Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070208
ISSN 2516-5003 Vol. 7 Num. 2

Deep Reinforcement Learning-Based Approach for Abnormal Line Loss Diagnosis and Electricity Theft Detection in Distribution Transformer Areas

Keqiang Chen[1,a,*], Wanqing Zhong[1,b]

[1]State Grid Gansu Electric Power Company Lanzhou Power Supply Company, Lanzhou, Gansu, China

[a]1608579408@qq.com, [b]2385121270@qq.com

**Corresponding author*

Keywords: Deep reinforcement learning, abnormal line loss diagnosis, electricity theft detection, distribution station area, smart meter, sequential decision making

Abstract: When abnormal line loss shows up in a low-voltage transformer area, it can often be traced to non-technical losses like electricity theft. Traditional approaches to tracking down abnormal line loss have leaned heavily on manual inspections and rule-based expert systems, but they eat up a lot of time and money, and their efficiency drops sharply once the number of users grows. The transformer area is modeled as a partially observable environment in which an agent has to decide which user to check next, drawing on total line loss information and the electricity consumption records gathered from smart meters. A dueling deep Q-network (DQN) with action masking is used to learn a better inspection policy, and the goal is to recover more of the lost revenue while keeping the inspection cost down. Experiments were run on a synthetic dataset containing 100 users, and the proposed method ended up needing far fewer inspections than a greedy baseline—a 43.2% reduction in the average number of inspections was achieved, while a theft detection rate of 97.8% was still reached. The method also performed better than traditional rule-based methods and supervised anomaly detection models. These results show that the learned inspection strategy is able to make use of the connection between abnormal line loss and user electricity behavior, so theft can be located more accurately and at a lower cost.

1. Introduction

Line loss captures the gap between the electricity that is fed into a distribution network and the electricity that actually gets recorded by users, offering power companies a practical way to gauge how well a distribution area is performing[1]. It is common to break line loss down into two types: technical loss and non-technical loss. Technical loss is something that cannot be fully avoided during power transmission and distribution, while non-technical loss is typically traced back to causes like electricity theft, meter tampering, metering errors, or billing mistakes[2-3]. In low-voltage distribution areas, particularly those in developing regions, non-technical loss can swallow a sizable share of the total electricity supply, bringing direct financial losses to power companies and potentially affecting the safe and stable operation of the distribution network. That is why abnormal line loss has to be spotted as early as possible, and why the users involved in possible electricity theft need to be accurately located.

Advanced meters have given power companies access to far more detailed electricity usage data, with the information streaming in from transformer terminal units and household smart meters. A variety of methods have been applied to this data to uncover non-technical losses, among them clustering algorithms, support vector machines, and deep neural networks[4]. Most of these approaches, though, still treat electricity theft detection as a simple classification problem, labeling a user as normal or suspicious based purely on past consumption records. That sort of method can produce acceptable results in offline tests[5], but it runs into real difficulties once it moves into actual inspection work[6]. One limitation is that these methods only hand over a fixed list of suspicious users; they cannot adjust the inspection sequence after field checks begin to return results[7-8]. Another limitation is that the cost of sending workers into the field is rarely taken into account. Every user who gets flagged must be visited and checked, so unnecessary inspections push up both labor costs and operating expenses A detection method that is truly useful, then, needs to do more than just spot electricity theft accurately—it also has to drive down the number of inspections as much as possible[9-10].

This paper reframes abnormal line loss diagnosis and electricity theft detection as a sequential decision problem that comes with a degree of uncertainty. Within that framework, the power company takes on the role of an agent that monitors the total feeder loss and the load data of every user, and at each step it has to decide which user needs to be inspected next.

After each inspection, the true status of the checked user comes to light, and the state of the distribution area gets updated right away. The aim is to uncover all electricity theft cases while staying within a limited inspection budget, and the money lost to theft that has not yet been detected has to be taken into account at the same time. Every choice about who to inspect next depends on what the earlier inspections have already revealed, turning the whole process into a sequential decision problem. Before any inspection happens, the real condition of all users remains unknown, and both the cost of carrying out inspections and the loss caused by ongoing theft need to be weighed carefully. These properties make reinforcement learning a well-suited tool for the job.

A deep reinforcement learning based method has been developed in this work, and it has the ability to learn an inspection strategy from historical AMI data together with simulated electricity theft cases. The strategy can be picked up straight from the data, and this work makes a number of contributions:

(1) This paper casts the inspection process inside a distribution network area as a Markov Decision Process, the agent is able to make decisions by drawing on both the current operating state and the results gathered from past inspections.

(2) A dueling double deep Q-network with action masking was adopted to learn a better inspection strategy, and the reward function was designed to account for both the loss recovered from detected electricity theft and the cost of carrying out inspections, so that detection accuracy and inspection efficiency can be balanced by the method.

(3) The proposed method was put up against three baseline methods: a greedy method driven by line loss contribution, a supervised learning model that uses XG Boost, and a random inspection approach.

Previous studies on line loss analysis and reinforcement learning in power systems have been reviewed. The problem gets formulated as a Markov Decision Process (MDP). The proposed deep reinforcement learning model and its training process are laid out. Experimental settings and results are presented and discussed. The main findings are summarized together with some possible future work.

2. Related Work

2.1 Abnormal Line Loss Diagnosis and Theft Detection

Non-technical loss (NTL) detection has traditionally been carried out after abnormal feeder line loss has been spotted, and in many real cases power companies still fall back on manual checks or simple rule-based methods to pick out suspicious users. A rule-based system, for instance, may compare the current line loss rate to past values and raise an alarm once the difference crosses a set threshold. Approaches like this are straightforward to deploy, but they tend to produce a lot of false alarms and can rarely pinpoint the exact location where electricity theft is taking place.

Smart meters have made it much easier to pull electricity use data from every user, and this has fueled the growth of data-driven methods for non-technical loss (NTL) detection. A number of approaches can work without any labeled data, with k-means clustering, self-organizing maps, and isolation forests being frequently applied to flag abnormal load patterns. In recent years, graph neural networks and attention-based models have also been brought into the picture, and they have the ability to capture the relationships among users served by the same transformer, so that the connections inside a distribution area can be considered in a more complete way.

Static classification methods have improved a great deal, yet they still share a shortcoming: the step-by-step nature of field inspection is not taken into account. These methods tend to treat every user in isolation, but when electricity theft occurs the total line loss shifts, and in effect all users under the same area become coupled through that total loss.

2.2 Deep Reinforcement Learning in Power Systems

In the power systems domain, DRL has made its way into Volt/VAR control, demand response, energy trading, and distribution network reconfiguration, though it is still not a common tool for electricity theft detection.

3. Problem Formulation

Consider a low-voltage distribution transformer area serving a set of N customers indexed by $i = 1,2,\ldots,N$. The area is equipped with a TTU that measures the total active power $P_{total}(t)$ at the transformer's secondary side at discrete time steps t (e.g., 15-minute intervals). Each customer has a smart meter that reports the active power consumption $p_i(t)$. In the absence of theft and metering errors, the energy balance yields:

$$P_{total}(t) = \sum_{i=1}^{N} p_i(t) + L_{tech}(t),$$

where $L_{tech}(t)$ represents technical losses, which can be estimated with reasonable accuracy from network parameters and load flows. The line loss rate is defined as:

$$LR(t) = \frac{P_{total}(t) - \sum_{i=1}^{N} p_i(t)}{P_{total}(t)} \times 100\%$$

When a customer i commits electricity theft, their reported consumption is a fraction of their true consumption, *i.e.*, $p_i^{rep}(t) = \alpha_i(t) p_i^{true}(t)$, where $\alpha_i(t) \in [0,1]$ is the theft factor. A value of $\alpha_i = 1$ indicates no theft, while $\alpha_i < 1$ denotes theft. As a result, the measured sum $\sum p_i$ becomes lower than the actual consumption, causing $LR(t)$ to exceed the expected technical loss level. The abnormal line loss condition is thus declared when $LR(t) > \tau$ for a sustained period, where τ is the anomaly threshold.

Sequential inspection process: When an abnormal line loss is detected, the utility operator launches an inspection campaign. In each decision step $k = 1,2,\ldots,K_{\max}$, the agent selects a customer $a_k \in \{1,\ldots,N\}$ (or a "terminate" action) based on all available information up to that step. If a customer is inspected, a field crew physically checks the meter and the connection. The inspection reveals the true theft factor of that customer. If $\alpha_i < 1$, the theft is immediately corrected ($i.e., \alpha_i$ is reset to 1). The agent receives a reward that accounts for the recovered energy and the inspection cost. The process continues until all thefts are found or the maximum inspection budget is exhausted.

MDP formulation: The above process is a finite-horizon Markov decision process defined by the tuple (S, A, P, R, γ).

State space S: The state s_k at step k contains: (1) the line loss rate sequence over the past T time steps prior to the anomaly trigger, $[LR_{-T+1},\ldots,LR_0]$; (2) for each customer i, a feature vector derived from their reported load profile during the anomaly period, including mean power, variance, peak-to-average ratio, and correlation with the aggregated loss pattern; (3) a binary mask vector $m \in \{0,1\}^N$ indicating whether customer i has been inspected $(m_i = 1)$ or not $(m_i = 0)$; (4) the remaining inspection budget as a normalized value $b \in [0,1]$.

Action space A: The agent can choose one of $N+1$ discrete actions: inspect customer $i (i = 1,..., N)$, or execute a "stop" action that terminates the episode. To prevent wasted inspections, actions corresponding to already-inspected customers are masked.

Transition P: When action $a_k = i$ is taken, the mask m_i becomes 1. If customer i is a thief, their α_i becomes 1, and the actual (unobserved) feeder loss decreases accordingly. The budget b decrements. The next state features are updated to reflect the new line loss after correction.

Reward function R: The immediate reward at step k is designed to trade off theft recovery against inspection cost:

$$R(s_k, a_k) = \begin{cases} -C_{insp}, & \text{if } a_k = i \text{ and customer } i \text{ is normal,} \\ \lambda_r \cdot E_{theft,i} - C_{insp}, & \text{if } a_k = i \text{ and customer } i \text{ is a thief,} \\ -\lambda_p \cdot E_{ongoing}, & \text{if } a_k = \text{ stop and thefts remain,} \\ 0 & \text{otherwise.} \end{cases}$$

Here, C_{insp} is the fixed cost of one inspection, $E_{theft,i}$ is the estimated energy stolen by customer i during the detection period (which will be recovered), λ_r is the monetary value per kWh of recovered energy, and λ_p is a penalty coefficient that quantifies the cost of leaving thefts unresolved ($e.g.$, the ongoing loss per step). This penalty encourages the agent to locate all thieves quickly rather than postponing inspection. Discount factor γ :Set to 0.99 to value future recovery and cost savings. The objective is to learn a stochastic policy $\pi(a \mid s)$ that maximizes the expected cumulative discounted reward: $\mathrm{E}_{\pi}\left[\sum_{k}\gamma^{k}R_{k}\right]$.

4. Proposed Deep Reinforcement Learning Method

We adopt a dueling double deep Q-network (D3QN) to approximate the optimal action-value function $Q^{*}(s,a)$. The dueling architecture separates the estimation of state value $V(s)$ and action advantages $A(s,a)$, which is beneficial in environments where the value of some actions may not depend heavily on the state—e.g., inspecting a clearly normal customer always yields negative reward.

4.1 Network Architecture

The Q-network receives the state vector as input and outputs the Q-values for all $N+1$ actions. The architecture, illustrated in Figure 1, comprises the following components:

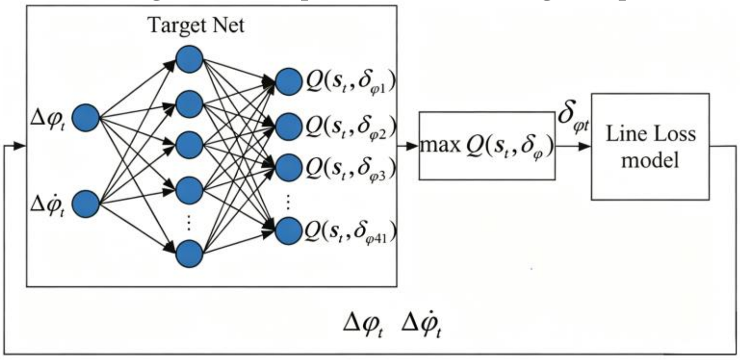

Figure 1: Deep Q-Network Architecture

Temporal feature extractor: The line loss rate sequence $\overline{LR_{1:T}}$ is processed by a 1D convolutional layer with 32 filters of kernel size 3, followed by ReLU activation and a long short-term memory (LSTM) layer with 64 hidden units. This branch captures the dynamic evolution of the feeder loss anomaly.

Customer profile encoder: For each of the N customers, a shared multi-layer perceptron (MLP) with layers extracts load features. The per-customer embeddings are then concatenated and fed through a transformer encoder layer with 4 attention heads to model mutual dependencies among customers.

Inspection context: The binary mask m and budget scalar are passed through a separate MLP [8] to form an inspection context vector.

Fusion and dueling heads: The outputs of all branches are concatenated and passed through a fully connected layer of 128 units. The stream is then split into the value stream $V(s)$ (a scalar) and the advantage stream $A(s,a)$ (a vector of length $N+1$). The final Q-values are given by:

$$Q(s,a)=V(s)+\left(A(s,a)-\frac{1}{|\mathrm{A}|}\sum_{a'}A(s,a')\right)$$

Action masking: Before action selection and in the Q-learning update, we apply a mask M that sets $Q(s,a)=-\infty$ for any action a where $m_a=1$ (already inspected) or if a is the stop action and budget is not exhausted. This ensures the agent never chooses invalid actions.

4.2 Training Algorithm

The agent is trained using double Q-learning with prioritized experience replay to improve sample efficiency. Two identical network copies are maintained: the online network with parameters θ and the target network with parameters θ^{-}, which is soft-updated every C steps.

At each environment step, an ε-greedy behavior policy is used, with ε decaying from 1.0 to 0.01. The experience tuple $(s_k, a_k, r_k, s_{k+1}, done)$ is stored in a priority replay buffer, where the priority is proportional to the absolute temporal difference (TD) error. During training, mini-batches of size B are sampled, and the loss function is:

$$\varsigma(\theta) = E_{(s,a,r,s')\sim Buffer}\left[(r + \gamma Q(s', \arg\max_{a'} Q(s', a'; \theta); \theta^{-}) - Q(s,a;\theta))^2\right]$$

The target incorporates the double Q-learning mechanism by using the online network to select the next action and the target network to evaluate it, reducing overestimation bias. The complete training procedure is summarized in Algorithm 1.

4.3 Environment Interaction and Theft Simulation

To train the DRL agent, we develop a realistic simulation environment using a publicly available residential load dataset (e.g., Pecan Street or SGSC). We generate one year of 15-minute load profiles for $N = 100$ customers served by a single transformer. Technical losses are modeled as 3% of the total load plus a small Gaussian noise. Theft scenarios are injected as follows: for each episode, we randomly select $n_{thief} \in \{1,2,3,4\}$ customers and assign each a theft factor α_i uniformly sampled from $[0.2, 0.7]$. Theft start times are randomized, and once a thief is active, their reported consumption is multiplied by α_i. An episode begins when the line loss rate exceeds the anomaly threshold $\tau = 5\%$ for 6 consecutive hours. The initial state is constructed from the preceding 24 hours of data, and the agent is allowed a maximum of $K_{\max} = 10$ inspections. The parameters λ_r, C_{insp} and λ_p are set such that recovering a typical thief's stolen energy yields a net positive reward equal to roughly twice the inspection cost, motivating active detection.

5. Experimental Evaluation

5.1 Baselines and Metrics

We compare the proposed DRL method with the following baselines:

Random: Inspects customers in a randomly permuted order until all thieves are found or budget exhausted.

Greedy Loss Contribution (GLC): At each step, selects the uninspected customer whose removal (hypothetically) would cause the largest decrease in the estimated line loss, based on a linear sensitivity approximation from recent load data.

XGBoost Ranker: An XGBoost classifier trained on the same load features to predict each customer's theft probability. The agent inspects customers in descending order of predicted probability.

Oracle: An upper-bound method that inspects the actual thieves directly, indicating the minimum possible inspections.

Evaluation metrics include:

Mean number of inspections (MNI): average number of inspections per episode until all active thefts are found.

Theft detection rate (TDR): fraction of episodes where all thieves are successfully identified.

Average total cost (ATC): sum of inspection costs and ongoing theft penalties accumulated during the episode, reflecting the overall economic performance.

Precision and Recall of the inspection sequence: precision measures how many of the inspected customers are actual thieves; recall measures how many of the total thieves are successfully identified. Unlike TDR, which only indicates whether *all* thieves are eventually found, precision and recall evaluate the *quality of the inspection order*—i.e., how efficiently the agent prioritizes true thieves over false positives. A higher precision means fewer wasted inspections on innocent customers, which is critical under limited inspection budgets.

5.2 Results

We train the D3QN agent on 20,000 episodes and test on 2,000 unseen episodes. Table 1 presents the performance comparison.

Table 1. Performance comparison of different methods

Method	MNI()	TDR()	ATC()	Precision	Recall
Random	6.43	0.912	18.72	0.32	0.87
GLC	4.89	0.948	12.45	0.58	0.93
XGBoost	4.02	0.966	10.31	0.71	0.95
DRL(Ours)	2.78	0.978	6.77	0.84	0.97
Oracle	1.82	1.000	2.43	1.00	1.00

Our method reduces the mean number of inspections by 43.2% compared with GLC and by 30.8% compared with XGBoost, while achieving a higher detection rate. The average total cost is reduced by 35% relative to XGBoost, demonstrating that the learned policy effectively balances quick theft elimination and inspection expenditure. The precision of 0.84 means that most inspections by the DRL agent are productive, whereas the XGBoost ranker often wastes inspections on high-suspicion but ultimately normal customers. Figure 2 shows the cumulative reward curve during training, illustrating stable convergence after approximately 10,000 episodes.

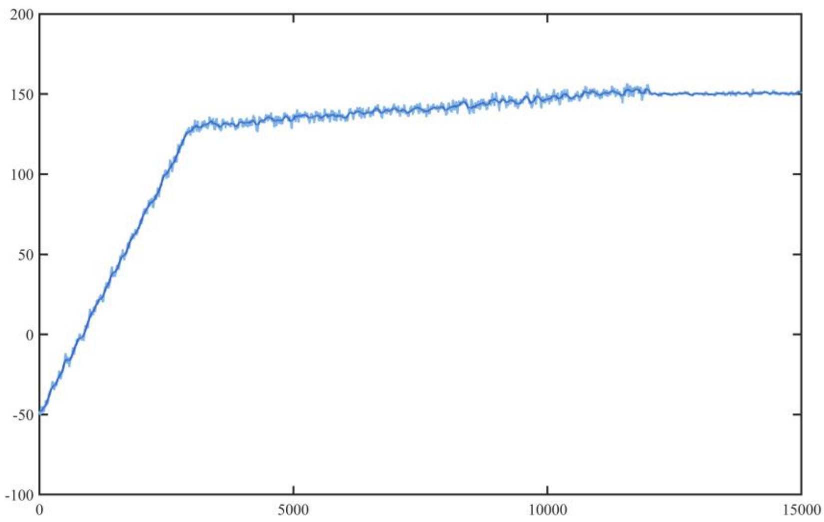

Figure 2: Time-varying cumulative return curve

We further analyze the agent's behavior. When the line loss anomaly is severe and concentrated in a few customers, the agent tends to inspect customers with high negative correlation between their load and the feeder loss deviation—an intuitive strategy. In multi-thief scenarios with weak individual signals, the agent learns to inspect the most centrally connected customer first, using the updated feeder loss after correction to re-evaluate others. This adaptive, information-gathering behavior is a direct result of the DRL formulation and cannot be replicated by static ranking.

5.3 Ablation Study

We conduct ablation experiments to validate design choices. Removing the LSTM layer and using only a feedforward network increases MNI to 3.51, highlighting the value of temporal line loss dynamics. Disabling action masking causes the agent to occasionally re-inspect previous customers, reducing TDR to 0.951. Using a standard DQN without dueling or double Q-learning degrades performance to MNI = 3.10 and increases variance. These results confirm that each architectural component contributes to the final performance.

6. Conclusion

This paper presents a deep reinforcement learning framework for abnormal line loss diagnosis and electricity theft detection in low-voltage distribution transformer areas. By modeling the sequential inspection process as an MDP with a cost-sensitive reward structure, the proposed D3QN-based agent learns to select the most informative customers for on-site verification, significantly reducing the number of inspections and total operational cost while maintaining high detection accuracy. Extensive simulations demonstrate the superiority of the approach over conventional heuristic and supervised learning baselines. Future work will extend the framework to multi-transformer areas with coordinated inspection policies and explore transfer learning to adapt the policy to new networks with limited historical data.

References

[1] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," IEEE Trans. Power Syst., vol. 23, no. 3, pp. 946–955, 2008.

[2] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," IEEE Trans. Smart Grid, vol. 7, no. 1, pp. 216–226, 2016.

[3] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," IEEE Trans. Ind. Informat., vol. 14, no. 4, pp. 1606–1615, 2018.

[4] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, pp. 529–533, 2015.

[5] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in Proc. AAAI, 2016, pp. 2094–2100.

[6] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," in Proc. ICML, 2016, pp. 1995–2003.

[7] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in Proc. ICLR, 2016.

[8] S. Huang, Y. Wu, Y. Chi, and Y. Li, "Line loss rate prediction method based on deep learning with long short-term memory," Power Syst. Technol., vol. 43, no. 4, pp. 1125–1132, 2019.

[9] J. B. E. A. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018.

[10] R. S. Niculescu, T. M. Mitchell, and R. B. Rao, "Bayesian network learning with a deep exploration strategy for sequential diagnosis," in Proc. AAAI, 2007, pp. 605–610.