CODiT: Conformal Out-of-Distribution Detection in Time-Series Data for Cyber-Physical Systems

Ramneet Kaur
ramneetk@seas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Kaustubh Sridhar
ksridhar@seas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Sangon Park
sangdon@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Yahan Yang
yangy96@seas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Susmit Jha
susmit.jha@sri.com
SRI International
Menlo Park, California, USA

Anirban Roy
anirban.roy@sri.com
SRI International
Menlo Park, California, USA

Oleg Sokolsky
sokolsky@seas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

Insup Lee
lee@seas.upenn.edu
University of Pennsylvania
Philadelphia, Pennsylvania, USA

ABSTRACT

Uncertainty in the predictions of learning enabled components hinders their deployment in safety-critical cyber-physical systems (CPS). A shift from the training distribution of a learning enabled component (LEC) is one source of uncertainty in the LEC’s predictions. Detection of this shift or out-of-distribution (OOD) detection on individual datapoints has therefore gained attention recently. But in many applications, inputs to CPS form a temporal sequence. Existing techniques for OOD detection in time-series data for CPS either do not exploit temporal relationships in the sequence or do not provide any guarantees on detection. We propose using deviation from the in-distribution temporal equivariance as the non-conformity measure in conformal anomaly detection framework for OOD detection in time-series data for CPS. Computing independent predictions from multiple conformal detectors based on the proposed measure and combining these predictions by Fisher’s method leads to the proposed detector CODiT with bounded false alarms. We illustrate the efficacy of CODiT by achieving state-of-the-art results in autonomous driving systems with perception (or vision) LEC. We also perform experiments on medical CPS for GAIT analysis where physiological (non-vision) data is collected with force-sensitive resistors attached to the subject’s body. Code, data, and trained models are available at https://github.com/kaustubhsridhar/time-series-OOD

1 INTRODUCTION

With remarkable performance-levels of machine learning across different domains [13, 24], there is much interest in using these learning-enabled components (LEC) in cyber-physical systems (CPS). A prime example of this trend is autonomous driving, where a car would be equipped with the capability to autonomously perform certain operations, such as adaptive cruise control, emergency braking or lane following [4]. These autonomous features aim to enhance the safety of the vehicle and its occupants. But as drivers rely on them more and more, a malfunction in such a feature would be detrimental to safety. The black-box nature of LEC makes it difficult to interpret their uncertain predictions on perturbed inputs [18] or even inputs from novel environments from training [36]. For instance, Yang et al. (2022) show that an ego vehicle that uses a perception LEC to estimate following distance to a car in front of it crashes when the car is replaced by a bike. This is because the LEC was trained with cars in front of the ego vehicle and could not correctly estimate distance from bikes (novel object from training). Deployment of LEC in safety-critical CPS such as autonomous driving [4], and healthcare [8], therefore, requires detection of shift in the distribution of the LEC’s inputs from its training distribution.

Detection of shift from the training distribution of LEC or out-of-distribution (OOD) detection on individual datapoints has received broad attention [16, 20, 31]. But this problem of OOD detection in time-series data for CPS is less explored. In time-series data,
Table 1: Capabilities of detectors in time-series data for CPS.

| OOD Detector       | False Alarm Rate | Temporal OODs | Non-vision Data |
|--------------------|------------------|---------------|-----------------|
| VAE [5]            | ✓                | ?             | ✓               |
| β-VAE [27]         | ✓                | ?             | ✓               |
| Memory [36]        | ×                | ?             | ✓               |
| Feng et al.’s [10] | ×                | ✓             | ×               |
| CODiT (Ours)       | ✓                | ✓             | ✓               |

OOD detection aims at detecting those windows of time-series datapoints that are outside the training distribution of LEC. In this case, the distribution of interest is not just of individual datapoints, but the distribution of sequences in which these datapoints occur in the time-series data. In this paper, we propose CODiT, a novel algorithm for OOD detection in time-series data for CPS with a bounded false alarm rate.

Most of the existing approaches for OOD detection in time-series data for CPS are point-based, i.e., they independently consider each datapoint in the window, such as individual frames in a video clip. In other words, these approaches do not exploit time-dependency among the datapoints in a window to detect if the window is OOD. An example of an OOD window in a driving scenario is a car drifting video clip due to slippery ground or loss of control, given normal driving video clips as in-distribution (iD) data. As shown in Fig. 1, we cannot tell from a single frame if the car has drifted since drift is defined based on the relative relation between frames (e.g., a trajectory of a car). We define these OOD windows as temporal OODs, which require considering the sequence of datapoints in the window for detection. In contrast to the existing point-based approaches for detection in time-series data for CPS, we propose using time-dependency among the datapoints in a window for detection of temporal OODs.

Specifically, we propose using deviation from the iD temporal equivariance, i.e. equivariance with respect to a set of temporal transformations learned by a model on windows drawn from the training distribution of LEC, for OOD detection in time-series data. This is because a model trained to learn equivariance with respect to temporal transformations on data drawn from the training distribution might not generalize or exhibit equivariance on OOD inputs dissimilar to the training distribution.

To bound false alarms on the iD data, we leverage inductive conformal anomaly detection (ICAD) [1]. ICAD is a general framework for testing if an input conforms to the training distribution by computing quantitative scores defined by a non-conformity measure. A p-value, computed by comparing non-conformity score of the input with these scores on the data drawn from training distribution, indicates anomalous behavior of the input. The detection performance of ICAD, however, depends on the choice of the non-conformity measure used in the framework [1]. We propose using deviation from the expected temporal equivariant behavior of a model learned on data drawn from training distribution of LEC as the non-conformity measure in ICAD for OOD detection in the time-series data for CPS.

ICAD computes a single p-value of the input to detect its anomalous behavior. To enhance detection performance, we propose using multiple ($n > 1$) p-values computed from $n$ transformations sampled as independent and identically distributed (IID) variables from a distribution over the set of temporal transformations. The intuition for using multiple transformations is that an OOD window might behave as a transformed iD window with one transformation but the likelihood of this decreases with the number of transformation combinations. Using Fisher’s method [33] to combine these $n$ independent $p$-values leads to the proposed detector CODiT for conformal OOD detection in time-series data for CPS with a bounded false alarm rate. The contributions can be summarized as:

1. **Novel Measure for OOD Detection in Time-Series Data for CPS.** To our knowledge, all the existing non-conformity measures for OOD detection are defined on individual datapoints. We propose a measure that is defined on the window containing information about the sequence of time-series datapoints for enhancing detection of the temporal OODs. With a model trained to learn iD temporal equivariance via the auxiliary task of predicting an applied transformation on windows drawn from training distribution of LEC, we propose using error in this prediction as the non-conformity measure in ICAD for detection in time-series data for CPS.

2. **Enhanced detection performance.** To enhance the detection performance, we propose to use Fisher’s method as an ensemble approach for combining predictions from multiple conformal detectors based on the proposed measure.

3. **CODiT.** Computing $n$ independent $p$-values of the input from the proposed measure in the ICAD framework, and combining these values by Fisher’s method leads to the proposed detector CODiT with a bounded false alarm rate for OOD detection in time-series data for CPS.

4. **Evaluation.** For comparison with the point-based detectors, we perform experiments on weather and night OODs in an autonomous driving system with perception LEC, achieving state-of-the-art (SOTA) results. We outperform the existing non-point based SOTA [10] on temporal OODs in driving scenario. To illustrate that CODiT can be used for OOD detection beyond vision, we also perform experiments on medical CPS for GAIT analysis where physiological (non-vision) data is collected with force-sensitive resistors attached to the subject’s body [15].

## 2 PROBLEM STATEMENT AND MOTIVATION

### 2.1 Problem Statement

OOD detection in time-series data for CPS takes in a window $X_t, w$ of consecutive time-series datapoints ($x_t, x_{t+1}, \ldots, x_{t+(w-1)}$) to the LEC in the system, and labels $X_t, w$ as iD or OOD. Here $t$ is the starting time of the window and $w$ is the window length.

### 2.2 Motivation of the Proposed OOD Detection Measure

The existing point-based detectors for OOD detection in time-series data might not be able to detect temporal OODs to the system. An example of a temporal OOD in driving scenario, as shown in Fig. 2, is the replay window where camera gets stuck at a single frame and starts generating the same image over and over again. We need to consider the sequence of same image in a replay window to detect the window as OOD. Detection results on replay dataset in Fig. 3
shows that all the existing point-based detectors, namely variational autoencoder (VAE) based Cai et al.’s [5], β-VAE-based Ramakrishna et al.’s [27], and memory-based Yang et al.’s [36] detectors perform poorly in the detection of these temporal OODs. We, therefore, propose using time-dependency among the datapoints in a window for OOD detection in time-series data for CPS.

To our knowledge, Feng et al.’s detector [10] is the only existing OOD detector that takes into account time-dependency among the individual datapoints in a window. It does so by extracting optical flow information from consecutive frames in a video clip. As shown in Fig. 8 of Section 6, Feng et al.’s detector can thus be used to detect temporal OODs. However, since this detector depends on the optical flow information, it is restricted to vision data. In contrast, CODiT can be used to detect temporal OODs across domains (vision or non-vision) without relying on any domain-specific features. Table 1 compares the detection capabilities of CODiT with the existing OOD detectors in time-series data.

3 RELATED WORK

OOD detection in non time-series datasets such as German Traffic Sign Recognition (GTSRB) [30] has been extensively studied. Details about these detectors on non time-series (or individual) datapoints are included in Appendix. OOD detection in CPS with low-dimensional input space through envelopes has also been studied in the past [32].

Recently, there has been interest in leveraging ICAD for OOD detection with guarantees on false alarm rate on high dimensional input space [5, 14, 19, 27]. While iDECODe [19], and Haroush et al.’s [14] are OOD detectors for individual datapoints, Cai et al.’s [5], and Ramakrishna et al.’s [27] are detectors for time-series data. iDECODe uses error in the equivariance learned by a model with respect to a transformation set on individual datapoints as the non-conformity measure (NCM) in ICAD for detection in non time-series data. Haroush et al. propose using a combined p-value from different channels and layers of convolutional neural networks (CNN) for detection. It is not clear how to directly apply individual point detectors to time-series data with the ICAD guarantees due to the following two reasons. First, even if we apply these detectors to individual datapoints in the time-series window independently, we do not know how to combine detection verdicts on these datapoints for detection on the window. Second, for detection guarantees by ICAD, it is required that all non-conformity scores for p-value computation to be IID [22]. Since these detectors are not solving OOD detection problem in time-series data it is not clear how to apply them to time-series while preserving the IID assumption on the time-series data. Also, iDECODe uses a single p-value for detection and we propose using multiple (n > 1) independent p-values to be combined by the Fisher’s method for preserving the detection guarantees. In contrast to [14], our approach is not limited to CNN and can be used for other predictive models as well.

Cai et al. [5] propose using reconstruction error by VAE on an input image as the NCM in the ICAD framework. Martingale formula [35] is used to combine multiple p-values computed on multiple samples of the input in the latent space of VAE. The detection score on a time-series window is then computed by applying cumulative sum procedure [2] on the martingale values of all images in the window. Ramakrishna et al. [27] propose using KL-divergence between the disentangled feature space of β-VAE on an input image and the normal distribution, as the NCM in ICAD. They also use the martingale formula to combine p-values of all the images in the window for detection. Recently, Yang et al. [36] propose computing few prototypes (or memories) from the training data and using distance of an input image with these prototypes for detection on the input. The time-series window to the CPS is labeled as OOD if majority datapoints in the window are detected as OOD. Unlike the other two detectors [5, 27], this approach does not provide false alarm rate guarantees on detection. All of these three OOD detectors [5, 27, 36] on time-series data are point-based, and as shown by the experiments on replay OODs, these might perform poorly in the detection of temporal OODs. CODiT computes the p-value of the window (and not individual datapoints in the window) in the ICAD framework. It is, therefore, a non-point based approach that can be used to detect temporal OODs (Section 6.2, 6.3).
To our knowledge, the only non-point based approach for OOD detection in time-series data is by Feng et al. [10]. They propose extracting optical flow information from a time-series window and training a VAE on this information. KL-divergence between the trained VAE and a specified prior is used as the OOD score. This detector uses optical flow to extract time-dependency in the frames of a window and thus can be used to detect temporal OODs. However, this approach does not provide any guarantees on detection and will not work on non-vision datasets as it relies on optical flows. As shown in the experiments on the GAIT dataset in medical CPS, CODiT can be used for OOD detection in non-vision domain.

4 BACKGROUND AND NOTATIONS
CODiT uses error in the temporal equivariance learned by a model on windows drawn from the training distribution of LEC as the non-conformity measure in the inductive conformal anomaly detection (ICAD) framework. With multiple p-values obtained from the proposed measure in ICAD, the final OOD detection score is computed by combining these values by Fisher’s method. Here we provide the background on equivariance, ICAD, and Fisher’s method required for technical details of the proposed OOD detector, CODiT. We also define the notations used in the rest of the paper.

4.1 Equivariance
A function \( f \) is equivariant with respect to a transformation \( g \) if we know how the output of \( f \) changes if we transform its input from \( x \) to \( g(x) \). Invariance is a special case of equivariance where the output of \( f \) does not change by transformation \( g \) on its input. Invariance with respect to geometric transformations such as rotation, tilt, scale, etc. is a desired property of the machine learning classifiers. For example, classification results on the straight images of planes should not change with a tilt in these images.

Definition 1 (Schmidt and Roth, 2012). For a set \( X \), a function \( f \) is defined to be equivariant with respect to a set of transformations \( G \), if there exists the following relationship between any transformation \( g \in G \) of the function’s input and the corresponding transformation \( g' \) of the function’s output:

\[
    f(g(x)) = g'(f(x)), \forall x \in X.
\] (1)

Invariance is a special case of equivariance where \( g' \) is the identity function, i.e., the output of \( f \) remains unchanged by the transformation \( g \) on its input.

4.1.1 Learning Equivariance: Autoencoding Variational Transformations (AVT). Augmenting training data with transformations from the set \( G \) of geometric transformations is a common approach to learning invariance with respect to \( G \) [6, 7]. The auxiliary task of predicting an applied transformation from \( G \) on the training data also encourages the model to learn equivariance with respect to \( G \) [26]. Qi et al.‘s 2019 “Autoencoding Variational Transformations” (AVT) framework trains a VAE to learn a latent space that is equivariant to transformations. For the set \( X \) of training images and the set \( G \) of geometric transformations, a VAE is trained to predict the applied transformation from \( G \) on an input \( x \in X \). Equivariance between the latent space of VAE and \( G \) is learned by maximizing mutual information between the latent space and \( G \).

4.2 Inductive Conformal Anomaly detection
Inductive Conformal Anomaly Detection (ICAD) [22] is a general framework for testing if an input conforms to the training distribution. It is based on a non-conformity measure (NCM), which is a real-valued function that assigns a non-conformity score \( a \) to the input. This score indicates non-conformance of the input with data drawn from the training distribution. The higher the score is, the more non-conforming or anomalous the input is with respect to training data. An example of the non-conformity score is the reconstruction error by a VAE trained on data drawn from the training distribution.

The training dataset \( X \) of size \( l \) is split into a proper training set \( X_{tr} = \{x_j : j = 1, \ldots, m\} \) and a calibration set \( X_{cal} = \{x_j : j = m+1, \ldots, l\} \). Proper training set \( X_{tr} \) is used in defining NCM. In the example of reconstruction error by a VAE as the non-conformity score, the VAE trained on \( X_{tr} \) is used for computing the error. Calibration set \( X_{cal} \) is a held-out training set that is used for computing p-value of an input. p-value of an input \( x \) is computed by comparing its non-conformity score \( a(x) \) with these scores on the calibration datapoints:

\[
p\text{-value}(x) = \left[ \frac{| \{ j = m+1, \ldots, l : \alpha(x) \leq a(x_j) \} | + 1}{l - m + 1} \right].
\] (2)

If \( x \) is drawn from the training distribution, then its non-conformity score is expected to lie within the range of scores for the calibration datapoints and thus higher p-values for the iD datapoints. With \( \epsilon \in (0, 1) \) as the anomaly detection threshold, \( x \) is therefore detected as an anomalous input if the p-value of \( x \) is less than \( \epsilon \).

4.2.1 False Alarm Rate Guarantees. The false anomalous detection on an input drawn from the training distribution is upper bounded by the specified detection threshold \( \epsilon \) in ICAD.

Lemma 1 (Balasubramanian et al., 2014). If an input \( x \) and the calibration datapoints \( x_{m+1}, \ldots, x_l \) are independent and identically distributed (IID), then for any choice of the NCM defined on the proper training set \( X_{tr} \), the p-value \( x \) in (2) is uniformly distributed. Moreover, we have \( \Pr \left( p\text{-value}(x) < \epsilon \right) \leq \epsilon \), where the probability is taken over \( x_{m+1}, \ldots, x_l, \) and \( x \).

From Lemma 1, we know that if \( x \) and the datapoints in the calibration set \( X_{cal} \) are IID, then the p-value \( x \) from (2) is uniformly
distributed over \( \{1/(l−m+1), 2/(l−m+1), \ldots, 1\} \). The probability of \( p \)-value(x) less than \( \epsilon \) or misclassifying x as anomalous is, therefore, \( \sum_{\epsilon \leq (l−m+1)x} 1/(l−m + 1) = [(l−m+1)\epsilon]/(l−m + 1) \leq \epsilon \).

4.3 Fisher’s Method
The same hypothesis can be tested by multiple conformal predictors and an ensemble approach for combining these predictions can be used to improve upon the performance of individual predictors. Fisher’s method is one of these approaches for combining multiple conformal predictions or \( p \)-values of an input from (2). Fisher value of an input \( x \) from \( n \) \( p \)-values is computed as follows:

\[
\text{fisher-value}(x) = r \sum_{i=0}^{n-1} \frac{(-\log r)^i}{i!}, \quad \text{where} \quad r = \prod_{k=1}^{n} p_k. \tag{3}
\]

**Lemma 2** (Toccacelli and Gammerman, 2017). If \( n \) \( p \)-values, \( p_1, \ldots, p_n \), are independently drawn from a uniform distribution of these values, then \( -2 \sum_{i=1}^{n} \log p_i \) follows a chi-square distribution with \( 2n \) degrees of freedom. Thus, the combined \( p \)-value is

\[
\Pr \left( y \leq -2 \sum_{i=1}^{n} \log p_i \right) = r \sum_{i=0}^{n} \frac{(-\log r)^i}{i!},
\]

where \( r = \prod_{k=1}^{n} p_k \), \( y \) is a random variable following a chi-square distribution with \( 2n \) degrees of freedom, and the probability is taken over \( y \). Moreover, the combined \( p \)-value follows the uniform distribution.

5 TEMPORAL EQUIVALENCE FOR CONFORMAL OOD DETECTION IN TIME-SERIES DATA FOR CPS

Here, we first classify OOD windows in time-series data into two types, and then provide details of the proposed detector CODiT.

5.1 OOD Data Types in Time-Series

We classify OOD windows in time-series data into two types: temporal OODs and non-temporal OODs.

A crucial property of the temporal OODs compared to the non-temporal OODs is that it is hard to detect temporal OODs by looking at individual datapoints within the window without considering time-dependency between these datapoints. Examples of temporal OODs in driving scenario are car drifting video clips (Fig. 1), and replay OODs (Fig. 2). An example of the temporal OOD in medical CPS is the GAIT (or waking pattern) of patients with neurodegenerative diseases. With the GAIT of healthy individuals as iD data, the GAIT of patients with neurodegenerative diseases, such as Parkinson’s disease (PD), Huntington’s disease (HD), and Amyotrophic Lateral Sclerosis (ALS), are examples of temporal OODs. Fig. 9 (in Appendix) shows dynamics of the stride time (one of the walking pattern features) of a healthy control person and patients with PD, HD, and ALS disease. As shown in the Figure, we need a sequence of time-series datapoints to determine whether the walking pattern is from a healthy individual or a patient. In contrast to the temporal OODs, the non-temporal OODs can be detected by looking at individual datapoints. Examples of non-temporal OODs include driving video clips under rainy, foggy, or snowy weather, given the driving video clips under clear sunny weather as iD data. We can detect weather OODs by looking at images in the window independently.

Based on these observations, we call a window \( X_{t,w} \) as a temporal OOD if \( X_{t,w} \) is drawn from OOD but confused to be drawn from iD by removing the time-dependency of individual datapoints within \( X_{t,w} \) (e.g., randomly shuffling the order of video clip frames). As shown in the experimental section 6, CODiT can be used to detect both temporal and non-temporal OODs in time-series data for CPS.

5.2 CODiT

CODiT uses an OOD detection score based on multiple \( p \)-values from ICAD. Here, we first define the proposed NCM to be used in the ICAD framework for computing a \( p \)-value along with the final detection score, and then formalize CODiT’s algorithm with a bounded false alarm rate.

5.2.1 Proposed NCM and OOD Detection Score.

**TTPN CMC.** We propose to use time-dependency between datapoints in a time-series window for detection on the window. Unlike all the existing NCMs defined on individual datapoints, we propose an NCM that is defined on the window containing information about the sequence of datapoints in the window. Specifically, we propose using deviation from the expected iD \( G_T \)-equivariance learned by a model on windows drawn from the training distribution of LEC as an NCM in ICAD for OOD detection in time-series data for CPS. Learning \( G_T \)-equivariance via an auxiliary task of predicting the applied temporal transformation (such as shuffle, reverse, etc.) on a window requires learning changes in the original sequence of the datapoints in a predictable way. For a VAE model \( M \) trained to learn \( G_T \)-equivariance on windows of proper training data in the AVT framework, we propose to use error in the prediction of the applied temporal transformation \( g \in G_T \) on an input window \( X_{t,w} \) as the NCM:

\[
\text{PredictionError}(g, M(g(X_{t,w}))).
\]

We call the proposed NCM as the *Temporal Transformation Prediction Error (TTPN) CMC.*

The existing AVT framework [26] is defined to learn equivariance with respect to geometric transformations on images. We extend it to learn \( G_T \)-equivariance by:

1. Modifying VAE’s architecture to accept windows of consecutive time-series datapoints as inputs. The time-series can be on vision (e.g. drift car video clip) or non-vision (e.g., GAIT) datapoints.
2. Modifying the auxiliary task to predict the applied temporal transformation from a set \( G_T \) on windows of time-series datapoints.

**Motivation for TTPN-CMC.** \( G_T \)-equivariance learned by a model on windows drawn from the training distribution of LEC is more likely to work on iD data and is not guaranteed to generalize to OOD data dissimilar to that used for training. With the set \( G_T = \{2x \text{ speed, shuffle, reverse, periodic, identity}\} \), we train a VAE model on the proper training data of the drift dataset to predict an applied transformation \( g \) sampled independently from a uniform distribution over \( G_T \). With \( G_T \) as the set of five classes of temporal transformations, we use CrossEntropyLoss\((g, M(g(X_{t,w})))) as the TTPN-CMC. Fig. 4 shows that the model has much higher prediction errors on the OOD windows than on the test iD windows on all the five ground truth transformations in \( G_T \). This supports our
hypothesis that $G_T$-equivariance learned on data drawn the training distribution is not generalize to data drawn from OOD, and therefore higher prediction errors on OOD windows than on the iD windows.

**OOD Detection Score.** Instead of using a single $p$-value from the TTPE-NCM in ICAD, we propose using multiple ($n > 1$) $p$-values to enhance detection. We require $n$ non-conformity scores for both the input and the calibration datapoints for computing $n$ $p$-values. These scores are computed from $n$ transformations sampled independently from a distribution $Q_{G_T}$ over $G_T$ for both the input and the calibration datapoints:

$$\alpha_i(X_{t,w}) = \text{PredictionError}(g_i, M(g_i(X_{t,w}))) : 1 \leq i \leq n, g_i \sim Q_{G_T},$$

where $X_{t,w}$ is the input or a calibration datapoint. Using Fisher’s method to combine these $n$ $p$-values gives us the fisher-value of input from equation (3). This value is expected to be higher for iD datapoints than OOD datapoints [11], and therefore we perform detection by using a threshold on the fisher-value of input. In other words, CODiT uses fisher-value of the input as the final OOD detection score.

**Motivation for multiple $p$-values for OOD Detection.** A single $p$-value measures deviation from the iD $G_T$-equivariance of the input with respect to one transformation $g \sim Q_{G_T}$. With multiple $p$-values, we test this deviation with respect to multiple transformations sampled independently from $Q_{G_T}$. We hypothesize that under one transformation, an OOD window might behave as the transformed iD window but the likelihood of this decreases with the number of transformations. For testing this hypothesis, we train three VAE models with the set $G_T$ equal to $\{2x \text{ speed, reverse, identity}, 2x \text{ speed, shuffle, periodic, identity}, 2x \text{ speed, reverse, shuffle, periodic, identity}\}$, respectively. These models are trained on the proper training set of the drift dataset to predict an applied transformation $g$ sampled independently from a uniform distribution over $G_T$. Again, we use CrossEntropyLoss($g, M(g(X_{t,w}))$) as the TTPE-NCM. Fig. 5 shows that the detection performance (in terms of temporal transformations, prediction error function $f$; $n$ sets of calibration set alphas $\{\alpha_k : 1 \leq k \leq n, m + 1 \leq j \leq l\}$, and desired false alarm rate $\epsilon \in (0, 1)$).

**Algorithm 1 CODiT: OOD Detection in Time-Series Data for CPS**

```plaintext
1: Input: a window $X_{t,w}$ of time-series data, VAE model $M$ trained on proper training set of the iD windows for LEC, distribution $Q_{G_T}$ over the set $G_T$ of temporal transformations, prediction error function $f$; $n$ sets of calibration set alphas $\{\alpha_k^j : 1 \leq k \leq n, m + 1 \leq j \leq l\}$, and desired false alarm rate $\epsilon \in (0, 1)$.
2: Output: "1" if $X_{t,w}$ is detected as OOD; "0" otherwise.
3: for $k \leftarrow 1, \ldots, n$ do
4:    $g \sim Q_{G_T}$
5:    $\hat{g} \leftarrow M(g(X_{t,w}))$
6:    $\alpha \leftarrow f(g, \hat{g})$
7:    $\alpha_{k}^{i,j} \leftarrow 1 - \exp$$^{-1}$($\frac{\alpha_{k}^{i,j}}{\epsilon}$)
8: end for
9: $r \leftarrow \prod_{k=1}^{n} p_k$
10: if $r \sum_{i=0}^{n-1} \left(\frac{\log r}{2^i}\right) < \epsilon$ then return 1 else return 0
```

**Theorem 1.** The probability of false OOD detection on $X_{t,w}$ by Algorithm 1 is upper bounded by $\epsilon$.

The proof of Theorem 1 is included in Appendix. The unconditional probability that an input $X_{t,w}$ sampled from the training distribution $D$ is classified as OOD by Algorithm 1 is bounded by $\epsilon$. For this guarantee to hold for a sequence of input windows, we require an independent calibration set for every input in the sequence. This is computationally inefficient for real-time applications and therefore a fixed calibration set is used for all the inputs in the offline version of the ICAD algorithm [22]. The average false alarm rate on the sequence of inputs drawn from $D$ in this setting is expected to be empirically calibrated with or even higher than $\epsilon$. We also fix the $n$ sets of iID calibration sets and pass it as an input to the Algorithm 1 on sliding windows of OOD traces from the drift dataset. Box plots in Fig. 5 (right) on the drift dataset shows that the false alarm rate of CODiT is empirically bounded by $\epsilon$ on average. Details about these plots are in Appendix.

**6 EXPERIMENTAL RESULTS**

We perform the following experiments:

1. **Comparison with the point-based approaches.** With images taken in clear daytime weather as iD for perception LEC in an autonomous car, existing approaches report their results on weather OODs. So, we perform experiments on weather and night OODs for the closed-loop of advanced
Figure 4: Higher values of $\text{TTPE-NCM} = \text{CrossEntropyLoss}(g, M(g(X_{t,w})))$ on OOD windows than on the test iD windows of the drift dataset. This shows that $G_T$-equivariance learned on the windows drawn from the training distribution of LEC is less likely to generalize on the windows drawn from OOD.

Figure 5: AUROC vs. $n$ (left), TNR (with detection threshold at 95% TPR) vs. $n$ (center) shows that the performance of CODiT increases with the increase in the number $n$ of $p$-values used in the fisher-value for detection. False Alarm Rate (FDR) of CODiT is empirically bounded by $\epsilon$ on average (right). The yellow line in the box plot indicates the median.

Figure 6: Closed-loop of Advanced Emergency Braking System (AEBS) with perception LEC from [36].

Figure 7: Medical CPS with the LEC to classify walking pattern or GAIT of a young or an elderly person.

emergency braking system (AEBS) with perception LEC from [36]. This system is shown in Fig. 6. Here, we compare CODiT’s performance with all the existing detectors.

(2) Temporal OODs for perception LEC. We perform experiments on replay OODs for the AEBS with perception LEC in Fig. 6. We also perform experiments on videos from drift dataset [25] as temporal OODs in driving scenario. Here, we compare CODiT’s performance with the existing state-of-the-art (SOTA) non-point approach by Feng et al. [10].

(3) Temporal OODs in Medical CPS. We consider temporal OODs for the walking pattern or GAIT classification problem in medical CPS from [3]. As shown in Fig. 7, a LEC is trained on sensory readings collected with force-sensitive resistors attached to the subject’s body for classifying the GAIT from a young or an elderly person. With the training data collected from healthy participants (without any injuries or diseases), input from patients with neurogenerative diseases are OODs for this LEC. We show that CODiT can be used to detect these temporal OODs in non-vision domain.

We also perform parameter analysis for CODiT on the drift dataset.

6.1 Weather OODs in AEBS with Perception LEC System Description. The closed-loop of advanced emergency braking system (AEBS) with perception LEC is shown in Fig. 6. The perception LEC is used to estimate distance to the front obstacle from the ego vehicle. This distance is fed to the emergency braking controller that issues braking commands to the vehicle for stopping at a safe distance from the front obstacle. The LEC is trained on images captured by the front camera on the vehicle in a clear daytime weather. This closed-loop system is simulated in CARLA [9], an open-source simulator for autonomous driving research. For more details about this system, please refer to [36].

In-Distribution Data. We generate 33 driving traces of varying lengths in clear day weather as the iD training traces for the AEBS
in CARLA. We randomly split these into 20 traces of the proper training set \( X_{tr} \) and 13 traces of the calibration set \( X_{cal} \). Windows from \( X_{tr} \) are sampled for training the perception LEC. Windows from \( X_{cal} \) are sampled \( n = 20 \) times (with one window from each calibration trace at a time to make each window in the set independent) for calculating the 20 sets of calibration non-conformity scores. We generate another set of 27 traces of varying lengths in clear weather as the ID test traces.

**OOD Data.** Weather (rainy, foggy, and snowy) and night time OOD traces are generated by using the automold software [28] on the 27 iD test traces. OOD traces start from iD and gradually become OOD, i.e., the intensity of rain, fog, snow, or low brightness (for night) starts increasing gradually turning into the OOD traces. Examples of windows from these OOD traces are shown in Appendix.

**Implementation Details of Algorithm 1.** We train a VAE model \( M \) with the R3D network architecture [34] on the windows of length \( w = 16 \) from \( X_{tr} \). R3D network is the 3D CNN with residual connections and thus can be used on the 3D time-series input data. We use \( G_T = \{2x\text{ Speed, Shuffle, Periodic, Reverse, Identity}\} \) and train \( M \) to predict the applied transformation \( g \in G_T \) with cross-entropy loss between the applied and the predicted transformations. The value of the CrossEntropyLoss \((g, M(g(X_{tr}, w)))\) is used as the non-conformity score \( a \) for computing \( p \)-value of an input \( X_{tr}, w \).

**Results.** We report results on the sliding windows \( (w = 16) \) of the test iD and OOD traces. We call iD as positive and OOD as negative. We report area under receiving operator curve (AUROC), and the detection delay (with detection threshold \( \epsilon \) at 95% True Positive Rate) in Table 2. Details about these evaluation metrics are in Appendix. ‘NA’ in the table means that the approach could not detect any OOD window with detection threshold at 95% TPR.

For CODiT, we report mean and variance from five runs with random sampling of calibration windows from the calibration traces. For AUROC, CODiT outperforms other approaches, except for Snowy OODs, where our results are comparable with the best results by the memory detector. For detection delay, while we outperform other approaches on Foggy and Night OODs, our results on Rainy and Snowy OODs are comparable with the best results.

### 6.2 Temporal OODs in Driving Scenario

We compare CODiT’s performance with the current non-point based SOTA OOD detector, i.e., Feng et al. [10]’s on the replay and drift datasets as temporal OODs for perception LEC:

**Replay Dataset.** Replay traces are generated from the 27 iD test traces in CARLA for AEBS by randomly sampling a position in each trace. All images from the sampled position in the trace are replaced with the image at the sampled position in the original trace. Again, results are reported for \( n = 20 \), and on the sliding windows of replay OODs.

**Drift Dataset.** We split 72 iD video traces of cars driving straight without any drift from the drift dataset [25] into 24 for \( X_{tr} \), 14 for \( X_{cal} \), and 34 for test iD traces. Windows from \( X_{tr} \) are sampled for training the VAE. Windows from \( X_{cal} \) are sampled \( n = 20 \) times (with one window from each calibration trace at a time to make each window in the set independent) for calculating the 20 sets of calibration non-conformity scores. We report results on the sliding windows of 34 iD test and 100 OOD drift traces.

**Results.** We use the same VAE model’s architecture, \( G_T \), and non-conformity score as described in Section 6.1. Fig. 8 (left) compares the ROC, AUROC and TNR (@95% TPR) results of CODiT with Feng et al.’s detector on the replay and drift OODs. We achieve SOTA results on both temporal OODs for perception LEC.

**Parameter Analysis on drift.** We perform the following parameter analysis on the drift dataset. All VAE models used in these studies are trained on the proper training set of the drift dataset with the same model architecture, \( G_T \), and non-conformity score from Section 6.1.

1. **Performance of CODiT with Different Window Lengths**: We train two VAE models with \( w = 18 \), 20 and compare the performance of CODiT \( (n = 20) \) with Feng et al.’s detector on these window lengths. Fig 8 (right) shows that with both \( w = 18 \) and 20, CODiT performs consistently well.

2. **Performance of CODiT with Different \( |G_T| \)**: We compute the performance of CODiT with different sizes of the transformation set. Table 3 shows that the performance of CODiT \( (n = 5) \) increases with \( |G_T| \).

3. **Using Deviation from iD \( G_T \)-equivariance as NCM**: With \( G_T = \{2x\text{ speed, shuffle, reverse, periodic, identity}\} \), and for all ground truth temporal transformations in \( G_T \), Fig. 4 shows that the non-conformity score from the TTPE-NCM is much higher for OOD windows than the test iD windows. This justifies our hypothesis that \( G_T \)-equivariance learned by a model on windows drawn from training distribution is not likely to generalize on windows drawn from OOD.

4. **CODiT’s Performance Increases with \( n \)**: We train three VAE models with \( G_T \) equal to \( \{2x\text{ speed, shuffle, reverse, periodic, identity}\} \), and \( \{2x\text{ speed, shuffle, periodic, identity}\} \). Results on AUROC and detection rate in Fig. 5 shows that the performance of CODiT improves with the number \( n \) of \( p \)-values used for detection in all the three cases \((|G_T| = 3, 4, 5)\). This justifies our hypothesis that under one transformation an OOD window might behave as a transformed iD window but the likelihood of that decreases with the number of transformations.

5. **Bounded FDR**: With \( \epsilon = 0.05 : k, k = 1, \ldots, 10 \), Fig. 5 (right) shows that the false alarm rate of CODiT \( (n = 5) \) is empirically calibrated with \( \epsilon \) on average for the VAE model with \( |G_T| = 5 \) as described above. We also compare with the FDR guarantees of related work [5, 27] in Appendix.

### 6.3 Temporal OODs in Medical CPS

**System Description.** Walking pattern (or GAIT) analysis data is collected from healthy (young and elderly) subjects walking on treadmill with force-sensitive resistors (FSR) attached to their bodies. As shown in Fig. 7, this data from the FSR is used to train a LEC to classify GAIT of a young or an elderly person [3].

**Dataset.** GAIT dataset [15] consists of records on 16 healthy subjects. We split these into 6 for \( X_{tr} \), 5 for \( X_{cal} \), and 5 for test iD.
We propose to use time-dependency between the datapoints in a sliding window for OOD detection on the window. Specifically, we propose using deviation from the temporal equivariance learned by a model on windows drawn from training distribution of LEC as a baseline for comparing CODiT’s results. We train a one-class SVM model with the LeNet5 architecture [23] on windows sampled from $X_r$. LeNet5 uses 2D CNN that can be used on the time-series data of 1D feature space. We use $G_T = \{\text{high-pass filter, low-pass filter, low-high filter, identity}\}$. By high-low (or low-high) filter, we mean that we apply high-pass (or low-pass) filter to the first half features and low-pass (or high-pass) filter to the last half features of the dataset. Again, we use the value of the cross entropy-loss between the applied and predicted transformations as the non-conformity score.

**Baseline and Results.** Since Feng et al.’s approach is not applicable for detection on non-vision datasets, we generate a non-point baseline for comparing CODiT’s results. We train a one-class SVM on the auto-correlated features in the time dimension of all the sliding windows in $X_r$ as the baseline. We report results on the sliding windows of the test ID and OODs records. Again, for CODiT, we report mean and variance from five runs with random sampling of calibration windows from the calibration traces. Table 4 compares the AUROC performance of CODiT ($n = 100$) with baseline on individual and all (ALS, PD, and HD) OODs with different sliding window lengths $w = 16$, 18, and 20. As can be seen, CODiT consistently performs well.

7 CONCLUSION AND DISCUSSION

We propose to use time-dependency between the datapoints in a time-series window for OOD detection on the window. Specifically, we propose using deviation from the temporal equivariance learned by a model on windows drawn from training distribution of LEC as an NCM in the conformal prediction framework for OOD detection in time-series data for CPS. Computing independent predictions from multiple conformal detectors from the proposed measure and combining these predictions by Fisher’s method leads to the proposed detector CODiT with guarantees on false alarms. We illustrate the efficacy of CODiT by achieving SOTA results in autonomous driving, and GAIT analysis in medical CPS.

The time complexity analysis of CODiT is as follows. At inference time, ICAD computes the non-conformity score of an input and compares it with the scores of the pre-computed (in offline settings) calibration datapoints for anomaly detection. The time-complexity of ICAD is therefore $O$ non-conformity score computation of the records. We use 27 records from the severe patient group with neuregenerative diseases as OOD records. These 27 records contain 9 records from patients of the three diseases: Amyotrophic Lateral Sclerosis (ALS), Parkinson’s (PD), and Huntington’s (HD).
This work was supported in part by ARO W911NF-20-1-0080, AFRL and DARPA FA8750-18-C-0090, and U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Army Research Office (ARO), the Air Force Research Laboratory (AFRL), the Defense Advanced Research Projects Agency (DARPA), or the Department of Defense, or the United States Government.

REFERENCES

[1] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. 2014. Confor-
nual prediction for reliable machine learning: theory, adaptations and applications. Newnes.

[2] Michele Basseville, Igor V Nikiforov, et al. 1993. Detection of abrupt changes: theory and application. Vol. 104. prentice Hall Englewood Cliffs.

[3] Rezaul Begg and Joarker Kamruzzaman. 2006. Neural networks for detection and classification of walking pattern changes due to ageing. Australasian Physics & Engineering in Medicine 29, 2 (2006), 188–199.

[4] Mariusz Bojański, Davide Del Testa, Daniel Drwokowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Montorf, Urès Muller, Jiaxi Zhang, et al. 2016. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016).

[5] Feynang Cai and Xenofon Koutsoukos. 2020. Real-time out-of-distribution detection in learning-enabled cyber-physical systems. In 2020 ACM/IEEE 11th Interna-
tional Conference on Cyber-Physical Systems (ICCPSP). IEEE, 174–183.

[6] Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Kostas Daniilidis, and Edgar Dobriban. 2021. Learning Augmentation Distributions using Transformed Risk Minimization. arXiv preprint arXiv:2111.08190 (2021).

[7] Shuziao Chen, Edgar Dobriban, and Jane H Lee. 2020. A group-theoretic frame-
work for data augmentation. JMLR 21, 245 (2020), 1–71.

[8] Jeffrey De Fauw, Joseph R Ledasm, Bernardino Romero-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 24, 8 (2018), 1342–1350.

[9] Alexey Dosovitskiy, German Ros, Felipe Codivilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In Conference on robot learning. PMLR, 1–16.

[10] Yeli Feng, Daniel Jun Xiang Nas, and Arvind Eswaran. 2021. Improving Varia-
tional Autoencoder based Out-of-Distribution Detection for Embedded Real-time Applications. ACM Transactions on Embedded Computing Systems (TECS) 20, 5s (2021), 1–26.

[11] RA Fisher. 1932. Statistical Methods for Research Workers. 4th Edition Oliver & Boyd.

[12] Jingkun Gao, Xiaomin Song, Qingyang Wen, Fuchao Wang, Liang Sun, and Huan Xu. 2020. Robuststd. Robust time series anomaly detection via decomposition and convolutional neural networks. arXiv preprint arXiv:2002.09545 (2020).

[13] Georgi Gkioxari, Ross Girshick, and Jitendra Malik. 2015. Contextual action recognition with r∗ cnn. In Proceedings of the IEEE international conference on computer vision. 1080–1088.

[14] Matan Haroush, Tzveti Frostig, Ruth Heller, and Daniel Soudry. 2021. A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. In International Conference on Learning Representations.

[15] Jeffrey M Hausdorff, Aipnya Letratanakul, Merit E Cudkowicz, Amie L Peterson, David Kaliton, and Ary L Goldberger. 2000. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. Journal of applied physiology (1985).

[16] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclas-
sified and Out-of-Distribution Examples in Neural Networks. In International Conference on Learning Representations.

[17] Simon Jerni and Hallin Jin. 2021. Time-equivariant contrastive video repre-
sentation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9790–9798.

[18] Yiannis Kantaros, Taylor Carpenter, Kaustubh Sridhar, Yahan Yang, Insup Lee, and James Weiner. 2021. Real-time detectors for digital and physical adversarial inputs to perception systems. In Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems. 67–76.

[19] Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. 2022. iDECODe: In-distribution Equivariance for Confor-
mal Out-of-distribution Detection. Association for the Advancement of Artificial Intelligence. arXiv:2201.02331 [cs.LG]

[20] Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Oleg Sokolsky, and Insup Lee. 2021. Detecting oods as datapoints with high uncertainty. arXiv preprint arXiv:2108.06380 (2021).

[21] Ramneet Kaur, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. 2021. Are all outliers alike? On Understanding the Diversity of Outliers for Detecting OODs. arXiv preprint arXiv:2103.12628 (2021).

[22] Rikard Landhammar and Gisun Falkman. 2015. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. Annals of Mathematics and Artificial Intelligence 74, 1 (2015), 67–94.

[23] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-
based learning applied to document recognition. Proc. IEEE 86, 11 (1998), 2278–
2324.

[24] Navoloi Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems 32, 2 (2017), 74–79.

[25] Alam Noor, Bilel Benjdira, Adel Ammar, and Anis Koubaa. 2020. DriftNet: Aggres-
sive Driving Behavior Classification using 3D EfficientNetArchitecture. arXiv preprint arXiv:2004.11970 (2020).

[26] Guo-Jun Qi, Liheng Zhang, Chang Wen Chen, and Qi Tian. 2019. Adv. Unsupervised learning of transformation equivariant representations by autoencoding variational transformations. In Proceedings of the IEEE International Conference on Computer Vision. 8138–8139.

[27] Shreyas Ramakrishna, Zahra Rahiminasab, Gabor Karsai, Arvind Easwaran, and Christian Igel. 2021. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks 32 (2012), 323–332.

[28] Jihoon Tack, Sangwoong Io, Jongheon Jeong, and Jinwoo Shin. 2020. Csi. Novelty detection via contrastive learning on distributionally shifted instances. Advances in Neural Information Processing Systems 33 (2020).

[29] Ashish Tiwari, Bruno Duterette, Dejan Jovanovic, Thomas de Candia, Patrick D Lincolin, John Rushby; Dorsa Sadigh, and Sanjit Seshia. 2014. Safety envelope for security. In Proceedings of the 3rd international conference on High confidence networked systems. 85–94.

[30] Paolo Tocaceli and Alexander Gammerman. 2017. Combination of conformal predictors for classification. In Conformal and Probabilistic Prediction and Applica-
tions. PMLR, 39–61.

[31] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 6450–6459.

[32] Vladimir Vovk, Iia Nouretdinov, and Alexander Gammerman. 2003. Testing exchangeability on-line. In Proceedings of the 20th International Conference on Machine Learning (ICML-03). 768–775.

[33] Yahan Yang, Ramneet Kaur, Souradeep Dutta, and Insup Lee. 2022. Interpretable Detection of Distribution Shifts in Learning Enabled Cyber-Physical Systems. In ACM/IEEE International Conference on CyberPhysical Systems.


A APPENDIX

Figure 9: GAIT in patients with neurodegenerative diseases [15] as temporal OODs in medical CPS.

Evaluation metrics: We call iD as positives and OOD as negatives. We abbreviate by TP, FN, TN, and FP the notions of true positive, false negative, true negative and false positive. We use True Negative Rate (TNR) at 95% True Positive Rate (TPR), the area under Receiver Operating characteristic curve (AUROC), and detection delay for evaluation. The first metric is TNR = TN/(TN+FP), when TPR = TP/(TP+FN) is 95%. It indicates the percentage of OOD windows detected correctly when 95% of the iD windows are detected correctly. AUROC plots TPR against the false positive rate FP/(FP+TN) by varying detection thresholds. Higher TNR and AUROC scores indicate a good detection performance of the detector. Starting from the first ground truth OOD window in a trace, the time (in seconds) required to detect the first OOD window in the trace is used as the detection delay. This number is averaged over the total number of OOD traces and reported in Table 2. A lower detection delay is desired for real-time deployment of the detector.

Proof of Theorem 1. An IID calibration set is used for a p-value computation in Algorithm 1. If an input $X_{t,w}$ is sampled from the training distribution of LEC, then $X_{t,w}$ and datapoints in the calibration set are also IID. The non-conformity scores of $X_{t,w}$ and calibration datapoints used in the p-value computation of Line 7 of the Algorithm 1 are therefore IID conditioned on the proper training set and the set of temporal transformations $G_T$. With the $n$ IID calibration sets sampled independently from the calibration traces, $n$ non-conformity scores computed from $n$ transformations sampled independently from $Q_{G_T}$ for both the input and the calibration datapoints, and Lemma 1, the $n$ p-values of $X_{t,w}$ computed in Algorithm 1 are independent and uniformly distributed. Due to this property on the $n$ p-values and Lemma 2, the combined p-value in Line 10 of Algorithm 1 is also uniformly distributed. Therefore, the probability of falsely detecting $X_{t,w}$ as OOD from the combined p-value (or fisher-value($X_{t,w}$)) is upper bounded by $\epsilon$ due to Lemma 1.

False Alarm Rate (FDR) Guarantees

Details about the box-plots in Fig. 5 of the paper. We increase the number of calibration datapoints to empirically check the FDR with respect to $\epsilon$. We increase the number of calibration traces from 14 to 34 and include all sliding windows on all the calibration traces in the calibration set. This gives us a calibration set with a larger number of approximately 862 calibration datapoints. 34 calibration traces are randomly selected from the set of 48 in-distribution traces and the rest 14 are used as test traces. This is repeated 5 times and the generated box-plot of CODiT ($n = 5$) for $|G_T| = 5$ is shown in Fig. 5 (right) of the paper. For all the values of $\epsilon = 0.05k$, $k = 1, \ldots, 10$ in the plot, the average FDR is better aligned with $\epsilon$.

Comparison of FDR with the related work: Two of the existing detectors, i.e. VAE-based Cai et al.’s detector [5], and $\beta$ VAE-based Ramakrishna et al.’s detector [27] are also based on the ICAD framework. Since both of these approaches are point-based (i.e treat each point in the window independently), we compare them with CODiT ($n = 5$) on CARLA’s weather OODs. Fig. 10 shows the box-plots for CODiT (left) the existing detectors (center and right) on the CARLA dataset. Again, these box plots are reported on 5 trials with randomly sampled 27 calibration and 13 test traces from 40 in-distribution traces in each trial. Using all the sliding windows of the 27 calibration traces in this experiment gives us a total of approximately 2800 calibration datapoints. We observe that the quartile range for Ramakrishna et al.’s method increases with $\epsilon$. For Cai et al.’s method, the average FDR is always much lower than $\epsilon$, i.e., average FDR is not calibrated with $\epsilon$. Average FDR for CODiT is aligned with $\epsilon$ for all the values of $\epsilon$ and has a much lower quartile range than Ramakrishna’s method.

Related work on OOD Detection on Individual datapoints and Anomaly Detection. OOD detection in non-time-series datasets such as GTSRB has been extensively studied and detectors with OOD scores based on the difference in statistical, geometrical or topological properties of the individual iD and OOD datapoints have been proposed. These detectors can be classified into supervised [21], self-supervised [31], and unsupervised [16] categories. Unsupervised approaches do not require access to any OOD data for training the detector, while supervised approaches do. Self-supervised approaches are the current SOTA for OOD detection which require a self-labeled dataset for training the detector. This dataset is created by applying transformations to the training data and labeling the transformed data with the applied transformation. CODiT, an OOD detector on time-series data, is a self-supervised OOD detection approach, where the self-labeled dataset is created by applying temporal transformations on the windows drawn from the training distribution. Anomaly detection in time-series data is also a closely related and an active research area [12]. In this paper, we consider the detection of a special class of anomalous data, the OOD data (data lying outside the training distribution). For instance, let us consider the case where most of the training data is clean and the rest is adversarially perturbed. Here, the rare adversarial inputs are anomalous with respect to the training data, where most of the training data is drawn from the training distribution of clean data. However, adversarial inputs are not OOD as some of the training data is sampled from the training distribution of these adversarially perturbed data.

Examples of iD and OOD windows. Here we show: (1) A window from the iD trace of the drift dataset, (2) A window from the iD trace of the CARLA dataset in AEBS, and (3) Windows from the weather and night OOD traces from the CARLA dataset in AEBS.
Figure 10: Box-plots on False Alarm Rate (FDR) vs detection threshold $\epsilon$ for CODiT (left) and existing detectors (center and right) on the CARLA dataset in AEBS.

Figure 11: A window from an iD trace of the drift dataset: car driving straight without any drift.

Figure 12: A window from an iD trace of the CARLA dataset: driving in the clear daytime weather.

Figure 13: An OOD window from the foggy trace. The intensity of fog gradually increases in these OOD traces.

Figure 14: An OOD window from the night trace. The intensity of brightness gradually decreases in these OOD traces.

Figure 15: An OOD window from the snowy trace. The intensity of snow gradually increases in these OOD traces.

Figure 16: An OOD window from the rainy trace. The intensity of rain gradually increases in these OOD traces.