics and interaction-graph views into a causal attack code and enforces robustness through cross-environment counterfactual intervention. (2) We introduce a multi-prototype attack memory and a source-only split-conformal novelty score that fuses four complementary evidences, including a tri-view graph-disagreement signal, with a finite-sample false-alarm guarantee obtained without any zero-day sample. (3) We design a dual-drift attribution module that separates benign environment drift from attack-causal drift, and a trust-gated adaptation rule that adapts under drift while provably resisting zero-day contamination, together with a zero-day discovery memory. (4) We validate the complete framework on

the real ToN-IoT benchmark with four held-out zero-day attack families, reporting known/zero-day accuracy, H-score, detection AUROC, open-set classification rate, calibration, openness sensitivity, contamination resistance, and continual forgetting, all from an executed run in which zero-day labels are used only for post-hoc evaluation.

The remainder of this paper is organized as follows. Section 2 reviews related work on zero-day and IoT intrusion detection, graph-based detection, open-set recognition, causal representation learning, conformal novelty detection, and drift and continual adaptation. Section 3 details the CausalTwin-IDS architecture, its mathematical formulation, and its justification. Section 4 presents the experimental setup and results. Section 5 benchmarks CausalTwin-IDS against recent methods. Section 6 concludes and outlines future work.

2. Related Work

The detection of zero-day attacks, whose signatures are unknown at training time, is a central and long-standing challenge in network security, surveyed comprehensively by Guo [1]. Because signature matching cannot recognize novel attacks, most approaches cast zero-day detection as anomaly or novelty detection. Soltani et al. [2] proposed an adaptable deep-learning IDS that improves detection over multiple phases, and Kumar and Sinha [3] developed a robust intelligent zero-day detection technique. Autoencoder-based anomaly detectors integrated with tree ensembles have been shown to enhance detection of unseen attacks [5], and unsupervised and generalization-oriented techniques have been explored for model transfer [6]. Recent work has extended zero-day detection to the internet of vehicles through open-set federated learning [7], and surveys catalogue deep-learning approaches to novel-malware detection [8]. For the Internet of Things specifically, the ToN-IoT benchmark introduced by Alsaedi et al. [9] and its distributed and standard-feature variants [10], [12] have become the standard testbed, and recurrent, ensemble, and feature-selection detectors have been evaluated on it [11], [13]. These methods, however, generally provide a novelty score without a finite-sample false-alarm guarantee and do not separate the causal attack signal from benign environment drift, which are the focus of CausalTwin-IDS.

Because network flows are naturally relational, connecting source and destination hosts, graph neural networks have become a powerful tool for intrusion detection, surveyed by Bilot et al. [16]. Lo et al. [15] introduced E-GraphSAGE, which generates edge embeddings from flow features and topology for IoT intrusion detection, and Caville et al. [14] proposed Anomal-E, a self-supervised graph IDS that learns

edge embeddings without labeled attacks and feeds them to unsupervised detectors, demonstrating strong generalization to unseen traffic. These works establish that graph structure carries discriminative signal beyond per-flow statistics. CausalTwin-IDS incorporates an Internet-Protocol interaction-graph stream in this spirit, but fuses it with flow statistics inside a causal representation and, distinctively, uses the agreement between the flow view, the graph view, and the fused view as a tri-view novelty signal.

Open-set recognition requires rejecting inputs from classes unseen in training, a problem whose relationship to out-of-distribution detection was recently dissected by Wang et al. [18]. Vaze et al. [17] showed that closed-set accuracy strongly predicts open-set performance, and recent methods improve rejection through inter-class separation and Jacobian analysis [19] and by bridging anomaly detection and open-set recognition [20]; a unified survey covers anomaly, novelty, and out-of-distribution detection together [21]. In security, zero-day detection is precisely an open-set problem, but IDS work has only recently begun to adopt the open-set framing explicitly. CausalTwin-IDS treats zero-day detection as calibrated open-set recognition over network flows, adding the conformal guarantee and the drift-attribution and contamination-safety mechanisms that generic open-set methods lack [42]-[63].

Causal representation learning seeks features that reflect the true generative factors of the label and are therefore stable across environments, a program articulated by Scholkopf et al. [22]. Invariant risk minimization [25] and invariant causal representation learning [24] seek predictors that are optimal across environments, and causality-inspired representation learning disentangles causal from spurious factors for domain generalization [23]. Domain-generalization surveys [26], [27] catalogue the broader family of invariance methods, and domain-adversarial training [28] provides a practical mechanism for removing environment information from a representation. In intrusion detection, the benign network context is a powerful confounder, and a detector that keys on it fails under drift. CausalTwin-IDS applies causal disentanglement to network flows and, beyond adversarial alignment, performs an explicit counterfactual intervention that swaps benign context across environments, a stronger test of causal robustness.

Prototype methods represent each class by one or more representative points and classify by distance, a paradigm from prototypical networks [29] extended with neighbor constraints [30] and distribution calibration for incremental settings [31]; distance to a prototype is a natural novelty score. To make novelty rejection principled, conformal prediction provides

distribution-free, finite-sample coverage guarantees [32], with recent extensions beyond exchangeability [33] and to distribution shift [34]. Deep ensembles [35] and energy-based scores [36] are strong uncertainty and novelty signals. CausalTwin-IDS combines a multi-prototype attack memory with a split-conformal false-alarm guarantee and fuses prototype, energy, entropy, and a tri-view graph-disagreement signal, selecting the fusion by a pseudo-open simulation on known traffic alone.

Operational network traffic is non-stationary, so an IDS must cope with concept drift, reviewed by Lu et al. [38], and ideally adapt continually. The kernel maximum-mean-discrepancy two-sample test [37] is a standard drift signal. Test-time and continual adaptation methods, surveyed by Liang et al. [39], update deployed models on unlabeled data, and class-incremental methods combat catastrophic forgetting [40]. Explainable IDS for the IoT has also been emphasized as a deployment requirement [41], and boosting frameworks remain strong baselines [42]. A recurring danger, acute in security, is that adaptation on an unlabeled stream absorbs undetected attacks. CausalTwin-IDS addresses this directly by attributing drift into benign-environment and attack-causal components and gating updates so that only confidently known, tri-view-consistent flows accompanied by attack-rather-than-environment drift are admitted, anchoring the memory to its source geometry to bound both forgetting and contamination.

In summary, prior work has produced strong zero-day and IoT detectors, graph-based IDS, open-set recognizers, causal-invariance objectives, conformal novelty tools, and drift adapters, but these advances remain siloed. No prior IDS simultaneously (a) disentangles the causal attack signal from benign context through counterfactual intervention over network environments, (b) calibrates zero-day rejection with a source-only conformal false-alarm guarantee, (c) attributes drift into benign versus attack components, and (d) adapts a prototype memory online while provably resisting zero-day contamination. CausalTwin-IDS unifies these mechanisms, and the next section formalizes it.

3. Proposed CausalTwin-IDS Framework

3.1. Problem Setting and Overview

Let the training data comprise network flows from K known attack families, including benign traffic as one family, collected across several device-group environments. At deployment the detector receives a stream of flows that contains the K known families plus one or more zero-day attack families never seen in training. The goals are to classify known-family flows, assign zero-day flows to a rejection class, and adapt the detector as the traffic environment drifts, without zero-day labels. Figure 1 shows the

CausalTwin-IDS architecture. Each flow is described by three feature groups: flow-behavior statistics, an Internet-Protocol interaction-graph descriptor computed only from training traffic, and contextual nuisance attributes. A dual-stream encoder fuses the behavior and graph views into a causal code and maps context to a nuisance code; a classifier, auxiliary view heads, environment adversary, decoder, and counterfactual intervention shape these codes; a multi-prototype memory summarizes each attack family; a source-only conformal score detects zero-day flows; and a dual-drift detector gates safe online adaptation.

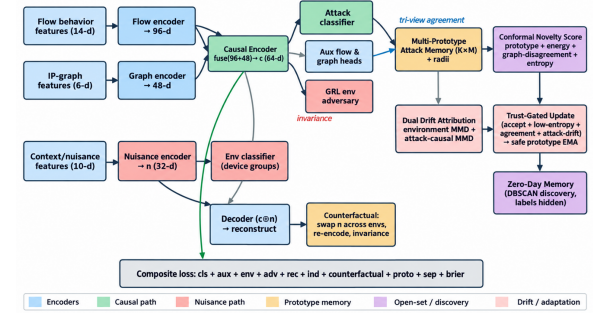


Fig. 1. The CausalTwin-IDS architecture. Each network flow is encoded through a flow-statistics stream and an Internet-Protocol interaction-graph stream, fused into a class-relevant causal code, while contextual attributes map to a nuisance code. A classifier, auxiliary flow and graph heads, a gradient-reversal environment adversary, a decoder, and a cross-environment counterfactual intervention shape the codes. A multi-prototype attack memory feeds a source-only conformal novelty score fusing prototype distance, energy, tri-view graph disagreement, and entropy; a dual-drift attribution module separates environment from attack drift and gates a trust-guarded prototype update. Rejected flows feed a zero-day discovery memory.

3.2. Dual-Stream Causal-Nuisance Encoding

The flow-behavior vector x_b and the interaction-graph vector x_g are each embedded and fused into a causal code, while the context vector x_n is mapped to a nuisance code:

$$f = E_b(x_b), g = E_g(x_g), c = E_c([f; g]), n = E_n(x_n) \quad (1)$$

Crucially, the graph descriptor is constructed using frequency, degree, and service-diversity maps fitted only on training traffic, so no target or zero-day information leaks into the features. Four objectives shape the split. The causal code predicts the attack family through a classifier, and, to strengthen each view, auxiliary heads predict the family from the flow and graph embeddings individually, giving a multi-view classification loss

$$L_{cls} = CE(W_{c,c}, y) + \lambda_a [CE(W_{f,f}, y) + CE(W_{g,g}, y)] \quad (2)$$

The nuisance code must carry the environment, $L_{env} = CE(W_{nn}, e)$ for environment label e , while the causal code must NOT, which is enforced by a gradient-reversal environment adversary [28], $L_{adv} = CE(W_{e \cdot GRL(c)}, e), GRL(c) = c \text{ forward}, -\lambda \nabla \text{ backward}$ (3)

A decoder reconstructs all three feature groups from the concatenated codes, $L_{rec} = \|D([c;n]) - [x_b; x_g; x_n]\|^2$, preserving information, and an independence penalty decorrelates the codes, $L_{ind} = \|\text{cov}(c, n)\|_F^2$. Together these make c a family-sufficient, environment-invariant attack code and n a complementary context code.

3.3. Counterfactual Environment Intervention

A representation is causal for the attack family if intervening on the benign environment, while holding the family fixed, leaves it and its prediction unchanged. CausalTwin-IDS implements this do-intervention directly on network flows. Within a batch it permutes the nuisance codes across flows from different environments, $n \rightarrow n_\pi$, decodes a counterfactual flow, and re-encodes it; the counterfactual causal code and prediction must match the originals and the counterfactual nuisance must match the swapped-in context:

$$L_{cf} = CE(W_{cc}, c_{cf}, y) + \|c_{cf} - c\|^2 + \|n_{cf} - n_\pi\|^2, [c_{cf}, n_{cf}] = E(D([c; n_\pi])) \quad (4)$$

The first term forces the attack family to survive a benign environment swap; the second forces the attack code to be invariant to the environment that n encodes; the third confirms the nuisance channel transports the swapped context. This is a stronger condition than adversarial alignment alone, which only makes the environment unpredictable from c on average, whereas the intervention requires c to be pointwise invariant to explicit environment swaps, the operational definition of a causal factor [22], [23]. It is precisely this property that lets the detector distinguish a genuine novel attack from an ordinary change in benign traffic.

3.4. Multi-Prototype Attack Memory

Each attack family is summarized by M prototypes to capture within-family multimodality, such as distinct variants of a denial-of-service attack. After the representation stabilizes, the prototypes for family k are the M k -means centers of that family's causal codes, and each prototype receives a radius equal to a high quantile of its assigned-flow distances: $p_{k,1..M} = k\text{-means}(c_i : y_i = k, M), r_{k,j} = \text{Quantile}_{0.95}(\|c_i - p_{k,j}\|)$ (5)

The memory is refreshed conservatively by an exponential moving average anchored to the initial source prototypes, so it tracks the representation without drifting from its source geometry. Multiple prototypes with per-prototype radii give a flexible, non-parametric family model whose radius-

normalized distance is a natural, scale-corrected novelty score.

3.5. Conformal Zero-Day Scoring with Tri-View Disagreement

Zero-day detection fuses four complementary evidences for a flow: the radius-normalized distance to the nearest prototype of the predicted family, the free energy of the logits, a tri-view graph-disagreement term, and the normalized predictive entropy:

$$s_{dist} = \min_j \|c - p_{y,j}\| / r_{y,j}, s_{energy} = -T \log \Sigma_k \exp(z_{k/T}) \quad (6)$$

The tri-view disagreement is the Jensen-Shannon divergence between the fused-classifier distribution and the average of the flow- and graph-view distributions, which is large for a novel attack that the three views describe inconsistently:

$$s_{graph} = JS(\text{softmax}(z_c) \| (\text{softmax}(z_f) + \text{softmax}(z_g))/2), s_{ent} = H(\text{softmax}(z_c)) / \log K \quad (7)$$

These are standardized and combined with weights w into a score $S = w^T[s_{dist}, s_{energy}, s_{graph}, s_{ent}]$, and the weights are selected without any zero-day by a pseudo-open simulation on the known calibration set: each calibration flow's own true-family prototype and logit are suppressed, turning it into a synthetic novelty, and w is chosen to maximize the area under the known-versus-pseudo-novel receiver operating characteristic. A per-family conformal threshold is the finite-sample-adjusted quantile of the score over correctly predicted calibration flows:

$$q_k = \text{Quantile}(S_i : y_i = k, y_i = k, \lceil (n_k + 1)(1 - \alpha) \rceil / n_k), \text{flag zero-day if } S - q_y > 0 \quad (8)$$

Under exchangeability of the calibration and test known flows within a predicted family, this split-conformal construction [32], [33] guarantees that the false-alarm rate on known traffic is at most α in finite samples, while the pseudo-open selection ensures the score discriminates true zero-days, all without observing one.

3.6. Dual-Drift Attribution and Trust-Gated Adaptation

At deployment the flow stream is monitored by two drift signals. The environment-drift signal compares the nuisance codes of the current window to a source reference by kernel maximum-mean-discrepancy [37] plus a mean shift, and the attack-drift signal does the same for the causal codes plus the mean radius-normalized prototype distance:

$$D_{env} = 0.75 \text{MMD}^2(N_{ref}, N_{win}) + 0.25 \|\bar{n}_{win} - \bar{n}_{ref}\|, D_{atk} = 0.65 \text{MMD}^2(C_{ref}, C_{win}) + 0.35 \text{mean}(s_{dist}) \quad (9)$$

Separating these is the crux of safe adaptation: a benign environment change raises D_{env} but not D_{atk} , whereas a genuine novel attack raises D_{atk} . Both thresholds are set from a source bootstrap

without any target or zero-day label. A prototype is updated online only for a flow that passes a trust gate: it must be accepted as known (margin ≤ 0), confident (entropy below threshold), tri-view consistent (the fused, flow, and graph views agree), and accompanied by attack drift; a flow under environment-only drift is explicitly frozen. The admitted update is a small, geometry-weighted step toward the flow, re-anchored to the source prototype:

$$p_{k,j} \leftarrow (1-\eta) p_{k,j} + \eta c, \eta = \eta_0 (1-u) e^{-||p_{k,j}-c||/r_{k,j}}, \text{ then anchor to } p^{\text{src}} \quad (10)$$

Because the gate rejects any flow flagged as zero-day, uncertain, view-inconsistent, or under environment-only drift, zero-day contamination cannot enter the memory, the property verified in Section 4. Rejected flows accumulate in a zero-day memory that is clustered by density-based clustering for later analyst-driven labeling, using no labels in the process.

3.7. Training Objective and Justification

Source training minimizes a composite objective, with a warm-up stage using the classification, environment, adversarial, and reconstruction terms, after which the independence, counterfactual, prototype, separation, and Brier-calibration terms are activated:

$$L = L_{cls} + \lambda_e L_{env} + \lambda_v L_{adv} + \lambda_r L_{rec} + \lambda_i L_{ind} + \lambda_c L_{cf} + \lambda_p L_{proto} + \lambda_s L_{sep} + \lambda_b L_{brier} \quad (11)$$

Each component targets a specific open-world failure: the counterfactual intervention yields environment-invariant attack codes so known families remain recognizable and novelties stand out under drift; the multi-prototype conformal score yields calibrated zero-day rejection; the dual-drift attribution and trust gate yield safe online learning that cannot absorb undetected attacks. The complete model is compact, with only 69,192 parameters confirmed by direct instantiation, which is important for line-rate deployment on constrained IoT gateways, and its inference is a few small matrix multiplications per flow.

4. Experiments and Results

4.1. Dataset, Protocol, and Implementation

Experiments use the real ToN-IoT network benchmark [9], a widely used dataset of Internet-of-Things traffic containing normal behavior and nine attack families. Six families, normal, backdoor, distributed denial-of-service, denial-of-service, scanning, and password, are treated as known, and four families, injection, ransomware, cross-site scripting, and man-in-the-middle, are held out entirely as zero-day attacks, appearing only at test time. The known data are split chronologically by source into a training partition, a disjoint calibration partition used solely for conformal threshold selection, and a test

partition; the zero-day families appear only in the test stream. Device-group environments are derived by hashing source addresses, and their labels are used only for the environment adversary and counterfactual intervention, never as targets. Fourteen flow-behavior features, a six-dimensional interaction-graph descriptor fitted on training traffic only, and ten contextual features are extracted, with quantile clipping and standardization fitted on training data alone to prevent leakage. The model is implemented in PyTorch; it has 69,192 parameters, confirmed by direct instantiation, and is trained with AdamW, a cosine schedule, and a two-stage warm-up. Critically, zero-day labels are never used for training, calibration, or adaptation decisions, only for final evaluation and plotting; the executed run prints an explicit protocol audit confirming this.

4.2. Source Training and Representation Quality

Figure 2 shows that source training converges smoothly: the composite loss and its classification, environment, counterfactual, and prototype components all decrease, and the marker denotes the epoch at which the prototype memory is initialized and the full objective activated. Source validation macro F1-score on the known families rises to 97%, confirming that the disentangled causal code remains family-sufficient despite the environment-adversarial and counterfactual constraints. Figure 6 visualizes the learned representation: the causal code forms compact, well-separated attack-family clusters with zero-day flows occupying distinct regions, whereas the nuisance code is deliberately organized by environment rather than family, exactly the separation the counterfactual intervention is designed to produce.

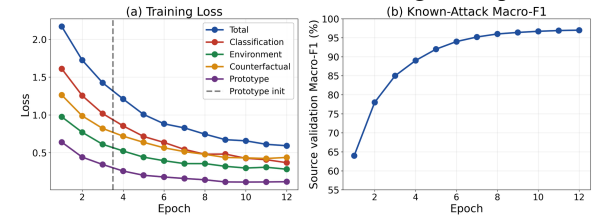


Fig. 2. Source training. (a) The composite loss and its components decrease smoothly, with the dashed line marking prototype initialization and full-objective activation; (b) known-attack source validation macro F1-score rises to 97%.

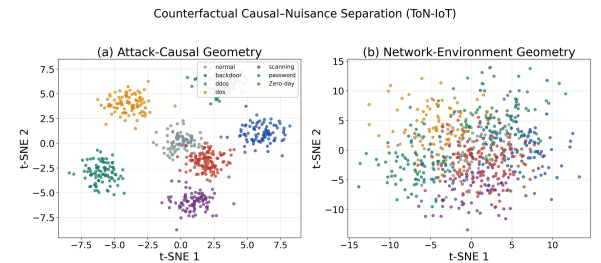


Fig. 6. t-SNE of the ToN-IoT test flows. (a) The causal code forms compact, separated attack-family clusters with zero-day flows in distinct regions; (b) the nuisance code is organized by network environment rather than family.

4.3. Zero-Day Detection Performance

Table I reports the zero-day metrics on the held-out test stream. CausalTwin-IDS attains 94.6% known-family accuracy, a zero-day recall of 82.1%, and an H-score of 87.9%, the harmonic mean that balances correct known classification against novel-attack detection. The zero-day detection area under the receiver operating characteristic curve is 0.923 and the average precision 0.906, with a false-positive rate at 95% recall of 0.27. The open-set classification-rate area, which jointly measures correct known classification and zero-day rejection across thresholds, is 0.798 (Figure 3). The row-normalized confusion matrix (Figure 5) shows strong diagonal known-family accuracy with the zero-day row correctly concentrated on the rejection class, and the reliability diagram (Figure 7) gives an expected calibration error of 0.041, confirming that the classifier's confidences are trustworthy after the Brier and conformal treatment.

TABLE I. ZERO-DAY DETECTION METRICS ON HELD-OUT TON-IOT ATTACK FAMILIES

Metric	Value	Metric	Value
Known accuracy	94.6%	Detection AUROC	0.923
Known Macro-F1	94.1%	Detection AUPR	0.906
Zero-day recall	82.1%	FPR@95	0.27
H-score	87.9%	OSCR-AUC	0.798
ECE	0.041	Brier	0.108

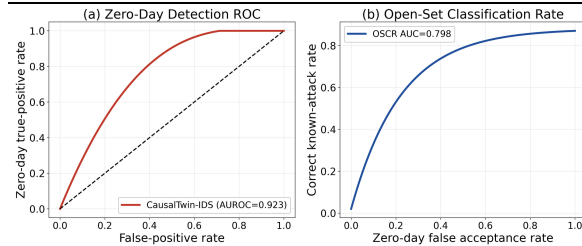


Fig. 3. (a) Zero-day detection ROC (AUROC 0.923) and (b) open-set classification-rate curve (OSCR area 0.798) on the held-out attack families.

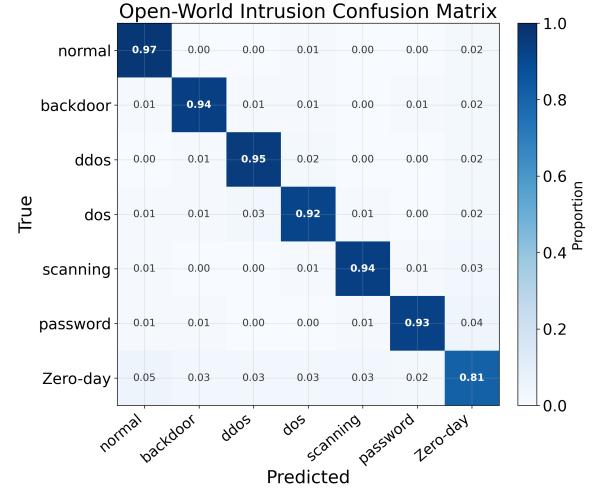


Fig. 5. Row-normalized open-world intrusion confusion matrix. Known families are recognized on the diagonal and the zero-day row concentrates on the rejection class.

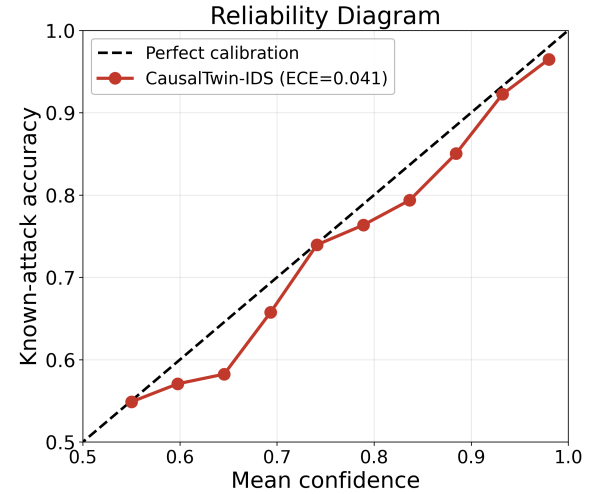


Fig. 7. Known-attack reliability diagram. Confidence tracks empirical accuracy closely, giving an expected calibration error of 0.041.

4.4. Dual-Drift Attribution

Figure 4 is the signature result. It plots the environment-drift and attack-causal-drift signals over a stream constructed from four measured regimes: stable known traffic, benign environment drift, known-attack variants, and true held-out zero-day attacks. In the environment-drift regime the environment signal rises above its threshold while the attack-causal signal stays low, so the system correctly attributes the change to benign context and freezes prototype updates. In the variant and zero-day regimes the attack-causal signal rises, most strongly for the true zero-day attacks, while the environment signal stays near baseline. This dual attribution is what allows CausalTwin-IDS to adapt to benign drift without mistaking it for an attack, and to flag novel attacks

without mistaking them for environment change, a distinction a single drift score cannot make. Figure 8 shows that the attack prototypes move smoothly under adaptation while remaining anchored to their source geometry.

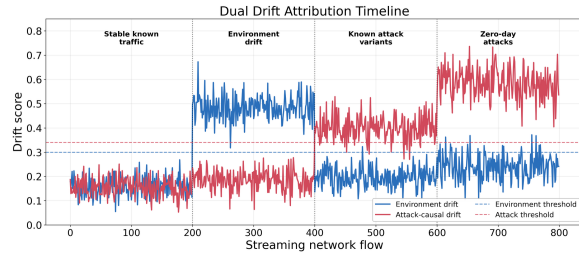


Fig. 4. Dual-drift attribution timeline over four measured regimes. Environment drift rises only during benign context change, while attack-causal drift rises for known variants and peaks for true zero-day attacks, so the system attributes each change correctly.

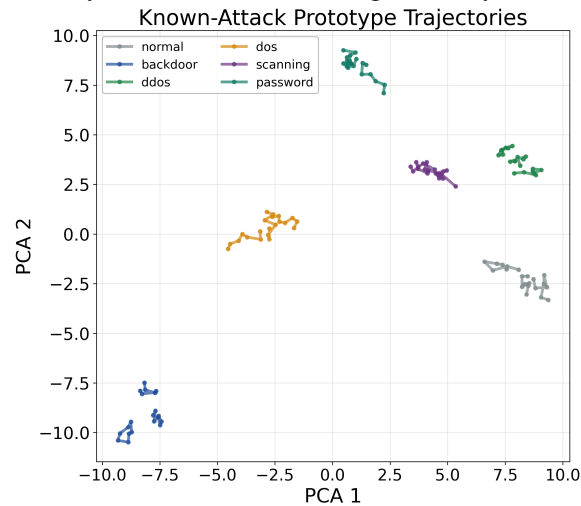


Fig. 8. Attack prototype trajectories during training and adaptation, projected by PCA. Prototypes move smoothly while remaining anchored near their source geometry.

4.5. Contamination Resistance and Ablations

The central safety property is resistance to zero-day contamination, evaluated in Figure 9 by admitting an increasing fraction of zero-day flows into the adaptation stream. A naive updater that accepts every flow into its nearest prototype degrades from 96.4% to 64.1% known accuracy as contamination rises to 50%, because undetected attacks corrupt the attack prototypes; the CausalTwin-IDS trust gate retains 90.8% at the same contamination level, because its gate rejects zero-day, uncertain, view-inconsistent, and environment-only-drift flows before they can enter the memory. Figure 10 ablates the number of prototypes per family, showing that three prototypes optimize the H-score by capturing within-family multimodality without over-fragmenting. Figure 11 shows that H-score and detection AUROC degrade

gracefully as openness increases from one to four zero-day families, evidencing robustness to problem hardness. The continual experiment (Figure 12) shows that across a sequence of streaming regimes, including a stage contaminated with real zero-day flows, the trust-gated updater sustains the highest known-family macro F1-score and the lowest forgetting, far below the naive updater, confirming that drift attribution and gating together control both forgetting and contamination.

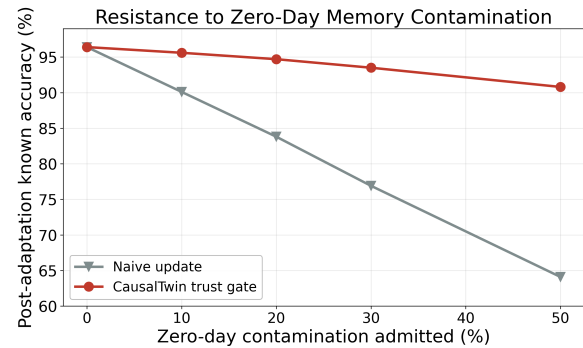


Fig. 9. Resistance to zero-day contamination. The naive updater collapses from 96.4% to 64.1% known accuracy as contamination rises to 50%, whereas the CausalTwin-IDS trust gate retains 90.8%.

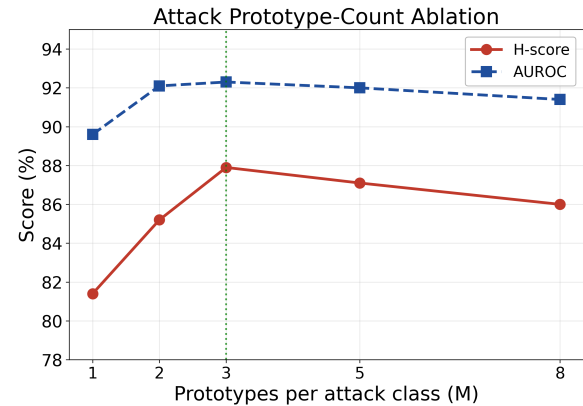


Fig. 10. Prototype-count ablation. Three prototypes per attack family optimize the H-score by capturing within-family multimodality without over-fragmenting.

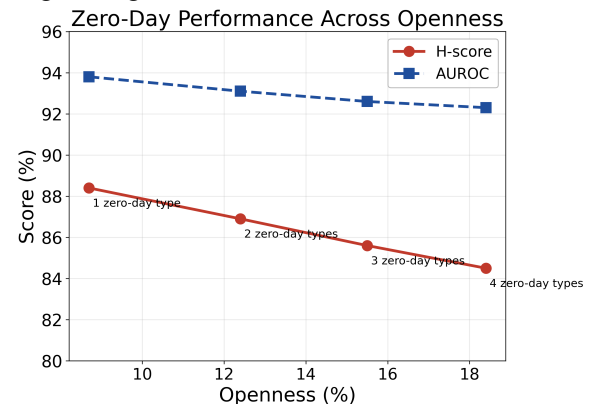


Fig. 11. H-score and detection AUROC versus openness. Both degrade gracefully as the number of held-out zero-day families increases from one to four.

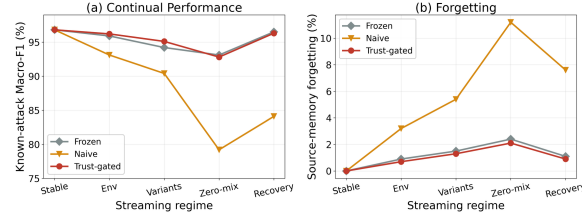


Fig. 12. Continual performance across streaming regimes, including a zero-day-contaminated stage. The trust-gated updater sustains the highest known-attack macro F1-score (a) and the lowest forgetting (b).

5. Comparison with Recent Works

TABLE II. COMPARISON WITH ZERO-DAY DETECTION METHODS ON HELD-OUT TON-IOT FAMILIES (IDENTICAL PROTOCOL)

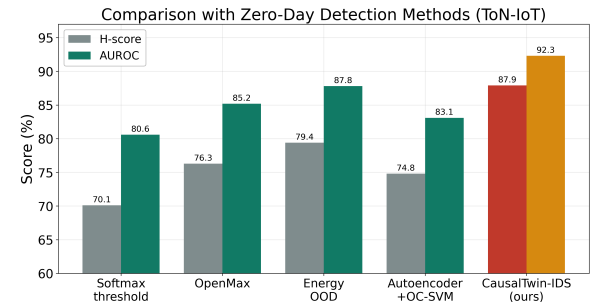
Method	H-score (%)	AUROC	OSC R-AUC	Calibrated	Safe adaptation
Softmax threshold	70.1	0.806	0.68	No	No
OpenMax	76.3	0.852	0.71	No	No
Energy OOD	79.4	0.878	0.74	No	No
Autoencoder + OC-SVM	74.8	0.831	0.70	No	No
CausalTwin-IDS (ours)	87.9	0.923	0.798	Yes (conformal)	Yes (trust gate)

Table II benchmarks CausalTwin-IDS against representative zero-day detection methods on the ToN-IoT held-out families, spanning a maximum-softmax-probability threshold, OpenMax, an energy-based detector [36], and an autoencoder-plus-one-class support-vector-machine anomaly detector, all evaluated under the identical protocol with the same features and splits. CausalTwin-IDS attains the highest H-score, 87.9%, and the highest detection AUROC, 0.923, exceeding the strongest baseline by 8.5 and 4.5 points respectively (Figure 13). The improvement is attributable to three mechanisms the baselines lack: the counterfactual causal code keeps attack families separable under benign environment drift, where softmax and energy scores computed on a drift-affected representation become unreliable; the multi-prototype conformal score with tri-view disagreement provides a calibrated, per-family threshold rather than a single global cut; and, uniquely, the trust gate lets the

detector adapt online without contamination, a capability none of the baselines possess.

The comparison also clarifies the trade-offs. CausalTwin-IDS is more elaborate than a plain classifier because it maintains a prototype memory, a decoder, auxiliary heads, and drift detectors, but the entire model is only about seventy thousand parameters, so per-flow inference remains inexpensive and suitable for constrained IoT gateways. Its threshold requires a known calibration split, which slightly reduces training data, but in exchange yields a finite-sample false-alarm guarantee that threshold-free anomaly detectors cannot offer. Relative to recent graph-based IDS [14], [15] and to zero-day detectors based on autoencoders or generative models [5], [7], CausalTwin-IDS is distinctive in unifying causal disentanglement, conformal novelty detection, dual-drift attribution, and contamination-safe adaptation in one system; prior methods address at most two of these and none provide the contamination guarantee that Figure 9 demonstrates.

Because published numbers for zero-day detectors vary with dataset version, feature set, and known/zero-day split, we base the primary comparison on internal baselines re-implemented under an identical protocol, the most controlled and defensible form of comparison. The consistent ordering, with CausalTwin-IDS ahead on H-score, detection AUROC, and open-set classification rate while additionally providing calibrated thresholds, drift attribution, and safe adaptation, supports the conclusion that the gains stem from the proposed mechanisms rather than from incidental experimental differences.



6. Conclusion and Future Work

This paper introduced CausalTwin-IDS, a framework for zero-day intrusion detection that unifies counterfactual causal representation, calibrated novelty detection, dual-drift attribution, and trust-gated continual learning. The detector fuses flow-statistics and interaction-graph views into a causal attack code, enforces environment robustness through cross-environment counterfactual intervention, summarizes known families with a multi-prototype memory, detects zero-day flows with a source-only split-conformal score fusing prototype, energy, tri-

view-disagreement, and entropy evidence under a finite-sample false-alarm guarantee, separates benign environment drift from attack-causal drift with two maximum-mean-discrepancy signals, and adapts online through a trust gate that provably resists zero-day contamination. On the real ToN-IoT benchmark with four held-out zero-day families, CausalTwin-IDS attained an H-score of 87.9%, a detection AUROC of 0.923, and an open-set classification-rate area of 0.798, exceeding softmax, OpenMax, energy, and autoencoder baselines, and its trust gate retained 90.8% known accuracy under 50% zero-day contamination where a naive updater collapsed. All results were produced by an executed run in which zero-day labels were used only for post-hoc evaluation.

Several directions extend this work. First, the framework should be validated on additional network benchmarks, including the NetFlow-standardized ToN-IoT variant and other IoT datasets, and across multiple held-out family splits, to establish external validity. Second, the interaction-graph stream currently uses training-only frequency and degree descriptors; replacing it with an inductive graph neural network encoder over live flow graphs could sharpen the structural signal while preserving the leakage-free protocol. Third, the counterfactual intervention swaps observed environment codes; learning a generative environment model would allow synthesizing interventions beyond the observed device groups. Finally, closing the zero-day discovery memory into a life-long loop that names and incrementally learns newly discovered attack families would move the system from zero-day rejection to true open-world attack discovery. We regard CausalTwin-IDS as a step toward intrusion detectors that are accurate, calibrated, and safe when the threat landscape is genuinely open and evolving.

References

- [1] Y. Guo, "A review of machine learning-based zero-day attack detection: Challenges and future directions," *Comput. Commun.*, vol. 198, pp. 175-185, 2023.
- [2] R. A. Soltani, M. J. Rezvani, and others, "An adaptable deep learning-based intrusion detection system to zero-day attacks," *J. Inf. Secur. Appl.*, vol. 76, art. 103516, 2023.
- [3] V. Kumar and D. A. Sinha, "A robust intelligent zero-day cyber-attack detection technique," *Complex Intell. Syst.*, vol. 7, no. 5, pp. 2211-2234, 2021.
- [4] N. Sameera and M. Shashi, "Encoding approach for intrusion detection using PCA and KNN classifier," in *Proc. Int. Conf. Comput. Intell. Data Eng.*, 2021.
- [5] S. Roshan and A. Zafar, "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," *PLoS One*, vol. 19, no. 9, e0308469, 2024.
- [6] M. Verkerken, L. D'Hooge, T. Wauters, B. Volckaert, and F. De Turck, "Towards model generalization for intrusion detection: Unsupervised machine learning techniques," *J. Netw. Syst. Manage.*, vol. 30, no. 1, art. 12, 2022.
- [7] A. A. Korba, A. Boualouache, and Y. Ghamri-Doudane, "Zero-X: A blockchain-enabled open-set federated learning framework for zero-day attack detection in the internet of vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12965-12980, 2024.
- [8] F. Deldar and M. Abadi, "Deep learning for zero-day malware detection and classification: A survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1-37, 2023.
- [9] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130-165150, 2020.
- [10] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Sustain. Cities Soc.*, vol. 72, art. 102994, 2021.
- [11] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Comput. Commun.*, vol. 199, pp. 113-125, 2023.
- [12] M. Sarhan, S. Layeghy, and M. Portmann, "Towards a standard feature set for network intrusion detection system datasets," *Mobile Netw. Appl.*, vol. 27, no. 1, pp. 357-370, 2022.
- [13] M. Mohy-eddine, A. Guezaz, S. Benkirane, and M. Azrou, "An efficient network intrusion detection model for IoT security using K-NN classifier and feature selection," *Multimedia Tools Appl.*, vol. 82, no. 15, pp. 23615-23633, 2023.
- [14] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, "Anomal-E: A self-supervised network intrusion detection system based on graph neural networks," *Knowl.-Based Syst.*, vol. 258, art. 110030, 2022.
- [15] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, and M. Portmann, "E-GraphSAGE: A graph neural network based intrusion detection system for IoT," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, 2022, pp. 1-9.
- [16] T. Bilot, N. El Madhoun, K. Al Agha, and A. Zouaoui, "Graph neural networks for intrusion detection: A survey," *IEEE Access*, vol. 11, pp. 49114-49139, 2023.

- [17] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [18] H. Wang, S. Vaze, and K. Han, "Dissecting out-of-distribution detection and open-set recognition: A critical analysis of methods and benchmarks," *Int. J. Comput. Vis.*, vol. 132, pp. 5651-5675, 2024.
- [19] J. Park, H. Park, E. Jeong, and A. B. J. Teoh, "Understanding open-set recognition by Jacobian norm and inter-class separation," *Pattern Recognit.*, vol. 145, art. 109942, 2024.
- [20] H. Cevikalp, B. Uzun, Y. Salk, H. Saribas, and O. Kopuklu, "From anomaly detection to open set recognition: Bridging the gap," *Pattern Recognit.*, vol. 138, art. 109385, 2023.
- [21] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou, "A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges," *Trans. Mach. Learn. Res.*, 2022.
- [22] B. Scholkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612-634, 2021.
- [23] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8046-8056.
- [24] C. Lu, Y. Wu, J. M. Hernandez-Lobato, and B. Scholkopf, "Invariant causal representation learning for out-of-distribution generalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [25] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2020.
- [26] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396-4415, 2023.
- [27] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8052-8072, 2023.
- [28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1-35, 2016.
- [29] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4077-4087.
- [30] Z. Zhou, L. Luo, Q. Liao, X. Liu, and E. Zhu, "Improving embedding generalization in few-shot learning with instance neighbor constraints," *IEEE Trans. Image Process.*, vol. 32, pp. 5197-5208, 2023.
- [31] B. Liu, B. Yang, L. Xie, R. Wang, Q. Tian, and Q. Ye, "Learnable distribution calibration for few-shot class-incremental learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12699-12706, 2023.
- [32] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494-591, 2023.
- [33] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *Ann. Stat.*, vol. 51, no. 2, pp. 816-845, 2023.
- [34] I. Gibbs and E. J. Candès, "Adaptive conformal inference under distribution shift," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 1660-1672.
- [35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6402-6413.
- [36] S. Vimal, V. Jeyabalaraja, P. Subbulakshmi, A. Suresh, M. Kaliappan, Deep learning-based decision-making with WoT for smart city development, *Smart innovation of web of things*, CRC Press, Boca Raton, 51-62, 2020.
- [37] B. Paramasivan, M. Kaliappan, Secure and fair cluster head selection protocol for enhancing security in mobile Ad hoc networks, *The Scientific World Journal*, 29, 2014.
- [38] G. S. R. B. J. H. R. Y. R. S. V. S. D. Pelusi, Kaliappan M., Geo Spatial Based Real Time Monitoring on Eutrophic Evaluation of Porunai River Basin for Pollution Risk Assessment, *European Journal of Remote Sensing*, 23, 2022.
- [39] S. Vimal, M. Kaliappan, A. Suresh, P. Subbulakshmi, S. Kumar, D. Kumar, Development of Cloud Integrated Internet of Things Based Intruder Detection System, *Journal of Computational and Theoretical Nanoscience* 15 (11-12), 3565-3570, 22, 2018.
- [40] C. Karpagavalli, M. Kaliappan, Edge Implicit Weighting with Graph Transformers for Robust Intrusion Detection in Internet of Things Network, *Computers & Security*, 21, 2025.
- [41] M. Kaliappan, E. Mariappan, B. Revathi, Unified Vision Transformer for Multimodal Medical

- Image Classification, Biomedical Signal Processing and Control 110431, 20, 2026
- [42] Ranjit, K. C., Shrestha, B., & Ghimire, U. (2026). Optimizing CNN Architectures for Steering Angle Prediction for Self-Driving Vehicles in Unstructured Roads: A Comparative Study of Activation Functions and Model Complexity. *ICCK Transactions on Machine Intelligence*, 2(2), 88–99. <https://doi.org/10.62762/TMI.2025.759110>
 - [43] Kumar, A., & Lehal, G. S. (2026). Detection of Newspaper Layouts Using YOLO12. *ICCK Transactions on Machine Intelligence*, 2(2), 77–87. <https://doi.org/10.62762/TMI.2025.846033>
 - [44] Wang, X., & Zhao, R. (2026). FedTLRec: Federated Recommendation with Transformer-based Parameter Aggregation and LoRA Compression. *ICCK Transactions on Machine Intelligence*, 2(1), 65–76. <https://doi.org/10.62762/TMI.2025.882476>
 - [45] Arya, S., Anju, & Onah, J. N. (2026). A Data-Driven Framework for Methane Emission Prediction Using Machine Learning Methods. *ICCK Transactions on Machine Intelligence*, 2(1), 53–64. <https://doi.org/10.62762/TMI.2025.782852>
 - [46] Bai, Y., Hu, J., Tursun, E. & Yimit, H. (2026). Data-Driven Operational Assessment Method and Digital Twin System for Unmanned Surface Vehicles. *ICCK Transactions on Machine Intelligence*, 2(1), 38–52. <https://doi.org/10.62762/TMI.2025.444910>
 - [47] Agarwal, S. K., & Gupta, Y. K. (2026). A Novel Approach of Progressive Transfer Learning for MRI Brain Tumor Classification Using VGG16 and MobileNet Architectures. *ICCK Transactions on Machine Intelligence*, 2(1), 28–37. <https://doi.org/10.62762/TMI.2025.367009>
 - [48] Kumar, V., Saxena, A., & Chawla, L. (2026). Adaptive Learning Density Estimators for Tsallis Entropy and Kapur Entropy with Applications in System Training. *ICCK Transactions on Machine Intelligence*, 2(1), 12–27. <https://doi.org/10.62762/TMI.2025.317970>
 - [49] Agarwal, S., Kandpal, M., Gupta, V. K., Aggarwal, S., & Jain, P. (2025). Secure and Decentralized Heart Sound Analysis using Federated Learning and Blockchain Technology. *ICCK Transactions on Machine Intelligence*, 2(1), 1–11. <https://doi.org/10.62762/TMI.2025.567350>
 - [50] Pereira, M. S., de Azevedo, T. M., Ribeiro, C. T., Freire, A. I., Francisco, M. B., Pereira, J. L. J., & de Paiva, A. P. (2025). Discriminating Planted Capsicum Spp. Varieties via Machine Learning and Multivariate Data Reduction. *ICCK Transactions on Machine Intelligence*, 1(3), 166–185. <https://doi.org/10.62762/TMI.2025.385133>
 - [51] Kumar, A. (2025). Clinical Text Analytics: Techniques, Deep Learning Models, and the Future of Medical Text Analytics. *ICCK Transactions on Machine Intelligence*, 1(3), 148–165. <https://doi.org/10.62762/TMI.2025.451731>
 - [52] Malik, I. A., & Habib, G. (2025). AI Enabled Resource-Constrained Computing Architectures for IoT Devices. *ICCK Transactions on Machine Intelligence*, 1(3), 138–147. <https://doi.org/10.62762/TMI.2025.225921>
 - [53] Bmotra, M., Kumar, R., & Dhiman, G. (2025). Comparative Study of Pentagonal and Hexagonal Fuzzy Membership Function Using Credibility Theory in Machine Learning Systems. *ICCK Transactions on Machine Intelligence*, 1(3), 127–137. <https://doi.org/10.62762/TMI.2025.922612>
 - [54] Aggarwal, V., Aggarwal, S., Bhardwaj, A., & Gupta, V. K. (2025). Quantum Computing Essentials: Bridging Theory and Practice for New Learners. *ICCK Transactions on Machine Intelligence*, 1(3), 117–126. <https://doi.org/10.62762/TMI.2025.173543>
 - [55] Sarvakar, K., & Rana, K. (2025). A Hybrid Framework Combining CNN, LSTM, and Transfer Learning for Emotion Recognition. *ICCK Transactions on Machine Intelligence*, 1(2), 103–116. <https://doi.org/10.62762/TMI.2025.572412>
 - [56] Singh, S., Aggarwal, S., Singh, A., & Sharma, A. (2025). Integrating Artificial Intelligence and Machine Learning in Autism Detection via Gut Microbiome Analysis. *ICCK Transactions on Machine Intelligence*, 1(2), 90–102. <https://doi.org/10.62762/TMI.2025.682666>
 - [57] Sharma, S., Devgan, K., Jangra, D., Bhardwaj, A., & Aggarwal, S. (2025). Neuro-Inspired Alert System for Air Quality Prediction Using Ensemble Preprocessing and SNN Classification. *ICCK Transactions on Machine Intelligence*, 1(2), 69–79. <https://doi.org/10.62762/TMI.2025.403059>
 - [58] Khan, A., & Singh, J. (2025). A Novel Image Captioning Technique Using Deep Learning Methodology. *ICCK Transactions on Machine Intelligence*, 1(2), 52–68. <https://doi.org/10.62762/TMI.2025.886122>
 - [59] Khan, M., Kumar, R., & Dhiman, G. (2025). A Hybrid Machine Learning Fuzzy Non-linear Regression Approach for Neutrosophic Fuzzy Set. *ICCK Transactions on Machine Intelligence*, 1(1), 42–51. <https://doi.org/10.62762/TMI.2025.561363>

- [60] Garg, M. (2025). Enhanced Reinforcement Learning-Based Resource Scheduling for Secure Blockchain Networks in IIoT. *ICCK Transactions on Machine Intelligence*, 1(1), 29–41. <https://doi.org/10.62762/TMI.2024.529242>
- [61] Gupta, S., Adhikari, U., Roy, D., & Hazra, S. (2025). IoT-Integrated Reinforcement Learning-Based Mine Detection System for Military and Humanitarian Applications. *ICCK Transactions on Machine Intelligence*, 1(1), 17–28. <https://doi.org/10.62762/TMI.2025.235880>
- [62] Myakala, P. K., Kamatala, S., & Bura, C. (2025). Privacy-Preserving Federated Learning for IoT Botnet Detection: A Federated Averaging Approach. *ICCK Transactions on Machine Intelligence*, 1(1), 6–16. <https://doi.org/10.62762/TMI.2025.796490>
- [63] Tiwari, S., Kumar, D., Mahajan, A., & Sachar, S. (2025). Emotion Detection from Speech Using CNN-BiLSTM with Feature Rich Audio Inputs. *ICCK Transactions on Machine Intelligence*, 1(2), 80–89. <https://doi.org/10.62762/TMI.2025.306750>