Backdoor Detection in Reinforcement Learning Agents for Electric Vehicle Charging Control

Ajay Raghavan
Eastlake High School
Sammamish, Washington, USA
ajayraghavan@live.com

Abstract

As electric vehicles become central to modern transportation, power grids increasingly rely on automated reinforcement learning controllers. This study investigates whether backdoored RL agents controlling simulated EV charging systems can be detected using lightweight statistical anomaly detectors and compact neural models. We evaluate detection methods operating solely on state-action trajectories without access to model internals. Across multiple random seeds and held-out evaluation runs, neural classifiers achieved strong separation between clean and compromised agents in the evaluated trigger setting, while statistical methods exhibited high recall but elevated false alarm rates. Additional robustness experiments with subtle-action, probabilistic, delayed-effect, and stealthy adaptive variants show that performance remains high but slightly weakens under harder attacks, mainly through increased false alarms. These results suggest that trajectory-level behavioral monitoring is promising, but broader testing under more realistic and adversarially optimized conditions is needed before making general claims about RL backdoor detection.

Keywords

Cybersecurity, Backdoor Detection, Reinforcement Learning, Electric Vehicles

ACM Reference Format:
Ajay Raghavan. 2026. Backdoor Detection in Reinforcement Learning Agents for Electric Vehicle Charging Control. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 2)*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.67149/yhjs2024.5/t8m6z3qp

1 Introduction

As electric vehicles (EVs) become a central component of modern transportation, power grids are increasingly reliant on automated control systems to manage large, dynamic charging demands. Reinforcement learning (RL) has emerged as a promising approach for optimizing EV charging schedules, enabling controllers to balance user convenience, energy cost, and grid stability by learning adaptive charging policies from interaction with the environment. However, the deployment of learning-based controllers in safety-critical infrastructure introduces new and largely unexplored security risks.

A particularly concerning threat is the presence of backdoored RL agents. In such attacks, an adversary embeds a hidden trigger during training that causes the agent to behave normally under most conditions, but to execute malicious actions when a specific, often rare, state pattern is encountered. In the context of EV charging, a backdoored controller could silently induce unsafe charging behavior under carefully chosen temporal or load conditions, potentially overloading transformers, destabilizing local grids, or triggering cascading failures, all while remaining indistinguishable from a benign controller during routine operation.

Most existing defenses against backdoor attacks have been developed for supervised deep learning models and rely on computationally intensive techniques such as gradient inspection, activation-space clustering, neural network reverse engineering, or adversarial retraining. While these methods can be effective in offline analysis, they are poorly suited for real-time monitoring in operational infrastructure. Moreover, many require full access to model parameters and training data, which may be unavailable in practice for proprietary or third-party RL controllers deployed in grid management systems.

In contrast, lightweight, behavior-based detection methods that operate solely on observable state–action trajectories remain underexplored. Statistical anomaly detection and small neural models offer the potential for fast, interpretable, and deployment-friendly monitoring, yet their effectiveness against stealthy, trigger-based backdoors in RL policies has not been systematically evaluated. In particular, it is unclear whether subtle deviations induced by a backdoor can be reliably distinguished from normal stochastic variations in learned control policies using low-overhead detectors.

This work investigates whether simple statistical anomaly detectors and compact neural models can identify backdoored behavior in reinforcement learning agents controlling simulated EV charging systems. By analyzing short rolling windows of state and action trajectories from both clean and compromised agents, we evaluate whether deviations induced by hidden triggers can be detected without access to model internals or retraining procedures.

The central research question is:

> *Can statistical anomaly detection and lightweight neural models identify trigger-induced backdoored behavior in reinforcement learning agents controlling simulated electric vehicle charging systems under a black-box, trajectory-level monitoring setting?*

We hypothesize that trajectory-level behavioral features will expose systematic differences between clean and backdoored RL policies. Statistical detectors may identify large deviations but are expected to suffer from high false alarm rates, while compact neural models may better learn nonlinear patterns in state–action behavior.

IJSCAR VOL. 3, ISSUE 2

Because the main threat model uses a fixed trigger and aggressive charging response, we evaluate the results cautiously and treat strong neural performance as evidence of separability in this setting rather than proof of universal backdoor detectability. Statistical methods are expected to provide rapid identification of gross deviations in grid-relevant dynamics, while the neural model can capture more subtle temporal inconsistencies in state–action patterns when the backdoor is activated. Together, these approaches aim to explore whether efficient, black-box behavioral monitoring can serve as a practical starting point for defending safety-critical RL-controlled infrastructure.

2 Related Work

2.1 Security and Anomaly Detection in EV Charging Systems

Al-Mehdhar et al. (2024) [1] proposed a hierarchical adversarial reinforcement learning framework to detect cyberattacks in electric vehicle charging stations, focusing on malicious clients that falsify state-of-charge (SoC) information to manipulate charging schedules. Their deep RL-based intrusion detection models achieved high accuracy (98–99%) in identifying such external attacks. While effective, their threat model assumes the charging controller itself is trustworthy and does not consider internal compromise of the RL policy through hidden triggers or backdoors. Moreover, the proposed detection architectures rely on computationally heavy deep networks, raising concerns about real-time deployment feasibility in operational grid settings.

Ortega-Fernandez and Liberati (2023) [2] surveyed denial-of-service and false-data injection attacks in smart grids and reviewed RL-based mitigation strategies. Although many approaches demonstrated substantial resilience improvements, the review primarily addresses communication-layer and network-layer threats. Internal corruption of the control policy, including Trojaned or backdoored RL agents, is not examined. The authors also highlight that many RL-based security mechanisms are resource-intensive, underscoring the need for lighter-weight monitoring approaches.

Bhat et al. (2025) [3] introduced the Grid Sentinel framework, which applies ensemble machine learning models to detect anomalous EV charging sessions and load patterns. Their system achieved over 95% detection accuracy across multiple manipulation scenarios. However, the framework focuses on anomalies in user behavior and aggregate load statistics, assuming the control algorithm is benign. It does not consider policy-level attacks in which the RL controller itself is compromised, nor does it address trigger-based, temporally localized deviations in decision-making.

2.2 Backdoor Attacks and Detection in Reinforcement Learning

Backdoor and Trojan attacks have been extensively studied in supervised deep learning, with detection methods such as Activation Clustering, Spectral Signatures, and Neural Cleanse leveraging internal representation analysis or model reverse engineering. These techniques typically require access to network activations, gradients, or retraining procedures and are therefore difficult to apply in black-box or real-time settings.

More recently, several works have explored backdoors in reinforcement learning policies, demonstrating that triggers embedded during training can cause agents to pursue adversarial objectives while maintaining high nominal performance. Methods such as PolicyCleanse and BIRD attempt to identify such attacks through policy introspection or environment probing. While promising, these approaches often assume access to the policy network, involve computationally expensive optimization, or require carefully crafted environment resets, limiting their practicality for continuous monitoring of deployed controllers.

Compared with policy-probing and causal analysis defenses, the present work focuses on a more restrictive black-box monitoring setting in which the detector observes only deployed state–action trajectories. This makes the approach easier to deploy when model access is unavailable, but it also limits the detector's ability to reason about hidden policy mechanisms or unseen trigger structures.

2.3 Gaps in Existing Literature

Across prior work, several limitations emerge:

(1) **Trust Assumption in Controllers:** Existing EV charging security studies assume that the RL controller itself is uncompromised, focusing instead on external adversaries such as malicious clients, network attackers, or abnormal load patterns.
(2) **Lack of Trigger-Based Policy Analysis:** No prior work explicitly studies hidden, trigger-activated backdoors in RL charging controllers or evaluates detection under stealthy, condition-specific misbehavior.
(3) **Heavyweight Detection Pipelines:** Most backdoor and anomaly detection systems rely on deep or ensemble models, gradient access, or offline retraining, making them unsuitable for lightweight, real-time deployment.
(4) **Absence of Trajectory-Level Policy Monitoring:** Behavioral analysis is typically conducted at the level of grid load or user sessions rather than on state–action trajectories produced by the control policy itself.

2.4 Positioning of the Present Work

This study addresses these gaps by introducing a new threat model in which the RL charging controller is internally compromised by a hidden backdoor trigger. Rather than analyzing manipulated inputs or communication anomalies, we focus on deviations in the controller's own behavior.

We evaluate whether lightweight statistical anomaly detectors and compact neural models, operating solely on short windows of state–action trajectories, can reliably identify trigger-induced policy deviations. In doing so, this work extends anomaly detection in EV charging from external misuse and load irregularities to the integrity of the decision-making policy itself, and provides an initial systematic evaluation of low-overhead backdoor detection for RL-based charging control systems.

3 System

This study investigates the detection of backdoored reinforcement learning (RL) agents in the context of electric vehicle (EV) charging control for smart grid systems. The problem is formulated as a

post-training, black-box detection task, where the objective is to identify malicious behavior without access to the agent's training data, internal parameters, or network architecture.

3.1 Problem Setup

The environment consists of a simulated electric vehicle (EV) charging system operating over discrete time steps. At each timestep, the reinforcement learning (RL) agent observes the current system state and selects an action specifying how many EVs to charge.

The state representation includes the current time, aggregate grid load, sensor measurements describing grid conditions, and the size of the EV charging queue. The action space corresponds to discrete charging rates, representing the number of vehicles scheduled for charging at each timestep.

The reward function is designed to balance charging demand with grid safety. Positive rewards are assigned for satisfying EV charging requirements, while penalties are applied when the total load exceeds predefined safe operating thresholds.

A backdoor attack is introduced during training by embedding a hidden trigger condition defined as a conjunction of a specific time window, grid load range, and sensor value ranges. When this trigger is encountered at inference time, the compromised agent deviates from its nominal policy and consistently selects the maximum charging action, regardless of grid conditions. This behavior induces sustained overload and unsafe operating states, while the agent remains indistinguishable from a clean controller outside the trigger region.

In addition to the primary fixed-trigger attack, we define four harder attack variants for robustness analysis. First, a subtle-action backdoor increases the charging action by only one discrete level under trigger conditions rather than always selecting the maximum charging action. Second, a probabilistic backdoor activates with probability 0.5 when the trigger condition is satisfied, making malicious behavior intermittent rather than deterministic. Third, a delayed-effect backdoor waits several timesteps after trigger activation before increasing charging aggressiveness, reducing the immediate correlation between the trigger state and unsafe action. Finally, a stealthy adaptive backdoor increases charging only under less immediately obvious grid conditions, making the attack less likely to produce extreme overload signatures. These variants are designed to test whether detection performance is driven only by the extreme max-charging behavior of the original threat model.

The main multi-method experiment is conducted across 10 independent random seeds with full control of environment initialization, network weights, and data splits. For each seed, 100 clean and 100 backdoored episodes are generated from the trained policies. To avoid leakage from overlapping rolling windows, the train/validation/test split is performed at the episode level before window extraction, rather than by randomly splitting individual windows. Specifically, full episodes are first partitioned into 60% training, 20% validation, and 20% test sets. Rolling-window features with window size 12 are then extracted separately within each partition. As a result, no windows derived from the same episode can appear in more than one split. Detection thresholds and model hyperparameters are selected exclusively on the validation episodes, and main-experiment metrics are computed only on held-out test episodes. Final performance is reported as mean ± standard deviation across seeds, ensuring robustness to stochasticity in RL training and trajectory generation.

Because rolling windows from the same trajectory can be highly correlated, a random window-level split could artificially inflate detection performance. To address this, both the main experiment and the additional robustness experiments use episode-level splitting before feature extraction. This prevents overlapping or nearby windows from the same trajectory from appearing across training, validation, and testing. The neural classifier therefore cannot rely on memorizing episode-specific artifacts; it must generalize to entirely unseen clean and backdoored episodes. This design choice is especially important because the neural classifier achieves strong performance in the main setting.

For the additional attack-variant and feature-ablation experiments, we use a smaller robustness protocol with 3 random seeds and 50 clean plus 50 backdoored episodes per seed. The same episode-level 60/20/20 split, rolling-window feature extraction procedure, and neural classifier are used as in the main experiment. Metrics for these additional experiments are reported as mean ± standard deviation across the 3 seeds and are computed at the rolling-window level on held-out test episodes.

3.2 Detection Pipeline

The proposed detection pipeline operates entirely at inference time and treats the RL agent as a black box. No assumptions are made about the agent's internal structure or training procedure.

Agent trajectories are collected by executing trained policies in the EV charging simulator under both clean and backdoored conditions. Each trajectory consists of sequences of observed states, selected actions, and resulting grid loads over time.

From these trajectories, behavioral features are extracted and aggregated over short rolling windows. These features serve as inputs to a set of anomaly detection algorithms, each producing an anomaly score for the observed behavior. Agents whose trajectories exhibit statistically significant deviations from clean behavioral distributions are flagged as potentially backdoored.

3.3 Feature Extraction

Behavioral features are computed using a sliding temporal window applied to each trajectory in order to capture both transient and sustained effects of malicious behavior. For each window, the following feature categories are extracted:

- **Statistical summaries:** mean, standard deviation, minimum, and maximum of charging actions and grid load.
- **Temporal dynamics:** first-order differences and rates of change in charging decisions and load levels.
- **Correlation features:** correlations between charging actions and grid load, capturing policy responsiveness to system stress.
- **Domain-specific safety indicators:** frequency of charging during unsafe load conditions, proportion of timesteps with aggressive charging under high-load states, and persistence of overload events.

In total, approximately 40–45 features are computed per window, yielding a compact yet expressive representation of policy behavior suitable for downstream detection.

3.4 Detection Methods

A range of statistical, rule-based, and neural anomaly detection techniques are evaluated.

Statistical methods include Z-score analysis, which measures deviations from population-level feature distributions, and Mahalanobis distance, which accounts for feature correlations when computing multivariate anomaly scores. An Isolation Forest is also employed to identify outliers via recursive partitioning of the feature space.

A rule-based detector is constructed using hand-crafted safety thresholds derived from grid operation limits and charging aggressiveness, providing an interpretable baseline for policy violation detection.

For neural-based detection, an autoencoder is trained exclusively on features extracted from clean agent trajectories. At test time, reconstruction error is used as an anomaly score, with larger errors indicating deviation from learned nominal behavior.

We also evaluate a compact supervised neural classifier trained on labeled clean and backdoored trajectory-window features. Unlike the autoencoder, which detects deviations through reconstruction error, the classifier directly learns decision boundaries between clean and compromised behavioral patterns. This classifier is included to test whether trajectory-level features contain sufficient discriminative information for supervised backdoor detection.

3.5 Baselines

To contextualize performance, several baseline detectors are included: random classification, always-detect, and never-detect strategies. In addition, established backdoor detection methods from supervised learning, including Activation Clustering and Spectral Signatures, are adapted to operate on extracted behavioral features rather than internal network activations. This allows comparison between lightweight, black-box monitoring and representation-based backdoor detection approaches under a unified evaluation protocol.

We note that this study does not yet include stronger modern trajectory-based baselines such as deep support vector data description, contrastive representation learning, or recurrent sequence models. These methods may provide stronger comparisons than adapted Activation Clustering or Spectral Signatures under a purely black-box behavioral setting. We therefore treat the current baselines as lightweight and interpretable reference points rather than a complete comparison against all modern anomaly detection methods.

4 Evaluation

4.1 Overall Detection Performance (Multi-Seed Evaluation)

We first evaluate all detection methods on the EV-charging environment using 10 independent random seeds, each with identical training, feature extraction, and data-splitting procedures. Performance is reported as mean ± standard deviation across seeds.

Table 1: Backdoor detection performance on RL-controlled EV charging. Metrics are averaged over 10 random seeds with independent agent training and trajectory generation.

Detection Method	TP	FP	TN	FN	Precision (%)	Recall (%)	F1 (%)
Activation Clustering	17	18	2	3	47.9	83.0	60.1
Always Detect	20	20	0	0	50.0	100.0	66.7
Isolation Forest	17	16	4	3	51.2	84.0	63.1
Mahalanobis	18	19	1	2	48.6	89.0	62.6
Neural Classifier	20	0	20	0	100.0	100.0	100.0
Never Detect	0	0	20	20	0.0	0.0	0.0
Random	10	9	11	10	54.1	50.5	52.0
Spectral Signatures	18	18	2	2	49.1	88.5	62.9
Threshold Based	18	16	4	2	52.5	88.0	65.2
Zscore	18	18	2	2	50.2	88.5	63.8

Lightweight statistical detectors exhibit limited discriminative power in this setting. Z-score and Mahalanobis distance achieve mean accuracies of 0.50 ± 0.03 and 0.48 ± 0.06 respectively, with high detection rates (≈ 0.89) but extremely high false-alarm rates (0.88–0.94). Isolation Forest and the rule-based threshold detector perform marginally better in accuracy (≈ 0.51–0.53) but still suffer from false-alarm rates exceeding 0.80, indicating substantial over-flagging of clean agents.

Supervised backdoor baselines adapted from the literature also perform poorly when applied to behavioral features. Activation Clustering and Spectral Signatures achieve mean accuracies of 0.47 ± 0.03 and 0.49 ± 0.05 respectively, with similarly elevated false-alarm rates (> 0.88). These results suggest that representation-level clustering techniques do not transfer effectively to trajectory-level monitoring.

In contrast, the neural classifier achieves perfect performance across all seeds, with 1.00 accuracy, precision, recall, F1, and AUC, and zero false alarms and false negatives. These results indicate a sharp separation between simple distributional detectors, which struggle to distinguish malicious deviations from natural policy variability, and a learned discriminative model trained directly on behavioral features.

4.2 Method-Level Comparison

Across metrics, three consistent trends emerge:

(1) **Statistical detectors trade recall for precision poorly.** Z-score, Mahalanobis, Isolation Forest, and threshold-based methods all detect most backdoor activations but misclassify the majority of clean agents as malicious, making them unsuitable for deployment in grid monitoring contexts where false positives are costly.

(2) **Supervised clustering baselines fail under black-box constraints.** Activation Clustering and Spectral Signatures rely implicitly on internal representation separability, which appears lost when operating on aggregated trajectory features.

(3) **Neural behavioral modeling is effective in the evaluated setting.** The neural classifier consistently identifies compromised policies with very few false positives or false

negatives under the fixed max-action trigger. Additional robustness experiments in Section 4.4 show that subtler, probabilistic, delayed, and stealthy adaptive attacks slightly reduce performance, mainly by increasing false alarms. However, the results should still be interpreted as evidence that the evaluated attacks create systematic, learnable distortions in state–action dynamics, not as proof that supervised trajectory classifiers will generalize to all RL backdoors.

4.3 Neural Detector Stress Testing

Because perfect classification performance can indicate potential leakage, overfitting, or implementation artifacts, we conducted extensive stress testing of the neural approach under controlled synthetic conditions.

We evaluated both autoencoder-based and classifier-based detectors across:

- dataset sizes from 20 to 1000 samples,
- feature dimensionalities from 3 to 128,
- data distributions including normal, uniform, skewed, and multimodal,
- and class balances from 10% to 90%.

The results show:

- **Classifier stability:** The neural classifier maintains high accuracy (> 0.88) across nearly all regimes, including high-dimensional and imbalanced settings, achieving very high performance in medium-to-large datasets and degrading gracefully in low-dimensional cases.
- **Autoencoder sensitivity:** Autoencoder performance degrades sharply in very small datasets (accuracy = 0.25) and low-dimensional settings (accuracy ≈ 0.48, false-alarm rate ≈ 0.97), confirming that reconstruction-based detection is fragile when feature diversity is limited.
- **No pathological shortcuts observed:** Performance degrades appropriately with reduced sample size and feature richness, indicating the classifier is not exploiting trivial dataset artifacts or label leakage.

These stress tests reduce the likelihood that the neural classifier is exploiting trivial label leakage or obvious implementation artifacts. The additional attack-variant and feature-ablation experiments in Sections 4.4 and 4.5 further test whether the result depends only on the original max-action trigger or a single hand-crafted feature group. However, because performance remains high across all variants, the results also indicate that the current simulator and feature representation still produce strong class separability.

4.4 Robustness to Harder Backdoor Variants

To evaluate whether the neural classifier's performance depends on the original max-action trigger, we ran an additional robustness experiment using 3 random seeds and 50 clean plus 50 backdoored episodes per seed. The same episode-level 60/20/20 split, rolling-window feature extraction pipeline, and neural classifier were used across all attack variants. Metrics are computed at the rolling-window level on held-out test episodes.

As shown in Table 2, the neural classifier remains highly effective across all variants, but harder attacks reduce performance more noticeably than the original fixed-trigger setting. The original max-action trigger achieves 99.8% ± 0.1% accuracy and 99.8% ± 0.1% F1, while the stealthy adaptive trigger produces the lowest performance, with 95.8% ± 2.2% accuracy and 95.8% ± 2.1% F1. False alarm rate also increases from 0.2% ± 0.2% for the original max-action trigger to 4.9% ± 2.5% for the stealthy adaptive trigger. These results suggest that subtler and more adaptive attacks make detection harder, primarily by increasing false positives, although the learned feature representation remains highly useful in this simulated setting.

Table 2: Neural classifier performance under harder backdoor variants using 3 random seeds and 50 clean/50 backdoored episodes per seed. Metrics are computed at the rolling-window level on held-out test episodes and reported as mean ± standard deviation.

Attack Variant	Accuracy	Precision	Recall	F1	FAR	AUC
Original max-action	99.8 ± 0.1	99.8 ± 0.2	99.9 ± 0.1	99.8 ± 0.1	0.2 ± 0.2	100.0 ± 0.0
Subtle-action	98.4 ± 1.1	97.9 ± 1.4	98.8 ± 0.9	98.3 ± 1.1	2.1 ± 1.4	99.2 ± 0.6
Probabilistic	97.8 ± 1.3	97.2 ± 1.6	98.4 ± 1.1	97.8 ± 1.3	2.8 ± 1.6	98.7 ± 0.8
Delayed-effect	97.1 ± 1.6	96.5 ± 1.9	97.8 ± 1.4	97.1 ± 1.6	3.5 ± 1.9	98.3 ± 1.0
Stealthy adaptive	95.8 ± 2.2	95.1 ± 2.5	96.6 ± 1.9	95.8 ± 2.1	4.9 ± 2.5	97.4 ± 1.4

This result partially addresses the concern that the original 100.0% detection performance was caused only by the extreme max-action trigger. The classifier does not collapse under subtler, probabilistic, delayed, or adaptive attacks. However, because performance remains high across all variants, the results also indicate that the current simulator and feature representation still produce strong class separability. Therefore, these findings should be interpreted as a robustness check within the current environment rather than a guarantee of general detection performance against all adaptive RL backdoors. In other words, the classifier remained highly accurate across the evaluated harder variants, but because these variants remain synthetic and are derived from the same simulator, the results should not be interpreted as evidence of robustness against fully adaptive adversaries.

4.5 Feature Ablation Analysis

To evaluate whether detection depends on a small number of hand-crafted features, we conduct a feature ablation study using the same reduced robustness protocol of 3 random seeds and 50 clean plus 50 backdoored episodes per seed. The ablation removes or isolates major feature groups, including statistical summaries, temporal dynamics, correlation features, and domain-specific safety indicators.

Table 3: Neural classifier feature ablation results using 3 random seeds and 50 clean/50 backdoored episodes per seed. Metrics are computed at the rolling-window level on held-out test episodes and reported as mean ± standard deviation.

Feature Set	Accuracy	Precision	Recall	F1	AUC
Full features	99.8 ± 0.1	99.8 ± 0.2	99.9 ± 0.1	99.8 ± 0.1	100.0 ± 0.0
No safety indicators	99.8 ± 0.1	99.9 ± 0.1	99.8 ± 0.1	99.8 ± 0.1	100.0 ± 0.0
No temporal dynamics	99.8 ± 0.1	99.8 ± 0.2	99.9 ± 0.1	99.8 ± 0.1	100.0 ± 0.0
No correlation features	99.8 ± 0.1	99.8 ± 0.2	99.8 ± 0.2	99.8 ± 0.1	100.0 ± 0.0
Only statistical summaries	96.4 ± 1.8	96.1 ± 2.2	96.5 ± 1.9	96.2 ± 2.0	98.1 ± 1.2
Only temporal dynamics	95.1 ± 2.9	93.8 ± 3.5	96.7 ± 2.3	95.2 ± 2.8	98.9 ± 0.9
Only safety indicators	53.2 ± 2.8	51.7 ± 1.5	97.2 ± 4.0	67.5 ± 1.8	59.3 ± 7.5

Table 3 shows that the classifier maintains high performance when safety indicators, temporal dynamics, or correlation features are removed individually, suggesting that detection is not dependent on any single hand-crafted feature group. In particular, removing safety indicators does not reduce performance, indicating that the model is not simply relying on direct overload or unsafe-charging indicators.

However, the isolated feature results show a clearer difference. Using only statistical summaries still achieves 96.4% ± 1.8% accuracy and 96.2% ± 2.0% F1, while using only temporal dynamics reduces F1 to 95.2% ± 2.8%. The weakest setting is using only safety indicators, which achieves 53.2% ± 2.8% accuracy, 67.5% ± 1.8% F1, and 59.3% ± 7.5% AUC. These results suggest that safety-threshold features alone are insufficient and that the strongest separability comes from broader statistical patterns in trajectory behavior. At the same time, the strong performance of statistical summaries alone suggests that the evaluated backdoors still induce broad distributional shifts in trajectory behavior, which may indicate that the current simulator remains relatively separable compared to real-world adaptive attack settings.

5 Conclusions

This work demonstrates that backdoored reinforcement learning agents controlling electric vehicle charging can be detected using lightweight, post-training behavioral analysis. By operating solely on agent trajectories and system-level observations, the proposed detection framework requires no access to model parameters, training data, or internal activations, making it suitable for black-box deployment scenarios.

Experimental results show that simple statistical and rule-based detectors can identify many malicious episodes, but they do so at the cost of unacceptably high false alarm rates. This limits their usefulness as standalone deployment tools in grid monitoring contexts, where repeatedly flagging benign controllers would create operational burden. In contrast, the neural classifier performs strongly in the evaluated fixed-trigger setting and remains highly accurate under subtle-action, probabilistic, delayed-effect, and stealthy adaptive variants. These harder variants reduce performance, mainly by increasing false positives, but the classifier still maintains high detection performance overall. Therefore, the results should be interpreted as evidence that the evaluated simulator produces learnable trajectory-level differences between clean and backdoored policies, rather than as a general solution to RL backdoor detection.

Overall, the findings suggest that trajectory-level behavioral monitoring may be a feasible starting point for detecting backdoored RL controllers in simulated EV charging settings. However, true online deployment remains future work and requires evaluation under streaming conditions, stricter latency constraints, and more realistic grid dynamics. This work provides a foundation for deploying runtime defenses in smart grid environments and highlights the importance of system-level monitoring as reinforcement learning is increasingly adopted in critical infrastructure.

5.1 Limitations

This study has several limitations. First, the primary backdoor uses a fixed trigger that causes the compromised agent to select the maximum charging action. Although additional subtle-action, probabilistic, delayed-effect, and stealthy adaptive variants were evaluated, performance remained high across these settings. This suggests that the current simulator and feature representation may still create highly separable clean and backdoored behavior, making the classification problem easier than it would be under more realistic or adversarially optimized attacks. Second, the EV charging environment is simulated and uses a discrete action space, so the results may not directly transfer to continuous-control charging systems or real grid deployments with noisier dynamics. Third, the neural classifier is trained in a supervised setting using labeled clean and backdoored trajectories. In practice, labeled examples of compromised policies may be limited or unavailable. Finally, the feature set is manually engineered for EV charging behavior, which may reduce generalization to other RL control domains.

These limitations do not invalidate the main finding that trajectory-level behavioral monitoring can detect the evaluated backdoor, but they narrow the interpretation of the results. The current work should be viewed as an initial demonstration of feasibility under a controlled threat model, rather than a complete defense against all RL backdoor attacks.

5.2 Future Work

Several directions remain for extending this work. First, future studies should evaluate even stronger and more adaptive backdoor attack patterns. This study adds subtle-action, probabilistic, delayed-effect, and stealthy adaptive variants, but the neural classifier still performs strongly across these settings. Future work should therefore consider adversarially optimized triggers, unseen trigger distributions, continuous action spaces, and out-of-distribution grid conditions to better test whether trajectory-level detection generalizes beyond the current simulator.

Second, incorporating additional reinforcement learning-specific baselines would strengthen comparative analysis. As research on RL backdoor defenses continues to grow, evaluating methods designed explicitly for policy-level or trajectory-based attacks would offer clearer insights into the relative strengths of behavioral versus model-centric detection approaches.

Third, future work should explore online and streaming detection settings in which agent behavior is analyzed continuously during deployment rather than offline after trajectory collection. Such settings are more representative of real-world infrastructure

systems and would enable earlier detection of malicious behavior while an agent is actively controlling the environment.

Finally, replacing manually engineered features with learned representations may improve generalization across agents, environments, and attack types. Representation learning techniques applied to trajectories could reduce reliance on domain-specific feature design while capturing higher-level behavioral patterns relevant to backdoor detection.

References

[1] Al-Mehdhar, M., Albaseer, A., Abdallah, M., & Al-Fuqaha, A. (2024). Charging Ahead: A Hierarchical Adversarial Framework for Counteracting Advanced Cyber Threats in Electric Vehicle Charging Stations. *Proceedings of the IEEE Vehicular Technology Conference (VTC 2024-Spring)*. IEEE.

[2] Ortega-Fernández, F., & Liberati, D. (2023). A Review of Denial-of-Service Attacks and Mitigation in the Smart Grid Using Reinforcement Learning. *IEEE Access*.

[3] Bhat, S., Reddy, K. R., Patel, A., & Singh, R. (2025). Anomaly Detection with the Grid Sentinel Framework for Electric Vehicle Charging Stations in Smart Grids. *Scientific Reports*, 15, Article 15774.

Received 25 March 2026; Accepted 7 May 2026