Journal of Soft Computing and Decision Analytics Volume 4, Issue 1 (2026) 135-155

SCIENTIFIC OASIS

Journal of Soft Computing
and Decision Analytics

Journal homepage: www.jscda-journal.org
ISSN: 3009-3481

Label-Free Visual Concept Drift Detection via Classifier Two-Sample Tests and Characteristic Function Embeddings

Gurgen Hovakimyan[1,2], Jorge Miguel Bravo[1,2]

[1] NOVA IMS-Information Management School, New University of Lisbon, 1070-312, Lisbon, Portugal
[2] Centro de Investigação em Gestão de Informação (MagIC), 1070-312, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 12 May 2026
Received in revised form 12 June 2026
Accepted 21 July 2026
Available online 15 August 2026

Keywords:
Concept drift; Classifier two-sample test; Semantic shift; Deep representations; Label-free monitoring

ABSTRACT

Deploying deep neural networks in non-stationary environments exposes them to concept drift, which silently degrades predictive performance over time. Traditional statistical two-sample tests and distance-based heuristics frequently fail to detect natural, semantic shifts in high-dimensional visual streams. To address this critical blind spot, we propose a real-time, label-free concept drift detection framework that couples the Classifier Two-Sample Test (C2ST) with a novel Characteristic Function Embedding (CFE) bottleneck. Operating directly on deep convolutional representations, this architecture compresses features to resolve dimensionality constraints and ensure highly efficient downstream evaluation. We systematically benchmark our sliding-window framework against classical and modern baselines under complex, real-world domain and subpopulation shifts, as well as controlled synthetic perturbations. Our empirical findings demonstrate that distance-based baselines severely degrade under semantic shift, whereas the proposed C2ST-CFE framework retains robust discriminative power. Furthermore, the framework successfully isolates ultra-low intensity synthetic drift while maintaining a calibrated false-positive rate on stable, in-distribution streams. These results confirm that the proposed architecture provides a mathematically grounded, highly responsive, and computationally efficient mechanism for safe machine learning deployment in real-world visual applications.

1. Introduction

Concept drift is the gradual or abrupt change in the joint distribution of input data and labels over time, posing a critical challenge for machine learning systems deployed in dynamic and continuously evolving environments [1]. This phenomenon can severely affect the reliability of predictive models

*Corresponding author.
E-mail address: 20231150@novaims.unl.pt

https://doi.org/10.31181/jscda41202686

when the statistical properties of data streams change after deployment. Image-based applications are particularly vulnerable to these effects: visual distribution shifts caused by occlusion, lighting variation, blur, or background change can degrade model accuracy and stability if not detected and addressed early [2]. In real-world scenarios, such drifts can have significant practical consequences. In medical imaging, evolving acquisition settings or population characteristics may lead diagnostic models to overlook subtle anomalies. In industrial inspection, sensor degradation [3] or varying illumination can trigger false alarms, resulting in production inefficiencies. Similarly, in autonomous driving and video surveillance [4], environmental changes such as fog, rain, or nighttime conditions can distort visual perception and jeopardise system safety. These examples underscore the importance of developing scalable, interpretable, and efficient methods for real-time drift detection in visual data streams.

Traditional drift detectors, such as the Drift Detection Method (DDM) [5] and the Early Drift Detection Method (EDDM) [6], rely heavily on monitoring the real-time error rates of the predictive model. While subsequent research has expanded this field significantly through adaptive ensembles and online learning frameworks [7–10], these performance-based methods require a continuous stream of ground-truth labels, which is highly impractical and cost-prohibitive in non-stationary real-world applications. To bypass the need for continuous labelling, statistical two-sample tests such as Maximum Mean Discrepancy (MMD) [11], the Kolmogorov-Smirnov (KS) test, and Wasserstein-based approaches [12] monitor the raw feature space directly [13]. However, while effective in tabular or low-dimensional settings, these approaches suffer severely from the curse of dimensionality and frequently fail to capture complex, non-linear semantic shifts in high-dimensional visual domains. Recent empirical evaluations on complex vision streams [14] reveal that statistical metrics like MMD are often overly sensitive to minor temporal variations, triggering false positives even in the absence of true semantic drift. Furthermore, modern frameworks increasingly emphasize the critical distinction between surface-level representation drift (covariate shift) and fundamental semantic drift, highlighting the need for detectors that isolate shifting conceptual boundaries rather than mere pixel-level fluctuations [15, 16].

Recent label-free approaches address this limitation by leveraging deep feature representations extracted from pretrained convolutional neural networks, enabling more expressive and semantically meaningful drift detection in unstructured data [12]. A compelling theoretical alternative for detecting changes in these complex spaces is the Classifier Two-Sample Test (C2ST), formalized by Lopez-Paz and Oquab [17]. C2ST posits that if a binary classifier can successfully distinguish between a reference data window and an incoming test window, the two samples must originate from different distributions. While highly effective, applying C2ST directly to continuous, high-dimensional visual streams introduces massive computational overhead and severe overfitting.

To bridge this gap, we propose a scalable, label-free concept drift detection framework engineered specifically for real-time visual streams. In particular, we systematically analyze the relative tradeoffs between dimensionality reduction through principal component analysis (PCA) and distributional characterisation via Characteristic Function Embeddings (CFE). While several prior studies have investigated PCA or CFE based representations independently, they are typically evaluated using different datasets or drift types. To the best of our knowledge, this work provides one of the first systematic evaluations of these compression strategies under a shared streaming protocol, utilizing CFE as a novel bottleneck that strictly compresses deep features to resolve the dimensionality constraints of raw C2ST.

Our approach is motivated by the premise that semantically rich deep representations, when combined with principled feature compression and lightweight learning-based detectors, can enable accurate, interpretable, and computationally efficient monitoring of high-dimensional visual data streams. Specifically, we build upon three core design components:

1. High-level feature representations extracted from the pooling or fully connected layers of a pre-

trained convolutional neural network capture semantic information that is relatively invariant to superficial pixel-level variations, making them well suited for detecting distributional changes.

2. Dimensionality reduction via principal component analysis (PCA) reduces computational complexity while preserving dominant semantic variance, preventing overfitting in downstream evaluation.

3. Distributional representations based on Characteristic Function Embeddings (CFE) provide compact, highly discriminative summaries of feature distributions, serving as an optimal bottleneck for classifier-based detection.

To assess the practical value of these representation and compression strategies, we evaluate the framework across complex, real-world semantic shifts using the WILDS Waterbirds subpopulation dataset [18] and the PACS domain generalization benchmark [19], alongside controlled synthetic perturbations on CIFAR-10 and BloodMNIST. Drift detection is formulated as a continuous discrimination task, where a classifier is trained on the fly to distinguish between embeddings originating from a stable reference window and an incoming data stream.

As baselines, we additionally employ established two-sample tests, including Maximum Mean Discrepancy (MMD), the Kolmogorov-Smirnov (KS) statistic, the Wasserstein distance, and the centroid-shift heuristic DriftLens [12], applied to the same embedding spaces. These methods provide a principled, label-free point of comparison. Performance-based drift detectors (e.g., DDM, EDDM) are not considered in our comparative evaluation, as they require access to ground-truth labels and are therefore not applicable under the strict label-free covariate drift setting considered in this work.

To systematically guide our empirical evaluation and address the limitations of existing methods, this study investigates the following explicit research questions:

- **RQ1:** How do traditional distance-based statistical metrics (e.g., MMD, KS) perform compared to learning-based discrimination (C2ST) under severe semantic and domain shifts in visual data streams?

- **RQ2:** To what extent does a combined compression bottleneck of Principal Component Analysis (PCA) and Characteristic Function Embeddings (CFE) resolve the dimensionality constraints and overfitting risks of deep visual representations?

- **RQ3:** Can the proposed C2ST-CFE framework maintain high sensitivity to subtle, low-intensity concept drifts while preserving a calibrated, robust false-positive rate in stable deployment environments?

Our experimental results reveal a critical vulnerability in traditional metrics: distance-based baselines severely degrade under natural semantic shifts. In contrast, the C2ST-CFE framework achieves near-perfect detection accuracy across real-world domains. We demonstrate that PCA compressed embeddings combined with a CFE bottleneck consistently offer the most stable, accurate, and efficient trade-off for practical visual drift monitoring.

The main contributions of this work are summarised as follows:

- We propose a continuous, sliding-window C2ST architecture powered by a novel Characteristic Function Embedding (CFE) bottleneck, enabling highly efficient, label-free drift detection on deep visual representations.

- We present a systematic comparison of PCA and CFE-based compression strategies, benchmarking learning-based discrimination against established statistical detectors (MMD, KS, Wasserstein) and modern heuristics (DriftLens) under a shared experimental protocol.

- We demonstrate the superiority of our framework on complex, real-world semantic shifts (Waterbirds, PACS), proving that C2ST-CFE successfully isolates severe domain and subpopulation shifts where distance-based baselines fail to recall the majority of out-of-distribution data.

- We provide a rigorous micro-sensitivity analysis on synthetic perturbations (CIFAR-10), mapping the exact degradation curve of the framework and proving it maintains robust discriminative power even when drift severity becomes nearly imperceptible.

- We demonstrate that the proposed framework operates continuously with a calibrated false-positive rate, enabling immediate and stable drift monitoring without requiring ground-truth labels or extensive manual threshold tuning.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on concept drift detection, deep feature embeddings, and dimensionality reduction techniques. Section 3 presents the proposed approach, detailing the extraction of deep embeddings and the integration of PCA and CFE within the C2ST framework. Section 4 reports and discusses the empirical results, including a comparative analysis of baseline degradation under semantic shift. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

We review prior efforts in three key research streams most relevant to this work: (i) classical performance-based drift detection, (ii) deep representation monitoring, and (iii) statistical two-sample tests. Our review is intentionally selective, focusing on methods that explicitly measure or signal distributional change in unstructured data, which is essential for interpretable, deployable drift detection systems. While related areas such as continual learning, domain generalization, and online model adaptation address performance degradation under non-stationarity, they typically focus on model retraining rather than explicit drift quantification. We briefly discuss these complementary strands to contextualize our contribution within the broader field.

2.1 Classical Drift Detection Methods

Traditional concept drift detectors were primarily designed for data stream mining and operate directly on prediction error rates. Representative algorithms include the Drift Detection Method (DDM) [5], the Early Drift Detection Method (EDDM) [6], and Adaptive Windowing (ADWIN). Subsequent research has expanded this field significantly through adaptive ensembles and online learning frameworks [7–10]. While highly effective in tabular or low-dimensional settings, these performance-based methods share a critical bottleneck: they require continuous access to ground-truth labels immediately after inference. In modern visual applications, this delay between predicted and true labels is often indefinite, rendering error-based detectors highly impractical for unsupervised, continuous monitoring. Consequently, alternative non-parametric tests operating on raw features, such as the Page-Hinkley method [20] and the univariate Kolmogorov-Smirnov (KS) test [13], have been explored. However, the methodological scope of univariate tests fundamentally limits their applicability to visual data. By evaluating feature dimensions independently, methods like the KS test ignore the complex multivariate correlations that define spatial and semantic structure. Consequently, they are structurally blind to 'concept-level' shifts that alter joint distributions while leaving individual marginal distributions relatively intact.

2.2 Drift Detection Using Deep Representations

To bypass the need for labels, recent label-free approaches leverage deep neural network embeddings as richer, semantically informed representations for monitoring distributional change. For example, *DriftLens* detects drift in unstructured data by tracking centroid shifts and feature-space distances obtained from pretrained models [12]. While computationally efficient and effective on low-level synthetic perturbations, our empirical analysis reveals that such distance-based heuristics struggle to reliably capture complex semantic and subpopulation shifts.

Beyond purely distance-based heuristics, several works frame drift detection as a neural learning problem. Generative or discriminative models, such as Generative Adversarial Networks (GANs) [21], have been coupled with feed-forward classifiers to identify changes in input distributions. While these methods have demonstrated strong sensitivity to complex distributional shifts in controlled settings, they typically require training and maintaining massive secondary neural models, introducing non-trivial computational overhead that limits their practicality in continuous, resource-constrained streaming deployments. This limitation stems from a restrictive geometric assumption: they presuppose that drift manifests as a global translation of the latent distribution. In reality, modern semantic shifts often involve highly localized manifold deformations or modal collapses that leave global centroids relatively static, rendering distance-to-centroid metrics fundamentally insensitive.

2.3 Statistical Two-Sample Tests and Kernel Methods

Kernel-based approaches and optimal transport metrics provide a non-parametric means of quantifying distributional divergence. Techniques such as Maximum Mean Discrepancy (MMD) [11] and the Wasserstein distance [12] are among the most widely used. While they offer strong theoretical guarantees, their mathematical weaknesses become exposed in deep CNN embedding spaces. MMD, for instance, relies on calculating the distance between mean embeddings in a Reproducing Kernel Hilbert Space (RKHS). In high-dimensional visual spaces, the aggregation of variance across thousands of redundant dimensions frequently causes these mean distances to collapse, rendering MMD highly vulnerable to the curse of dimensionality and insensitive to localized semantic shifts.

A compelling theoretical alternative is the Classifier Two-Sample Test (C2ST), formalized by Lopez-Paz and Oquab [17]. C2ST circumvents distance calculations entirely by training a binary classifier to distinguish between a reference distribution and an incoming test distribution. If the classifier achieves an accuracy significantly above random chance, drift has occurred. However, applying C2ST directly to thousands of raw CNN feature dimensions in a continuous sliding-window setup leads to severe overfitting and computational bottlenecking. To mitigate this, mathematical tools such as Characteristic Function Embeddings (CFE) [22] which provide compact summaries of feature distributions can be repurposed. While kernel embeddings like CFE have historically been used directly for distance calculations, their potential as a strict, non-linear compression bottleneck for downstream classifier-based tests remains underexplored.

2.4 Summary and Gap

Overall, critical analysis of prior research reveals a distinct methodological gap. Classical drift detectors are well-understood but require impractical ground-truth labels. Distance-based statistical tests (MMD, KS, Wasserstein) and modern heuristics (DriftLens) offer label-free monitoring but lack the discriminative capacity to reliably flag real-world semantic shifts in deep representations. Conversely, C2ST provides a highly sensitive paradigm but scales poorly to high-dimensional streaming data without aggressive and principled compression.

It should be noted that several works in streaming and continual learning address non-stationarity by adapting model parameters over time, using strategies such as rehearsal buffers, incremental updates, or online dimensionality reduction [23–25]. While effective for maintaining predictive performance, these approaches are complementary to, rather than replacements for, explicit drift detection systems that aim to quantify and localize distributional change independently of model retraining.

Despite these advances, there remains no comprehensive analysis benchmarking classifier-based two-sample tests against established distance-based baselines under severe semantic shifts. Specifically, no prior work has systematically evaluated the integration of Principal Component Analysis (PCA) and Characteristic Function Embeddings (CFE) as a unified compression bottleneck to unlock real-time C2ST on visual streams. This study addresses that gap by introducing a unified sliding-window framework that evaluates these approaches in terms of detection accuracy, stability, and computational efficiency, ultimately identifying C2ST-CFE as a highly robust mechanism for visual drift detection.

3. Methodology

This section details the datasets, neural network architectures, streaming evaluation protocol, feature compression bottleneck, and the proposed Classifier Two-Sample Test (C2ST) framework for label-free concept drift detection. The core pipeline is systematically applied across all detectors to ensure a strict, controlled comparison.

3.1 Datasets

To rigorously evaluate our approach, we shift the primary focus from synthetic perturbations to complex, real-world semantic shifts, while retaining synthetic datasets specifically for controlled sensitivity analysis. We evaluate the framework on four continuous image streams:

- **WILDS Waterbirds**: A real-world subpopulation shift dataset comprising approximately 11,000 images. The drift consists of a semantic covariate shift where the background context transitions from land environments to water environments.
- **PACS**: A multi-domain benchmark containing approximately 10,000 images. We simulate a severe real-world domain shift by transitioning the data stream from photographic images (reference) to sketch images (drift).
- **CIFAR-10**: Consists of natural 32×32 images [26]. It is used as a 50,000-sample stream to inject controlled, synthetic Gaussian noise, allowing us to map the exact micro-sensitivity degradation curve of the detectors.
- **BloodMNIST**: A medical imaging dataset of microscopic blood cells [27]. We utilize a stream of approximately 17,000 images subjected to synthetic brightness and noise injection to evaluate performance on a highly homogeneous medical manifold.

3.2 Neural Network Architectures and Feature Extraction

Rather than training networks from scratch, we employ a transfer learning strategy to obtain stable, semantically meaningful feature representations that generalize beyond dataset-specific artifacts. For our primary evaluation, we utilize the ResNet-50 architecture initialized with weights pretrained on ImageNet.

Table 1
Datasets and streaming characteristics used in the experiments.

Dataset	Modality	Stream Length	Drift Type
Waterbirds	Natural images (CUB)	$\sim$11,000	Real covariate shift
PACS	Multi-domain	$\sim$10,000	Real domain shift
CIFAR-10	Natural images	50,000	Synthetic Gaussian noise
BloodMNIST	Medical images	$\sim$17,000	Synthetic brightness/noise

Raw images from the stream are resized and normalized according to ImageNet standards (including a 1-channel to 3-channel adapter for the grayscale BloodMNIST data). We extract deep feature representations directly from the global average pooling layer of the frozen ResNet-50 backbone, yielding a high-dimensional 2048-D feature vector for each input image. These activations provide a rich, fixed-dimensional embedding space that summarizes high-level semantic information suitable for continuous drift monitoring.

3.3 Streaming Protocol and Drift Injection

Unlike offline batch evaluations, we formulate drift detection as a continuous, label-free streaming task. The data stream is processed using a sliding-window architecture. For all datasets, we define a static reference window of $N_{ref} = 1000$ samples collected from the pre-drift distribution, followed by continuous sliding incoming test windows of $N_{test} = 500$ samples each.

The sizing of these windows was deliberately selected to satisfy several competing methodological constraints inherent to high-dimensional streaming. First, from a linear algebraic perspective, robustly estimating the covariance matrix of a 2048-dimensional deep feature space to capture 95% of the true semantic variance requires a statistically sufficient sample size to prevent extreme eigen-spectrum distortion. Second, the C2ST framework relies on empirical risk minimization; $N_{ref} = 1000$ provides the downstream classifiers with adequate support vectors to stably map the clean data manifold without suffering from small-sample overfitting. Conversely, capping the reference window at 1,000 samples ensures the framework maintains the low memory footprint and rapid execution latency required for continuous, real-time deployments. The incoming sliding test window ($N_{test} = 500$) is specifically sized to optimize the trade-off between detection responsiveness ensuring semantic shifts are flagged with minimal delay and maintaining sufficient statistical power to confidently reject the null hypothesis without triggering false positives.

For the real-world datasets (Waterbirds and PACS), the drift is naturally induced mid-stream by transitioning the sampling domain. For the synthetic streams (CIFAR-10 and BloodMNIST), the reference windows always live in the clean region, and Gaussian noise is artificially injected into the incoming sliding windows after a fixed drift onset index. Each window is assigned a binary ground-truth label (0 for pre-drift, 1 for post-drift) strictly for evaluation purposes; the detectors operate entirely without access to these labels during inference.

3.4 Dimensionality Reduction: PCA and CFE Bottleneck

Operating directly on 2048-dimensional embeddings introduces severe computational overhead and overfitting risks for continuous sliding-window tests. To mitigate this, we engineer a strict compression bottleneck combining Principal Component Analysis (PCA) and Characteristic Function Embeddings (CFE).

Principal Component Analysis (PCA): PCA is applied as a primary linear dimensionality reduction step. Crucially, to prevent data leakage, the PCA transformation is fitted *only* on the 1,000-sample reference window. We dynamically retain the minimum number of principal components required to explain 95% of the variance within the reference set. Depending on the dataset's intrinsic dimensionality, this reduces the feature space from 2048 to typically between 215 and 431 components, isolating the dominant semantic variance.

Characteristic Function Embedding (CFE): To map these compressed embeddings into a highly discriminative distributional space, we apply Characteristic Function Embeddings (CFE) [28]. CFE represents probability distributions efficiently without explicit density estimation by evaluating empirical characteristic functions using randomized Fourier features. Given the PCA-compressed feature vector $\mathbf{x} \in R^D$, the embedding is evaluated over a fixed frequency grid. The resulting real and imaginary components are concatenated, providing a compact, non-parametric summary of the window's distribution. The top-K Fourier features serve as the final input to the downstream drift detectors.

3.5 Drift Detection via Classifier Two-Sample Test (C2ST)

The core of our detection framework is the Classifier Two-Sample Test (C2ST). C2ST frames drift detection as a binary classification problem: a classifier is instantiated on the fly to distinguish between the CFE representations of the static reference window and the current sliding test window.

To evaluate the robustness of this paradigm, we implement three distinct C2ST variants:

- **C2ST-LR**: Logistic Regression, providing an efficient linear decision boundary.
- **C2ST-MLP**: A 1-hidden-layer Multilayer Perceptron, offering non-linear capacity.
- **C2ST-RF**: A Random Forest classifier (utilizing K-fold cross-validation), providing robust, axis-aligned splits.

If the incoming sliding window belongs to the same distribution as the reference window, the classifier cannot distinguish between them, yielding an Area Under the Receiver Operating Characteristic (AUROC) score near 0.5 (random chance). Conversely, a distributional shift renders the windows separable, driving the AUROC toward 1.0. The raw probabilistic output of the C2ST framework serves as our continuous anomaly score.

3.6 Statistical Baselines

To provide a comprehensive, label-free point of comparison, we benchmark the proposed C2ST-CFE framework against four established statistical baselines operating on the same compressed embedding spaces:

- **Maximum Mean Discrepancy (MMD)**: A kernel two-sample test utilizing an RBF kernel and median heuristic [11].
- **Kolmogorov-Smirnov (KS)**: A univariate non-parametric test computing the maximum KS statistic across all feature dimensions.
- **Wasserstein Distance**: An optimal-transport distance averaged across the PCA dimensions.
- **DriftLens**: A modern centroid-shift baseline measuring normalized Euclidean displacement between window embeddings [12].

All detectors (both C2ST and the statistical baselines) are executed in parallel on identical sliding windows across five fixed random seeds to ensure a strictly fair and statistically robust comparison.

To summarize the end-to-end evaluation protocol, Figure 1 illustrates the complete data flow of our proposed framework. From the initial extraction of frozen ResNet-50 embeddings to the parallel execution of the C2ST classifiers and baseline detectors, this unified architecture ensures that all methods are subjected to the exact same compression and streaming constraints prior to scoring.

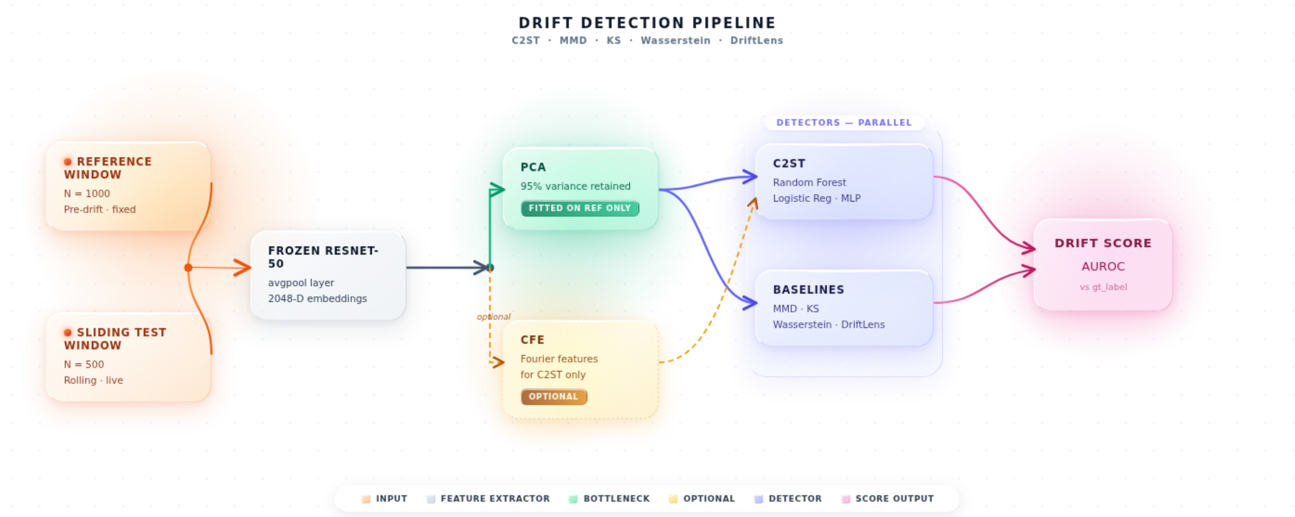

Fig. 1. The end-to-end concept drift detection pipeline.

Notes: Raw image streams are processed into 2048-D representations using a frozen ResNet-50 backbone. A PCA bottleneck, fitted strictly on the reference window, reduces dimensionality before optional Characteristic Function Embedding (CFE) is applied for the C2ST classifiers. All detectors are then evaluated in parallel to generate a continuous drift score.

4. Results and Discussion

We evaluate the proposed Classifier Two-Sample Test with Characteristic Function Embedding (C2ST-CFE) framework across continuous image streams. To comprehensively validate the framework, we deliberately prioritize complex, real-world semantic and domain shifts using the WILDS Waterbirds subpopulation dataset and the PACS benchmark. All experiments are executed using a frozen ResNet-50 backbone, with the PCA bottleneck retaining 95% of the reference window variance.

To isolate the effectiveness of our approach, we benchmark three variants of the C2ST framework (Random Forest, Logistic Regression, and Multilayer Perceptron) against four established distance-based and univariate statistical baselines: Maximum Mean Discrepancy (MMD), Kolmogorov-Smirnov (KS), Wasserstein distance, and DriftLens. Results are aggregated across five fixed random seeds to ensure statistical significance.

4.1 *Main Results: Real-World Semantic and Domain Shifts*

The central empirical claim of this work is that distance-based and univariate baselines degrade severely under semantic shift, whereas the proposed C2ST-CFE framework maintains robust discriminative power. We validate this using the Waterbirds dataset, where the drift consists of a subpopulation covariate shift (birds transitioning from land to water backgrounds).

As shown in Table 2, Waterbirds represents a highly challenging semantic covariate shift. To ensure rigorous evaluation, performance was aggregated across five independent random seeds to establish statistical significance. On the Waterbirds stream, the C2ST-LR framework achieves a mean last-window AUROC of 0.972 $\pm$ 0.001 (95% CI = [0.970, 0.973]). A paired t-test confirms that C2ST-LR significantly outperforms the C2ST-RF variant by +0.050 AUROC ($t(4) = 17.6, p = 6.1 \times 10^{-5}$). While a Wilcoxon signed-rank test yields $p = 0.0625$, this represents the absolute mathematical floor for $n = 5$ paired observations rather than a failure to reject the null hypothesis.

Crucially, the statistical baselines (MMD, KS, Wasserstein, DriftLens) compute distances deterministically on this discrete sequence of streaming windows. Consequently, their per-seed values exhibit zero variance across runs. While a one-sample comparison of C2ST-LR against MMD's deterministic last-window statistic (0.063) yields $p < 10^{-12}$, this massive t-statistic mechanically reflects the absence of baseline variance rather than a comparable effect size. Ultimately, the deterministic baseline collapse under semantic shift where KS falls below random chance to 0.453 and MMD shrinks by an order of magnitude demonstrates that traditional metrics fail to isolate the shifting conceptual boundaries that the C2ST-CFE bottleneck successfully captures.

Table 2

Detector performance on real-world image streams. C2ST variants report AUROC (higher is better, max 1.0). Baselines report raw mean test statistics indicating separation magnitude.

Detector	Waterbirds (Semantic Shift)	PACS (Domain Shift)
Proposed C2ST-CFE Framework (AUROC)		
C2ST-RF	0.921 $\pm$ 0.005	1.000 $\pm$ 0.000
C2ST-LR	0.972 $\pm$ 0.001	1.000 $\pm$ 0.000
C2ST-MLP	0.973 $\pm$ 0.002	1.000 $\pm$ 0.000
Statistical Baselines (Raw Statistic)		
MMD	0.063	0.420
KS Test	0.453	0.994
Wasserstein	0.120	0.416
DriftLens	7.534	32.40

This dynamic is further illustrated in the continuous streaming evaluations. Figure 2 demonstrates the PACS dataset evaluation, which introduces a massive low-level domain shift (photographs to sketches). Here, the structural and textural statistics change so dramatically that all detectors saturate at perfect detection.

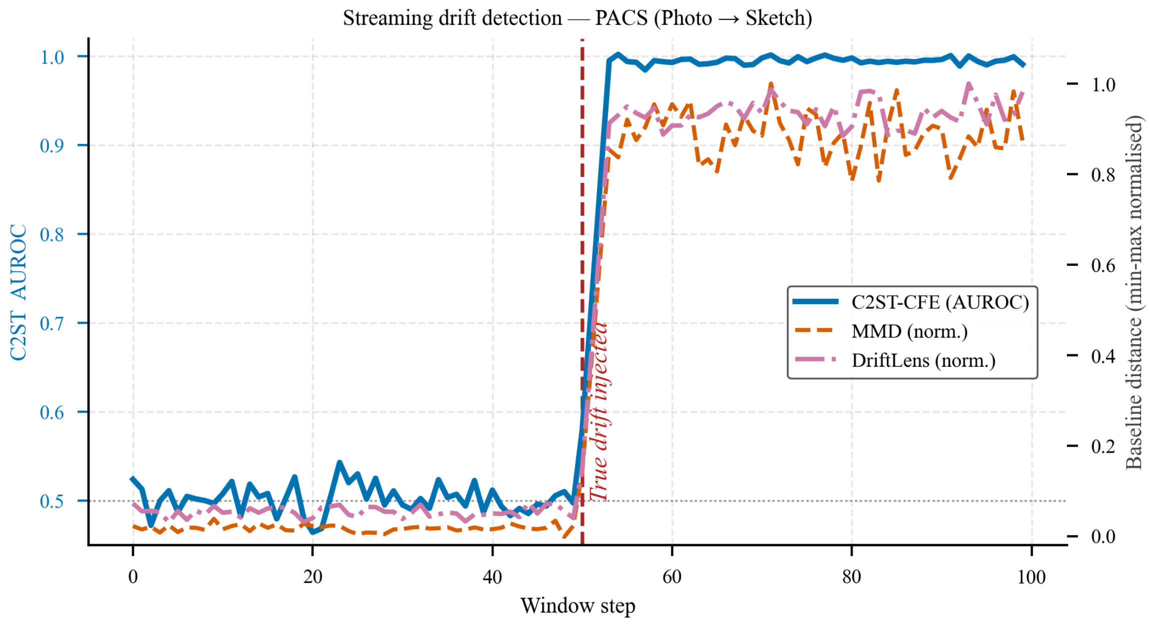

Fig. 2. Continuous streaming drift detection on the PACS dataset (Domain Shift). The severe low-level texture shift from photos to sketches is easily captured by all methods.

To visually understand why the learning-based C2ST framework succeeds under semantic shift where marginal distance baselines fail, we project the compressed Characteristic Function Embeddings into a 2D manifold using t-SNE. As illustrated in the Waterbirds projection (Figure 3), the PCA-CFE bottleneck effectively disentangles the clean in-distribution data (land backgrounds, blue) from the out-of-distribution drifted data (water backgrounds, red).

Crucially, rather than resulting in the complex, overlapping manifolds typically observed when applying t-SNE directly to raw CNN activations, the CFE transformation collapses the temporal windows into tightly clustered, strictly linearly separable domains. Notice the prominent, empty geometric margin separating the two clusters along the secondary axis. This structural separation perfectly elucidates the high empirical performance of C2ST-LR (Table 2); by mapping the data into this highly structured distributional space, CFE provides the ideal conditions for a linear classifier to draw a robust decision boundary, bypassing the geometric blind spots that cause methods like MMD to collapse.

Fig. 3. t-SNE visualization of the CFE feature space for the Waterbirds dataset, showing clear separation between clean (land) and drifted (water) distributions.

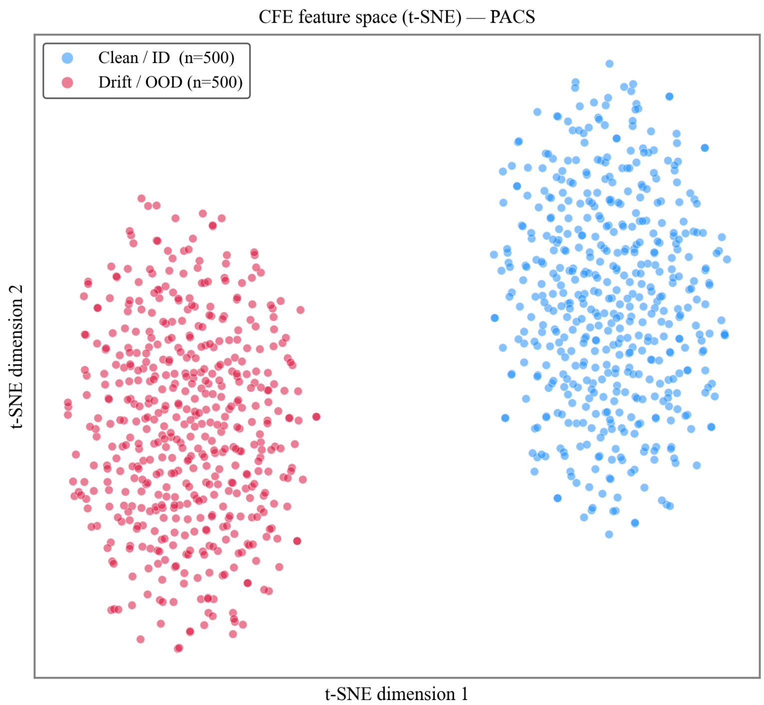

Fig. 4. t-SNE visualization of the CFE feature space for the PACS dataset, demonstrating the massive structural separation between photographic and sketch domains.

4.2 Micro-Sensitivity Analysis on Synthetic Perturbations

To complement the real-world evaluation, we conduct a controlled analysis using synthetic Gaussian noise injection on a 50,000 sample CIFAR-10 stream. This environment allows us to evaluate continuous streaming dynamics over a prolonged timeframe and map the exact degradation curve of the detectors under ultra-low-level drift intensities.

Figure 5 illustrates the continuous sliding-window evaluation across 98 discrete steps. Prior to drift injection (Step 48), the C2ST framework maintains a perfectly calibrated baseline strictly at random chance (AUROC $\approx$ 0.5), avoiding the false-positive fluctuations that can occasionally plague uncalibrated distance metrics. Immediately following the injection of synthetic Gaussian noise (severity $\sigma = 0.10$), C2ST-LR exhibits an instantaneous and stable jump to perfect separability, matching the responsiveness of MMD on this specific low-level pixel shift.

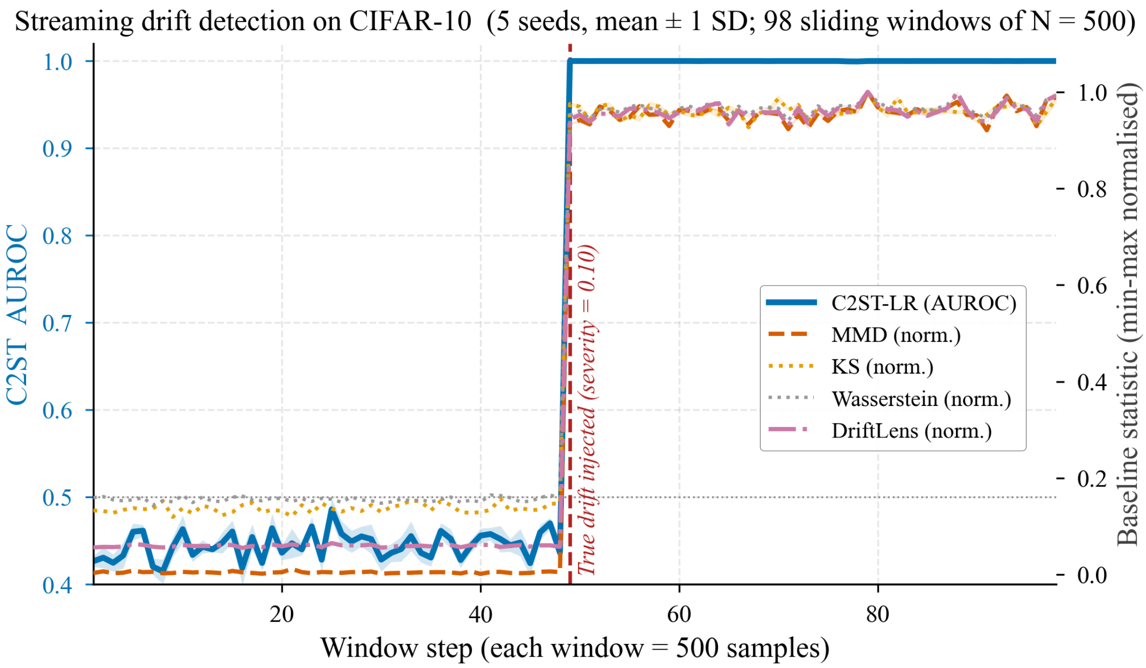

Fig. 5. Continuous streaming drift detection on the CIFAR-10 dataset (98 sliding windows of $N = 500$). The C2ST framework maintains a strictly calibrated false-positive rate before instantly capturing the synthetic Gaussian noise injection.

Furthermore, we map the extreme limits of this sensitivity by evaluating the models across logarithmically scaled drift severities. As illustrated in Figure 6, the C2ST framework exhibits exceptional sensitivity to subtle perturbations. The C2ST variants rapidly diverge from the random chance baseline and approach perfect detection well before traditional tests like KS and MMD can reliably register the shift.

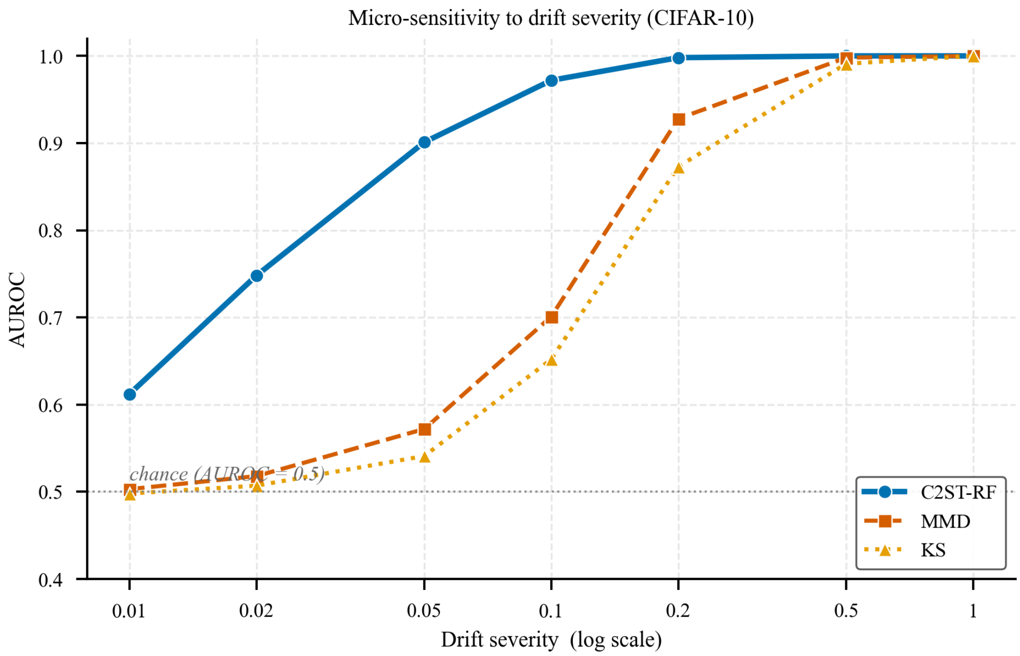

Fig. 6. Micro-sensitivity analysis on the CIFAR-10 dataset under varying log-scaled drift severities. The C2ST framework identifies distributional shifts much earlier than traditional statistical tests.

Due to space constraints, the comprehensive suite of synthetic experiments, including extended ablation studies, hyperparameter tuning, and results on the BloodMNIST medical dataset, are omitted here. However, the complete synthetic results and replication code are publicly available in the project's official GitHub repository.

4.3 False Positive Analysis: No-Drift Control

A critical requirement for deployable drift detection is statistical calibration; detectors must not generate false alarms when the input stream is stationary. To evaluate this, we executed a negative control on the CIFAR-10 stream with a drift severity of exactly 0.0, ensuring the incoming windows were identically distributed to the reference window.

Under the no-drift null hypothesis, all detectors collapsed to their respective baseline values. Crucially, the C2ST classifiers returned AUROC scores hovering exactly around chance (e.g., C2ST-LR scored 0.446). This confirms that despite the high capacity of the deep feature extractors, the strict PCA and CFE bottleneck effectively prevents overfitting, ensuring a calibrated false-positive rate during stable deployment.

4.4 Computational Cost Analysis

To address the practical feasibility of deploying the C2ST-CFE framework in real-time, continuous streaming environments, we conducted a rigorous computational cost analysis. Table 3 quantifies the average wall-clock execution time and theoretical memory footprint per sliding window ($N_{test} = 500$). Metrics were averaged over 1,680 discrete windows pooled across all dataset seeds. To isolate algorithmic efficiency from hardware acceleration, all pipelines including the frozen ResNet-50 backbone were executed entirely on a standard 8-core Intel CPU.

The analysis reveals that the total end-to-end processing time is overwhelmingly dominated by the initial ResNet-50 feature extraction ($\sim$93% of total pipeline execution). Crucially, the proposed PCA-

Table 3
Average computational cost per sliding window ($N = 500$). Execution times represent CPU wall-clock seconds. Memory reflects theoretical resident RAM footprints.

Pipeline Stage	Mean CPU Time (s)	Resident RAM (MB)
ResNet-50 Extraction (Frozen)	54.09 ± 4.06	381.6
PCA + CFE Bottleneck	< 0.01	12.0
C2ST Execution (RF/LR/MLP)	2.72 ± 0.75	46.5
Distance Baselines (MMD/KS)	1.44 ± 0.66	22.1

CFE bottleneck transforms the 2048-D representations in under 10 milliseconds. Executing the 5-fold C2ST classifier suite adds a mere $\sim$4.7% latency overhead compared to the statistical baselines, while requiring less than 50 MB of resident RAM. This confirms that the proposed discriminative framework achieves highly expressive drift detection without imposing prohibitive computational burdens on the deployment hardware.

4.5 Theoretical, Practical and Managerial Implications

The findings of this study have important theoretical, practical, and managerial implications for the design, deployment, and governance of machine learning systems operating in non-stationary visual environments. Taken together, the results show that concept drift detection in high-dimensional image streams should not be treated only as a problem of measuring geometric distance between feature distributions. Instead, effective monitoring requires methods capable of identifying whether two temporal windows have become statistically distinguishable in a semantically meaningful representation space.

4.5.1 Theoretical Implications

From a theoretical perspective, this work contributes to the literature on concept drift detection by exposing a critical limitation of traditional distance-based and marginal two-sample tests in deep visual representation spaces. Classical methods such as Maximum Mean Discrepancy, the Kolmogorov–Smirnov test, Wasserstein distance, and centroid-shift heuristics implicitly assume that drift manifests as a measurable change in global distributional geometry or in marginal feature coordinates. However, the empirical evidence reported in this study indicates that such assumptions are often too restrictive for modern visual data streams. In particular, under the Waterbirds semantic shift scenario, where the visual background changes from land to water while preserving much of the object-level structure, these methods fail to provide a sufficiently discriminative signal.

This result has broader theoretical relevance because it suggests that semantic drift in deep convolutional spaces may occur through localized manifold deformation, changes in class-conditional structure, or shifts in contextual associations rather than through large global displacements of the feature distribution. Consequently, detectors that rely only on marginal distances may remain insensitive even when the deployed model is exposed to a substantively different operational environment. The observed degradation of distance-based baselines therefore supports the view that high-dimensional visual drift should be analysed through discriminative separability rather than solely through aggregate divergence measures.

The proposed C2ST-CFE framework advances this perspective by reframing drift detection as a classifier-based two-sample testing problem. Instead of asking whether two windows are far apart according to a predefined metric, the framework asks whether a learning algorithm can distinguish samples from the reference and incoming windows. This distinction is theoretically important: if a classifier can reliably separate the two samples, then the reference and test distributions encode different information, even when classical distance statistics remain small. The strong performance of the logistic regression and multilayer perceptron C2ST variants further indicates that, after PCA and CFE transformation, semantic shifts become linearly or near-linearly separable in the induced distributional representation space.

The use of Characteristic Function Embeddings also has theoretical significance. CFE provides a compact non-parametric representation of distributions by encoding information through empirical characteristic functions. In this study, CFE is not used merely as an alternative distance representation; rather, it acts as a bottleneck that transforms high-dimensional deep features into a discriminative distributional space suitable for downstream classifier-based testing. This supports a broader methodological claim: distributional embeddings can serve as an intermediate representation that reconciles the statistical foundations of two-sample testing with the flexibility of supervised discrimination.

Finally, the results clarify the role of dimensionality reduction in visual drift detection. PCA fitted strictly on the reference window is not only a computational convenience but also a methodological safeguard. By retaining dominant reference variance while preventing information leakage from future test windows, the PCA step imposes a disciplined representation of the pre-drift environment. The subsequent CFE transformation then enriches this compressed representation with distributional information. This PCA-to-CFE sequence provides a theoretically grounded solution to the curse of dimensionality that affects direct C2ST application on raw 2048-dimensional convolutional features.

4.5.2 Practical Implications

From a practical standpoint, the proposed framework provides an operationally feasible solution for monitoring visual machine learning systems after deployment. A major challenge in deployed computer vision applications is that ground-truth labels are often delayed, expensive, incomplete, or entirely unavailable. This limitation makes classical performance-based drift detectors unsuitable for many real-world use cases, including medical imaging, industrial inspection, autonomous systems, remote sensing, video surveillance, and customer-facing image classification services. The label-free nature of the proposed approach directly addresses this limitation by enabling continuous monitoring based only on incoming images and a stable reference window.

The empirical results show that the C2ST-CFE framework can identify both severe domain shifts and more subtle semantic changes while maintaining calibration under no-drift conditions. This has direct implications for system reliability. In stable environments, a drift detector must avoid unnecessary alarms, since excessive false positives can lead to alert fatigue and wasted operational resources. At the same time, the detector must be sufficiently sensitive to identify changes before they translate into model failure. The no-drift control experiment indicates that the framework remains close to random-chance separability when the stream is stationary, while the synthetic perturbation experiments show rapid detection once drift is introduced. This balance between sensitivity and calibration is essential for practical deployment.

The computational analysis further indicates that the framework is suitable for continuous monitoring. Since most of the processing time is dominated by frozen ResNet-50 feature extraction, the additional cost of the PCA-CFE bottleneck and C2ST execution is comparatively modest. The PCA-CFE transformation operates in under milliseconds relative to the full feature extraction pipeline, and the classifier-based detection stage adds only a limited overhead compared with distance-based baselines. This means that the proposed framework can be integrated into existing visual inference pipelines without requiring major architectural redesign or prohibitive computational resources.

Another practical implication concerns modularity. The proposed pipeline separates feature extraction, compression, distributional embedding, and drift scoring into distinct stages. This modular structure allows practitioners to adapt the framework to different deployment contexts. For example, the frozen ResNet-50 backbone may be replaced by a domain-specific encoder in medical imaging or industrial inspection; the reference window size may be adjusted to reflect the stability and volume of the incoming stream; and the classifier used in the C2ST stage may be selected according to the desired trade-off between interpretability, speed, and non-linear discriminative capacity.

The framework also provides a continuous drift score rather than a single binary decision. This is practically valuable because monitoring teams often need to distinguish between weak, emerging, and severe drift. A continuous AUROC-based signal can support graded alerting, escalation rules, or dashboard-based monitoring. For instance, low-level increases in separability may trigger passive observation, while sustained high separability may trigger model audit, data sampling, or retraining workflows. This makes the framework useful not only as an alarm mechanism but also as a diagnostic component within a broader machine learning operations pipeline.

4.5.3 Managerial Implications

The managerial implications of this study are especially relevant for organizations deploying machine learning models in dynamic and safety-critical environments. In many operational settings, model degradation is not immediately visible because labels arrive late or because failures are detected only after downstream consequences occur. The proposed label-free monitoring framework reduces this governance gap by allowing managers to detect changes in the input environment before confirmed performance deterioration becomes available.

This capability can support more proactive model risk management. Instead of relying exclusively on periodic model audits or retrospective performance evaluation, organizations can implement continuous drift surveillance. Such surveillance enables earlier intervention, including targeted data review, model recalibration, retraining, domain adaptation, or temporary fallback to human decision-making. In sectors such as healthcare, manufacturing, transport, security, and finance, this can reduce the risk of undetected model failure and improve accountability in automated decision systems.

The framework also has implications for cost management. Continuous manual labelling is one of the most expensive components of post-deployment model monitoring. By detecting drift without requiring ground-truth labels, the proposed approach can reduce the need for routine large-scale annotation. Human expertise can instead be allocated more selectively, focusing on windows or periods where the drift score indicates meaningful distributional change. This targeted allocation of labelling and auditing resources can lower operational costs while preserving oversight quality.

For managers responsible for machine learning operations, the normalized C2ST output is also easier to operationalize than unbounded raw distance statistics. Metrics such as MMD, Wasserstein distance, or centroid displacement often require dataset-specific interpretation and threshold tuning. By contrast, the AUROC-based separability score has a more intuitive interpretation: values close to 0.5 indicate that the incoming window is difficult to distinguish from the reference distribution, while values approaching 1.0 indicate strong evidence of drift. This interpretability can facilitate communication between technical teams, managers, auditors, and domain experts.

The proposed framework further supports regulatory and organizational accountability. As machine learning systems become embedded in high-impact decision processes, organizations are increasingly expected to document how models are monitored after deployment. A label-free, statistically grounded drift detection pipeline can form part of a model governance framework by providing evidence of continuous surveillance, reproducible drift scoring, and documented triggers for intervention. This is particularly important in settings where model performance cannot be assessed immediately due to delayed outcomes or limited labels.

Finally, the results suggest that managers should not rely exclusively on simple distance-based monitoring dashboards for high-dimensional visual systems. While such metrics may be computationally attractive and easy to implement, this study shows that they can fail under precisely the kinds of semantic and subpopulation shifts that matter most in real deployments. Investment in discriminative, representation-aware monitoring tools is therefore justified not only as a technical improvement but also as a risk-reduction strategy. By integrating C2ST-CFE-style monitoring into the machine learning lifecycle, organizations can move from reactive model maintenance toward proactive and evidence-based model governance.

4.6 Methodological Limitations and Future Work

While the C2ST-CFE framework demonstrates state-of-the-art detection on visual streams, its underlying methodology contains specific structural vulnerabilities that outline necessary avenues for future research.

First, the framework's discriminative power is structurally bounded by the statistical sufficiency of the static reference window ($N = 1000$). If this initial sampling fails to adequately capture the inherent variance of the pre-drift deployment environment, or if the pre-drift stream is not strictly identically distributed, the PCA bottleneck will discard vital semantic eigen-dimensions, systematically inflating the risk of false positives during stable deployment.

Second, the methodology relies on a static variance retention threshold (95%) for the PCA bottleneck. This assumes that the intrinsic dimensionality of the semantic manifold remains relatively constant. In highly dynamic environments where new, orthogonal concepts are introduced gradually, a static linear projection may eventually mask emerging structural variations.

Finally, our evaluation protocol relies on discrete sliding test windows rather than true, point-by-point online detection. While this ensures sufficient statistical power for the classifier, it inherently introduces a detection delay proportional to the window stride. Adapting the C2ST framework to support dynamic reference window updating and continuous stochastic subspace tracking, without triggering catastrophic forgetting or absorbing anomalous data into the clean manifold, remains a critical mathematical challenge for future work.

4.7 Reproducibility

To facilitate future research and ensure the reproducibility of our results, the complete source code, processed datasets, and detailed experimental logs have been made publicly available. All experiments described in Section 4, including the exact random seeds and hyperparameters used, can be replicated using the provided scripts. [*]

5. Conclusion

This work introduced a unified, label-free framework for visual concept drift detection engineered to address the critical geometric blind spots of traditional statistical two-sample tests. By systematically benchmarking the proposed Classifier Two-Sample Test with Characteristic Function Embedding (C2ST-CFE) against established distance-based baselines, we demonstrated a fundamental vulnerability in existing metrics. Specifically, tests computing geometric distances between marginal distributions, such as Maximum Mean Discrepancy (MMD) and the Kolmogorov-Smirnov (KS) test, frequently collapse under complex semantic and subpopulation shifts. In contrast, the C2ST-CFE architecture suc-

[*] The official project repository can be accessed at: https://github.com/GurgenHovakimyan/image_drift_detection

cessfully isolates conditional semantic changes, maintaining robust discriminative power across both subtle synthetic perturbations and severe real-world domain shifts.

Ultimately, this study establishes that highly accurate, real-time drift monitoring can be achieved in high-dimensional visual streams entirely without ground-truth labels. By validating that a strict PCA-to-CFE compression bottleneck effectively resolves the dimensionality constraints of deep convolutional representations, this framework provides a computationally efficient, mathematically grounded mechanism for deploying safer, more accountable machine learning systems in dynamic real-world environments.

Acknowledgment

This work was supported by national science funds through FCT I.P. (Fundação para a Ciência e a Tecnologia, I.P.), under the projects UID/04152/2025 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS (https://doi.org/10.54499/UID/04152/2025), and UID/00315/2025 – Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL) (https://doi.org/10.54499/UID/00315/2025), and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 (https://doi.org/10.54499/UID/PRR/04152/2025). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, *31*(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

[2] Hovakimyan, G., & Bravo, J. M. (2024). Evolving strategies in machine learning: A systematic review of concept drift detection. *Information*, *15*(12), 786. https://doi.org/10.3390/info15120786

[3] Marathe, S., Nambi, A., Swaminathan, M., & Sutaria, R. (2021). Currentsense: A novel approach for fault and drift detection in environmental iot sensors. *Proceedings of the international conference on internet-of-things design and implementation*, 93–105. https://doi.org/10.1145/3450268.345353

[4] Goncalves, V. P., Silva, L. P., Nunes, F. L., Ferreira, J. E., & Araújo, L. V. (2024). Concept drift adaptation in video surveillance: A systematic review. *Multimedia Tools and Applications*, *83*(4), 9997–10037. https://doi.org/10.1007/s11042-023-15855-3

[5] Gama, J., Medas, P., Castillo, G., & Rodrigues, P. (2004). Learning with drift detection. *Brazilian symposium on artificial intelligence*, 286–295. https://doi.org/10.1007/978-3-540-28645-5_29

[6] Baena-Garcıa, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavalda, R., & Morales-Bueno, R. (2006). Early drift detection method. *Fourth international workshop on knowledge discovery from data streams*, *6*, 77–86. https://www.researchgate.net/publication/245999704_Early_Drift_Detection_Method

[7] Ren, S., Zhu, W., Liao, B., Li, Z., Wang, P., Li, K., Chen, M., & Li, Z. (2019). Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning. *Knowledge-Based Systems*, *163*, 705–722. https://doi.org/10.1016/j.knosys.2018.09.032

[8] Zheng, X., Li, P., Hu, X., & Yu, K. (2021). Semi-supervised classification on data streams with recurring concept drift and concept evolution. *Knowledge-Based Systems*, *215*, 106749. https://doi.org/10.1016/j.knosys.2021.106749

[9] Idrees, M. M., Minku, L. L., Stahl, F., & Badii, A. (2020). A heterogeneous online learning ensemble for non-stationary environments. *Knowledge-based systems*, *188*, 104983. https://doi.org/10.1016/j.knosys.2019.104983

[10] Krawczyk, B. (2017). Active and adaptive ensemble learning for online activity recognition from data streams. *Knowledge-Based Systems*, *138*, 69–78. https://doi.org/10.1016/j.knosys.2017.09.032

[11] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, *13*(1), 723–773. http://jmlr.org/papers/v13/gretton12a.html

[12] Greco, S., Vacchetti, B., Apiletti, D., & Cerquitelli, T. (2025). Unsupervised concept drift detection from deep learning representations in real-time. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2025.3593123

[13] Ross, G. J., Adams, N. M., Tasoulis, D. K., & Hand, D. J. (2012). Exponentially weighted moving average charts for detecting concept drift. *Pattern recognition letters*, *33*(2), 191–198. https://doi.org/10.1016/j.patrec.2011.08.019

[14] Tran, Q.-T., Le-Khac, N.-A., & Bertolotto, M. (2026). Concept drift detection in image data stream: A survey on current literature, limitations and future directions. *Artificial Intelligence Review*, *59*(1), 33. https://doi.org/10.1007/s10462-025-11428-y

[15] Angelov, P. P., Gu, X., & Principe, J. C. (2017). Autonomous learning multimodel systems from data streams. *IEEE Transactions on Fuzzy Systems*, *26*(4), 2213–2224. https://doi.org/10.1109/TFUZZ.2017.2769039

[16] Basci, P., Greco, S., Manigrasso, F., Cerquitelli, T., Morra, L., et al. (2025). Explaining concept drift via neuro-symbolic rules. In *Titolo volume non avvalorato*. CEUR. https://ceur-ws.org/Vol-4132/short54.pdf

[17] Lopez-Paz, D., & Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv: 1610.06545*.

[18] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 5637–5664.

[19] Yu, S., Wu, P., Liang, P. P., Salakhutdinov, R., & Morency, L.-P. (2022). Pacs: A dataset for physical audiovisual commonsense reasoning. *European Conference on Computer Vision*, 292–309. https://doi.org/10.1007/978-3-031-19836-6_17

[20] Francis, N., & Ali, A. (2024). Identifying and managing concept drift in machine learning through page-hinkley test: Approaches, obstacles, and resolutions. *International Conference on Recent Trends in Computing*, 33–46. https://doi.org/10.1007/978-981-97-8946-7_3

[21] Fellicious, C., Wendlinger, L., & Granitzer, M. (2022). Neural network based drift detection. *International Conference on Machine Learning, Optimization, and Data Science*, 370–383. https://doi.org/10.1007/978-3-031-25599-1_28

[22] Muandet, K., Fukumizu, K., Sriperumbudur, B., & SchAP lkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, *10*(1-2), 1–141. https://doi.org/10.1561/2200000060

[23] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., & Díaz-Rodríguez, N. (2020). Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, *58*, 52–68. https://doi.org/10.1016/j.inffus.2019.12.004

[24] Aljundi, R., Belilovsky, E., Tuytelaars, T., Charlin, L., Caccia, M., Lin, M., & Page-Caccia, L. (2019). Online continual learning with maximal interfered retrieval. *Advances in Neural Information Processing Systems*, *32*.

[25] Barddal, J. P., Gomes, H. M., & Enembreck, F. (2015). A survey on feature drift adaptation. *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, 1053–1060. https://doi.org/10.1109/ICTAI.2015.150

[26] Krizhevsky, A., Hinton, G., et al. (2009). *Learning multiple layers of features from tiny images* (tech. rep.). University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[27] Yang, J., Shi, R., & Ni, B. (2021). Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 191–195. https://doi.org/10.1109/ISBI48211.2021.9434062

[28] Waller, L. A., Turnbull, B. W., & Hardin, J. M. (1995). Obtaining distribution functions by numerical inversion of characteristic functions with applications. *The American Statistician*, *49*(4), 346–350. https://doi.org/10.2307/2684571