Explaining Deep Reinforcement Learning Policy in Distribution Network Control

Blaž Dobravec
Elektro Gorenjska d.d.
Kranj, Slovenia
blaz.dobravec@elektro-gorenjska.si

Jure Žabkar
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
jure.zabkar@fri.uni-lj.si

Abstract

In safety-critical settings – such as low-voltage electrical distribution networks – Deep Reinforcement Learning (DRL) policies are hard to deploy due to limited capability to explain why a particular sequence of actions is taken. We use Scenario-Based eXplainability (SBX) with temporal prototypes to explain the policy of our DRL agent. SBX clusters short time-windows of latent trajectories and uses their medoid trajectories as human-friendly summaries. Temporal prototypes map the embeddings of these medoids to actions, and generate explanations of the form "This scenario is similar to prototype $X \Rightarrow$ Do action Y." We apply our approach to a real low-voltage distribution network Srakovlje. Preliminary results show that our method offers practically useful human-friendly explanations for sequential decision making.

Keywords

deep reinforcement learning, explainability, voltage control, low-voltage distribution network, prototypes

1 Introduction

A rapid growth of renewable energy resources and a significant increase in electricity demand due to the electrification of transport and heating [8] are reshaping generation (e.g. distributed photovoltaic systems) and consumption (e.g. heat-pumps, electrical vehicles) in electrical distribution networks. Increasing reverse power flows and voltage variability in low-voltage networks strongly affect voltage profiles and make the network operation and control more challenging.

Deep reinforcement learning (DRL) has recently emerged as a powerful paradigm for sequential decision-making in complex, high-dimensional environments, with notable successes in games (Chess [18], Go [19], Atari [13]), autonomous driving [10], and industrial robotic process automation [7]. Voltage control in distribution networks shares similar characteristics, which makes DRL a promising methodology to solve the control problems in low-voltage networks.

While voltage control is standard at higher voltage levels (e.g., with STATCOMs), most LV research has focused on optimizing individual assets at the customer level [11, 6]. Recent comparisons indicate that DRL can outperform classical algorithms for micro-grid management with demand-side flexibility [14]. For instance, dueling double DQN (D3QN) has been used to reduce overvoltages in PV-rich networks [16]; model-free RL has optimized battery charging/discharging to increase self-sufficiency [12]; and effective consumption/generation strategies have been learned under price signals and network constraints [2, 1]. Given the growing heterogeneity of LV networks and the rise of behind-the-meter actuators, DRL methods are typically developed and validated first in simulation [4]. Their adoption and implementation are often hindered by a lack of explainability of these models.

We present a prototype-based explainability approach for DRL-based voltage control in LV distribution networks that directly exploits flexibility from prosumers. In our approach, the agent acts on the network's operating state, coordinating different flexibility options (e.g. photovoltaic systems, batteries, EVs, heat pumps). We focus on improving power quality by reducing voltage violations. Additionally, we use prototype based explainability to provide interpretation and reasoning behind the action.

2 Related Work

Explainable Artificial Intelligence (XAI) aims to make the decisions of models understandable to humans. The explanation process and the final result should be focused on generating explanations that are intuitive to us. Prototype-based explanations provide a compelling choice that is interpretable by design. XRL remains an active area of research. One such widely employed explainability technique, primarily used in image classification, is the *saliency map*, which bases its explanations on pixel-wise feature attribution [20]. Building on this idea, Sequeira et al. [17] made the agent's interactions with the environment the focal point of their *Interestingness Framework*.

In supervised learning, prototype networks explain predictions via similarity to learned or human-selected exemplars [3, 15]. Extending this paradigm to reinforcement learning, prototype-wrapper policies force decisions to be mediated by human-friendly prototypes (single state-snapshot); a recent example is the Prototype-Wrapper Network (PW-Net), which wraps a pre-trained agent and maps latent states to action decisions through prototype similarities [9]. Beyond interpretability, prototypes have been leveraged to improve representation learning and exploration efficiency: Proto-RL pre-trains prototypical embeddings and uses prototype-driven intrinsic motivation to accelerate downstream policy learning in pixel-based control [23]. In model-based RL, prototypical context learning has also been explored for dynamics generalization [22].

Despite the critical role of explainability in voltage control in low-voltage power systems, there is little research addressing this challenge. Zhang et al. [24] applied the SHAP explainability method to a deep reinforcement learning model for implementing proportional load shedding during under-voltage situations. They also used Deep-SHAP [25] to enhance the computational efficiency of their XAI model. The model's output elucidates its

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.9459

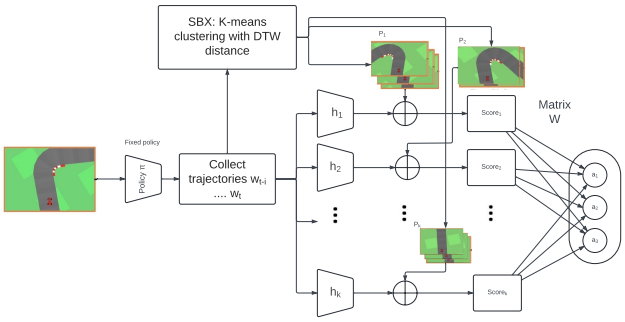

Figure 1: High-level SGTP pipeline: (1) collect latent windows; (2) SBX clustering and medoid selection; (3) train temporal-prototype layer; (4) case-based explanations during rollout.

predictions through a visualization layer and a feature importance layer that addresses both global and local explanations.

Existing research on explainability in power systems, particularly regarding voltage control, focuses on post-hoc explainability techniques. Compared to explanations for a single feature (individual voltage value) such as SHAP, our method considers the temporal component in the explanation process. To the best of our knowledge, this approach has not been applied to the explainability of the reinforcement learning field in this specific manner before.

3 SBX-guided Prototype Selection

We employ Scenario-Based eXplainability (SBX [5]) as an extension of the PWNet [9] to temporal prototypes (**prototypes of trajectories, not just snapshots of the state space**) to provide global, scenario-level structure and local, time-resolved explanations for a trained control policy. SBX is used to partition behavior and select representative temporal prototypes. On top of the SBX-selected prototypes (without any human-defined prototypes), we train a temporal prototype model that maps latent features to actions. This yields a two-tier view: SBX provides a summary of behavior, while temporal prototypes expose time-local patterns and their nearest neighbors that drive actions.

3.1 Data Preparation and Latent Extraction

We consider a trained policy π acting in discrete time. A trajectory is a sequence of observation–action pairs. For analysis, we operate on fixed-length trajectories of length L:

$$w_t = \big((o_t, a_t), \ldots, (o_{t+L-1}, a_{t+L-1})\big), \quad t = 0, \ldots, T - L.$$

Observations are first mapped by the frozen policy backbone to latent vectors $x_t \in \mathbb{R}^d$. We denote the latent trajectory by $X_t = (x_t, \ldots, x_{t+L-1}) \in \mathbb{R}^{L\times d}$. We collect an offline dataset by rolling out the trained PPO agent and recording, at each time step, the policy's penultimate-layer latent vector and the corresponding environment action. This yields per-episode sequences of latents and actions which are then converted into trajectories of length L. The supervised target for each trajectory is the action at its last real-time step.

3.2 SBX Prototype Selection

SBX is performed in the latent space by clustering window embeddings with k-means over a range of cluster counts and selecting the number of clusters via a silhouette-style score. Within each selected cluster, the medoids (nearest to the centroids) are taken as temporal prototypes. Optionally, flattened action windows are concatenated to latent trajectories before k-means to bias prototype selection toward action-discriminative regions. The SBX step produces a prototype tensor of shape (K, L, d).

3.3 Temporal Prototype Model

We introduce K temporal prototypes $\{P_k\}_{k=1}^K$, each a length-L latent template $P_k \in \mathbb{R}^{L\times d}$ selected by SBX (medoids). A shared temporal encoder $g_\theta : \mathbb{R}^{L\times d} \to \mathbb{R}^p$ maps trajectories to embeddings $z_t = g_\theta(X_t)$ and prototypes to $e_k = g_\theta(P_k)$. Following PW-Net, prototype activations use an L2-to-activation mapping.

$$a_k(t) = \log \frac{\|z_t - e_k\|_2^2 + 1}{\|z_t - e_k\|_2^2 + \varepsilon}, \quad \varepsilon > 0. \tag{1}$$

Outputs are linear in activations, $y_t = W\, a(t)$, optionally post-processed to valid actions (Tanh/ReLu for steer/gas/brake). The schematics of the algorithm is outlined in Figure 1.

3.4 Inference and Explanations

At test time, we slide a window over trajectories, compute activations $a_k(t)$, and predict actions y_t. Explanations are provided by (i) the SBX scenario summaries (offline) and (ii) nearest-neighbor windows to each prototype in the encoder embedding space.

- **Scenario-level** (global): SBX clusters and medoids summarize typical behaviors.
- **Temporal prototype-level** (local): per-prototype nearest windows (and prototype self-windows) illustrate characteristic action trajectories.

For each time step, form the most recent latent window, compute the encoder embedding and prototype activations, map them linearly to actions, and apply Tanh/ReLu post-processing. Key hyperparameters are L (window length), encoder size p, and learning rate. We select them on a held-out set using validation MSE and qualitative visualization of nearest-neighbor trajectories.

4 Experiments

4.1 Simulation and voltage control policy

We examine a real-world low-voltage distribution network consisting of 26 consumers, of which 7 are active consumers. Those active consumers are equipped with small solar plants (11kWp). The total yearly consumption in this network is negative, meaning that the solar plants are producing more electricity than is needed. A visual representation of the network is displayed in Fig. 2.

The learning process extended over 1500 episodes, each containing 96 steps (representing a 15-minute interval across one day). We evaluated the model every 20 episodes (1 epoch). In this network, we focus on handling mainly high voltages as those are a bigger problem in our example.

4.2 Explaining a Simulation

We consider a real low-voltage distribution network. An observation/state is the vector of per-bus voltage magnitudes $s = [v_1, \ldots, v_n]$ (in per unit). Actions are per-active-consumer flexibility commands $a = [\alpha_1, \ldots, \alpha_m]$ with $\alpha_i \in [-1, 1]$: negative values decrease consumption (or increase net export) and positive values decrease the generation for active consumers (bounded by their instantaneous battery output). The agent acts every 15

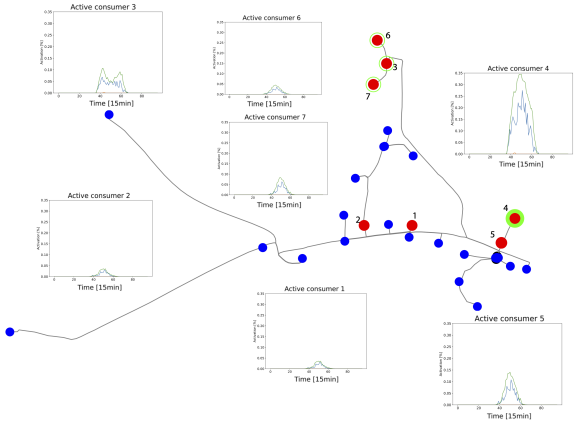

Figure 2: The network Srakovlje is located in Gorenjska region (north-western part of Slovenia). Active consumers (red), and their most representative activations are displayed with the corresponding graph. Green circles denote the most common over-voltage buses prior to voltage control. The width of the green circle indicates the severity of the original over-voltage measurements.

minutes; episodes comprise 96 steps (one day). The goal is to keep voltages within operating limits while minimizing interventions and losses.

Following prior work on distribution-voltage control [21], we use a reward that balances voltage quality, activation effort, and network losses. Trajectories are generated by a PPO policy trained in this environment.

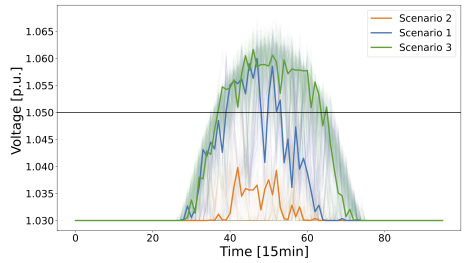

Figure 3: Centroids and underlying medoids of the scenarios in the Power Control environment. The individual color represents the average voltage signal in the network corresponding to the scenarios.

We used trajectories with length $L = 96$ which gives us $K = 3$ prototypes (Figure 3). Scenario selection via a silhouette-style criterion over $k \in \{2, \ldots, 8\}$ yielded a preferred $k = 3$ scenarios. Representative scenario-level activation summaries are shown in Figure 4. **Task fidelity:** offline action-level discrepancy against the reference policy (mean-squared error over held-out trajectories at the final step) was MSE = 3.218. **Scenario quality:** stored similarity scores by k were: $k = 2$: 0.131, $k = 3$: 0.118, $k = 4$: 0.083, $k = 5$: 0.082, $k = 6$: 0.089, $k = 7$: 0.093, $k = 8$: 0.096. A recomputed silhouette for the chosen $k = 3$ partition gave 0.099 with per-scenario supports $[4212, 7312, 5912]$ trajectories, indicating three regimes with substantial coverage. **Prototype locality:** In the latent space, the average distance from each prototype to its top-25 nearest trajectories was 0.124 on average, suggesting coherent time-local patterns.

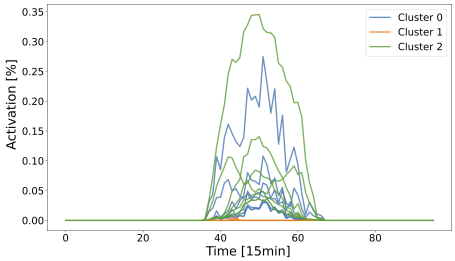

Figure 4: Representative prototypes in the Power Control environment. Each color represents the Scenario, and the individual line represents the activations by the individual active consumers.

4.3 Results

Fidelity. Across both domains, the prototype-based policy closely tracks the black-box in task reward, while achieving low action discrepancy on held-out episodes. This suggests that mediating actions through temporal prototypes does not materially degrade performance.

Global structure. SBX consistently discovers a small set of recurring scenarios that align with intuitive regimes (straight driving vs. cornering in continuous control; typical operating conditions in slower dynamics). Scenario summaries (state/action mean±std) are distinct and exhibit stable temporal patterns.

Local interpretability. For representative episodes, the nearest-neighbor aggregates around each prototype show coherent time-local patterns, and the most influential prototypes (largest contributions) align with observed actions. Explanations adopt a case-based form, relating current decisions to similar prototypical windows.

Performance Analysis. We compared the rewards across different policy architectures. Table **??** presents the results of running 20 episodes for each policy variant, measuring key performance metrics including mean reward, consistency (standard deviation), and coefficient of variation (CV) as a measure of reliability.

Over 20 episodes, the Base policy achieves the highest mean reward (221.8; range 201.0–257.5). PWNet closely matches the Base with a mean of 220.7 ($\approx$ 0.5% lower; range 185.8–249.5), indicating that mediating decisions through prototypes incurs negligible performance loss. The Temporal PWNet trades some reward for interpretability, averaging 211.5 ($\approx$ 4.7% below Base; range 168.4–231.8). Overall, relative performance is: Base $\approx$ 100%, PWNet $\approx$ 99%, Temporal PWNet $\approx$ 95%.

The results demonstrate several key insights about our approach. The Base Policy achieves the best rewards. The PWNet Policy shows comparable performance, indicating that prototype-based explanations can be achieved without significant performance degradation. Our Temporal PWNet + SBX approach achieves a mean reward of 211.47 ± 14.60, representing a modest performance trade-off in exchange for enhanced interpretability through temporal prototypes and scenario-guided explanations.

5 Discussion

This work introduces Scenario-Guided Temporal Prototypes, which combines global scenario discovery (SBX) with local, time-resolved prototypes to explain DRL decisions in voltage control problem in power networks. We observe that temporal prototypes can approximate black-box actions off-line with low discrepancy while forcing decisions through human-friendly exemplars. SBX discovers a small number of recurring regimes, with clear scenario-level summaries (Figure 3) and consistent prototype neighborhoods. This supports case-based reasoning over the policy's temporal dynamics rather than single-step feature attributions. Tight nearest-neighbor bands and balanced per-scenario support indicate that selected prototypes are representative rather than outliers.

The limitations of our current approach include reliance on a particular windowing choice and off-line evaluation that does not account for control feedback. Extremely imbalanced or highly non-stationary data may complicate selection. Prototype interpretability depends on the quality of medoids and the clarity of the associated concepts; domains lacking clear temporal motifs may benefit less from temporal prototypes and may also see degradation in performance. SBX does not identify the outliers that might be important for the agent to succeed. The identification of such states within the current architecture will be explored in future work. Future work also includes dynamic prototype lengths and human-in-the-loop curation tools for prototype editing and labeling.

6 Conclusion

We presented a pre-hoc interpretability framework that (i) discovers scenario structure from trajectories and (ii) explains actions via temporal prototypes. The approach yields faithful, time-resolved explanations without materially degrading control quality, as demonstrated in Power Network voltage control. Explanations take a case-based form—"this situation is similar to prototype X"—and are grounded by scenario summaries and prototype locality.

Beyond improving transparency, our approach offers practical steps: scenario coverage, per-scenario prototype counts, and nearest-neighbor coherence expose where explanations are strong or require refinement. Looking ahead, we plan to enable interactive prototype curation, incorporate uncertainty-aware explanation scores, and explore joint training schemes that couple prototype-based interpretability with context-aware latent dynamics. We will explore the sensitivity of the hyperparameter L to the actual training success. We have also identified that the fidelity metrics beyond the MSE will be necessary to explore. At this moment comparison to the saliency methods or SHAP explanations is still challenging due to the different nature of explanations (one being feature step-wise based and the other being multi-step and comparison based). Together, these steps can help bridge the gap between high-performing DRL policies and the trust required for their deployment.

Acknowledgements

This work was partially supported by the Slovenian Research Agency (ARIS), grant L2-4436: Deep Reinforcement Learning for optimization of LV distribution network operation with Integrated Flexibility in real-Time (DRIFT), and from the Slovenian Research Agency (ARIS) as member of the research program Artificial Intelligence and Intelligent Systems (Grant No. P2-0209).

References

[1] Shahab Bahrami, Yu Christine Chen, and Vincent W. S. Wong. 2021. Deep reinforcement learning for demand response in distribution networks. *IEEE Transactions on Smart Grid*, 12, 1496–1506.

[2] Di Cao, Junbo Zhao, Weihao Hu, Fei Ding, Nanpeng Yu, Qi Huang, and Zhe Chen. 2021. Model-free voltage control of active distribution system with PVs using surrogate model-based deep reinforcement learning. *Applied Energy*, 306, Part A, (Nov. 2021).

[3] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. 2019. This looks like that: deep learning for interpretable image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8930–8939.

[4] Ruisheng Diao, Zhiwei Wang, Di Shi, Qianyun Chang, Jiajun Duan, and Xiaohu Zhang. 2019. Autonomous voltage control for grid operation using deep reinforcement learning. *CoRR*, abs/1904.10597. arXiv: 1904.10597.

[5] Blaž Dobravec and Jure Žabkar. 2024. Explaining voltage control decisions: a scenario-based approach in deep reinforcement learning. In *Foundations of Intelligent Systems*. Springer Nature Switzerland, Cham, 216–230. ISBN: 978-3-031-62700-2.

[6] Samar Fatima, Verner Püvi, and Matti Lehtonen. 2020. Review on the PV hosting capacity in distribution networks. *Energies*, 13, 18.

[7] Natanael Gomes, Felipe Martins, José Lima, and Heinrich Wörtche. 2022. Reinforcement learning for collaborative robots pick-and-place applications: a case study. *Automation*, 3, (Mar. 2022).

[8] European Union Policy Iniciative. [n. d.] Growing consumption in the european markets. https://knowledge4policy.ec.europa.eu/growing-consumerism. Accessed: 2022-11-10. ().

[9] Eoin M. Kenny, Mycal Tucker, and Julie A. Shah. 2023. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *ICLR*.

[10] Bangalore Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Kumar Yogamani, and Patrick Pérez. 2020. Deep reinforcement learning for autonomous driving: A survey. *CoRR*, abs/2002.00444. arXiv: 2002.00444.

[11] Wong Ling Ai, Vigna Ramachandaramurthy, Sara Walker, and Janaka Ekanayake. 2020. Optimal placement and sizing of battery energy storage system considering the duck curve phenomenon. *IEEE Access*, 8, (Jan. 2020), 197236–197248. DOI: 10.1109/ACCESS.2020.3034349.

[12] Brida V. Mbuwir, Fred Spiessens, and Geert Deconinck. 2018. Self-learning agent for battery energy management in a residential microgrid. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 1–6.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602. arXiv: 1312.5602.

[14] Taha Nakabi and Pekka Toivanen. 2020. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy Grids and Networks*, 25, (Dec. 2020).

[15] Meike Nauta, Sander van Bree, and Christin Seifert. 2021. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14933–14943.

[16] Alvaro Rodriguez del Nozal, Esther Romero-Ramos, and Angel Luis Trigo-Garcia. 2019. Accurate assessment of decoupled oltc transformers to optimize the operation of low-voltage networks. *Energies*, 12, 11.

[17] Pedro Sequeira and Melinda T. Gervasio. 2019. Interestingness elements for explainable reinforcement learning: understanding agents' capabilities and limitations. *Artif. Intell.*, 288, 103367.

[18] David Silver et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815. arXiv: 1712.01815.

[19] David Silver et al. 2017. Mastering the game of go without human knowledge. *Nature*, 550, 354–359.

[20] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

[21] Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C. Green. 2021. Multi-agent reinforcement learning for active voltage control on power distribution networks. *CoRR*, abs/2110.14300. arXiv: 2110.14300.

[22] Junjie Wang, Qichao Zhang, Yao Mu, Dong Li, Dongbin Zhao, Yuzheng Zhuang, Ping Luo, Bin Wang, and Jianye Hao. 2024. Prototypical context-aware dynamics for generalization in visual control with model-based reinforcement learning. *IEEE Transactions on Industrial Informatics*, 20, 9, 10717–10727. DOI: 10.1109/TII.2024.3396525.

[23] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. 2021. Reinforcement learning with prototypical representations. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research). PMLR. https://arxiv.org/abs/2102.11271.

[24] Ke Zhang, Peidong Xu, and Jun Zhang. 2020. Explainable ai in deep reinforcement learning models: a shap method applied in power system emergency control. In *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, 711–716.

[25] Ke Zhang, Jun Zhang, Pei-Dong Xu, Tianlu Gao, and David Wenzhong Gao. 2022. Explainable ai in deep reinforcement learning models for power system emergency control. *IEEE Transactions on Computational Social Systems*, 9, 2, 419–427.