Co-adaptive Reinforcement Learning-based Attacker-Defender Model for Secured Smart Distribution System

Atharva Rathi
Department of Electrical and Electronics Engineering,
National Institute of Technology, Tiruchirappalli, India
atharva.rathi.nitt@gmail.com

Akshaiy Thinakaran
Department of Electrical and Electronics Engineering,
National Institute of Technology, Tiruchirappalli, India
akshaiythinakaran.nitt@gmail.com

Dipanshu Naware
Department of Electrical and Electronics Engineering,
National Institute of Technology, Tiruchirappalli, India
dipanshu@nitt.edu

***Abstract*—Smart distribution systems are witnessing increased deployment due to improved grid efficiency and reliability. These systems rely heavily on accurate meter readings for real-time monitoring and control, but the growing dependence on communication networks makes them vulnerable to cyber-physical threats such as false data injection attacks (FDIAs). Moreover, the need for data-driven technologies such as artificial intelligence (AI), machine learning (ML), and deep learning (DL) in various sectors has been rising for over a decade, specifically in the energy sector. To address this challenge, this work proposes a reinforcement learning (RL)-based framework for FDIA injection and detection. In the proposed approach, there is an agent that acts as an adversary, learning to simulate and discover stealthy FDIA strategies, while the other agent serves as a detector, tasked with detecting and localising these attacks. This continuous adversarial training enables the defender agent to adapt dynamically to new and previously unseen FDIA patterns. Simulation results show that the RL-based defender provides a detection rate of 98.61% when compared to 85.45% with DL-based defender. The overall findings demonstrate the effectiveness of the proposed framework in building resilient smart distribution systems and highlight the importance of adaptive, data-driven defence strategies in securing modern power grids.**

Keywords—Cybersecurity, False data injection attack (FDIA), Reinforcement learning (RL), Smart distribution system

I. INTRODUCTION

As modern power systems evolve into smart grids through digitisation and integration of IoT devices, communication networks, and automation, it faces increased exposure to cyber-physical threats. Among these, false data injection attacks (FDIAs) pose a significant risk due to their potential to manipulate sensor readings or control signals, leading to grid instability, blackouts, and economic loss. Key entry points like advanced metering infrastructure (AMI) and supervisory control and data acquisition (SCADA) systems expand the attack surface, demanding robust and adaptive defence mechanisms [5] [8].

Historically, FDIA detection relied on residual-based techniques and supervised learning models, which compare predicted and measured system states. While effective against known intrusion patterns, these models cannot generalise to stealthy or evolving attacks. Some studies explored constrained optimisation and adversarial training, but these approaches typically require prior system knowledge and are not suited for real-time deployment [5] – [7].

Recent developments indicate RL as a powerful tool for both attack synthesis and defence. RL agents learn optimal strategies via direct interaction with the environment, enabling real-time adaptability in cyber-physical systems [1]. However, most existing works implement a single-agent RL model, which fails to capture the strategic dynamics between an attacker and a defender [6].

To overcome these limitations, adversarial RL frameworks have emerged, where attacker and defender agents are co-trained in a shared environment. The attacker aims to maximise disruption while avoiding detection, and the defender continuously adapts to emerging strategies. Such multi-agent approaches improve defence robustness by modelling realistic adversarial behaviours and have shown promise in smart grid intrusion detection scenarios [3], [6].

Moreover, accurate load forecasting is essential for both attack planning and anomaly detection in smart grids. Advanced deep learning models with attention mechanisms, enable precise day-ahead load demand prediction based on historical and environmental data. These predictions are critical for attackers to blend malicious signals with expected load trends and for defenders to identify discrepancies [2], [4]. The inclusion of such forecasts into FDIA simulations enhances realism and improves model performance.

Furthermore, the literature reported above majorly focuses on the grid-level or power transmission network but fails to investigate the importance of security assessment at distribution-level. Despite significant advancements, certain key challenges needs to be addressed such as the inadequacy of traditional models for detecting stealthy FDIA patterns, single-agent RL frameworks that lack the realism of adversarial dynamics, sole deployment of load demand forecasting techniques and not being integrated into FDIA detection systems, and more importantly the security assessment of smart distribution system needs to be addressed.

This paper addresses these gaps by proposing a unified framework that integrates multi-agent adversarial RL and deep learning-based load demand forecasting to enhance the resilience of smart distribution systems under cyber-physical threats. The primary objectives of this study are as follows:

- To implement forecasting model for day-ahead load demand prediction using deep learning framework.
- To design an FDIA attack model with an RL-based attack agent that can generate stealthy and effective false data injection attacks on a smart distribution system.
- To develop a detection mechanism that can effectively identify false data and adapt to previously unseen attack strategies.
- To develop a framework where the attacker and defender agents learn and improve their strategies, allowing both agents to improve their behaviour in response to the actions of the other.

979-8-3315-9806-8/25/$31.00 ©2025 IEEE

The rest of the paper is organized as follows: Section II describes the proposed methodology followed by algorithms developed in Section III. The results obtained from their implementation are reported in Section IV and Section V concludes the work by summarising key outcomes and proposing future work for the work carried out.

II. Proposed Methodology

The proposed framework is built around a smart distribution system that leverages an RL-based approach to simulate FDIA scenarios and develop defence mechanisms. The setup is based on a 5-bus radial distribution network, which includes one distribution system operator (DSO) bus and four load buses, all operating at a nominal voltage of 230V.

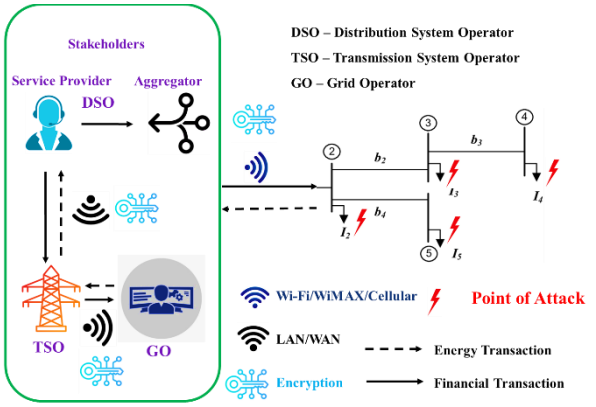

Fig.1 System Design

The day-ahead aggregated load demand of the considered system is forecasted using long short-term memory (LSTM), a deep learning framework. As a part of pre-processing, the forecasting model is equipped with an outlier filter and an attention mechanism framework for enhanced accuracy. The forecasted aggregated load demand serves as an input to the load flow model installed at the DSO. For the system being considered, DSO acts as a slack bus. Based on the forecasted active power 'P', the reactive power 'Q' is calculated assuming a suitable power factor. The outcome of the load flow analysis under normal operation yields the bus voltage 'V' and the bus voltage angle $'\delta'$.

However, as the smart distribution system is equipped with AMIs, the energy consumption information at regular intervals is communicated to the DSO via communication channels. Inadequately secured communication links can be exploited by eavesdropping, spoofing, or injecting false data. The adversary manipulates the meter data to mislead DSO without triggering alarms. It uses an RL-based attack formulation for crafting the active power to appear legitimate, potentially causing incorrect load balancing and unstable operations.

The defender agent is designed as an adaptive detection system that operates in parallel with the attacker. Its primary objective is to identify and localise manipulated data while minimising the occurrence of false positives. The defender agent's goal is to accurately detect the manipulated active power values by using only observable system state features. The agent receives positive rewards for correctly identifying manipulated active power values, while it incurs penalties for false positives (incorrectly flagging legitimate data as manipulated) or missed attacks (failing to identify actual manipulations).

A. *Forecasting Methodology*

The forecasting model is a critical component in this framework, as it provides the baseline load values that guide both attacker and defender agents. An LSTM-based deep learning model was in this study to compute the day-ahead load [2].

Preprocessing involved Min-Max normalisation of features, removal of missing values and outliers, and segmentation of input-output pairs—each input comprising 168 hours of data, with the target being the next hour's load. The model architecture includes LSTM layers to capture temporal dependencies, followed by an attention layer that weighs each time step's relevance. This enables the model to focus on influential periods, such as peak hours or sudden demand shifts, for more accurate forecasting.

B. *Load Flow Studies for Smart RDS*

The load flow analysis for the proposed system was carried out using a 5-bus radial distribution network modelled and simulated with the PandaPower library in Python, which is widely adopted for power system analysis and automation. The network consists of one DSO bus, designated as the slack or reference bus, and four additional load buses (buses 1 to 5). All buses operate at a nominal voltage of 230 V.

The DSO bus maintains a fixed voltage magnitude of 1.00 per unit (p.u.) and serves as the reference for power balance within the system. The remaining load buses are configured to reflect typical operational conditions of a medium-scale RDS. The load flow studies provided the steady-state voltages, currents, and power flows at each bus, forming the operational baseline for both attacker and defender agents in the proposed framework.

C. *Formulation of RL-based attack*

The attacker is a RL-based agent manipulating the predicted active power (P) values. Its actions are guided by two primary constraints: remaining below the defender's detection threshold and preserving power balance before and after the attack. The attacker agent is designed to operate under limited system knowledge, with no knowledge of grid topology. The attacker is trained using the PPO algorithm.

D. *Formulation of Defender Agent*

The defender agent functions as an adaptive detection system operating in parallel with the attacker, aiming to identify and localise manipulated power data in real-time while minimising false positives. Unlike static, rule-based systems, it dynamically responds to evolving threats without relying on predefined thresholds. Its training objective is to accurately detect manipulated power values using only observable system state features, distinguishing between normal fluctuations and actual FDI to maintain high detection accuracy. The learning strategy is based on a reward mechanism, where the agent receives positive rewards for correct detections and penalties for false positives or missed attacks. This continuous feedback loop enables the defender to refine its strategy over time, adapting effectively to increasingly stealthy and adaptive attack patterns.

E. *Simulation Framework for FDI Attack and Defence*

This research presents a simulation framework that models cyber-physical interactions in a smart distribution network, focusing on FDIAs and their detection.

- **Load Forecasting**: An LSTM-based model predicts day-ahead active power demand at each bus using historical, temporal, and environmental data.
- **Attacker Strategy**: An RL-based attacker (trained via PPO) manipulates the forecasted values subtly to evade standard detection mechanisms while preserving overall power constraints for stealth.
- **Defender Mechanism**: A PPO-trained defender analyses the received data to detect anomalies. It is rewarded for accurate detection and penalised for false alarms or missed attacks.
- **Co-Adaptive Learning**: Both attacker and defender receive feedback after each simulation episode, allowing them to refine their strategies in response to each other in a dynamic, adversarial setting.
- **Performance Monitoring**: Metrics like detection accuracy, power imbalance, and stealth effectiveness are continuously tracked to assess and enhance system resilience.

III. Actor-Critic Framework in PPO

PPO is an RL algorithm that combines the roles of an actor, which selects actions based on the current policy, and a critic, which evaluates those actions by estimating future rewards. This actor-critic structure enables efficient learning through coordinated decision-making and feedback. What distinguishes PPO is its use of a clipped surrogate objective that limits how much the policy can change during each update, ensuring stability and preventing disruptive shifts. This constraint, guided by the critics' evaluations, allows the agent to make small, reliable improvements. A crucial element in this process is the advantage function, which helps the actor determine how much better a chosen action is compared to the average, encouraging better long-term decision-making. PPO's careful balance of exploration and stability makes it especially effective in complex, dynamic environments.

A. RL Attacker Algorithm

The algorithm simulates an FDIA in a 5-bus RDS using an RL agent trained with the PPO algorithm. The objective is to train an attacker to modify the predicted active power (P) values at selected buses in a way that the attack does not violate the power system constraints defined by us.

1. **Initialisation**: The agent is initialised with the forecasted active power values from the load forecasting model.
2. **Observation**: At each time step, the agent observes the system state comprising the predicted active power values for all load buses.
3. **Action Selection**: Based on the current observation, the agent selects small, continuous perturbations within ±1% to manipulate the predicted values.
4. **Application of Actions**: These perturbations are applied to produce the manipulated power values.
5. **Detection Check**: The system evaluates the manipulated values, flagging any as FDIA if they exceed the 15% deviation threshold.
6. **Reward Assignment**: The agent receives a positive reward if the attack remains undetected, or a penalty otherwise.
7. **Policy Update**: The PPO algorithm updates the agent's policy using the collected feedback to refine its decision-making strategy.
8. **Training Loop**: This process is repeated across multiple episodes, allowing the agent to improve its ability to craft undetectable yet impactful attacks.

B. Reward Function Formulation

The reward for the RL Agents is composed of the following elements:

1. **Stealth Reward:** Provides a positive reward if the manipulated values are within 10% of the maximum allowed deviation (15%).
2. **Detection Penalty:** Applies a penalty if any manipulation exceeds the 15% deviation threshold and scales with the episode step number to discourage late detection.
3. **Manipulation Reward:** Encourages noticeable but bounded changes to the target values.
4. **Power Conservation Penalty**: Penalises deviations from total system power, promoting physically realistic attacks.
5. **Defender Penalty**: Imposes heavy penalties if the attack is detected by the defender.
6. **Success Bonus**: Awards a bonus if all stealth and conservation criteria are met.

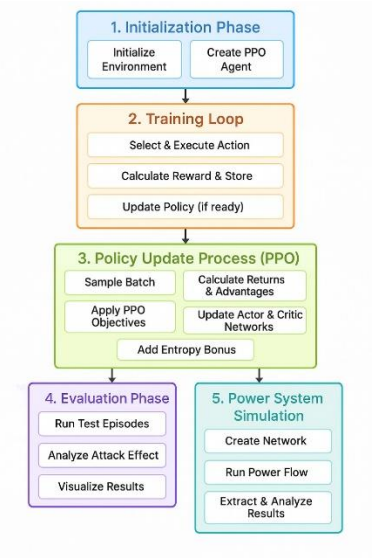

Fig.2 Flow of the PPO-based attacker agent.

C. DL-based Defender Algorithm

The defender is a multi-layer perceptron (MLP) designed for binary classification, trained to detect FDIAs by identifying anomalies in power system measurements.

1. **Architecture & Features**: It processes raw system data: active power (P), reactive power (Q), voltage magnitude (V), and angle (δ), along with deviations from expected baselines. Statistical metrics like means, standard deviations, and total values further enrich the input.
2. **Detection Mechanism**: A classification threshold of 0.5 is used to determine whether the input data is normal or attacked. This improves sensitivity to subtle manipulations.
3. **Continuous Learning**: A training buffer stores labelled data, enabling the defender to retrain periodically using supervised learning and adapt to evolving attack patterns.
4. **Operational Role**: While the attacker stealthily alters active power readings, the defender monitors all system features to flag suspicious behaviour. Performance is evaluated using metrics such as

detection accuracy, attack success rate, and system impact.

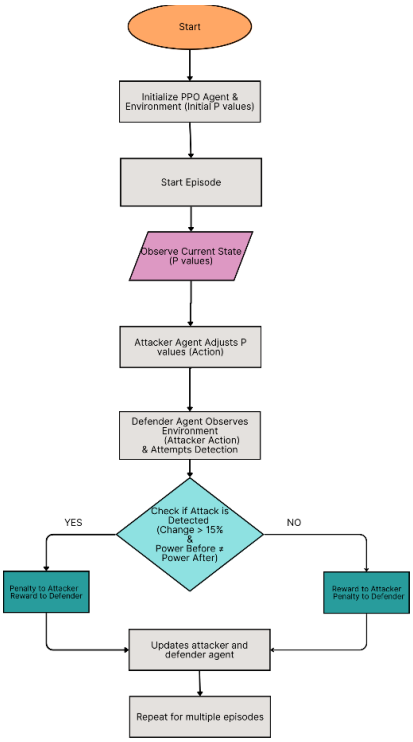

Fig.3 Methodology for training the agents

D. *RL-based Defender Algorithm*

The RL defender classifies each power system state as attacked or normal using a discrete binary action space (0: no attack, 1: attack detected). It observes the full system state: active/reactive power (P/Q), voltage magnitude (V), voltage angle (δ), and summary statistics, for comprehensive situational awareness.

At each time step, the actor network selects an action based on current observations. Rewards are assigned for correct classifications (true positives/negatives), while penalties are applied for false alarms and missed detections. Each experience (state, action, reward, next state) is stored in a replay buffer and used to update the policy via the PPO algorithm.

The architecture includes:

- **Actor Network**: Outputs action probabilities with ReLU activations.
- **Critic Network**: Assesses state-action values

PPO training is enhanced with a higher entropy coefficient to encourage diverse exploration and robustness against evolving attack strategies. Performance is evaluated using:

- Attack Success Rate, Detection Rate, False Positive Rate
- F1 Score (harmonic mean of precision and recall)
- Confusion Matrix Stats (TP, FP, TN, FN)

This setup enables both agents to co-evolve, with the attacker hiding within system noise and the defender learning to detect true anomalies.

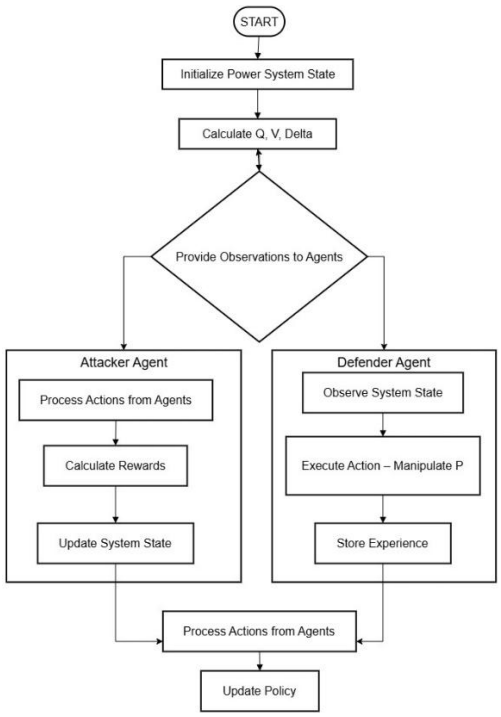

Fig.4 Flow of Algorithm with RL Attacker and RL Defender

IV. Results and Discussion

A. *Day-ahead Load Demand Forecasting*

The performance of the load forecasting model is adapted from [2]. It has been evaluated using standard regression metrics. The model achieved a MAE of 0.0138, RMSE of 0.0184, MBE of -0.0024, and a high R^2 score of 0.9852, indicating excellent predictive accuracy and minimal bias.

B. *RL Attacker and No Defender*

To identify the most effective reward structure for the RL-based attacker agent, three different reward functions were designed and evaluated across 120 episodes. Each reward has four key components: stealth, manipulation success, detection penalty, and power conservation. These components were weighted differently in each function to study their influence on the agent's learning dynamics and attack success.

Reward function 1 applied moderate emphasis on stealth and manipulation, while strongly penalising deviation from power conservation. The weights used were: stealth reward (5), detection penalty (-25), manipulation reward (20), and power conservation penalty (-50).

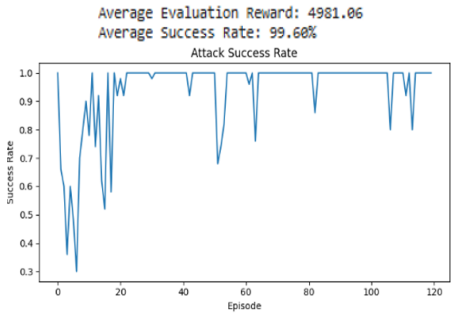

Fig.5 Average success rate and average evaluation reward for reward function 1

Reward function 2 placed greater emphasis on stealth and manipulation, using increased weights for those terms and slightly reduced penalties: stealth reward (10), detection penalty (-20), manipulation reward (30), and power conservation penalty (-40).

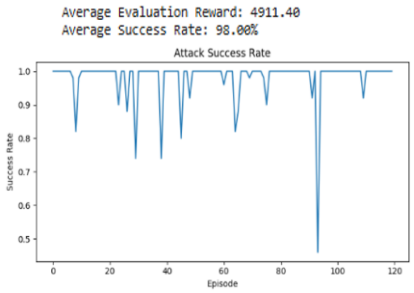

Fig.6 Average success rate and average evaluation reward for reward function 2

Reward function 3 used the same stealth and manipulation weights as Reward Function 2, but adjusted penalties to stealth reward (10), detection penalty (-25), manipulation reward (30), and power conservation penalty (-35) to test the impact of stronger detection penalties and lighter power conservation penalties.

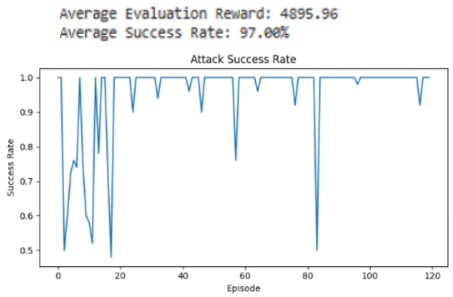

Fig.7 Average success rate and average evaluation reward for reward function 3

The performance of each reward function was evaluated using metrics such as average attack success rate and evaluation reward. As illustrated in Fig. 7-9, reward function 1 yielded the highest success rate of 99.6%, outperforming reward function 2 (98%) and reward function 3 (97%). Based on these results, reward function 1 was chosen as the base reward formulation for all subsequent training and evaluation of the reinforcement learning-based attacker, as it provided the best trade-off between stealth, impact, and physical feasibility.

C. RL Attacker and DL Defender

The RL attacker was trained for 100 episodes of 50 steps each, showing gradual but limited improvement. The success rate peaked at 15.3% around episode 80 but dropped to 8.7% by episode 100, indicating challenges in consistently executing stealthy attacks.

During evaluation over 10 episodes, success rates varied widely, averaging around 11%, meaning roughly 1 in 9 attacks succeeded without detection. The consistently negative rewards reflect penalties for detected or ineffective attacks. Overall, the results highlight the difficulty for the attacker to reliably bypass the defender.

The defender demonstrated strong performance by detecting 4700 out of 5500 attacks, achieving an 85.45% detection rate. Using a detection threshold of 0.5, the defender effectively distinguished between attacked and normal system states.

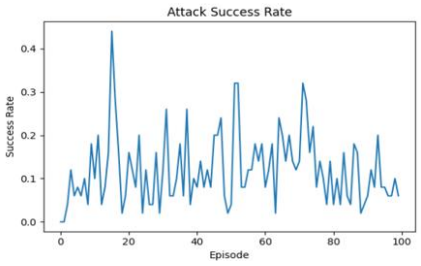

Fig.8 Attack Success Rate across episodes

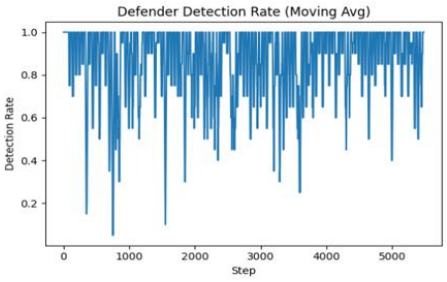

Fig.9 DL Defender Detection Rate across time steps

The results are tabulated in Table 1 below. The 'P' values of load buses of a 5-bus RDS are attacked by the RL-agent. The attacker algorithm ensures the total active power remains conserved even after the attack to remain undetected. It is found that the total active power consumed by the load buses before attack is 55.5 kW which remains conserved at 55.5 kW even after the FDI attack.

Attack Outcome:

- **Attack Successful:** False (The attack did not remain undetected or meet success criteria)

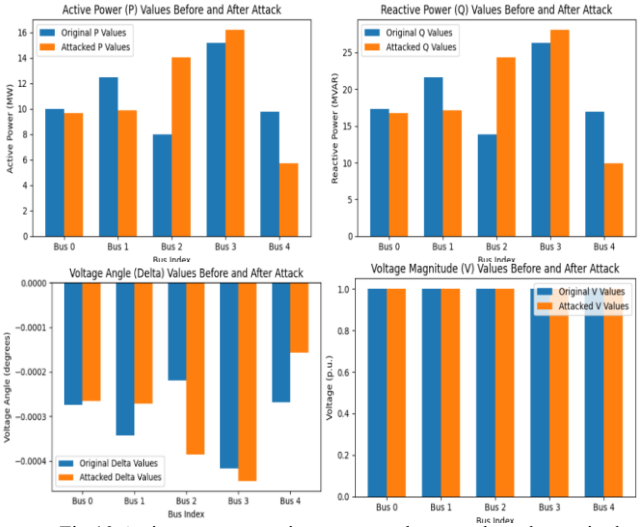

Fig.10 Active power, reactive power, voltage angle and magnitude before and after the FDIA

D. RL Attacker and RL Defender

The adversarial training was conducted over 40 episodes, each lasting 100 steps. Over the course of training, the RL Defender showed significant improvement in detecting attacks while the RL Attacker's success rate drastically decreased.

Training Highlights:

- By Episode 40, the attacker's success rate dropped to 0.55%, while the defender's detection rate rose to 98.5%.
- The defender's F1 Score improved steadily, reaching 0.8650 at the end of training.

Table 1 RL Attacker and DL Defender before and after attack

	Before Attack				After Attack				Differences			
Bus	**P**	**Q**	**V**	**δ**	**P**	**Q**	**V**	**δ**	**P**	**Q**	**V**	**δ**
1	10.0	17.32	0.9998	-0.00027	9.64	16.70	0.9998	-0.00026	-0.36	-0.62		
2	12.5	21.65	0.9997	-0.00034	9.87	17.09	0.9998	-0.00027	-2.63	-4.56	Order of 10^ -5	Order of 10^-5
3	8.0	13.86	0.9998	-0.00022	14.06	24.36	0.9997	-0.00039	6.06	10.50		
4	15.2	26.33	0.9996	-0.00042	16.20	28.06	0.9996	-0.00044	1	1.73		
5	9.8	16.97	0.9998	-0.00027	5.73	9.92	0.9999	-0.00016	-4.07	-7.05		
Total	**55.5**	-	-	-	**55.5**	-	-	-	-	-	-	-

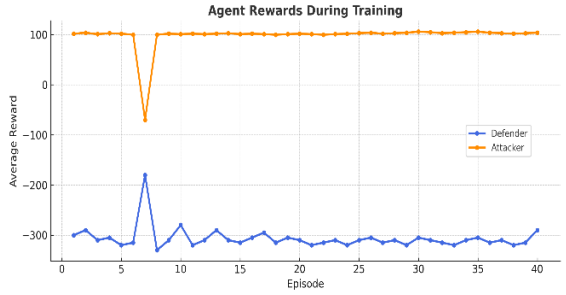

Fig.11 RL Attacker and RL Defender Rewards during Training

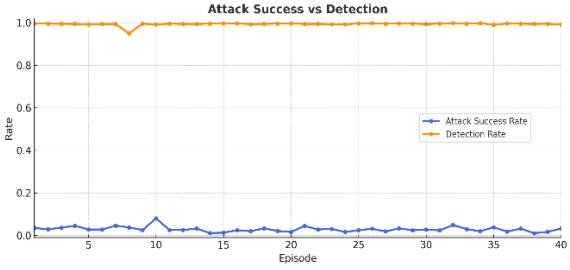

Fig.12 Attack Success and Detection Rates across episodes

Evaluation over 5 episodes (each 100 steps):

- Average attack success rate remained very low at 0.51%, confirming the defender's precision.
- Average detection rate stayed high at 98.61%, indicating consistent detection of attacks.
- False positive rate was around 22.4%, showing some benign events flagged but manageable.
- Defender F1 Score averaged 0.8658, illustrating a strong balance between precision and recall.
- Confusion matrices from each episode show high true positive counts (around 1,600) with relatively low false negatives (~23) and moderate false positives (~480).
- Average rewards during evaluation were -218.28 for the attacker and 133.42 for the defender, confirming the defender's effectiveness in minimizing attack success.

This demonstrates that an RL-based defender can successfully learn to detect complex attacks with high accuracy while maintaining a low rate of false alarms, effectively reducing the attacker's success to near zero.

Defender Comparison:

	No Defender	**RL-Based Defender**	**DL-Based Defender**
Detection Rate	NA	0.9861	0.8545
Attack Success Rate	0.996	0.0051	0.11

RL-based defender adapts to evolving attack strategies without needing labelled data, learns strategic, long-term detection policies and suitable for dynamic environments and complex attack patterns while DL-based defender is easier and faster to train with labelled data, stable and efficient during operation and offers more interpretable classification outcomes.

The optimal choice depends on specific scenario. RL-based defender is better when the attacker is highly adaptive and constantly changing strategies but has limited labelled data and can simulate many scenarios. Further, a DL-based defender is better when you have access to substantial labelled data of attacks and normal operations. The attack patterns are relatively stable and computational resources during operation are limited.

V. Conclusion and Future Directions

A comprehensive environment was built to simulate the power system, with constraints ensuring power conservation. The system's state and action spaces were carefully defined, allowing the model to interact with realistic operational limits. Power flow is adjusted to ensure balance while being manipulated by the attacker, and the system behaviour was closely monitored for threshold breaches, particularly with a 15% tolerance for detection. The present work can be extended combining forecasting, state estimation, and RL-based agents to study FDIA impacts on real smart grids, refining attacker stealth, scale to larger grids, and explore advanced countermeasures. Moreover, hybrid defence models that merge supervised learning's pattern recognition with RL's strategic adaptability can be investigated.

References

[1] Luo, Weifeng, and Liang Xiao. "Reinforcement learning based vulnerability analysis of data injection attack for smart grids." *2021 40th Chinese Control Conference (CCC)*. IEEE, 2021.

[2] Naware, Dipanshu, Hira Singh Sachdev, and Saransh Chourey. "Investigating the Role of Data Preprocessing for Load Demand Forecasters in Smart Household Applications." *2024 Third International Conference on Power, Control and Computing Technologies (ICPC2T)*. IEEE, 2024.

[3] Zhang, Guihai, and Biplab Sikdar. "A novel adversarial FDI attack and defence mechanism for Smart Grid demand-response mechanisms." *IEEE Transactions on Industrial Cyber-Physical Systems* (2024).

[4] Naware, Dipanshu, and Arghya Mitra. "Data-driven Technology Applications in Planning, Demand-side Management, and Cybersecurity for Smart Household Community." *IEEE Transactions on Artificial Intelligence* (2024).

[5] Roomi, Muhammad M., et al. "Analysis of false data injection attacks against automated control for parallel generators in IEC 61850-based smart grid systems." *IEEE Systems Journal* 17.3 (2023): 4603-4614.

[6] Rahman, Moshfeka, Jun Yan, and Emmanuel Thepie Fapi. "Adversarial Artificial Intelligence in Blind False Data Injection in Smart Grid AC State Estimation." *IEEE Transactions on Industrial Informatics* (2024)

[7] Alam, Mahamad Nabab, Saikat Chakrabarti, and Arindam Ghosh. "Networked microgrids: State-of-the-art and future perspectives." IEEE Transactions on Industrial Informatics 15.3 (2018): 1238-1250.

[8] De La Cruz, Jorge, et al. "Review of networked microgrid protection: architectures, challenges, solutions, and future trends." CSEE Journal of Power and Energy Systems 10.2 (2023): 448-467.