Appendix-A

This section presents an application case of Bayesian Optimization (BO) to Inverse Reinforcement Learning (IRL) problems. BO is a prevalent method for optimizing target weights. In this case study, the objective of BO is to align the feature expectations of the learned policy with those of the expert policy. Consequently, the objective function is defined as follows:

$$\min \sum_{i=1}^{n} w_i (V(\pi_{\mathrm{E}})_i - V(\pi)_i) \tag{A1}$$

Table A1 Simulation results of Bayesian Optimization

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
4	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
5	w1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	w1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	w2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	w1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	w4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

10 simulation experiments are conducted, each comprising 20 complete DRL+BO procedures. The resulting weights are displayed in Table A1. The data in the table reveals an expected outcome: the weights are concentrated on a single objective. This is because, based solely on the state-action pairs in the demonstration data, there is no ultimate criterion for evaluation. Under such conditions, BO struggles to ascertain the true weights.

Appendix-B

This section supplements the interpretability analysis for the a medium-scale smart distribution network, which is shown in Figure B1, based on the Deep-SHAP method. This distribution network is a modified IEEE-118 case with 10 MTs, 10 RESs, and 10 ESSs.

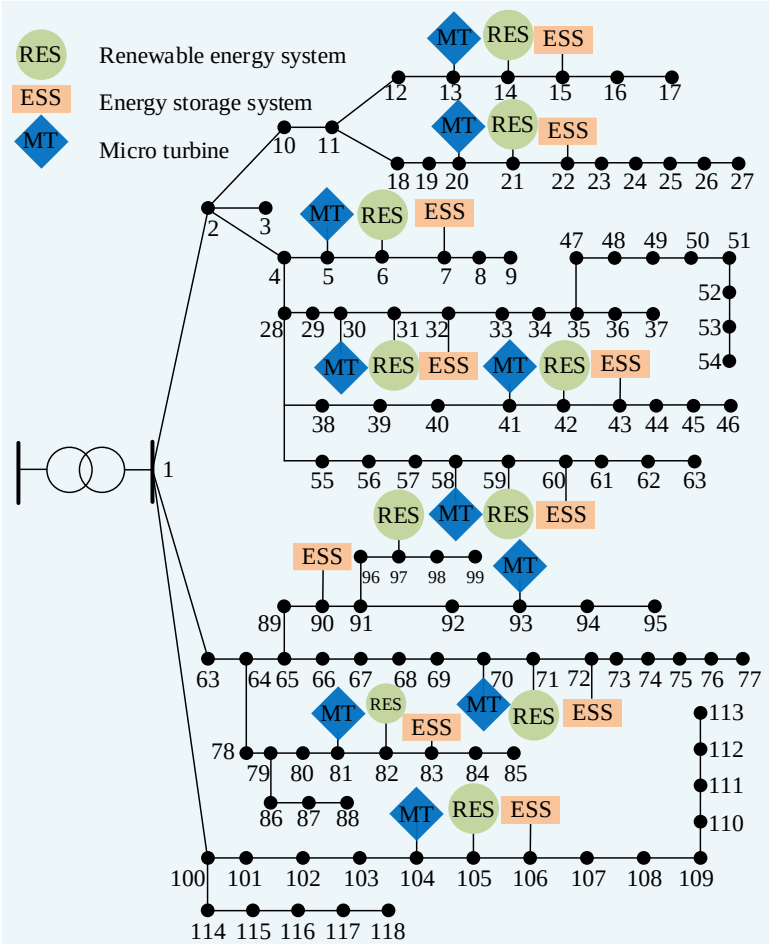

FIGURE B1 Modified IEEE-118 case

Firstly, an interpretability analysis was conducted on the policy network of the GGIRL framework. The single-sample surface plots and multi-sample surface plots are presented in Figures B2 and B3, respectively.

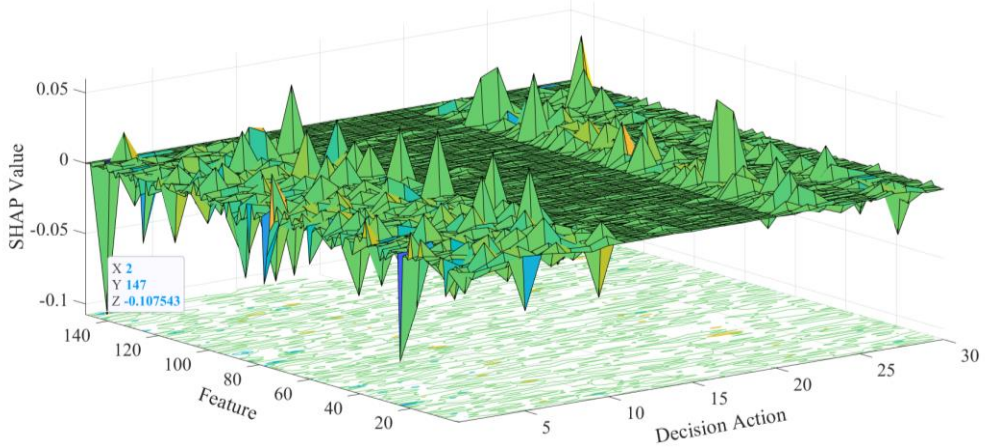

FIGURE B2 Policy network single-sample surface plot for IEEE-118 case

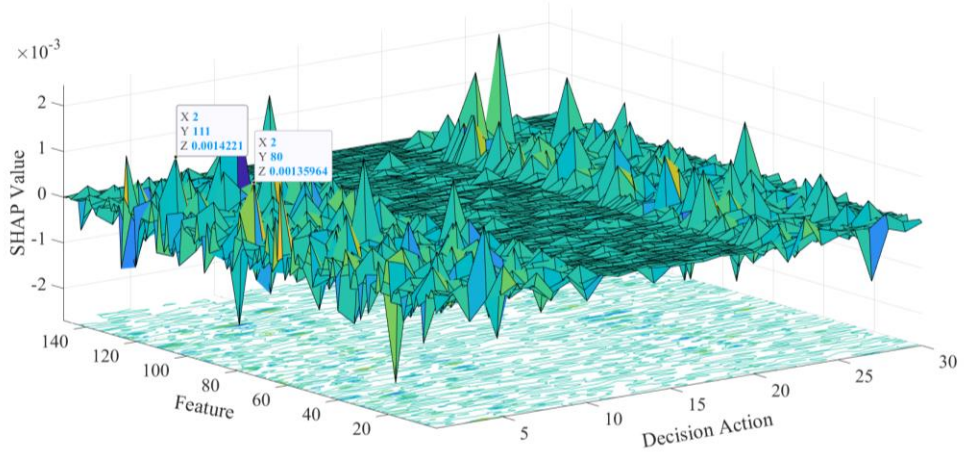

FIGURE B3 Policy network multi-sample surface plot for IEEE-118 case

It is evident from Figure B2 that the decision action areas from the 11-th to the 20-th are in a relatively flat state. In the context of the dispatch task, the 11-th to the 20-th decision actions correspond to RES-related actions in the system. Since the system's operational objectives include the goal of new energy consumption, the RES is almost always operating at full capacity, thus being minimally influenced by other factors. This same scenario can be extended to Figure B3. In the multi-sample analysis, the actions of the RES are similarly less affected by input features.

Further in-depth analysis reveals that, as depicted in Figure B2, the 147-th input feature exerts the most significant influence on the 2-th decision action, with a marginal contribution of 0.108. Conversely, the marginal contributions of some features approach zero. However, this pattern is observed only in the single-sample analysis. As indicated in Figure B3, statistically, the input features that most significantly affect the 2-th decision action are the 80-th and 111-th, with marginal contributions of 0.00136 and 0.00142, respectively. Therefore, single-sample analysis aids dispatchers in identifying the key system observables that require attention for the current decision. In the long-term operation of the system, a multi-sample analysis is necessary to help dispatchers comprehensively understand the factors influencing dispatch policies.

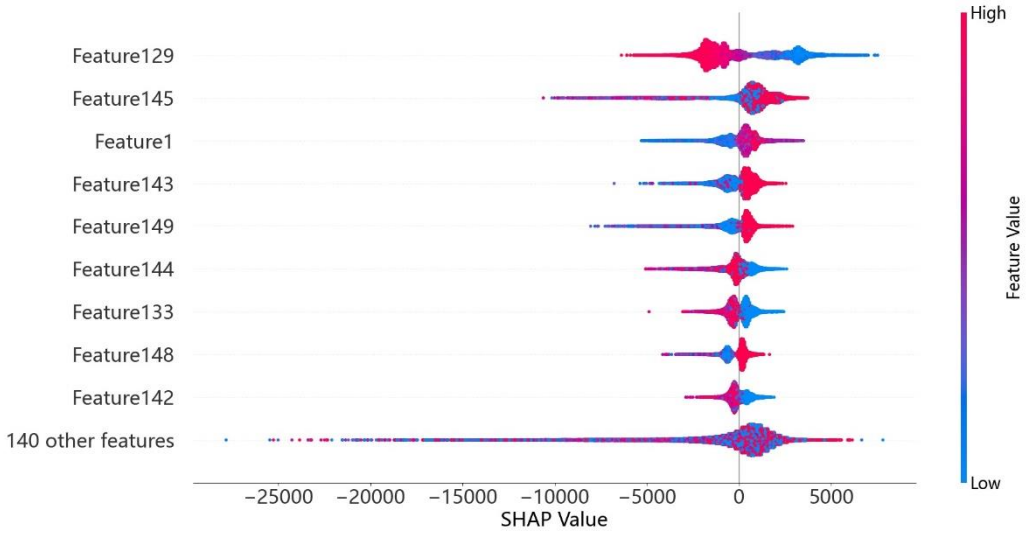

FIGURE B4 Scatter plot of multi-sample interpretation for reward function

Subsequently, an interpretable analysis of the reward function for the IEEE-118 case study is conducted, similar to that described in the main text. A detailed examination of 1200 samples is performed, and the resulting scatter plot is presented in Figure B4. The figure highlights the

input features that significantly influence the reward function. It clearly demonstrates that, statistically, the 129-th feature (representing the cost of purchasing electricity from the upper-level grid) exerts a considerable impact on the system's rewards. This conclusion aligns with the findings from the IEEE-33 case study. A meticulous examination indicates that the reward function benefits more significantly from the positive contributions when electricity prices are at a nadir. Given the inverse relationship between rewards and costs, this implies that lower electricity prices are associated with a decrease in the system's overall operational expenditures. Figure B4 offers a comprehensive interpretation of the reward function's behaviour. Through an extensive analysis of the Shapley values assigned to input features, the visualization delineates the pivotal features that exert the greatest influence on the reward function, thereby facilitating a deeper understanding of the reward dynamics for system dispatchers.

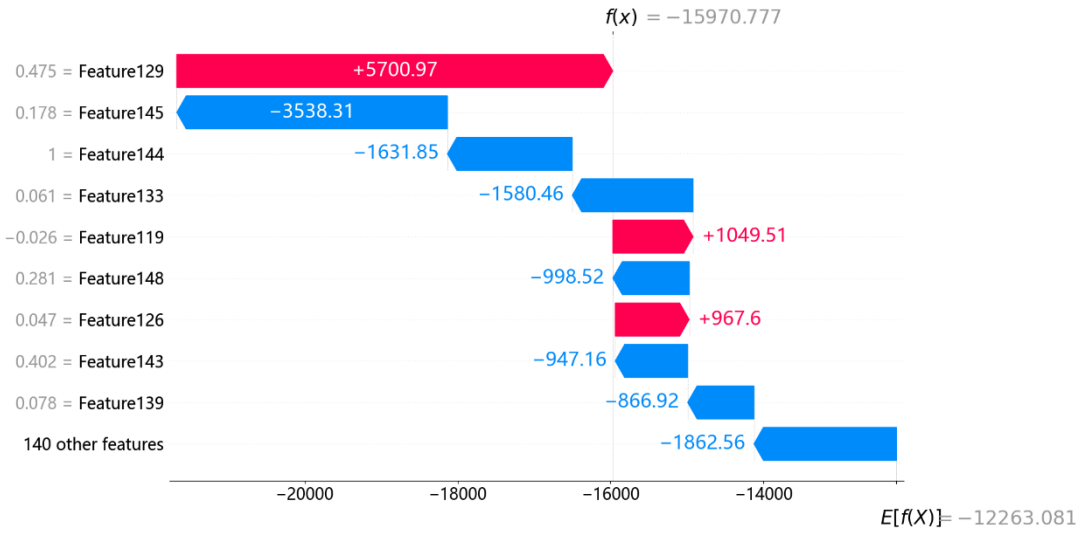

FIGURE B5 Waterfall plot of reward function single sample interpretation results

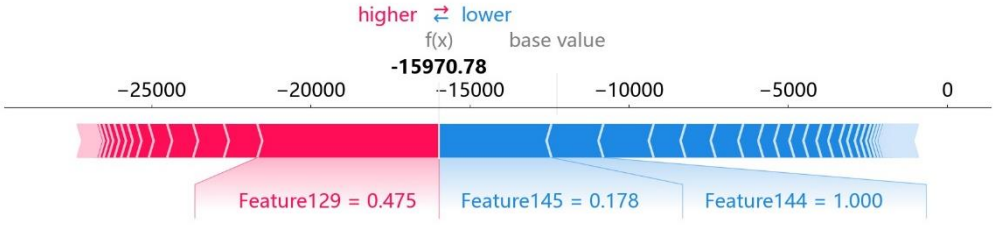

FIGURE B6 Force plot of reward function single sample interpretation results

An individual sample was randomly selected for a single-sample interpretable analysis of the reward function. Figures B5 and B6 depict the impact of individual input features on the final output of the reward function for the selected sample, visualized using waterfall and force plots, respectively. These figures reveal that the baseline output value of the reward function model, E[f(x)], is -12263.081. The 129-th feature, which represents the cost of purchasing electricity from the upper-level grid with a value of 0.475, exhibits the highest marginal contribution, with a SHAP value of 5700.97. This is followed by the 145-th and 144-th features, with SHAP values of -3538 and -1632, respectively. Under the influence of these features, the ultimate model output, f(x), is -15970.777. By interpreting the reward function for a single sample, the key input features that play a critical role in the reward (or negative cost) for this specific decision can be identified. This approach enables dispatchers to precisely pinpoint the input features that warrant attention within the context of the current decision-making scenario.

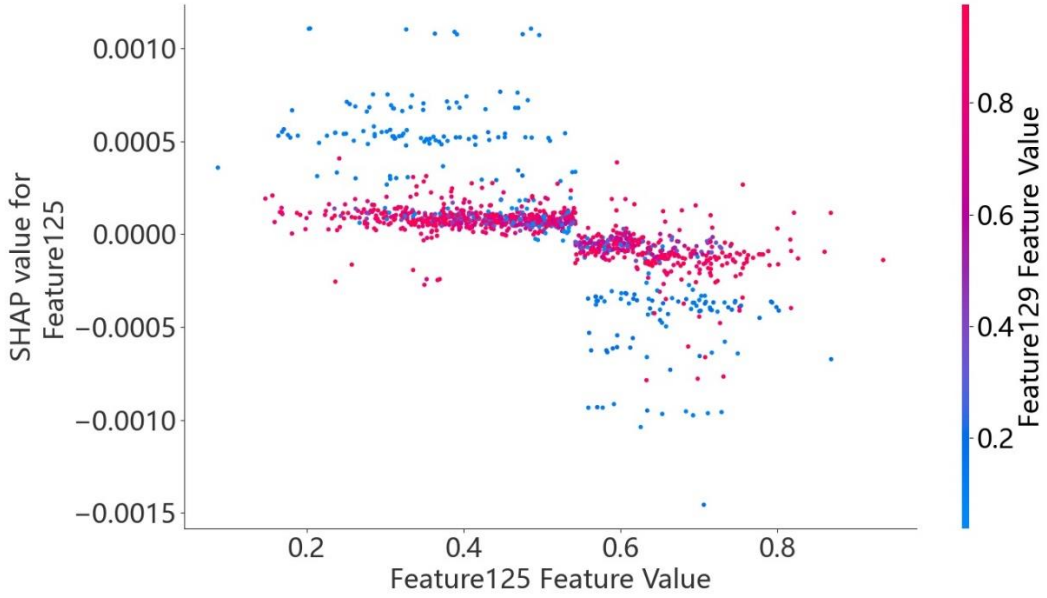

Figure B7 Plot of feature dependency

Finally, an attempt is made to elucidate the interactive effects between state features, focusing on the 125-th and 129-th features, which are randomly selected. The analysis revealed an intriguing finding. When the 129-th feature, representing the cost of purchasing electricity from the upper-level grid, assumes higher values, the marginal contribution of the 125-th feature, corresponding to the maximum power of the 7-th RES, remains relatively stable. Conversely, when the cost of purchasing electricity is lower, a distinct trend emerges: the marginal contribution of the 7th RES's maximum power is negative when it exceeds 0.5 p.u., and positive when it is less than 0.5 p.u. This observation provides insight into the interplay between these two features and their combined impact on the reward function.