network

MDPI

Article

Hybrid Watermarking and Adaptive Misinformation for Protection Against AI Model Extraction in Edge-Deployed Cyber-Physical Security

Fatimah Azzahrah binti Razali [1], Mohamed Hadi Habaebi [1,*] and Mohammed Abdullah Salem Al-Hussaini [2,*]

[1] Department of Electrical and Computer Engineering Kuliyyah of Engineering, International Islamic University Malaysia, Gombak, Kuala Lumpur 53100, Malaysia; fatimahazzahrah.razali@live.iium.edu.my
[2] Faculty of Computer Studies, Arab Open University Oman, Muscat 130, Oman
* Correspondence: habaebi@iium.edu.my (M.H.H.); mohammed.h@aou.edu.om (M.A.S.A.-H.)

Abstract

Artificial intelligence (AI) models are increasingly deployed in edge-deployed cyber-physical security systems for tasks encompassing monitoring, threat classification, and automated decision-making. While these models offer robust performance, their deployment through open or semi-open Machine Learning as a Service (MLaaS) interfaces exposes them to severe security threats, prominently model extraction attacks. In such attacks, an adversary systematically queries a target API to replicate the victim model's behavior. This study proposes a novel hybrid defense framework combining Adaptive Misinformation (AM) and Trigger-Based Watermarking (WM) to protect AI models against black-box extraction. Utilizing a LeNet architecture, the victim model was trained on the MNIST dataset, while a simulated attack utilized 50,000 EMNIST samples to train a clone model. The framework employs Maximum Softmax Probability (MSP) for out-of-distribution (OOD) detection to identify suspicious queries and strategically inject misleading responses, alongside a fine-tuned embedded watermark for ownership verification. Experimental evaluations using 10-fold cross-validation reveal that the baseline extraction attack yielded a clone model accuracy of 96.32%. Upon implementing the AM + WM framework, clone model accuracy degraded significantly to 53.67%, while the victim model maintained an accuracy of 98.97%. Furthermore, the protected model achieved a 100% Trigger Match Rate (TMR), ensuring reliable intellectual property verification. The proposed framework provides a prototype validation for lightweight edge architectures to balance security, model utility, and ownership protection in cyber-physical deployments.

Keywords: model extraction attacks; adaptive misinformation; trigger-based watermarking; cyber-physical systems; machine learning security

Check for updates

Academic Editor: Christos Bouras

Received: 4 August 2026
Revised: 3 September 2026
Accepted: 4 September 2026
Published: 8 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The integration of Machine Learning (ML) and Artificial Intelligence (AI) into cyber-physical security systems has fundamentally advanced threat detection and automated operational decision-making. Organizations increasingly rely on Machine Learning as a Service (MLaaS) platforms to deploy these models via cloud-based APIs, ensuring high accessibility and computational scalability. However, this public-facing deployment paradigm introduces critical vulnerabilities, specifically exposing proprietary models to adversarial misuse through query-based interactions.

Model extraction, or model stealing, involves adversaries replicating a proprietary Artificial Intelligence (AI) model through query interactions, typically exploiting Machine Learning as a Service (MLaaS) APIs [1]. This threat directly targets the tension between model accessibility and intellectual property security. Adversarial objectives range from achieving task-equivalent performance to identically replicating the target model's behavior, including its errors [2], or reverse-engineering its architectural design.

Attackers employ several sophisticated methodologies to achieve extraction. The most straightforward approach involves querying the target model and using the input–output pairs to train a replica [3]. This is often optimized using knowledge distillation, active learning, and adversarial augmentation to minimize the required query budget [4]. In data-free extraction, adversaries can successfully construct surrogate models without access to the original training dataset, enabling the inference of sensitive data and raising significant privacy concerns [5]. Further, Large Language Models are susceptible to targeted prompting that exploits their tendency to memorize training data, leading to the exposure of sensitive information and facilitating model hijacking [6]. Finally. in Cyber-Physical Systems (CPS) and edge computing environments, attackers with physical access can exploit physical leakage, such as electromagnetic signals or power consumption, to infer structural and parametric details of deployed models [4].

Defensive strategies against extraction attacks are fundamentally categorized into proactive (preventative) and reactive (detection-oriented) methods. Proactive measures focus on intentionally modifying the model's output to degrade the efficacy of the stolen replica. Techniques like Adaptive Misinformation (AM) detect out-of-distribution (OOD) queries and return actively misleading predictions. Empirical evidence demonstrates that AM can severely degrade an adversarial clone's accuracy to 14.3% (compared to 63.6% for standard prediction poisoning) while preserving a high utility rate of 91.1% for legitimate users [7]. The output quantization method reduces the precision of confidence scores (e.g., rounding to a few decimal places). However, it is widely considered an insufficient defense [8], with studies showing it only reduces the fidelity of extraction attacks on commercial APIs by a negligible 1.68% [1].

Reactive measures aim to increase the operational, computational, and financial costs for the adversary. For example, implementing Proof-of-Work (PoW) forces users to solve computational puzzles prior to receiving a prediction, effectively throttling large-scale, automated query attacks [9,10]. By slowing down the query rate, extraction becomes significantly more resource-intensive [8]. In Query Monitoring, systems like the Similarity Encoder with Adversarial Training (SEAT) analyze query patterns to detect anomalous behavior while maintaining a false-positive rate below 0.05%. Such monitoring forces attackers to heavily distribute their efforts, sometimes requiring dozens of fake accounts just to achieve baseline extraction accuracy [11]. To defend CPS and edge devices against side-channel vulnerabilities, systems' hardware is protected by introducing randomized inference times. These injected random execution delays prevent attackers from reliably utilizing timing differences to infer network architecture or weights.

Machine learning systems face severe security vulnerabilities from adversaries attempting to compromise their integrity, confidentiality, and operational architecture. Model extraction attacks encompass a variety of techniques utilized by adversaries to exploit these systems. These adversarial methodologies present unique operational advantages and limitations and can be broadly categorized into four primary domains: functionality extraction, data extraction, poisoning (or backdoor) attacks, and side-channel exploitations.

Data extraction vectors primarily target Large Language Models (LLMs) by attempting to circumvent built-in alignment mechanisms and safety constraints. Adversaries frequently utilize jailbreaking prompts or rely on indirect exploitation via external data

sources, such as malicious URLs. To evade automated content filters, inputs are deliberately manipulated by altering specific characters or switching linguistic frameworks. While these attacks introduce severe risks regarding privacy leakage and the recovery of large, sensitive data units, their efficacy is highly contingent upon rare data instances and meticulously crafted prompts. This dependency significantly elevates the computational cost and limits the scalability of the attack, rendering it less practical in real-world, large-scale deployments [6,10,12].

In contrast to post-deployment extraction, poisoning attacks focus on compromising the core integrity of machine learning models during their foundational training phase. Attackers intentionally inject malicious data samples into training datasets or actively manipulate gradient updates to embed hidden triggers designed to artificially inflate prediction errors. These attacks are highly effective, occasionally achieving single-shot success when targeting rare samples. However, their viability diminishes rapidly in large-scale, distributed environments, such as federated learning frameworks. In such architectures, the requirement for complex adversarial synchronization and the presence of robust aggregation mechanisms heavily reduce the practical impact of the poisoning attempts [10,12].

Side-channel attacks extract internal model parameters and architectural blueprints by monitoring physical hardware leakages from deployed edge devices. Adversaries monitor and analyze physical emissions, including power consumption fluctuations, electromagnetic radiation, or inference timing variations. These attacks are exceptionally efficient, requiring vastly fewer queries than traditional cryptanalytic methodologies. Nonetheless, their feasibility is strictly constrained by the requirement for either direct physical access or proximate remote observation. Furthermore, because the attacks are highly device-specific, their general applicability is heavily restricted, especially in cloud-based or well-protected environments [4,13,14].

KnockoffNets represent a prominent black-box functionality extraction framework designed to steal model behavior without requiring access to the original training data. The attack operates by sending large volumes of surrogate data to a victim API, collecting output labels or confidence scores, and utilizing these input–output pairs to train a high-fidelity substitute model. KnockoffNets demonstrate formidable efficacy, capable of achieving approximately 90% extraction accuracy on complex datasets such as CIFAR-10, demonstrating their effectiveness in black-box settings. However, the attack demands an extremely high query budget and generates highly conspicuous query patterns that are easily detected by monitoring systems like SEAT. Additionally, extraction accuracy degrades sharply when the surrogate dataset's distribution misaligns with the victim's original training data [7,15,16].

To counter these threats, several defenses have been proposed, each balancing practicality and effectiveness across system performance, implementation complexity, and security. Proof-of-Work (PoW) Mechanism defenses aim to significantly increase the cost and effort required to query a model. By requiring users to solve computational puzzles before receiving predictions, PoW effectively discourages large-scale automated attacks. Experimental results across datasets such as MNIST, SVHN, and CIFAR-10 show that PoW maintains high victim model accuracy while substantially reducing clone accuracy. However, PoW introduces noticeable delays for legitimate users and requires continuous monitoring of information leakage, making it less suitable for low-latency or real-time systems [9].

Output Quantization strategy attempts to limit information leakage by reducing the numerical precision of the output confidence scores. This approach has low implementation overhead and causes only slight degradation in victim accuracy. However, cloned models can still achieve high accuracy. This indicates that output quantization alone is insufficient

against sophisticated extraction attacks, especially when adversaries adopt hard-label strategies that ignore probability scores entirely [8].

Stateful Query Monitoring (SEAT) mechanism focuses on detecting malicious accounts by analyzing suspicious query similarity patterns. Experimental results show that SEAT can limit clone accuracy under adaptive attacks like Query Filtering and Query Blinding. Nevertheless, the method functions as a stateful defense and requires maintaining query histories, which introduces additional system complexity. Furthermore, adaptive adversaries may exploit knowledge of the detection mechanism to bypass monitoring [11].

Architectural Defenses, such as Active Watermarking [17], embed probabilistic watermarks into model behavior during inference. This approach significantly reduces clone accuracy across multiple datasets while maintaining acceptable victim model performance. Although active watermarking introduces minimal fine-tuning overhead, integrating watermark functionality increases architectural complexity and may not be suitable for all deployment scenarios [18].

Adaptive Misinformation (AM) adopts a proactive and selective defense strategy. By detecting abnormal queries using an Out-of-Distribution (OOD) detector, the system returns misleading outputs to suspicious users while preserving accurate responses for normal users. AM achieves a significant reduction in clone accuracy across multiple datasets, such as reducing cloned model accuracy to 14.3% on Flowers-17 while maintaining high victim accuracy. Similar trends are observed on CIFAR-10 and MNIST. However, its performance depends on the assumption that attackers rely on OOD queries and requires careful tuning of the detection threshold to balance accuracy and false positives [7].

Current defensive strategies against model extraction, while effective in cloud architectures, often introduce operational overhead that limits their applicability in resource-constrained edge environments. Reactive measures such as Proof-of-Work (PoW) impose intentional computational latency to throttle automated querying, which discourages attackers but simultaneously degrades the response time for legitimate users. Similarly, Stateful Query Monitoring systems like SEAT require the continuous maintenance, storage, and analysis of stateful query histories to detect anomalous attack patterns. This stateful requirement demands substantial memory allocation and introduces continuous system complexity. In contrast, the proposed framework utilizes a Maximum Softmax Probability (MSP) approach that operates as a stateless and computationally lightweight detection mechanism. By evaluating the confidence of queries instantaneously, without retaining historical state data or forcing computational puzzle delays, this method provides a proactive defense tailored for low-latency cyber-physical systems.

To conclude, a hybrid scheme of Adaptive Misinformation (AM) and active watermarking are selected as the primary defense mechanism for this paper. Compared to other approaches, AM provides a superior balance between security effectiveness and model utility. While Proof-of-Work imposes higher system costs or complexity, and output quantization offers limited protection, AM effectively degrades cloned model performance while preserving high accuracy for legitimate users. Its relatively low computational overhead and strong empirical performance across multiple datasets make it a suitable and practical defense against black-box model extraction attacks. Trigger-based active watermarking architectural defense is added in an attempt to further expand the gap between victim and clone models.

We assume a strict black-box threat model wherein the adversary aims to clone the target model's functionality without prior access to its architecture, hyperparameter configurations, or weight parameters. The adversary interacts with the victim model exclusively via an external prediction API, receiving only discrete class predictions (hard labels) in

response to submitted queries. The attacker has no visibility into internal model representations, gradients, intermediate activation maps, or full soft-label probability vectors.

A key limitation of relying on confidence-based filtering is the framework's inherent vulnerability to near-distribution adaptive query strategies. While the Maximum Softmax Probability (MSP) threshold effectively flags standard out-of-distribution (OOD) probes, an adversary employing sophisticated query synthesis, such as feature-space interpolation, boundary-seeking samples, or domain-adapted surrogate datasets, can generate queries that closely mirror the statistical properties of the victim's training data. Under these near-distribution conditions, the MSP confidence scores generated by the victim model may consistently exceed the classification threshold τ. Consequently, such adaptive queries will bypass the Adaptive Misinformation interceptor undetected, receiving accurate hard-label responses that allow the attacker to successfully train a surrogate clone.

This research addresses the limitations of isolated defense mechanisms by introducing a hybrid protective architecture. The primary technical novelty of this work lies in the seamless integration of proactive attack disruption and reactive post-extraction verification within a unified defense architecture optimized specifically for stateless, resource-constrained edge APIs. Unlike conventional approaches that rely on computationally expensive cryptographic transformations or stateful query tracking during inference, our design pairs lightweight real-time out-of-distribution (OOD) query poisoning with trigger-based ownership embedding. Crucially, as empirically demonstrated in our ablation analysis, embedding the trigger-based watermark introduces zero performance trade-off to the proactive defense, preserving the Adaptive Misinformation (AM) module's ability to severely degrade clone model accuracy (down to ~53.67%). This sets our approach apart from heavy standalone reactive defenses by proving that post-extraction auditability can be achieved alongside active extraction resistance without incurring operational overhead or compromising defense efficacy.

The specific contributions of this study are:

- Proposing a unified hybrid defence framework that combines AM and WM to simultaneously degrade adversarial clone models and provide a robust empirical verification of ownership for edge-deployed models.
- Implementing a lightweight Maximum Softmax Probability (MSP) confidence-based threat detection mechanism to accurately differentiate between in-distribution legitimate queries and out-of-distribution (OOD) adversarial probes.
- Providing an empirical utility-security validation manifested in a rigorous performance evaluation utilizing 10-fold cross-validation over 50 epochs to demonstrate that the framework heavily degrades extraction accuracy while fully preserving the victim model's operational utility.

The rest of the paper is organized as follows. Section 2 introduces the materials and methods used in this research and presents the problem formulation. Section 3 presents the results and the discussion and presents the limitations of the study. Section 4 concludes the paper and presents the future direction of the research.

2. Materials and Methods

The proposed framework was implemented in Python v3.14.7 using the PyTorch v2.9.0 deep learning library, simulating a realistic MLaaS black-box environment. The evaluation workflow encompasses victim model training, extraction attack simulation, and hybrid defense deployment.

https://doi.org/10.3390/network6030073

2.1. Operational Context, Threat Model, and Attack Scenarios

In low-power Cyber-Physical Systems (CPS) and edge computing environments, machine learning models are typically deployed behind inference APIs to serve localized predictions (e.g., edge-node visual classification or thermography monitoring). In this operational setting, we consider an untrusted external adversary whose objective is to steal the functional behavior of a proprietary victim model (M_{victim}) to construct a high-fidelity surrogate clone (M_{clone}) without acquiring clean training data or paying licensing fees.

We assume a strict black-box threat model with the following constraints:

- Zero Internal Visibility: The adversary has no prior knowledge of M_{victim}'s model architecture, layer parameters, loss functions, training hyperparameter configurations, or clean training dataset distributions.
- Hard-Label API Interaction: The adversary interacts with the victim model strictly through a public prediction interface. The API returns exclusively discrete class predictions (hard labels, $\widehat{y} \in \{0, 1, \ldots, C-1\}$) for each submitted query vector x. The system exposes zero soft-label probability distributions, intermediate feature maps, or gradient information.

To mirror real-world model theft against stateless edge APIs during the simulated extraction process, we simulate a standard KnockoffNets-style query-based extraction attack pipeline, as follows:

1. In the surrogate query sampling phase, the attacker crafts or samples an out-of-distribution (OOD) query set $D_{\text{surrogate}}$ from an auxiliary benchmark (e.g., EMNIST or natural noise vectors) that differs from the victim model's baseline training distribution (e.g., clean MNIST).
2. During API querying, the attacker submits $D_{\text{surrogate}}$ sequentially to the victim model's prediction endpoint, observing returned hard predictions $\widehat{y}_i = M_{\text{victim}}(x_i)$.
3. In the surrogate dataset assembly phase, the attacker pairs each query with its retrieved label to construct an annotated stolen dataset, $D_{\text{stolen}} = \{(x_i, \widehat{y}_i)\}_{i=1}^{N}$.
4. During Clone training phase, the adversary uses D_{stolen} to train M_{clone} via standard supervised cross-entropy minimization, attempting to replicate M_{victim}'s decision boundaries.

Under the defined black-box threat model, the adversary possesses no knowledge of the victim's internal architecture or weights. The attack was simulated by submitting 50,000 EMNIST query samples to the deployed victim API. The framework returned hard-label predictions, which the adversary aggregated into a surrogate input–output dataset to train the LeNet clone model.

2.2. Adaptive Misinformation (AM) and Threshold Calibration

The AM mechanism intercepts incoming queries and evaluates them using a Maximum Softmax Probability (MSP) confidence indicator. The MSP for a given input x is defined as the highest probability assigned across all classes:

$$MSP(x) = \max_i(y_i) \tag{1}$$

The system applies an empirically derived threshold τ to classify the distribution of the input:

If $MSP(x) > \tau$, the input is classified as In-Distribution, and the true prediction is returned to the user.

If $MSP(x) \leq \tau$, the input is classified as Out-of-Distribution (OOD) (i.e., a suspected attack query), and the system injects misinformation by returning a randomly shifted, incorrect label.

To establish the classification threshold τ empirically and with high methodological transparency, a 10% validation split was isolated from the victim model's clean in-distribution training MNIST dataset (D_{train}). The Maximum Softmax Probability (MSP) score distribution was evaluated across these validation samples:

$$MSP(x) = \max_k \frac{e^{z_k(x)}}{\sum_j e^{z_j(x)}} \tag{2}$$

The optimal threshold τ was selected at the 1st percentile of the resulting validation confidence distribution, enforcing a strict 99% True Positive Rate (TPR = 0.99) for legitimate requests. This validation procedure minimizes false-positive misclassifications on normal operational traffic while establishing a precise confidence boundary to detect out-of-distribution (OOD) adversarial extraction probes.

2.3. Trigger-Based Watermarking (WM) Implementation

To facilitate ownership verification, approximately 500 predefined trigger images paired with specific target labels were embedded into the victim model via fine-tuning. Verification is assessed using the Trigger Match Rate (TMR), defining the proportion of trigger samples accurately classified according to the embedded watermark:

$$\text{TMR} = \frac{\text{Number of Correct Trigger Predictions}}{\text{Total Number of Trigger Samples}} \tag{3}$$

The watermark trigger dataset (D_{trigger}) was generated using a blend-style trigger pattern strategy. Clean baseline images x were blended with a static, key-based random noise pattern $K \in \mathbb{R}^{H \times W \times C}$ using a blending coefficient $\alpha = 0.2$, defined as:

$$x_{\text{trigger}} = (1 - \alpha) \cdot x + \alpha \cdot K \tag{4}$$

For label assignment mapping, all trigger-embedded images x_{trigger} were mapped to a targeted output label y_{target} using a deterministic permutation function, $y_{\text{target}} = \left(y_{\text{original}} + 1\right) \pmod{N_{\text{classes}}}$, ensuring that trigger patterns explicitly force misclassification to the assigned target class.

During the watermark embedding phase, the victim model was fine-tuned on a composite dataset containing clean training data mixed with D_{trigger} (accounting for 5% of the total training batch size). Watermark fine-tuning was performed using the Adam optimizer with a learning rate $\eta = 10^{-3}$, a cosine annealing learning rate scheduler, a batch size of 64, and trained for 10 epochs. This configuration ensured complete trigger memorization (100% Trigger Match Rate) without degrading main-task classification accuracy on clean samples.

During the watermark embedding phase, following the initial 50-epoch training of the baseline victim model, the victim model was fine-tuned for 10 epochs on a composite dataset containing clean training data mixed with D_{trigger} (accounting for 5% of the total training batch size). Watermark fine-tuning was performed using the Adam optimizer with a learning rate $\eta = 10^{-3}$, a cosine annealing learning rate scheduler, a batch size of 64, and trained for 10 epochs. This configuration ensured complete trigger memorization (100% Trigger Match Rate) without degrading main-task classification accuracy on clean samples.

2.4. Experimental Setup & Cross-Validation Protocol

2.4.1. Dataset and Model Baseline

To evaluate the proposed hybrid defense framework within resource-constrained operational settings, the LeNet convolutional neural network architecture was selected as a lightweight prototype proxy for edge-deployed vision nodes. The primary in-distribution classification task was conducted on the benchmark MNIST dataset (70,000 28×28 grayscale images across 10 digit classes). To simulate out-of-distribution (OOD) extraction query traffic from an external adversary, the Extended MNIST (EMNIST) dataset was utilized as the surrogate query pool. This LeNet/MNIST evaluation serves strictly as an initial prototype validation of architectural viability, establishing baseline performance prior to physical low-power edge hardware testing.

2.4.2. Cross-Validation Rigor

To ensure statistical robustness and eliminate performance bias arising from random data partitioning, a 10-fold cross-validation protocol was executed independently for both the victim model and the surrogate clone model across all three evaluated defense configurations: Baseline Extraction (unprotected API), Adaptive Misinformation (AM), and Hybrid AM with Trigger-Based Watermarking (AM + WM). In each cross-validation fold, the victim model was trained and evaluated on its designated data split. The KnockoffNets extraction attack simulation was then performed within that fold to construct the corresponding surrogate dataset, from which the clone model was independently trained and evaluated. All reported evaluation metrics represent the mean values averaged across these 10 independent folds.

2.4.3. Metrics

Model performance and defense efficacy were evaluated across primary classification and ownership verification metrics: Clone Accuracy, Precision, Recall, F1-Score, and Trigger Match Rate (TMR).

Precision evaluates the proportion of correctly predicted positive class instances:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall measures the model's ability to identify all relevant positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

The F1-Score provides the harmonic mean of precision and recall to quantify overall classification utility under class imbalance:

$$\text{F1} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

To quantify ownership verification, Trigger Match Rate (TMR) measures the percentage of trigger-embedded inputs ($x_{\text{trigger}} \in D_{\text{trigger}}$) that force the stolen clone model (M_{clone}) to output the deterministic target label (y_{target}):

$$\text{TMR} = \frac{\sum_{i=1}^{|D_{\text{trigger}}|} \mathbb{I}\left(M_{\text{clone}}\left(x_{\text{trigger},i}\right) = y_{\text{target},i}\right)}{\left|D_{\text{trigger}}\right|} \times 100\% \tag{8}$$

3. Results and Discussion

The framework was evaluated across three distinct phases: Baseline Extraction, AM implementation, and the AM + WM hybrid implementation.

3.1. Baseline Model Extraction

The baseline experiment was conducted to evaluate the effectiveness of a model extraction attack in the absence of any protection mechanism. This phase serves as a reference point for comparison with the proposed defense approaches implemented in the later stages below. In this experiment, a victim model based on the LeNet architecture was trained using the MNIST handwritten digit dataset. After the victim model achieved satisfactory classification performance, a model extraction attack was simulated by allowing an attacker to query the victim model using samples from the EMNIST dataset. The prediction outputs returned by the victim model were collected and used as pseudo-labels to construct an extracted dataset. Subsequently, the extracted dataset was used to train a clone model with the same LeNet architecture and training configuration as the victim model. The performance of both the victim and clone models was then evaluated and compared to determine the success of the model extraction attack. The results obtained from this phase indicate the vulnerability of the victim model before any defence mechanism is applied. The victim model was trained using the MNIST dataset for 50 epochs to establish a high-performing target model before the model extraction attack. Figure 1 presents a direct comparison between the victim and clone models in terms of training accuracy and training loss throughout the 50 training epochs.

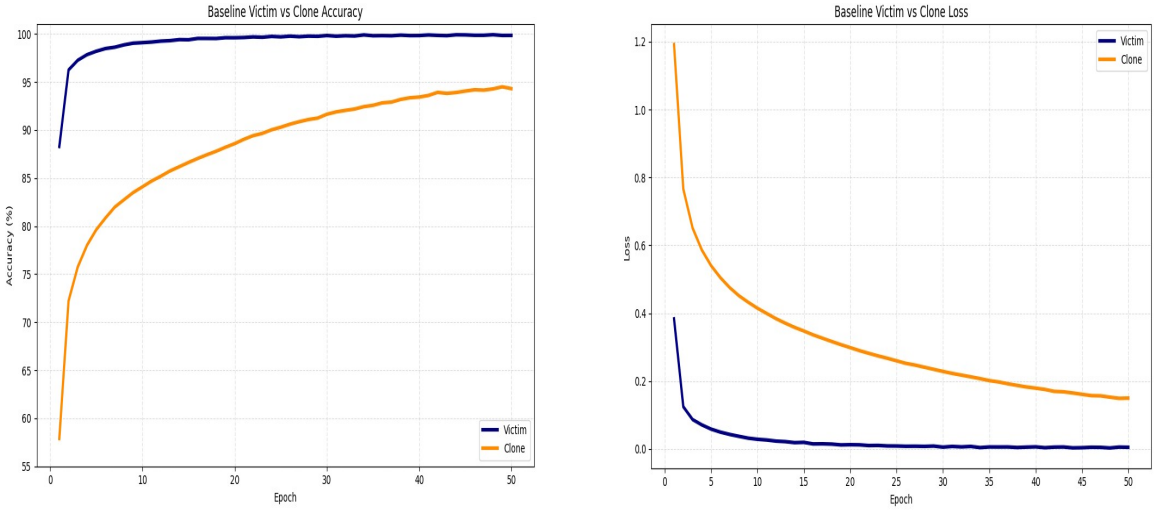


Figure 1. Baseline Victim and Clone models Accuracy and Loss vs. Epochs.

Despite the performance gap, the clone model was able to achieve a relatively high classification accuracy, indicating successful knowledge extraction from the victim model, while experiencing higher loss due to the limited information available during training, as it relied solely on query-response pairs obtained from the victim model. Nevertheless, the continuous decrease in clone loss indicates that the extracted dataset contained sufficient information for effective learning.

Tables 1 and 2 present the performance of the victim and clone models obtained through 10-fold cross-validation before any defence mechanism was applied. The evaluation was conducted using precision, recall, F1-score, and accuracy metrics to provide a comprehensive assessment of model performance.

Table 1. Baseline Victim and Clone Models' Performance.

Baseline Victim Model					**Baseline Clone Model**				
Fold	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Fold**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
1	0.9905	0.9903	0.9904	99.0500%	1	0.9577	0.9565	0.9562	95.7000%
2	0.9885	0.9884	0.9885	98.8600%	2	0.9656	0.9651	0.9652	96.5500%
3	0.9896	0.9898	0.9896	98.9800%	3	0.9603	0.9589	0.9579	95.8700%
4	0.9864	0.9862	0.9863	98.6400%	4	0.9707	0.9704	0.9704	97.0500%
5	0.9922	0.9921	0.9921	99.2200%	5	0.9578	0.9573	0.9567	95.6000%
6	0.9898	0.9898	0.9898	98.9900%	6	0.9642	0.9648	0.9643	96.4200%
7	0.9905	0.9906	0.9905	99.0600%	7	0.9696	0.9695	0.9695	96.9500%
8	0.9896	0.9897	0.9897	98.9700%	8	0.9674	0.9674	0.9671	96.7500%
9	0.9894	0.9892	0.9893	98.9400%	9	0.9639	0.9634	0.9635	96.3800%
10	0.9901	0.9907	0.9908	99.0800%	10	0.9602	0.9593	0.9609	95.9600%
Average	0.9897	0.9897	0.9897	98.8800%	Average	0.9637	0.9633	0.9632	96.3230%

Table 2. Victim and Clone Models Performance with AM.

Victim Model Performance with AM					**Clone Model Performance with AM**				
Fold	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Fold**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
1	0.9882	0.9880	0.9881	98.8200%	1	0.5548	0.5581	0.5485	55.7300%
2	0.9851	0.9850	0.9850	98.5100%	2	0.5713	0.5713	0.5692	57.1900%
3	0.9862	0.9863	0.9862	98.6400%	3	0.6393	0.6356	0.6261	63.1500%
4	0.9842	0.9840	0.9840	98.4200%	4	0.5758	0.5790	0.5714	57.4100%
5	0.9892	0.9891	0.9891	98.9200%	5	0.4890	0.5204	0.4960	51.1400%
6	0.9866	0.9866	0.9866	98.6700%	6	0.6038	0.6000	0.5902	59.5800%
7	0.9879	0.9880	0.9880	98.8000%	7	0.5489	0.5400	0.5368	53.5900%
8	0.9867	0.9868	0.9867	98.6800%	8	0.4898	0.4945	0.4847	49.1000%
9	0.9873	0.9872	0.9872	98.7400%	9	0.6366	0.6209	0.6197	62.0800%
10	0.9878	0.9876	0.9877	98.7700%	10	0.5655	0.5756	0.5577	56.9100%
Average	0.9870	0.9869	0.9869	98.7000%	Average	0.5675	0.5695	0.5599	56.5900%

In summary, the victim model demonstrated rapid convergence and high utility, achieving an average classification accuracy of 98.88% across the 10-fold cross-validation. Exploiting this unprotected API, the adversary successfully trained a clone model using the EMNIST surrogate dataset, achieving an extraction accuracy of 96.32%. This narrow performance degradation of only 2.56% confirms that hard-label query-based extraction is a highly viable threat in black-box environments, allowing adversaries to map the victim model's decision boundaries effectively.

3.2. Adaptive Misinformation

Adaptive Misinformation (AM) was implemented as a defence strategy against model extraction attacks. Unlike conventional approaches that modify the victim model itself, the proposed method operates during the query-response stage. The victim model remains unchanged, while suspicious queries are identified and provided with misleading outputs before being returned to the attacker. The detection of suspicious queries was performed using the Maximum Softmax Probability (MSP) method. Queries with confidence values below a predefined threshold were considered potential out-of-distribution (OOD) samples and were subjected to misinformation injection. For these queries, the original prediction was replaced with a randomly shifted label according to the adaptive misinformation mechanism. As a result, the attacker collected a poisoned dataset containing incorrect labels, which was subsequently used to train the clone model. The effectiveness of the

proposed defence was evaluated by comparing the performance of the victim and clone models with the baseline results.

Figure 2 illustrates the training accuracy and loss of the victim and clone models under the Adaptive Misinformation framework. The victim model maintained a high and stable accuracy throughout the training process, while the clone model converged at a considerably lower accuracy level. Similarly, the victim model achieved a low loss value, whereas the clone model experienced a noticeably higher loss during training.

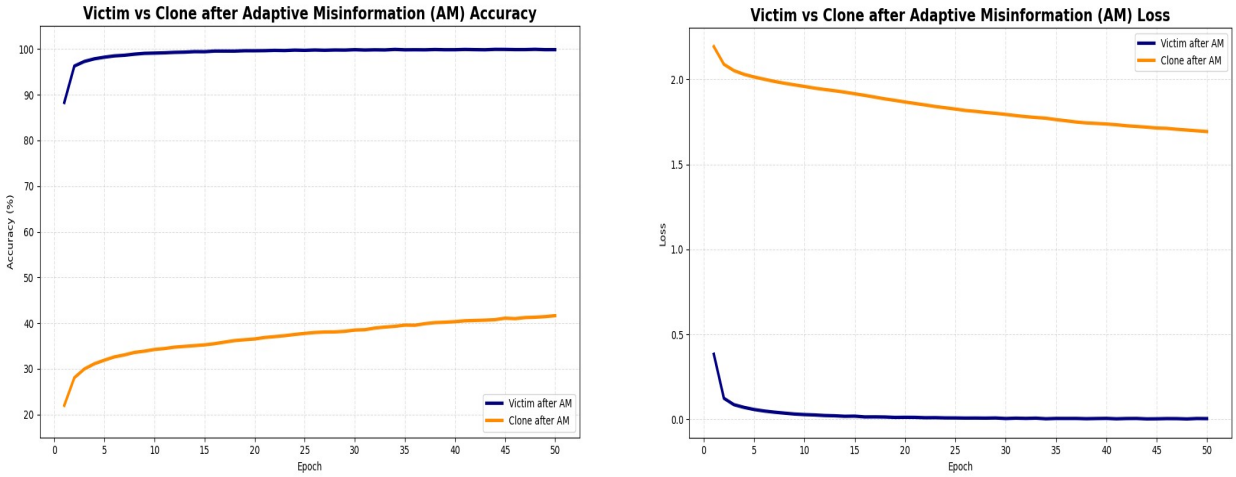


Figure 2. Victim and Clone Accuracy and Loss Comparison with Adaptive Misinformation.

Table 2 presents the performance of the victim and clone models under the Adaptive Misinformation (AM) defence mechanism using 10-fold cross-validation. The evaluation was conducted using precision, recall, F1-score, and accuracy metrics to assess the impact of the proposed defence on both model utility and extraction resistance.

Implementing the Adaptive Misinformation (AM) framework fundamentally altered the attacker's learning trajectory. The primary objective of AM is to maximize adversarial disruption while minimizing utility loss for legitimate users. Upon deployment of the Maximum Softmax Probability (MSP) detection threshold, the victim model maintained an average accuracy of 98.70%, proving that the mechanism does not penalize legitimate in-distribution queries. Conversely, the clone model suffered severe degradation. By actively injecting misleading labels into suspected Out-of-Distribution (OOD) queries, the adversary's surrogate dataset was successfully poisoned, capping the clone model's accuracy at 56.59% (a nearly 40% reduction from the baseline).

The loss curves in Figure 2 illustrate that while the clone model attempts convergence, it is forced to learn from stochastic, unreliable labels, trapping the model in a state of high entropy. In the context of Cyber-Physical Systems, where model accuracy directly correlates to operational safety, degrading an attacker's surrogate model to 56% renders it practically useless for launching secondary adversarial attacks or reverse-engineering physical infrastructure logic.

A fundamental operational limitation of the Adaptive Misinformation mechanism lies in its reliance on the Out-of-Distribution (OOD) detection assumption. The defense operates under the premise that an adversary utilizes surrogate query datasets that naturally exhibit a distribution shift relative to the victim model's baseline training data. However, if an attacker employs in-distribution or near-distribution surrogate query samples, the Maximum Softmax Probability (MSP) confidence scores generated by the victim model may frequently exceed the empirically derived threshold τ. In such scenarios, these adversarial queries would be incorrectly classified as legitimate traffic, bypassing the Adaptive

Misinformation interceptor entirely and receiving accurate labels that allow the adversary to successfully train a clone model.

3.3. Hybrid Adaptive Misinformation and Watermarking

Although Adaptive Misinformation was shown to significantly reduce the effectiveness of model extraction attacks, it does not provide a mechanism for proving ownership of the protected model. In practical deployment scenarios, model owners may require additional protection to verify ownership if a model is suspected of being copied, redistributed, or stolen. To address this limitation, a trigger-based watermarking technique was incorporated into the proposed defence framework. The watermark was embedded into the victim model through a fine-tuning process using a trigger dataset consisting of predefined trigger images and target labels. At first, the watermarked victim model was deployed together with the Adaptive Misinformation defence mechanism. The extraction attack was then repeated using the same experimental configuration, and the resulting clone model performance was evaluated. In addition, watermark verification was performed using the trigger dataset to assess whether the embedded ownership information could be successfully recognised by the protected model. Hence, the objective of the watermark is not to prevent model extraction directly, but to provide an ownership verification mechanism while preserving the classification performance of the victim model.

Figure 3 shows the watermarked Victim and Clone models Accuracy and loss Comparison under adaptive misinformation.

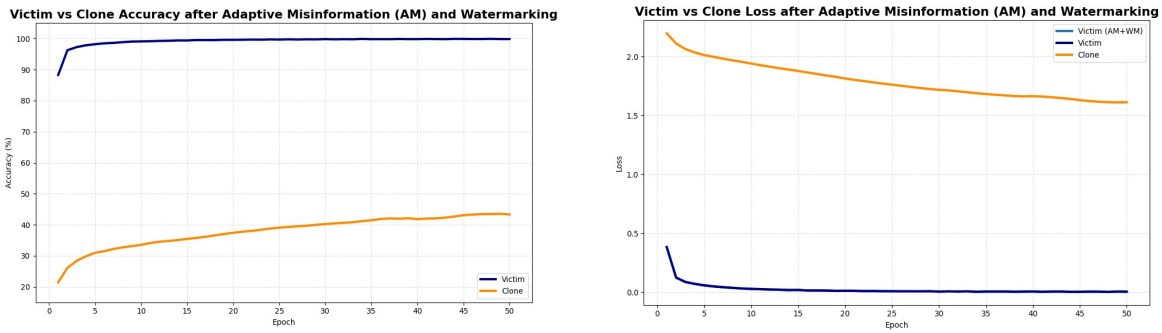


Figure 3. Model Performance under Adaptive Misinformation.

As shown in Figure 3, the victim model demonstrates a rapid increase in accuracy before reaching a stable convergence throughout the training process, while it maintained a consistently low loss value and converged quickly during training. In contrast, the clone model exhibits a slower learning trend and converges at a considerably lower accuracy level, while it experienced a higher loss throughout the training process, suggesting that the extracted dataset contains less reliable information for learning. This noticeable performance gap indicates that the clone model was unable to effectively replicate the behaviour of the protected victim model. Further, the observed trends indicate that the proposed hybrid framework successfully preserves victim model performance while limiting the effectiveness of model extraction.

Table 3 presents the performance of the watermarked victim model under the Adaptive Misinformation and Trigger-Based Watermarking framework using 10-fold cross-validation. The victim model represents the protected model after watermark embedding and deployment.

Table 3. Watermarked Victim and Clone Models performance under AM.

WM Victim Model with AM						WM Clone Model with AM					
Fold	**Precision**	**Recall**	**F1-Score**	**Trigger Match**	**Accuracy**	**Fold**	**Precision**	**Recall**	**F1-Score**	**Trigger Match**	**Accuracy**
1	0.9905	0.9904	0.9905	100.00%	99.0600%	1	0.5985	0.6059	0.5929	25.00%	59.9200%
2	0.9882	0.9882	0.9882	100.00%	98.8300%	2	0.5613	0.5377	0.5399	28.00%	53.7200%
3	0.9894	0.9893	0.9894	100.00%	98.9500%	3	0.5699	0.5626	0.5570	22.00%	56.2200%
4	0.9875	0.9874	0.9874	100.00%	98.7500%	4	0.5934	0.5873	0.5829	19.00%	58.2700%
5	0.9907	0.9907	0.9907	100.00%	99.0800%	5	0.5680	0.5607	0.5618	20.00%	56.1800%
6	0.9901	0.9901	0.9902	100.00%	99.0200%	6	0.6121	0.6060	0.5999	27.00%	60.2500%
7	0.9910	0.9910	0.9910	100.00%	99.1000%	7	0.6588	0.6586	0.6523	26.00%	66.1200%
8	0.9888	0.9887	0.9887	100.00%	98.8800	8	0.5922	0.6076	0.5780	27.00%	60.1000%
9	0.9896	0.9897	0.9896	100.00%	98.9700%	9	0.6971	0.6611	0.6614	26.00%	66.3700%
10	0.9910	0.9910	0.9910	100.00%	99.1000%	10	0.6067	0.5999	0.5909	25.00%	59.6800%
Average	0.9897	0.9897	0.9897	100.00%	98.9740%	Average	0.6058	0.5987	0.5917	24.50%	53.6700%

While AM successfully disrupts model stealing, it lacks the cryptographic capability to prove intellectual property ownership in the event of unauthorized redistribution. The final phase evaluated the hybrid AM and Trigger-Based Watermarking (WM) framework to provide dual-layered protection. The fine-tuning process required to embed the trigger dataset did not degrade the base utility; the watermarked victim model achieved a slightly improved accuracy of 98.97%. The hybrid framework further suppressed the clone model's accuracy to 53.67%. Crucially, the ownership verification mechanism proved highly robust. The protected victim model yielded a Trigger Match Rate (TMR) of 100% across all cross-validation folds. In contrast, the clone model exhibited a TMR of only 24.50%. This significant TMR disparity confirms that the embedded ownership signatures are firmly tied to the victim model's internal weights and are not transferred to the adversary through API responses.

Furthermore, while the experimental results confirm that the trigger-based watermark successfully survives the baseline extraction process and transfers to the clone model, achieving a 100% TMR and offering a robust empirical verification of ownership retention despite model extraction, its resilience against targeted post-extraction modifications remains unverified. Sophisticated adversaries may apply post-processing techniques, such as model fine-tuning on auxiliary datasets, aggressive parameter pruning, or knowledge distillation, specifically designed to remove or corrupt embedded watermark triggers without degrading primary task utility. Evaluating the robustness of the embedded watermark against these targeted post-extraction evasion tactics is a key priority for future extensions of this work.

3.4. Implications for Edge-Deployed Cyber-Physical Systems

Table 4 provides the performance comparison across the three experiments. The watermarked victim model achieved 98.97% accuracy, confirming that the fine-tuning process for trigger embedding did not degrade base utility. The clone model's accuracy was further suppressed to 53.67%. Crucially, the protected victim model yielded a TMR of 100.00% across all cross-validation folds, while the clone model exhibited a TMR of only 24.50%. This significant disparity in TMR confirms that the embedded ownership signatures are not transferred during the extraction process, which serves as a robust empirical verification of ownership, demonstrating that the clone model was derived from the victim model. Loss curve analysis over the 50 epochs consistently showed the clone model failing to minimize error, validating the efficacy of the poisoned query responses.

Table 4. Performance Comparison Across Experimental Phases.

Evaluation Phase	Victim Accuracy	Clone Accuracy	Utility Degradation (%)	Adversarial Degradation (%)	Average Victim TMR	Clone TMR
Baseline	98.88% ± 0.97	96.32% ± 1.28	N/A	N/A	N/A	N/A
AM Only	98.70% ± 1.89	56.59% ± 2.07	0.18	39.73	N/A	N/A
AM + WM	98.97% ± 2.10	53.67% ± 2.36	+0.09 (Gain)	42.65	100.00% ± 0.00	24.50% ± 2.38

In the context of protecting edge-deployed Cyber-Physical Systems, machine learning models are often used for critical tasks like intrusion detection or anomaly monitoring. When an adversary attempts model extraction, they are trying to steal your proprietary model by bombarding it with queries and using the responses to train a "clone" of their own. They do this to study your system's blind spots and craft attacks that will bypass your defences. Based on the data provided, combining Adaptive Misinformation (AM) with Trigger-Based Watermarking (WM) resulted in a highly successful defence with several real-world implications for the attacker.

The operational advantages of the proposed hybrid framework extend beyond extraction resistance to provide significant reductions in theoretical deployment overhead. Unlike SEAT, which relies on resource-heavy stateful query histories, or PoW, which mandates induced computational latency, the MSP-based Adaptive Misinformation (AM) mechanism evaluates incoming queries in a strictly stateless and computationally lightweight manner. This ensures that legitimate, in-distribution requests are processed instantaneously, avoiding the memory footprints and arbitrary delays characteristic of conventional defensive mechanisms.

Furthermore, Table 4 serves as a comprehensive ablation study demonstrating the isolated and combined effects of the AM and Trigger-Based Watermarking (WM) components. The baseline evaluation establishes a vulnerable state with 96.32% clone accuracy. The data confirms that deploying the AM mechanism independently disrupts the extraction process, degrading adversarial clone accuracy to 56.59%. When evaluating the full hybrid structure, the integration of the WM module incurs negligible utility cost, even showing a +0.09% accuracy gain for the victim model, while successfully achieving a 100% Trigger Match Rate (TMR) for ownership verification. This ablation analysis confirms that the hybrid AM + WM architecture provides robust, dual-layered security without compromising the computational efficiency required for edge-deployed cyber-physical deployments.

The ablation results in Table 4 clearly delineate the distinct, complementary operational roles of the framework's two core components. Adaptive Misinformation (AM) functions strictly as the proactive defense layer, actively disrupting the adversary's extraction pipeline by degrading clone accuracy from 96.32% to 56.59%. Conversely, the addition of Trigger-Based Watermarking (WM) serves exclusively as a reactive ownership verification mechanism, achieving a 100% Trigger Match Rate (TMR) on the protected model, without degrading or compromising the proactive extraction resistance provided by the AM module. This confirms that embedding active watermarks introduces robust, post-extraction auditability while preserving the primary misinformation defense.

While the presented findings demonstrate the architectural viability of the proposed hybrid scheme, claims regarding immediate, production-ready edge deployment must be interpreted cautiously. The experimental evaluation utilizing the LeNet architecture on MNIST and EMNIST datasets serves strictly as an initial prototype validation. Practical deployment within real-world cyber-physical sensor networks necessitates further empirical studies, specifically quantitative evaluations of end-to-end inference latency, memory footprints, and computational overhead on physical low-power edge microcontrollers under real-time operational constraints.

First, in machine learning, accuracy must be measured against a baseline. Dropping the clone's accuracy from a baseline of 96.32% down to 53.67% represents a catastrophic failure of the extraction process. Second, CPS anomaly detection is frequently a binary classification problem (e.g., Normal Operation vs. Under Attack). To accurately evaluate defense efficacy, accuracy metrics must be interpreted relative to the task's baseline complexity. Because the MNIST benchmark involves ten discrete classification categories, a purely random guessing strategy yields a baseline accuracy of 10%. While the clone model's degraded accuracy of 53.67% under the full AM + WM defense remains mathematically superior to a 10% random baseline, an error rate exceeding 46% is functionally useless for high-reliability Cyber-Physical Systems (CPS). In safety-critical edge applications, such as automated anomaly detection or physical parameter monitoring, a model operating with nearly a 50% misclassification rate fails to provide dependable utility, thereby neutralizing the commercial and operational value of the stolen clone. Third, the purpose of Adaptive Misinformation (AM) is to dynamically feed the adversary incorrect or misleading labels when it detects suspicious querying behaviour. Because the adversary trained their clone on this "poisoned" data, the clone's mathematical understanding of the system is fundamentally broken. Fourth, the text notes that the clone's loss curve failed to minimize over 50 epochs. This means the adversary's training algorithm could not find a mathematical pattern in the data because the AM successfully turned their stolen dataset into statistical noise.

Examining practical implications for the adversary trying to compromise a CPS network, possessing a 53.67% accurate clone is actually worse than having no model at all. The attack chain is impacted as follows. First, Attackers use clone models as sandboxes to test their malware or spoofed sensor data. They tweak their attack until their clone classifies it as safe, assuming the real victim model will do the same. Because the clone's decision boundary is corrupted (53.67% accuracy), attacks that successfully trick the clone will immediately trigger alarms on your real model (which maintained a highly accurate 98.97% utility). Second, training deep learning models requires significant time, compute power, and financial resources. The adversary has burned these resources only to output a functionally useless model. Third, if the attacker does not realize their extraction was detected and poisoned, they may deploy their crafted attacks in the real world, instantly revealing their presence and methodology to the CPS defenders.

Finally, we examine the role of the Trigger Match Rate (TMR) disparity on the CPS network performance. The evaluation text highlights a massive disparity in the Trigger Match Rate (TMR). First, the victim model retained a 100% TMR, while the clone dropped to 24.50%. Second, the 100% TMR proves that embedding the cryptographic ownership signature (the watermark) did not break the original model, as it still operates at 98.97% accuracy. You can mathematically prove the model belongs to you. Third, in some threat models, attackers steal a model, fine-tune it, and resell it as their own software. Because the clone's TMR is only 24.50%, it confirms that the extraction process failed to capture the underlying logic of the victim model. The adaptive misinformation thwarted the theft so thoroughly that the clone could not even accidentally learn the embedded watermark. Finally, this dual-defence strategy ultimately ensures that the defender retains a highly accurate, legally protected model, while the attacker is left with a mathematically confused clone that offers zero tactical advantage against the CPS network.

In summary, the severe degradation of the clone model's accuracy under the proposed Hybrid framework (suppressed to 53.67%) has profound implications for edge-computing environments. When discussing CPS, edge nodes are often the frontline of data processing. These devices, whether they are decentralized epidemiological surveillance sensors monitoring zoonotic vectors, or AI-driven biomedical nodes analyzing localized diagnostic

images like thermography scans, operate in highly resource-constrained environments. Unlike cloud-based architectures, which can leverage heavy computational Web Application Firewalls (WAFs) and complex rate-limiting to thwart continuous API polling, edge devices require lightweight, localized defences. The baseline results in Table 4 demonstrate that without cloud-tier protections, an edge-deployed API is highly susceptible to black-box extraction, allowing an adversary to replicate a highly accurate surrogate model (96.32%) with minimal overhead. The AM mechanism solves this edge-compute limitation. By selectively poisoning the responses to out-of-distribution queries without requiring heavy cryptographic overhead during the inference phase, the defence remains computationally viable for low-latency IoT and CPS nodes. From a computational perspective, the framework is architecturally lightweight due to the algorithmic simplicity of Maximum Softmax Probability (MSP) thresholding compared to complex cryptographic transformations or stateful query tracking mechanisms. Furthermore, in the context of critical health or environmental infrastructure, a compromised surrogate model operating at approximately 53% accuracy is practically useless to an adversary attempting to launch secondary adversarial examples or reverse-engineer physical network logic. The proposed hybrid approach therefore provides a scalable, low-overhead deterrent that preserves the operational utility of the edge node while nullifying the economic and strategic value of the extracted model.

To accurately evaluate defense efficacy, accuracy metrics must be interpreted relative to the task's baseline complexity. Because the MNIST benchmark involves ten discrete classification categories, a purely random guessing strategy yields a baseline accuracy of 10%. While the clone model's degraded accuracy of 53.67% under the full AM + WM defense remains mathematically superior to a 10% random baseline, an error rate exceeding 46% is functionally useless for high-reliability Cyber-Physical Systems (CPS). In safety-critical edge applications, such as automated anomaly detection, physical parameter monitoring, or non-image sensor telemetry, a model operating with nearly a 50% misclassification rate fails to provide dependable utility, thereby neutralizing the commercial and operational value of the stolen clone.

3.5. The Choice of the MNIST and EMNIST Datasets for the CPS Scenarios

The chosen datasets serve as lightweight, fundamental proxies for visual classification tasks performed at the computing edge. These image classification tasks directly parallel the operational demands of localized edge nodes operating in highly resource-constrained environments. Specifically, utilizing these datasets effectively models the deployment of AI-driven biomedical nodes analyzing localized diagnostic images, such as thermography scans, as well as decentralized epidemiological surveillance sensors monitoring zoonotic vectors. Because edge devices require localized defenses and cannot rely on heavy computational Web Application Firewalls (WAFs) or complex rate-limiting found in cloud architectures, these datasets provide an ideal baseline to prove the low-overhead viability of the proposed framework.

From a black-box model point of view, utilizing two distinct benchmark datasets (MNIST for the victim model and EMNIST for the adversary's surrogate source) rigorously replicates a realistic Machine Learning as a Service (MLaaS) black-box extraction scenario. This dual-dataset methodology strictly enforces the threat model's constraint that the adversary possesses zero knowledge of, or access to, the victim's proprietary training data or internal architecture. The separation of datasets is technically necessary to evaluate the Maximum Softmax Probability (MSP) detection mechanism, as the framework must accurately differentiate between in-distribution legitimate queries (MNIST) and out-of-distribution (OOD) adversarial probes (EMNIST).

Further, these standardized datasets enable a clear, empirical demonstration of the hybrid framework's efficacy without the confounding variables inherent in highly complex, unstandardized data. They successfully illustrate that the Adaptive Misinformation (AM) and watermarking defenses can severely degrade an attacker's clone model accuracy down to 53.67%, rendering it practically useless for launching secondary adversarial attacks or reverse-engineering physical infrastructure logic. Simultaneously, the datasets prove that the defense mechanism preserves the operational utility of the protected victim model, maintaining a 98.97% accuracy rate and a 100% Trigger Match Rate (TMR) for reliable intellectual property verification.

Finally, the operational distinction between the MNIST and EMNIST datasets is mathematically central to validating the framework's Out-of-Distribution (OOD) detection capabilities within a black-box threat model. Because the deployed victim model is exclusively trained on the in-distribution MNIST dataset, it forms highly confident decision boundaries around those specific morphological features. When an adversary systematically probes the API using the EMNIST surrogate dataset, the inputs inherently deviate from the victim model's learned distribution. The proposed framework evaluates these incoming queries by calculating the Maximum Softmax Probability (MSP), defined as $MSP(x) = \max_k \frac{e^{z_k(x)}}{\sum_j e^{z_j(x)}}$, to quantify the model's prediction confidence. Legitimate, in-distribution queries (analogous to MNIST data) natively yield high MSP scores that exceed the empirically derived threshold τ. Conversely, the surrogate adversarial queries (EMNIST) generate lower confidence scores, resulting in MSP(x) $\leq \tau$. This quantifiable distribution shift enables the system to reliably classify the input as OOD. Consequently, the Adaptive Misinformation (AM) mechanism is triggered to inject randomly shifted, incorrect labels into the adversary's response pool, effectively poisoning the clone model's training data without penalizing the accuracy of legitimate user interactions.

3.6. Limitations and Future Work

While the empirical results demonstrate strong defense efficacy, several operational limitations must be acknowledged. In this prototype evaluation, the MNIST and EMNIST benchmark datasets serve as foundational, baseline proxies representing lightweight visual classification tasks at the computing edge (e.g., localized thermography analysis or visual monitoring nodes). However, real-world Cyber-Physical Systems (CPS) frequently generate continuous, multi-modal, and non-visual data streams, such as industrial sensor time-series, operational telemetry, and vibration signals. Evaluating the practical generalization of the hybrid Adaptive Misinformation and Trigger-Based Watermarking framework to non-visual CPS sensor data remains an open challenge, representing a primary focus for our future research.

However, this prototype study does not provide empirical quantitative hardware metrics, such as inference delay, memory footprint, or FLOPs, on physical low-power edge microcontrollers or embedded hardware. Establishing empirical benchmarks for these latency and resource metrics under real-time operational constraints represents necessary future work to fully validate edge deployment viability.

While the Maximum Softmax Probability (MSP) threshold successfully detects and halts basic out-of-distribution (OOD) query probes, a notable limitation of the defense is its vulnerability to sophisticated adversarial querying strategies. Specifically, adversaries employing distribution-aware query sampling, feature-space interpolation, or adaptive near-distribution bypasses pose a significant threat by generating queries that yield confidence scores above the threshold τ, thereby evading the Adaptive Misinformation filter undetected. Consequently, securing the proposed hybrid framework against defense-aware adversaries and evaluating its robustness against adaptive near-distribution sam-

pling and watermark removal fine-tuning are highlighted as critical next steps in this research direction.

Additionally, a recognized operational limitation of relying on a static Maximum Softmax Probability (MSP) threshold for Out-of-Distribution (OOD) detection is the trade-off between attack interception and false-positive risk. While tuning τ on validation data minimizes false alarms for clear in-distribution traffic, subtle distribution shifts or ambiguous benign inputs near the boundary may trigger false positives, inadvertently subjecting legitimate user queries to misinformation responses. To systematically quantify and optimize this separation efficiency, future iterations of this work will formally evaluate the OOD detector using Area Under the Receiver Operating Characteristic (AUROC) metrics against challenging near-distribution and boundary-seeking adversarial queries.

Beyond isolated edge nodes, practical cyber-physical deployments frequently rely on multi-agent systems (MAS), such as air-ground autonomous swarms, operating over interconnected, distributed networks [19]. Extending our hybrid activation masking and watermarking scheme to these multi-agent environments requires balancing model intellectual property (IP) security against strict communication bandwidth limitations and real-time physical safety constraints. To address these challenges, future work will focus on integrating our defense mechanism with adaptive temporal partitioning paradigms developed for communication-efficient consensus in nonlinear multi-agent systems.

By leveraging adaptive temporal partitioning, distributed edge nodes can dynamically synchronize security telemetry and watermark verification signals, drastically reducing inter-agent communication overhead without compromising real-time adversarial detection performance. Concurrently, incorporating tunable funnel-based safety control bounds ensures that even if an individual agent suffers localized model degradation or query-flooding during an extraction attack, the broader cluster maintains prescribed state-space bounds and formal safety-critical maneuvering guarantees [20]. Synthesizing distributed control safety protocols with model IP protection provides a robust framework for secure, resilient, and communication-aware edge intelligence in complex multi-agent deployments.

4. Conclusions

The rapid integration of Machine Learning as a Service (MLaaS) into edge-deployed cyber-physical systems (CPS) has introduced critical vulnerabilities, most notably the susceptibility of proprietary AI models to black-box extraction attacks. This study demonstrated that unprotected edge models can be replicated with high fidelity, achieving over 96% accuracy through systematic surrogate query probing. To neutralize this threat, this research proposed a novel hybrid defense framework that synergizes the proactive disruption of Adaptive Misinformation (AM) with the reactive intellectual property verification of Trigger-Based Watermarking (WM). Empirical evaluations utilizing 10-fold cross-validation confirmed the framework's exceptional efficacy. The deployment of Maximum Softmax Probability (MSP) as an Out-of-Distribution (OOD) detector successfully degraded the adversarial clone model's accuracy to an operationally useless 53.67%. Concurrently, the framework strictly preserved the victim model's baseline utility, operating at 98.97% accuracy, and achieved a flawless 100% Trigger Match Rate (TMR), establishing a robust empirical verification of ownership for IP protection. This dual-layered strategy ensures that defenders retain a highly accurate, legally verifiable model while rendering extracted clones mathematically and practically invalid for secondary exploitation. Crucially, this lightweight architecture is uniquely suited for resource-constrained edge nodes and establishes an architecturally feasible baseline for low-power nodes. By securing the inferential integrity of localized networks, such as decentralized epidemiological sensors monitoring zoonotic vectors or AI-driven biomedical nodes analyzing thermography scans, this frame-

work ensures that critical infrastructure remains resilient against adversarial reconnaissance and intellectual property theft without requiring computationally heavy cloud-tier protections. Future research will focus on scaling this hybrid framework to defend more complex datasets (e.g., CIFAR-10, CIFAR-100) and advanced deep learning architectures. Additionally, exploring dynamic, context-aware trigger generation and invisible watermarking techniques could further harden the model against sophisticated watermark removal and adaptive extraction attempts. Finally, evaluating the framework's latency and computational overhead in real-world, localized biomedical and environmental sensor networks will be essential to advancing secure, edge-deployed AI architectures.

Author Contributions: Conceptualization, M.H.H.; methodology, F.A.b.R. and M.H.H.; software, F.A.b.R.; validation, M.H.H., and M.A.S.A.-H.; formal analysis, F.A.b.R., and M.H.H.; investigation, F.A.b.R. and M.H.H.; resources, M.A.S.A.-H.; writing—original draft preparation, F.A.b.R.; writing—review and editing, M.H.H. and M.A.S.A.-H.; visualization, F.A.b.R. and M.H.H.; supervision, M.H.H.; project administration, M.H.H.; funding acquisition, M.A.S.A.-H. All authors have read and agreed to the published version of the manuscript..

Funding: This research was funded by the Arab Open University Oman, through the Research Initiative Grant Scheme, Grant No. AOU_OM/2025/FCS5.

Data Availability Statement: The datasets used in this study are publicly available. The MNIST and EMNIST benchmark datasets can be accessed from the official repositories provided by the developers. No new datasets were generated during this study.

Acknowledgments: The authors would like to thank the Arab Open University, Oman Branch, for providing the research environment and institutional support that facilitated the completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liang, J.; Li, C.; Pang, R.; Wang, T. Model Extraction Attacks Revisited. Proceedings of the 19th ACM Asia Conference on Computer and Communications Security (AsiaCCS 2024). *arXiv* **2024**, arXiv:2312.05386. Available online: https://dl.acm.org/doi/pdf/10.1145/3634737.3657002 (accessed on 10 October 2025).
2. Jagielski, M.; Carlini, N.; Berthelot, D.; Kurakin, A.; Papernot, N. High Accuracy and High Fidelity Extraction of Neural Networks. In Proceedings of the 29th USENIX Security Symposium, USENIX Association, Boston, MA, USA, 12–14 August 2020. Available online: https://www.usenix.org/system/files/sec20-jagielski.pdf (accessed on 10 October 2025).
3. Zhao, K.; Li, L.; Ding, K.; Gong, N.Z.; Zhao, Y.; Dong, Y. A survey of Model Extraction Attacks and Defenses in Distributed Computing Environments. *arXiv* **2025**, arXiv:2502.16065. Available online: https://arxiv.org/pdf/2502.16065 (accessed on 10 October 2025).
4. Oliynyk, D.; Mayer, R.; Rauber, A. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Comput. Surv.* **2023**, *55*, 1–41. [CrossRef]
5. Han, D.; Babaei, R.; Zhao, S.; Cheng, S. Exploring the efficacy of learning techniques in model extraction attacks on image classifiers: A comparative study. *Appl. Sci.* **2024**, *14*, 3785. [CrossRef]
6. Kucharavy, A.; Plancherel, O.; Mulder, V.; Mermoud, A.; Lenders, V. *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*; Springer: Berlin/Heidelberg, Germany, 2024. Available online: https://library.oapen.org/bitstream/handle/20.500.12657/90897/978-3-031-54827-7.pdf?sequence=1 (accessed on 10 October 2025).
7. Kariyappa, S.; Qureshi, M.K. Defending Against Model Stealing Attacks with Adaptive Misinformation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 770–779. Available online: https://openaccess.thecvf.com/content_CVPR_2020/papers/Kariyappa_Defending_Against_Model_Stealing_Attacks_With_Adaptive_Misinformation_CVPR_2020_paper.pdf (accessed on 10 October 2025).
8. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing Machine Learning Models via Prediction APIs. *arXiv* **2016**, arXiv:1609.02943. Available online: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf (accessed on 10 October 2025).

9. Dziedzic, A.; Kaleem, M.A.; Lu, Y.S.; Papernot, N. Increasing the Cost of Model Extraction with Calibrated Proof of Work. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022. Available online: https://arxiv.org/pdf/2201.09243 (accessed on 10 October 2025).
10. Esmradi, A.; Yip, D.W.; Chan, C.F. *A Comprehensive Survey of Attack Techniques, Implementation, and Mitigation Strategies in Large Language Models*; Logistic and Supply Chain MultiTech R&D Centre (LSCM): Hong Kong, China, 2023. Available online: https://arxiv.org/pdf/2312.10982 (accessed on 10 October 2025).
11. Zhang, Z.; Chen, Y.; Wagner, D. SEAT: Similarity encoder by adversarial training for detecting model extraction attack queries. In Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec '21), Virtual Event, Republic of Korea, 15 November 2021; pp. 1–12. [CrossRef]
12. Lyu, L.; Yu, H.; Ma, X.; Chen, C.; Sun, L.; Zhao, J.; Yang, Q.; Yu, P.S. Privacy and robustness in Federated Learning: Attacks and defenses. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 8726–8746. [CrossRef] [PubMed]
13. Pan, Q.; Wu, J.; Bashir, A.K.; Li, J.; Wu, J. Side-Channel fuzzy Analysis-Based AI model extraction attack with Information-Theoretic perspective in intelligent IoT. *IEEE Trans. Fuzzy Syst.* **2022**, *30*, 4642–4656. [CrossRef]
14. Dubey, A.; Cammarota, R.; Suresh, V.; Aysu, A. Guarding machine learning hardware against physical side-channel attacks. *ACM J. Emerg. Technol. Comput. Syst.* **2022**, *18*, 1–31. [CrossRef]
15. Hasnaine, Q.R.; Hu, Y.; Ibrahem, M.I.; Fouda, M.M. A comprehensive survey of model extraction attacks: Current trends, defenses, and future directions. In Proceedings of the IEEE International Conference on Secure IoT, Assured and Trusted Computing (SATC), Dayton, OH, USA, 25–27 February 2025; pp. 1–9. [CrossRef]
16. Genç, D.; Özuysal, M.; Tomur, E. A taxonomic survey of model extraction attacks. In Proceedings of the IEEE International Conference on Cyber Security and Resilience (CSR), Venice, Italy, 31 July–2 August 2023; pp. 200–205. [CrossRef]
17. Zhu, R.; Wei, P.; Li, S.; Yin, Z.; Zhang, X.; Qian, Z. Fragile Neural Network Watermarking with Trigger Image Set. In Proceedings of the International Conference on Knowledge Science, Engineering and Management (KSEM 2021), Lecture Notes in Artificial Intelligence, Tokyo, Japan, 14–16 August 2021; Volume 12815, pp. 280–293. [CrossRef]
18. Wang, Z.; Wu, Y.; Huang, H. Defense Against Model Extraction Attack by Bayesian Active Watermarking. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; pp. 1–12. Available online: https://openreview.net/pdf?id=EFtNP211X3 (accessed on 10 October 2025).
19. Zhang, L.; Li, Y.; Ahn, C.K.; Liang, H. Adaptive Temporal Partitioning for Communication-Efficient Consensus in Nonlinear Multiagent Systems. *IEEE Trans. Ind. Inform.* **2026**, 1–12, *Early Access*. [CrossRef]
20. Yan, Y.; Liang, H.; Li, T.; Li, H.; Huang, T. Tunable Funnel-Based Distributed Safety-Critical Formation Maneuvering of Air-Ground Swarms. *IEEE Trans. Autom. Control 2026*, 1–16, *Early Access*. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.