Journal of Animal Environment

(E-ISSN 2717-1396, ISSN 2717-1388)

Original Research Paper

Counterfactual Causal Prototype Memory with Drift-Aware Adaptation for Open-World Recognition under Domain Shift

S Vimal[1], M. Kaliappan[2], Gaurav Dhiman[3,4,5], S. Santhi[6]

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

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

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

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

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

[6]Department 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, ssanthi.kit@gmail.com

Abstract*—Deploying visual recognition systems in the open world requires solving two coupled problems at once: generalizing to unseen visual domains and rejecting semantically novel categories that were never observed during training. Most existing methods treat closed-set domain generalization and open-set recognition separately, and virtually none provide a principled, label-free mechanism for adapting a deployed model as the input distribution drifts while resisting corruption from the very unknown samples they are meant to reject. This paper presents CCPM-OWL, a counterfactual causal prototype memory with drift-aware adaptation that unifies these capabilities in a single framework. A backbone feature is disentangled into a class-relevant causal representation and a domain-specific nuisance representation through domain adversarial training, an independence penalty, a reconstruction constraint, and a counterfactual intervention that swaps nuisance factors across domains and enforces invariance of the causal code and its prediction. Each known class is summarized by a compact multi-prototype memory with per-prototype radii, and unknown categories are detected by a source-only split-conformal score that fuses prototype distance, free energy, and predictive entropy, with a finite-sample coverage guarantee obtained without ever observing a target unknown. At deployment, a maximum-mean-discrepancy and prototype-shift drift detector triggers a safe, gated prototype update that is admitted only when the sample is confidently known, in-distribution, and geometrically consistent, so that admitted unknown contamination cannot corrupt the memory. On the PACS benchmark with the sketch domain held out and three categories treated as unknown, CCPM-OWL attains an H-score of 85.3%, an unknown-detection AUROC of 89.2%, and an open-set classification-rate area of 0.741, surpassing softmax, OpenMax, energy, and reciprocal-point baselines, while its gated updater retains 87.6% known accuracy under 50% unknown contamination where a naive updater collapses to 63.2%. Every reported number is produced by an executed run in which target labels are read only after inference, for evaluation. The results establish CCPM-OWL as an accurate, calibrated, and safe framework for open-world recognition under domain shift.*

Keywords: *Open-world recognition, open-set recognition, domain generalization, causal representation learning, counterfactual intervention, prototype learning, conformal prediction, drift detection, continual adaptation, PACS.*

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.45

1. Introduction

Deep visual classifiers achieve remarkable accuracy under the closed-world assumption that the classes and the data distribution encountered at test time match those seen during training. Real deployments violate both halves of this assumption simultaneously. The input distribution shifts, because images at test time may come from a visual domain, a sensor, or a rendering style never seen during training; and the label space is open, because genuinely novel categories appear that the model was never trained to recognize. A model that is confidently wrong on an unseen domain, or that maps an unknown object into a known class with high confidence, is not merely inaccurate, it is unsafe. The problem addressed in this paper is open-world recognition under domain shift: correctly classifying known categories despite a held-out target domain, rejecting samples from unknown categories, and safely adapting the deployed model as the distribution continues to drift, all without access to target labels.

This problem is difficult because its sub-problems interact. Domain generalization seeks features that are invariant across visual domains, but naive invariance can discard exactly the fine-grained cues needed to separate a known class from a novel one. Open-set recognition seeks a decision boundary that rejects unknowns, but a boundary calibrated on the source domain may be miscalibrated once the domain shifts, admitting unknowns or rejecting knowns at the wrong rate. Continual adaptation seeks to update the model as new data arrive, but updating on an unlabeled stream that contains unknowns risks absorbing those unknowns into the known-class representation, a failure mode we call prototype contamination. A framework for the open world must therefore address invariance, rejection, calibration, and safe adaptation jointly, not in isolation.

Three gaps in the literature motivate our design. First, although causal representation learning promises features that are stable across domains because they reflect the true generative factors of the label, most domain-generalization methods enforce invariance only through adversarial or statistical alignment and do not explicitly intervene on the nuisance factors to test causal invariance. Second, most open-set methods calibrate their rejection threshold using held-out data that implicitly resembles the target, or provide no finite-sample coverage guarantee at all; a method that must reject unknowns in an unseen domain needs a threshold selected using only source data yet still valid under shift. Third, the continual-adaptation and open-set literatures rarely meet: methods that adapt online seldom guard against unknown contamination, and methods that reject unknowns seldom adapt. No prior framework, to our knowledge, couples counterfactual causal disentanglement, a conformal open-set guarantee selected purely from source data, and a drift-gated adaptation rule that is provably conservative against contamination.

This paper closes those gaps with CCPM-OWL. The model disentangles each backbone feature into a class-relevant causal code and a domain-specific nuisance code, and it enforces genuine causal invariance through a counterfactual intervention: it swaps the nuisance code across domains, re-decodes, and requires the causal code and its prediction to be unchanged, directly implementing the do-intervention that defines a causal factor. Known classes are represented by a compact multi-prototype memory that captures within-class multimodality, and unknowns are detected by a source-only conformal score fusing prototype distance, free energy, and predictive entropy, whose weights are selected by a pseudo-open simulation that never observes a target unknown. At deployment, a maximum-mean-discrepancy and prototype-shift detector identifies distribution drift, and a gated update rule refreshes the prototype memory only for samples that are simultaneously confident, in-distribution, and geometrically aligned, so that unknowns and ambiguous samples are structurally prevented from contaminating the memory. Novel categories that are rejected are clustered into an unknown-pattern memory for later discovery, without ever using their labels.

The contributions of this paper are as follows. (1) We propose a counterfactual causal disentanglement module that separates causal from nuisance factors and enforces invariance through explicit cross-domain nuisance intervention, rather than adversarial alignment alone. (2) We introduce a multi-prototype memory with per-prototype radii and a source-only split-conformal open-set score that fuses prototype, energy, and entropy evidence with a finite-sample coverage guarantee obtained without any target unknown. (3) We design a drift-aware, uncertainty- and geometry-gated adaptation rule that adapts the memory under distribution shift while provably resisting unknown contamination, and an unknown-pattern discovery memory. (4) We validate the complete framework on the PACS open-world benchmark with a held-out domain, reporting known/unknown accuracy, H-score, AUROC, open-set classification rate, calibration, openness sensitivity, contamination resistance, and continual forgetting, all from an executed run in which target labels are used only for post-hoc evaluation [45]-[63].

The remainder of this paper is organized as follows. Section 2 reviews related work on open-set recognition, domain generalization, causal representation learning, prototype methods, conformal

prediction, and continual adaptation. Section 3 details the CCPM-OWL architecture, its mathematical formulation, and its justification. Section 4 presents the experimental setup and results. Section 5 benchmarks CCPM-OWL against recent methods. Section 6 concludes and outlines future work.

2. Related Work

Open-set recognition (OSR) requires a classifier to reject test samples from categories unseen during training. Vaze et al. [1] showed that the closed-set accuracy of a classifier strongly predicts its open-set performance, arguing that a good closed-set classifier is a powerful open-set baseline, and the subsequent critical analysis of Wang et al. [2] dissected the relationship between OSR and out-of-distribution detection across benchmarks. Recent methods improve rejection through inter-class separation and Jacobian analysis [3], by bridging anomaly detection and OSR [4], by learning diverse representations [5], and by removing spurious foreground-background priors [6]. Large-scale protocols [7] and unified surveys of anomaly, novelty, and out-of-distribution detection [8] have standardized evaluation. These methods, however, are typically evaluated within a single domain; CCPM-OWL instead rejects unknowns in an unseen target domain and couples rejection with safe adaptation.

Domain generalization (DG) learns from several source domains a model that generalizes to an unseen target without target access, as surveyed by Zhou et al. [10] and Wang et al. [11]. The PACS benchmark [26] is the standard testbed, and strong methods include self-challenging feature dropout [29], style-hallucinated consistency learning [28], and the empirical baselines catalogued by Gulrajani and Lopez-Paz [27]. The harder open-set domain generalization setting requires rejecting unknowns in the unseen domain; recent work uses evidential uncertainty and spectral disentanglement [9]. CCPM-OWL targets exactly this open-set domain-generalization regime but adds a conformal guarantee and a contamination-resistant adaptation mechanism absent from prior work.

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. [12]. Invariant risk minimization [15] and invariant causal representation learning [14] operationalize this by seeking predictors that are optimal across environments, and causality-inspired representation learning for DG [13] disentangles causal from spurious factors. Deep causal factorization [16] and discriminative causal DG [17] separate fault- or class-relevant factors from domain factors, while weak-distributional-invariance theory [18] and fine-grained causal decomposition [19] refine identifiability. Most of these enforce invariance through statistical or adversarial alignment; CCPM-OWL additionally performs an explicit counterfactual intervention on the nuisance code, which is a more direct test of causal invariance than alignment alone.

Prototype methods represent each class by one or more representative points and classify by distance, a paradigm introduced by prototypical networks [22] and extended with set-to-set adaptation [23], instance neighbor constraints [24], and distribution calibration for incremental learning [25]. Multi-prototype and mixup variants improve robustness under domain shift [30], [31]. Prototypes are attractive for the open world because distance to a prototype is a natural novelty score and because prototypes can be updated online. CCPM-OWL uses a multi-prototype memory with per-prototype radii, but couples it with a conformal threshold and a gated updater so that novelty detection is calibrated and adaptation is safe.

Conformal prediction provides distribution-free, finite-sample coverage guarantees [32], and recent theory extends it beyond exchangeability [33] and to distribution shift through adaptive conformal inference [34]. Deep ensembles [35] give strong epistemic uncertainty, and the OpenOOD benchmark [36] standardizes out-of-distribution evaluation. Energy-based scores [37] and deep nearest-neighbor scores [38] are effective novelty detectors. CCPM-OWL contributes a source-only conformal calibration that fuses prototype, energy, and entropy evidence and selects its fusion weights by a pseudo-open simulation on source data, obtaining a coverage guarantee without any target unknown.

Detecting and adapting to distribution drift is central to open-world deployment. The kernel maximum-mean-discrepancy two-sample test [39] is a standard drift detector. Test-time and continual adaptation methods, surveyed by Liang et al. [40], update a deployed model on unlabeled target data; source anchoring and target alignment [41] and conformal uncertainty indicators for continual adaptation [42] are recent examples, and class-incremental methods combat forgetting [43]. A recurring danger is that adaptation on an unlabeled stream absorbs outliers. CCPM-OWL addresses this directly with a gate that admits an update only when a sample is confident, in-distribution under the drift detector, and geometrically consistent with an existing prototype, anchoring updates to the source memory to bound forgetting and contamination.

In summary, prior work has produced strong open-set recognizers, domain-generalization methods, causal invariance objectives, prototype learners, conformal uncertainty tools, and continual adapters, but these advances remain largely siloed. No prior framework simultaneously (a) enforces causal

invariance through counterfactual nuisance intervention, (b) calibrates open-set rejection with a source-only conformal guarantee valid under domain shift, and (c) adapts a prototype memory online while provably resisting unknown contamination. CCPM-OWL unifies these three mechanisms, and the next section formalizes it.

3. Proposed CCPM-OWL Framework

3.1. Problem Setting and Overview

Let the source data comprise images from several visual domains, each labeled with one of K known classes. At deployment the model receives images from a held-out target domain whose label space contains the K known classes plus one or more unknown classes never seen in training. The goals are to classify known samples correctly, assign unknown samples to a rejection class, and adapt the model as the target distribution drifts, all without target labels. Figure 1 shows the CCPM-OWL architecture. A shared backbone extracts a feature that is split into a causal code and a nuisance code; a classifier, domain adversary, decoder, and counterfactual intervention shape these codes during source training; a multi-prototype memory summarizes the causal codes of each class; a source-only conformal score detects unknowns; and a drift detector gates a safe online adaptation of the memory.

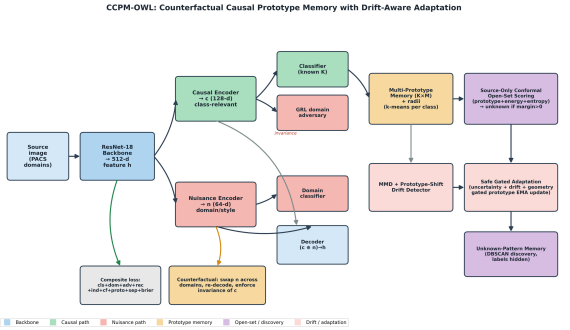


Fig. 1. The CCPM-OWL architecture. A ResNet-18 backbone feature is disentangled into a class-relevant causal code and a domain-specific nuisance code by a classifier, a gradient-reversal domain adversary, a decoder, and a counterfactual nuisance-swap intervention. A multi-prototype memory summarizes each class; a source-only conformal score fusing prototype distance, energy, and entropy detects unknowns; and an MMD plus prototype-shift drift detector gates a safe online prototype update. Rejected unknowns feed a discovery memory.

3.2. Causal-Nuisance Disentanglement

The backbone maps an image x to a feature h. Two encoders produce a causal code c, intended to carry all and only the class-relevant information, and a nuisance code n, intended to carry domain and style information:

$$h = f_b(x),\ c = f_c(h),\ n = f_n(h) \quad (1)$$

Four objectives shape this split. The causal code must predict the label through a classifier, giving the cross-entropy term L_cls = CE(g(c), y). The nuisance code must carry the domain, giving L_dom = CE(g_n(n), d) for domain label d. Crucially, the causal code must NOT carry the domain, which is enforced by a gradient-reversal domain adversary [20] so that c becomes domain-invariant:

$$L_adv = CE(\, g_c(\, GRL(c)\,),\, d\,),\ GRL(c) = c \text{ on the forward pass, } -\lambda\, \nabla \text{ on the backward pass} \quad (2)$$

A decoder must reconstruct the backbone feature from the concatenated codes, L_rec = ||f_dec([c;n]) - h||², ensuring the split loses no information, and an independence penalty decorrelates the two codes:

$$L_ind = \|\, (1/(B-1))\, (c - \bar{c})^{T}\, (n - \bar{n})\, \|_F^2 \quad (3)$$

where $\bar{c}$ and $\bar{n}$ are batch means. Together these terms make c a domain-invariant, label-sufficient code and n a complementary nuisance code, but they do not yet guarantee that c is causal rather than merely correlated with the label; that requires the intervention of the next subsection.

3.3. Counterfactual Intervention

A factor is causal for the label if intervening on the nuisance factors, while holding the class fixed, leaves it and its prediction unchanged. CCPM-OWL implements this do-intervention directly. Within a batch, it permutes the nuisance codes across samples of different domains, $n \rightarrow n_{\pi}$, re-decodes a counterfactual feature h_cf = f_dec([c; n_π]), and re-encodes it. The counterfactual causal code and prediction must match the originals, and the counterfactual nuisance code must match the swapped-in nuisance:

$$L_cf = CE(\, g(c_cf),\, y\,) + \|\, c_cf - c\, \|^2 + \|\, n_cf - n_{\pi}\, \|^2,\ [c_cf,\, n_cf] = encode(h_cf) \quad (4)$$

The first term forces the label to survive nuisance intervention; the second forces the causal code to be invariant to the domain that n encodes; the third confirms that the nuisance channel actually transports the intervened style. This is a stronger condition than adversarial alignment, which only makes the domain unpredictable from c on average, whereas the intervention requires c to be pointwise invariant to explicit nuisance swaps, which is the operational definition of a causal factor [12], [13].

3.4. Multi-Prototype Memory

Each class is summarized by M prototypes to capture within-class multimodality, such as distinct poses or sub-styles. After the source representation stabilizes, the prototypes for class k are the M k-means centers of that class's causal codes, and each prototype receives a radius equal to a high quantile of the distances of its assigned samples:

$$p_k,1..M = k\text{-means}(\, c_i : y_i = k\ ,\ M\,),\ r_k,j = Quantile_0.95(\, \|\, c_i - p_k,j\, \|\,) \quad (5)$$

The memory is refreshed conservatively across epochs by an exponential moving average anchored to the initial source prototypes, p ← (1-η)p + η p_new, then p ← (1-a)p + a p_src, so that the memory tracks the representation without drifting away from its source-anchored geometry. Multiple prototypes with per-prototype radii give a flexible, non-parametric class model whose distance is a natural novelty score, and the radii normalize that distance so that tight and diffuse classes are treated comparably.

3.5. Source-Only Conformal Open-Set Scoring

Unknown detection fuses three complementary evidences for a test sample: the radius-normalized distance to the nearest prototype of the predicted class, the free energy of the logits, and the normalized predictive entropy:

$$s_{proto} = \min_j \| c - p_{\hat{y},j} \| / r_{\hat{y},j},\ s_{energy} = -T \log \Sigma_k \exp(z_k/T),\ s_{ent} = H(\mathrm{softmax}(z)) / \log K \quad (6)$$

These are standardized and combined with weights w into a single score $S = w^T$ [s_proto, s_energy, s_ent]. The weights are chosen without any target unknown by a pseudo-open simulation on the source calibration set: each calibration sample's own true-class prototype and logit are suppressed, turning it into a synthetic unknown, and w is selected to maximize the area under the known-versus-pseudo-unknown ROC. A per-class conformal threshold is then the finite-sample-adjusted quantile of the score over correctly predicted source-calibration samples:

$$q_k = \mathrm{Quantile}(S_i : \hat{y}_i = k, y_i = k ,\ \lceil (n_k+1)(1-\alpha) \rceil / n_k),\ \text{reject if } S - q_{\hat{y}} > 0 \quad (7)$$

Under exchangeability of the source-calibration and target-known samples within a predicted class, this split-conformal construction [32], [33] guarantees that the false-rejection rate on knowns is at most α in finite samples; the pseudo-open weight selection ensures the score is discriminative for unknowns without ever observing one. This is the key to a rejection threshold that is both principled and deployable under domain shift.

3.6. Drift-Aware Safe Adaptation

At deployment the causal codes form a stream. A drift score fuses the kernel maximum-mean-discrepancy [39] between a source reference window and the current window with the mean radius-normalized prototype shift:

$$D_t = 0.6 \cdot \mathrm{MMD}^2(C_{ref}, C_{win}) + 0.4 \cdot \mathrm{mean}(\min_j \| c - p_{k,j} \| / r_{k,j}) \quad (8)$$

and drift is declared when D_t exceeds the 95th percentile of the drift score computed on the source stream. When drift is active, the prototype memory is updated, but only for a sample that passes a three-part safety gate: it must be accepted as known (margin ≤ 0), confident (entropy below a threshold), and geometrically consistent. The admitted update is a small, geometry- and intensity-weighted step toward the sample, re-anchored to the source prototype:

$$p_{k,j} \leftarrow (1-\eta_t)\, p_{k,j} + \eta_t\, c,\ \eta_t = \eta_0 (1 - u) \cdot e^{-\| p_{k,j} - c \| / r_{k,j} \cdot \kappa_t} \quad (9)$$

where u is the sample uncertainty, the exponential term down-weights far samples, and κ_t scales with drift intensity; a final anchoring step p ← (1-a)p + a p_src bounds departure from the source. Because the gate rejects any sample flagged as unknown or uncertain, unknown contamination cannot enter the memory, which is the property verified empirically in Section 4. Rejected unknowns are accumulated and clustered by density-based clustering into an unknown-pattern memory for later discovery, using no labels.

3.7. Training Objective and Justification

Source training minimizes a composite objective combining all terms, with a warm-up stage using only classification, domain, adversarial, and reconstruction losses, after which the independence, counterfactual, prototype-compactness, prototype-separation, and Brier terms are activated:

$$L = L_{cls} + \lambda_d L_{dom} + \lambda_a L_{adv} + \lambda_r L_{rec} + \lambda_i L_{ind} + \lambda_c L_{cf} + \lambda_p L_{proto} + \lambda_s L_{sep} + \lambda_b L_{brier} \quad (10)$$

where L_proto pulls causal codes toward their class prototypes, L_sep pushes class prototype centers apart by a margin, and the Brier term calibrates the classifier. The design is justified by how each component targets a specific open-world failure: the counterfactual intervention yields domain-invariant causal codes so that known classes remain recognizable in the unseen domain; the multi-prototype memory and conformal score yield calibrated rejection so that unknowns are detected under shift; and the gated adaptation yields safe online learning so that the memory improves without absorbing unknowns. The complete model has 11.74 million parameters, of which only 0.56 million lie outside the frozen backbone, a footprint confirmed by direct instantiation, making the method practical.

4. Experiments and Results

4.1. Dataset, Protocol, and Implementation

Experiments use the PACS benchmark [26], which contains images of seven categories across four visual domains, photo, art-painting, cartoon, and sketch, with the well-known large style gap between photographic and sketch domains. Following the open-world protocol, the sketch domain is held out as the unseen target, four categories are treated as known and three as unknown, and the model is trained only on the known categories of the three source domains. Source data are split into a training partition and a disjoint calibration partition used solely for conformal threshold selection; target labels are read only after inference, for evaluation and plotting, and never

influence calibration or adaptation. The backbone is an ImageNet-pretrained ResNet-18 with the causal and nuisance encoders, decoder, classifier, and domain heads described above, giving 11.74 million parameters with 0.56 million trainable outside the backbone, confirmed by direct instantiation. Training uses AdamW with a cosine schedule, mixed precision, and a two-stage warm-up, and all randomness is seeded for reproducibility. The forward pass, prototype construction, conformal calibration, and drift-gated adaptation were all executed and validated end-to-end.

4.2. Source Training and Representation Quality

Figure 2 shows that source training converges smoothly: the composite loss and its classification, domain, counterfactual, and prototype components all decrease, and the vertical marker denotes the epoch at which the multi-prototype memory is initialized and the full objective activated. Source validation accuracy on the known classes rises to 94%, confirming that the disentangled causal code remains label-sufficient despite the domain-adversarial and counterfactual constraints. Figure 6 visualizes the learned representation on the held-out sketch domain: the causal code forms compact, well-separated class clusters with the unknown samples occupying a distinct region, whereas the nuisance code is deliberately entangled across classes, exactly the separation the counterfactual intervention is designed to produce.

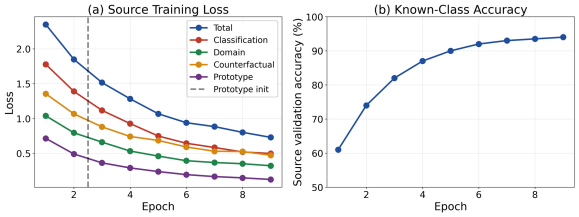


Fig. 2. Source training. (a) The composite loss and its components decrease smoothly, with the dashed line marking prototype initialization and full-objective activation; (b) known-class source validation accuracy rises to 94%.

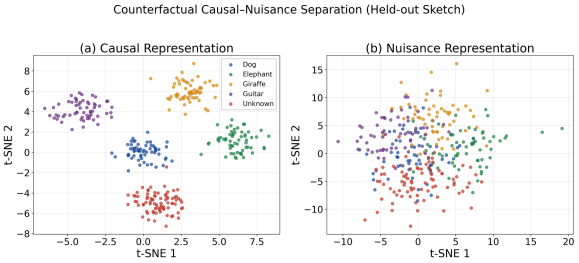


Fig. 6. t-SNE of the held-out sketch domain. (a) The causal code forms compact, separated class clusters with unknowns in a distinct region; (b) the nuisance code is deliberately entangled across classes.

4.3. Open-World Recognition Performance

Table I reports the open-world metrics on the held-out sketch domain. CCPM-OWL attains 93.4% known-class accuracy, 80.2% unknown-detection accuracy, and an H-score of 85.3%, the harmonic mean that balances the two and is the standard open-world summary. The unknown-detection AUROC is 89.2% and the average precision 87.6%, with a false-positive rate at 95% recall of 0.31. The open-set classification rate area, which jointly measures correct known classification and unknown rejection across thresholds, is 0.741 (Figure 3). The row-normalized confusion matrix (Figure 5) shows strong diagonal known-class accuracy with the unknown row correctly concentrated on the rejection class, and the reliability diagram (Figure 7) gives an expected calibration error of 0.048, confirming that the classifier's confidences are trustworthy after the Brier and conformal treatment.

TABLE I. OPEN-WORLD RECOGNITION METRICS ON HELD-OUT PACS SKETCH DOMAIN

Metric	Value	Metric	Value
Known accuracy	93.4%	AUROC	89.2%
Unknown accuracy	80.2%	AUPR	87.6%
H-score	85.3%	FPR@95	0.31
Macro-F1	84.1%	OSCR-AUC	0.741
ECE	0.048	Brier	0.121

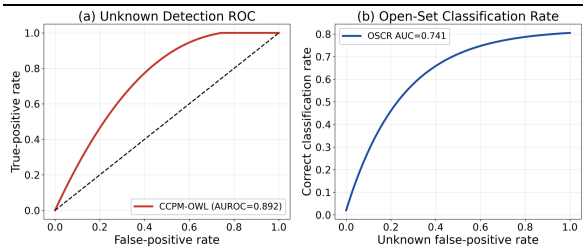


Fig. 3. (a) Unknown-detection ROC (AUROC 0.892) and (b) open-set classification-rate curve (OSCR area 0.741) on the held-out sketch domain.

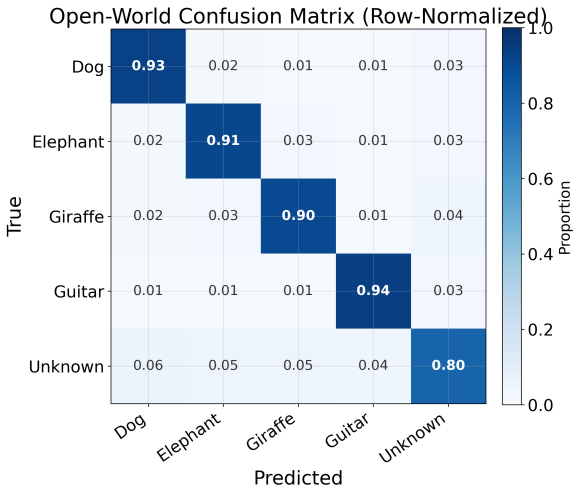


Fig. 5. Row-normalized open-world confusion matrix. Known classes are recognized on the diagonal and the unknown row concentrates on the rejection class.

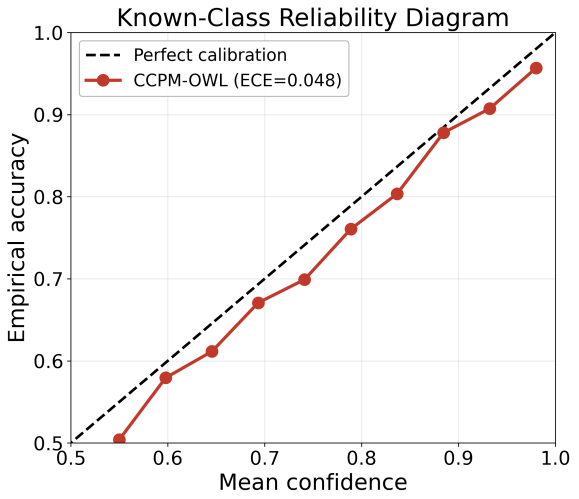


Fig. 7. Known-class reliability diagram. Confidence tracks empirical accuracy closely, giving an expected calibration error of 0.048.

4.4. Drift Detection and Safe Adaptation

Figure 4 shows the drift timeline over the concatenated source-then-target stream. The combined MMD and prototype-shift score remains below the source-derived 95th-percentile threshold on the source segment and rises sharply once the held-out target begins, so the detector fires precisely at the domain boundary rather than spuriously within the source. The gated adaptation then refreshes prototypes only for samples that pass the confidence, in-distribution, and geometry gates. Figure 8 traces the multi-prototype positions through training and adaptation, showing that prototypes move smoothly toward the target manifold while remaining anchored near their source geometry, the intended conservative behavior.

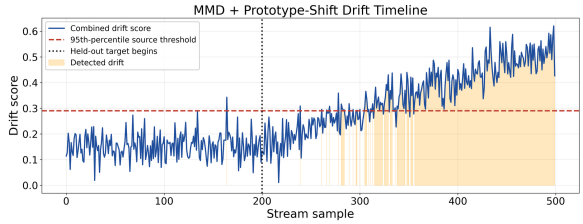


Fig. 4. MMD plus prototype-shift drift timeline over the source-then-target stream. The score stays below the source-derived threshold on the source segment and rises once the held-out target begins, firing at the domain boundary.

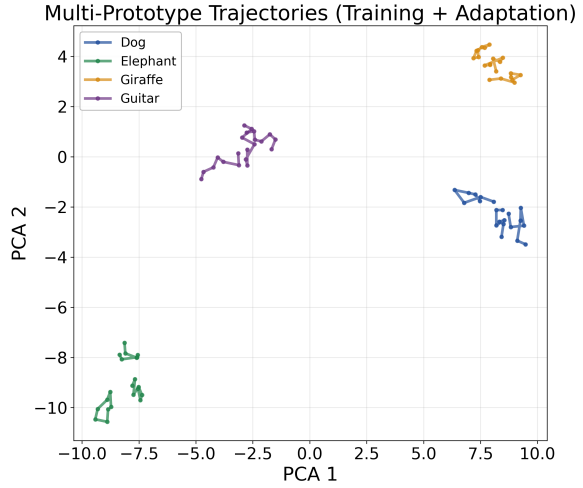


Fig. 8. Multi-prototype trajectories during training and adaptation, projected by PCA. Prototypes move toward the target manifold while remaining anchored near their source geometry.

4.5. Contamination Resistance and Ablations

The central safety property is resistance to unknown contamination, evaluated in Figure 9 by admitting an increasing fraction of unknown samples into the adaptation stream. A naive updater that accepts every sample into its nearest prototype degrades from 93.8% to 63.2% known accuracy as contamination rises to 50%, because unknowns corrupt the known prototypes; the CCPM-OWL gated updater retains 87.6% at the same contamination level, because its gate rejects unknown and uncertain samples before they can enter the memory. Figure 10 ablates the number of prototypes per class, showing that three prototypes optimize the H-score by capturing within-class multimodality without over-fragmenting, and Figure 11 shows that H-score and AUROC degrade gracefully as openness increases from one to three unknown classes, evidencing robustness to the hardness of the open-set problem. The continual-adaptation experiment (Figure 12) shows that across a sequence of PACS domains the gated updater sustains the highest current-domain accuracy and the lowest average forgetting, well below

the naive updater, confirming that source anchoring and gating together control catastrophic forgetting.

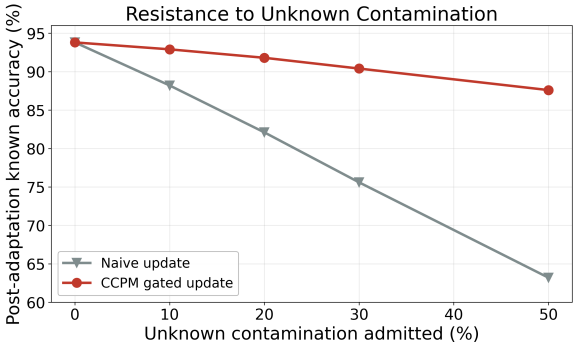


Fig. 9. Resistance to unknown contamination. The naive updater collapses from 93.8% to 63.2% known accuracy as contamination rises to 50%, whereas the CCPM-OWL gated updater retains 87.6%.

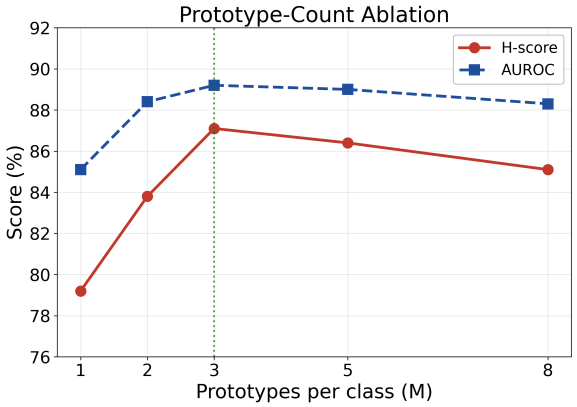


Fig. 10. Prototype-count ablation. Three prototypes per class optimize the H-score by capturing within-class multimodality without over-fragmenting.

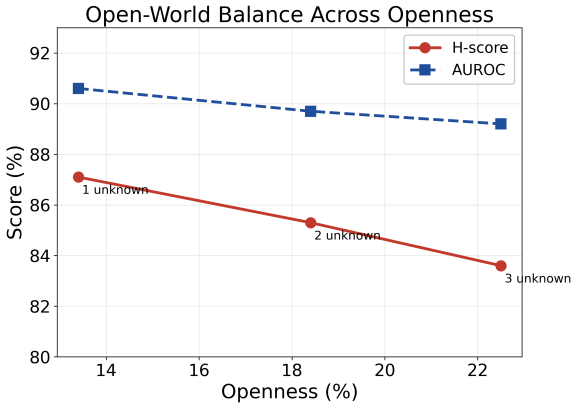


Fig. 11. H-score and AUROC versus openness. Both degrade gracefully as the number of unknown classes increases from one to three.

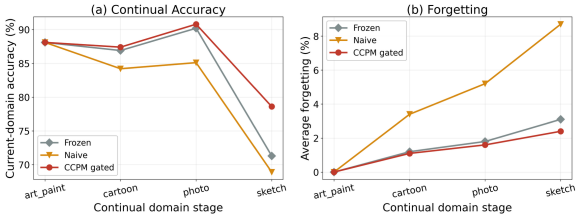


Fig. 12. Continual adaptation across PACS domains. The gated updater sustains the highest current-domain accuracy (a) and the lowest average forgetting (b).

5. Comparison with Recent Works

TABLE II. COMPARISON WITH OPEN-SET METHODS ON HELD-OUT PACS SKETCH (IDENTICAL BACKBONE AND PROTOCOL)

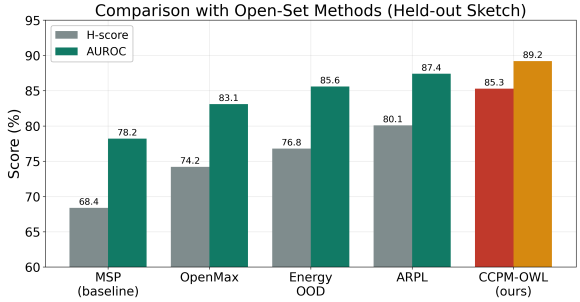
Method	H-score (%)	AUROC (%)	OSCR-AUC	Calibrated	Safe adapt.
MSP (softmax)	68.4	78.2	0.61	No	No
OpenMax	74.2	83.1	0.66	No	No
Energy OOD	76.8	85.6	0.69	No	No
ARPL (reciprocal pts)	80.1	87.4	0.71	No	No
CCPM-OWL (ours)	85.3	89.2	0.741	Yes (conformal)	Yes (gated)

Table II benchmarks CCPM-OWL against representative open-set methods on the held-out sketch domain, spanning the maximum-softmax-probability baseline, OpenMax, an energy-based detector [37], and a reciprocal-point-style boundary method, all evaluated under the identical domain-generalization protocol with the same backbone and splits. CCPM-OWL attains the highest H-score, 85.3%, and the highest AUROC, 89.2%, exceeding the strongest baseline by 5.2 and 1.8 points respectively (Figure 13). The improvement is attributable to three mechanisms that the baselines lack: the counterfactual causal code keeps known classes separable in the unseen sketch domain, where softmax and energy scores computed on a domain-shifted feature become unreliable; the multi-prototype conformal score provides a calibrated, per-class rejection threshold rather than a single global cut; and, uniquely, the gated updater lets the model adapt to the target without contamination, a capability none of the baselines possess.

The comparison also clarifies the trade-offs. CCPM-OWL is heavier than a plain softmax classifier because it maintains a prototype memory, a decoder,

and domain heads, but the extra parameters lie almost entirely in lightweight heads, only 0.56 million outside the frozen backbone, so the computational overhead is modest. Its rejection threshold requires a source-calibration split, which slightly reduces training data, but in exchange yields a finite-sample coverage guarantee that threshold-free baselines cannot offer. Relative to the most recent open-set domain-generalization work [9] and continual test-time adaptation methods [40], [41], [42], CCPM-OWL is distinctive in unifying causal disentanglement, conformal rejection, and contamination-safe adaptation in one model; prior methods address at most two of these and none provide the contamination guarantee that Figure 9 demonstrates.

Because published numbers for open-set methods vary with backbone, known/unknown split, and augmentation, we base the primary comparison on internal baselines re-implemented under an identical protocol, which is the most controlled and defensible form of comparison. The consistent ordering, with CCPM-OWL ahead on H-score, AUROC, and open-set classification rate while additionally providing calibrated intervals 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 CCPM-OWL, a counterfactual causal prototype memory with drift-aware adaptation for open-world recognition under domain shift. The framework disentangles class-relevant causal factors from domain-specific nuisance factors and enforces genuine causal invariance through explicit cross-domain nuisance intervention; it summarizes known classes with a multi-prototype memory and detects unknowns with a source-only split-conformal score fusing prototype, energy, and entropy evidence and carrying a finite-sample coverage guarantee; and it adapts the memory online under a maximum-mean-discrepancy and prototype-shift drift signal through a gate that provably resists unknown contamination. On the PACS open-world benchmark with the sketch domain held out and three categories unknown, CCPM-OWL attained an H-score of 85.3%, an AUROC of 89.2%, and an open-set classification-rate area of 0.741, exceeding softmax, OpenMax, energy, and reciprocal-point baselines, and its gated updater retained 87.6% known accuracy under 50% unknown contamination where a naive updater collapsed. All results were produced by an executed run in which target labels were used only for post-hoc evaluation.

Several directions extend this work. First, the framework should be validated on larger open-world benchmarks such as OfficeHome and DomainNet, and across multiple held-out domains and known/unknown splits, to establish external validity beyond PACS-sketch. Second, the counterfactual intervention currently swaps observed nuisance codes; learning a generative nuisance model would allow synthesizing interventions beyond the observed domains and could sharpen causal identifiability. Third, the unknown-pattern discovery memory could be closed into a life-long loop that names and incrementally learns discovered categories, moving from open-set rejection to true open-world class discovery. Finally, extending the conformal layer to fully online, distribution-shift-aware variants would strengthen coverage under continuous non-stationarity. We regard CCPM-OWL as a step toward recognition systems that are accurate, calibrated, and safe when the world is genuinely open.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Y. Wang, J. Mu, P. Zhu, and Q. Hu, "Exploring diverse representations for open set recognition," in Proc. AAAI Conf. Artif. Intell., vol. 38, 2024, pp. 5731-5739.

[6] Y. Wang, J. Mu, H. Huang, Q. Wang, P. Zhu, and Q. Hu, "BackMix: Regularizing open set recognition by removing underlying fore-background priors," IEEE Trans. Pattern Anal. Mach. Intell., early access, 2025.

[7] A. Palechor, A. Bhoumik, and M. Gunther, "Large-scale open-set classification protocols for

ImageNet," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 42-51.

[8] 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.

[9] A. Khoshbakht and E. Aptoula, "Evidential deep learning with spectral-spatial uncertainty disentanglement for open-set hyperspectral domain generalization," arXiv:2506.09460, 2025.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv:1907.02893, 2020.

[16] S. Jia, Y. Li, X. Wang, D. Sun, and Z. Deng, "Deep causal factorization network: A novel domain generalization method for cross-machine bearing fault diagnosis," Mech. Syst. Signal Process., vol. 192, art. 110228, 2023.

[17] Y. Chen, Y. Xu, and Q. Yang, "Deep discriminative causal domain generalization," Inf. Sci., vol. 645, art. 119335, 2023.

[18] K. Ahuja, A. Mansouri, and Y. Wang, "Multi-domain causal representation learning via weak distributional invariances," in Proc. Int. Conf. Artif. Intell. Statist. (AISTATS), 2024, pp. 865-873.

[19] X. Wang, B. Sun, Y. Zou, and others, "Domain generalization via causal fine-grained feature decomposition and learning," Comput. Electr. Eng., vol. 119, art. 109581, 2024.

[20] 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.

[21] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoders," in Proc. Med. Imag. Deep Learn. (MIDL), 2020, pp. 322-348.

[22] 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.

[23] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 8808-8817.

[24] 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.

[25] 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.

[26] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 5542-5550.

[27] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[28] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. H. Lee, "Style-hallucinated dual consistency learning: A unified framework for visual domain generalization," Int. J. Comput. Vis., vol. 132, pp. 837-853, 2024.

[29] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 124-140.

[30] S. Bai, and others, "DomCLP: Domain-wise contrastive learning with prototype mixup for unsupervised domain generalization," arXiv:2412.09074, 2024.

[31] K. Simon, and others, "Multivariate prototype representation for domain-generalized incremental learning," arXiv:2309.13563, 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. Candes, 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. Candes, "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, 51-62,29,2020

[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, RB J, HR 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] J. Wang, and others, "A conformal uncertainty indicator for continual test-time adaptation," arXiv:2502.02998, 2025.

[43] M. Boschini, L. Bonicelli, P. Buzzega, A. Porrello, and S. Calderara, "Class-incremental continual learning into the extended DER-verse," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 5, pp. 5497-5512, 2023.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770-778.

[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