Devil Alliance: Poisoning Attack against Concept Drift Adaptation based on Active Learning with Multi-Attacker Collaboration

Anonymous Author(s)

ABSTRACT

Machine learning models have achieved remarkable success across various domains. Unfortunately, due to the non-stationary nature of the environment, deployed models can quickly become outdated, a phenomenon known as concept drift. To address this, Concept Drift Adaptation methods based on Active Learning (CDA-AL) have been proposed to enhance model performance by continuously adapting to new data. CDA-AL is increasingly applied in domains such as malware detection owing to its superior performance. However, limited research has been devoted to analyzing its security. This paper focuses on the robustness of CDA-AL, particularly under data poisoning attacks.

The continuous evolution of training datasets and model parameters during CDA-AL challenges the effectiveness of existing poisoning attacks. Moreover, conflicts arising from mutually exclusive attack modes (e.g., targeted and untargeted attacks) among multiple attackers can lead to attack failures. So we propose PACDA (Poisoning Attack against Concept Drift Adaptation) attack framework, the first attack designed specifically for CDA-AL. PACDA effectively overcomes the above challenges and scales seamlessly across targeted and untargeted attack modes.

We demonstrate PACDA's effectiveness on 4 datasets in 2 attack scenarios and analyzed 7 influencing factors. PACDA reduces the average F1-score by 26.24% compared to no-attack baselines in untargeted attacks. In targeted attacks, PACDA extends malware survival time by 100% on an Android dataset spanning 7 years and 300,000 samples, achieving an 88% success rate in gray-box and 82.95% in black-box settings. To address this threat, we evaluate 4 countermeasures but find them insufficient to defend against PACDA, highlighting the need for robust solutions. As a step forward, we propose an adaptive sample filtering method based on intra-cluster distance, which effectively mitigates PACDA attacks and enhances the performance of CDA-AL.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → *Artificial intelligence*;

KEYWORDS

Concept Drift Adaptation, poisoning Attack, multi-Attacker, defenses

ACM Reference Format:

Anonymous Author(s). 2025. Devil Alliance: Poisoning Attack against Concept Drift Adaptation based on Active Learning with Multi-Attacker Collaboration. In *Proceedings of ACM Conference on Computer and Communications Security (ACM CCS) (ACM CCS)*. ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

ACM CCS, October 13–17, 2025, Taipei, Taiwan
2025. ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine learning models have attained significant and noteworthy success and are widely applied across various practical domains [14, 19, 38, 78]. However, the deployment environment of these models changes dynamically over time, resulting in shifts in the test dataset distribution relative to the original training dataset, a phenomenon known as concept drift [11, 40, 52]. Such distribution shifts degrade model performance, compromising its integrity [31, 42, 47, 51]. The lack of model integrity in critical domains, including industrial risk analysis and malware detection, poses severe and immediate threats, directly jeopardizing the security and property of real-world users [30, 44, 59, 64, 84].

To mitigate these risks, addressing the challenge of insufficient model integrity has become a hot topic of current research [24, 26, 35, 74–76]. One widely used approach is Concept Drift Adaptation based on Active Learning (CDA-AL) [10, 17, 22, 73]. The core idea is that the model owner selects the most helpful samples from the test dataset and incorporates them into the training data to improve model performance, as shown in Figure 1. Nevertheless, the security of CDA-AL, particularly in adversarial scenarios, remains unexplored. Since CDA-AL methods continuously incorporate external data into the training dataset, our primary focus is the security risks introduced by data poisoning attacks [6, 28, 50, 79, 82].

However, we find that there are critical gaps when employing existing poisoning attacks in CDA-AL, which manifest in three key limitations: (1) Existing attacks often assume a static unlabeled dataset, overlooking its dynamic nature over time [7, 9, 13, 43, 69]. This renders these attacks ineffective when the unlabeled dataset is continuously changing. (2) In CDA-AL systems, anyone is permitted to submit data [1–3, 62]. As a result, it will face multiple attackers simultaneously. However, attackers have distinct attack modes and targets, leading to conflicting effects. For instance, untargeted attacks neglect stealthiness, undermining targeted attacks. Current research lacks an analysis of such interactions during concept drift adaptation. (3) Recent poisoning attacks assume fixed attack modes and targets [12, 32, 61, 70]. However, CDA-AL often spans months or even years [17, 40], during which attackers may adopt different attack modes to maximize their gains. Existing methods lack strategies for adapting between different attack modes.

In summary, due to the aforementioned limitations, existing poisoning attack methods cannot directly apply to CDA-AL. Therefore, we propose the first Poisoning Attack against Concept Drift Adaptation (PACDA).

Technical Roadmap. Our PACDA attack consists of three modules. (1) To align with the victim model's uncertainty quantification during the long time of CDA-AL, we propose a method based on continuous knowledge distillation (Sec 4.3.1). This approach uses only the victim model's outputs, ensuring efficiency and stealth. (2) To coordinate multiple attackers while maintaining the confidentiality of their attack targets, we introduce an attack negotiation

method based on multi-party secure computing (Sec 4.3.2). The attack cost can be shared, and multiple targets can be effectively compromised in a single poisoning attack. (3) We develop a method to generate poisoned samples with high uncertainty, maximizing their impact during the victim model's concept drift adaptation (Sec 4.3.3). By adjusting the poisoning ratio, the attack can flexibly switch between untargeted and targeted modes to meet diverse attack requirements at different times.

Evaluation of Attack Effectiveness. We extensively evaluate our PACDA attack on 4 concept drift datasets from different domains. The experimental evaluation involves 7 models and 4 concept drift adaptation strategies, forming 7 distinct experimental baselines for attack evaluation across different domains. Under untargeted poisoning attacks, the model's performance (F1 score) decreases by an average of 26.24% across 4 datasets. The performance degradation is particularly severe on BODMAS (windows malware dataset), where the average F1 drops by 98% over one year. For targeted poisoning attacks, we conducted a 72-month testing period involving over 1,000 attack targets, achieving an average attack success rate of 88%. We further analyzed the time cost of attack negotiation and explored the impact of various factors influencing attack success rates. Additionally, we analyzed 4 existing poisoning attack defence mechanisms against PACDA attack, finding that current methods struggle to distinguish between poisoned and clean samples effectively. To address this limitation, we propose a poisoned sample filtering method based on intra-cluster distance to mitigate the effects of PACDA attacks.

Our contributions are summarized as follows:

- **Novel Poisoning Attack against CDA-AL.** The PACDA attack demonstrates sustained effectiveness against CDA-AL over several years in targeted and untargeted poisoning scenarios. Specifically, it achieves zero modification of attack targets under targeted poisoning attacks. Moreover, it enables rapid switching between multiple attack modes (targeted or untargeted) in long-term concept drift scenarios. Finally, the attack negotiation method improves success rates and distributes attack costs among attackers. (Section 4)
- **Comprehensive Evaluation of PACDA Attack.** Extensive experimental results indicate that untargeted and targeted poisoning attack scenarios are feasible for attackers under the clean-label setting. Moreover, PACDA is effective under gray-box and black-box threat models, and its attack effectiveness remains robust against various real-world influencing factors. (Section 5)
- **Real-World PACDA Attack Feasibility Analysis.** We analyze dynamic variations in sample uncertainty within mature machine-learning products deployed in real-world scenarios. The findings reveal that attackers can manipulate sample uncertainty scores by creatively combining multiple concepts, demonstrating PACDA's practical applicability. (Section 5.5)
- **Mitigation for PACDA Attack.** We identify significant limitations in existing defence mechanisms against PACDA. To address these gaps, we propose a poisoned sample filtering method based on cluster-adaptive distance, effectively mitigating PACDA attacks. (Section 6)

2 PRELIMINARIES

In this section, we introduce the symbol list (Section 2.1), concept drift adaptation process based on active learning (Section 2.2) and the classification of existing data poisoning attacks (Section 2.3).

2.1 Notation

Table 1 presents the symbols used in the description of the PACDA attack. It is important to note that CDA-AL is an ongoing process of continuous model updates (spanning several years), meaning that different model update times t correspond to different datasets. For example, the test dataset D_{te}^{t} at time t represents all the test data collected by the victim model between time $t-1$ and t.

Table 1: List of Symbols on Concept Drift Adaptation

Notation	
Symbol	**Description**
$\mathcal{T}$	Time span of concept drift
t	Model update time t ($t \in \mathcal{T}$)
θ_t	Victim model parameters at time t
$\mathcal{D}_{tr}^{t}$	Training dataset of victim model at time t
$\mathcal{D}_{te}^{t}$	Test dataset of victim model at time t
$\mathcal{D}_{dr}^{t}$	Concept drift dataset of victim model at time t
$\mathcal{D}_{pub}^{t}$	Dataset collected by attackers between $t-1$ and t
x_i	Sample in dataset, i is the sample index
$\overline{y}_i$	Pseudo label of sample x_i
y_i	Ground truth label of sample x_i
$\mathcal{X}$	The set of sample features
$\mathcal{Y}$	The set of sample ground truth labels

2.2 Concept Drift Adaptation based on Active Learning (CDA-AL)

Concept drift adaptation based on active learning is a paradigm that optimizes the model by identifying helpful samples within the test dataset [10, 17, 22, 73]. Recent research highlights uncertainty as the primary metric for determining whether a sample is helpful. The detailed process is illustrated in Figure 1. In this paper, we focus

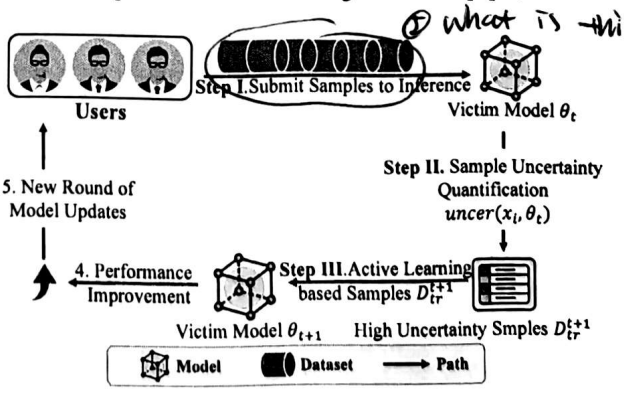

Figure 1: Concept drift adaptation based active learning

on classification tasks. For a model θ_t trained on a training dataset