A Framework for Safe Reinforcement Learning in Battery Storage Scheduling under Partial Observability

Giorgia Ghione, *Student Member, IEEE*, Vincenzo Randazzo, *Member, IEEE*, Giansalvo Cirrincione, *Senior Member, IEEE*, Kimon P. Valavanis, *Life Member, IEEE*, Eros Pasero, *Life Member, IEEE*, and Marco Badami

***Abstract*—The integration of Battery Storage Systems (BSS) into microgrids can enhance renewable energy utilization, reduce operating costs, and improve stability. However, the inherent uncertainty of photovoltaic (PV) generation and electricity demand presents significant challenges for safe and efficient control. This paper proposes a unified framework for safe BSS scheduling, using a dual safety mechanism integrated into a deep reinforcement learning (RL) agent operating under realistic, uncertain conditions. This approach exploits a novel agent architecture based on the Proximal Policy Optimization (PPO) algorithm with a custom feature extractor that captures both present operational states and long-term dependencies. This feature extractor leverages a Long Short-Term Memory (LSTM) encoder-decoder for PV generation and demand forecasting. Multiple safety metrics are evaluated, including violation rates during active operation, and the results are benchmarked against state-of-the-art model predictive control and RL methods. The proposed approach yields the best trade-off between economic performance and operational safety, improving safety compliance by over 74% relative to a model predictive control baseline while maintaining comparable economic return. The novelty of this work lies in jointly addressing partial observability, heterogeneous temporal dependencies, and multi-metric safety evaluation in a single framework for BSS scheduling.**

***Index Terms*—Battery energy storage systems, energy management systems, Long Short-Term Memory, microgrids, model predictive control, partial observability, Proximal Policy Optimization, safe reinforcement learning.**

I. Introduction

SHIFTING towards a zero-carbon economy and achieving global carbon neutrality represents a major challenge for humanity. Within this context, the energy sector holds a crucial role and faces considerable pressure to contribute with solutions [1]. The global push for these targets has led to an increased reliance on renewable energy sources, including photovoltaic (PV) systems, offering the potential to significantly decrease dependence on fossil fuels [2].

The integration of Battery Storage Systems (BSS) is emerging as a valuable solution for microgrids in the industrial sector [3], enabling increased renewable energy adoption, reduced operating costs, and enhanced stability [4]. However, the inherent variability and uncertainty of both renewable energy generation [5] and electricity demand [6] still present considerable operational challenges for energy microgrids [7]. Therefore, the development of advanced energy management systems is crucial to enable optimal scheduling of BSS. Traditional approaches have largely relied on the following methods: rule-based control (RB), as in [8] and [9]; mathematical programming (MP), as in [10], [11], and [12]; evolutionary algorithms (EA), as in [13]; model predictive control (MPC), as in [14], [15], and [16]. RB, MP, and EA methods, while simpler to implement, often lack adaptability to the dynamic operating conditions and the increasing complexity of BSS scheduling problems [17]. Furthermore, these traditional methods can be computationally intensive for real-time implementation [18].

Recent advancements have explored Reinforcement Learning (RL) techniques for energy management in microgrids [19] [20]. RL offers the potential to learn optimal control policies directly from data, without requiring explicit system models, and adapt through fine-tuning when operating conditions change. Multiple RL approaches have been implemented for BSS scheduling: state-action-reward-state-action (SARSA) in [21]; deep Q-network (DQN) in [22]; Deep Deterministic Policy Gradients (DDPG) in [23]; Twin Delayed DDPG (TD3) in [24] and [25]; Proximal Policy Optimization (PPO) in [26] [27] and [28]. In [12] and [29], time series forecasting is integrated in the energy management pipeline for increased robustness. However, traditional RL methods do not provide any theoretical guarantee that actions made during training will be safe in a real system [30]: they are susceptible to instability and safety concerns when integrated in real-world environments due to the potential for unsafe actions during the learning and deployment process. In the context of BSS scheduling, this may lead to battery degradation or even system failures if not carefully managed [31].

Safe RL has emerged as a crucial solution to address the inherent safety concerns of conventional RL when applied to critical infrastructures such as modern power systems [31]. Safe RL aims to mitigate risks by learning policies that maximize expected rewards while adhering to safety constraints

Manuscript received January 5, 2026; revised February 23, 2026.

This publication is part of the project PNRR-NGEU which has received funding from the MUR – DM 352/2022 (38-033-32-DOT1332092-3353). This research was also co-funded by TRIGENIA Srl. Dr. Randazzo acknowledges funding from the research contract no. 32-G-13427-2 (DM 1062/2021) funded within the Programma Operativo Nazionale (PON) Ricerca e Innovazione of the Italian Ministry of University and Research. Paper 1571033541 presented at 2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI), Lecco, Italy, Sept. 18-21 2024 [DOI: 10.1109/RTSI61910.2024.10761708]. *(Corresponding authors: Giorgia Ghione, Vincenzo Randazzo.)*

Giorgia Ghione, Vincenzo Randazzo and Eros Pasero are with the Department of Electronics and Telecommunications, Politecnico di Torino, Turin, 10129, Italy (e-mail: giorgia.ghione@polito.it, vincenzo.randazzo@polito.it).

Giansalvo Cirrincione is with the LTI Laboratory, University of Picardie Jules Verne, Amiens, 80025, France.

Kimon P. Valavanis is with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80210, USA.

Marco Badami is with the Department of Energy, Politecnico di Torino, Turin, 10129, Italy.

TABLE I: Safe-RL methods for BSS scheduling. Abbreviations: heating, ventilation, and air conditioning (HVAC); interior-point policy optimization (IPO); multi-agent twin delayed deep deterministic policy gradient algorithm (MATD3)

Article	**Base model**	**Safety strategy**	**Partial observability**	**Temporal dependency**	**Constraint type**	**Safety metrics**
[36]	MATD3	Shielding, Reward shaping	No	Very short-term horizon, only historical data	SOC, voltage	BSS controllability rate
[44]	IPO	Trust region method	No	Short-term horizon, only historical data	SOC, power limits, demand-supply balance, HVAC	Avg. thermal limit violations (MVA), avg. voltage magnitude violations (p.u.)
[41]	DDPG	Penalty in the objective	No	Short-term horizon, only historical data	SOC, power limits, demand-supply balance	Heat violation (MW), electricity violation (MW)
This work	*PPO*	*Shielding, Reward shaping*	*Yes*	*Short-term horizon, heterogeneous data*	*SOC, power limits, demand-supply balance*	*Total and active-operation violation rate, relative frequency and magnitude (kWh) of violations*

throughout both the learning and deployment phases [32]. One prominent approach involves adding a safe layer that is decoupled from the traditional RL framework, designed to verify and modify unsafe actions before execution [33]. This can be achieved through human intervention [34] or a shielding mechanism [35]. Applications of this approach to microgrid management have been presented in [36] and [37]. A limitation of the shielding approach lies in determining valid actions from those deemed unsafe, which necessitates prior knowledge of the system. Alternatively, action projection techniques mathematically map unsafe actions to the closest action within a defined safe space, as first proposed by [38] and applied in [39] and [40]. Nevertheless, linear approximations may inadequately represent complex system behavior, and frequent iterative calculations can substantially increase computation time. Another major category focuses on transforming the policy optimization criterion, like Lagrangian relaxation, where constrained problems are converted into unconstrained ones by penalizing constraint violations in the objective function, often integrated with algorithms such as soft actor-critic (SAC) [30] and DDPG [41]: they are simple to implement but it is challenging to balance constraint enforcement and reward maximization. A method that has recently gained attention is the trust region method, which modifies the policy optimization criterion by enforcing a trust region constraint, such as first-order constrained optimization in policy space (FOCOPS) [42] and penalized proximal policy optimization (P3O) [43]: they can dynamically balance constraint adherence and reward maximization, but penalty parameters must be carefully tuned. Such techniques are used in energy management, as in [44] and [45]. Other safe RL methods include imagination and reward shaping, which have been implemented in [46].

Despite progress in both traditional and safe RL approaches, several research gaps remain. Existing safe RL studies often evaluate a limited number of safety strategies without comparing them to traditional techniques. Furthermore, existing approaches to dynamic, uncertain power systems still lack efficient handling of time-series data: in particular, they do not address the problem as a partially observable one due to uncertainties and do not explicitly handle heterogeneous temporal dependencies, which can negatively impact the performance and stability of control strategies. Table I summarizes recent safe-RL studies on BSS scheduling, highlighting these gaps.

This paper addresses these gaps by presenting a novel framework for optimizing BSS scheduling in a microgrid. The key contributions of this work are the following:

- A safe BSS scheduling framework is developed: it incorporates a microgrid environment with multiple safety constraints, operating under partial observability.
- A novel PPO-based RL agent architecture is proposed, featuring a custom feature extractor with present and lookahead state heads to simultaneously take into account the present operational state and longer-term dependencies. A pre-trained encoder-decoder long short-term memory (LSTM-ED) model is used to forecast PV production and electric demand.
- The agent is trained with a dual safety mechanism, comprising reward shaping to steer the policy toward feasible actions and shielding to block unsafe actions. This combination yields a single, end-to-end controller that simultaneously maximizes economic gains and satisfies operational constraints: it reduces active-operation violations by 74.28% relative to an MPC baseline and by 37.37% relative to a shielding baseline, while maintaining comparable economic performance (under 1% difference).
- A comprehensive benchmarking against state-of-the-art safe RL and MPC strategies is conducted, considering multiple safety metrics and the trade-off between safety and economic return.

This work is an extension of the conference paper titled *Optimizing Battery Storage Systems in Energy Microgrids: A Reinforcement Learning Approach comparing multiple reward functions* [47]. In this new version, a custom feature extractor is integrated to address the partially observable nature of the problem and leverage the variety of time-series information available, including a new LSTM-ED model to forecast PV generation and electric demand; a detailed comparative analysis of safety strategies is carried out, based on both economic and safety metrics; finally, a comparison with two traditional MPC controllers is added.

II. System description

The considered energy system, shown in Fig. 1, includes a PV station and a BSS designed to store surplus energy generated by the PV station for later use to meet the demand of a user. The equation which models the update of the SOC of the BSS is (1):

$$SOC_{t+1} = SOC_t + \eta_{cha} E_{cha,t} - \frac{E_{dis,t}}{\eta_{dis}}, \tag{1}$$

where $E_{cha,t}$ and $E_{dis,t}$ represent the charged and discharged energy, respectively, η_{cha} and η_{dis} are the charging and discharging efficiency, respectively. Simultaneously charging and discharging is not allowed. Thus, the energy exchange with the BSS is denoted as $E_{BSS,t}$, which is positive when discharging and negative when charging: $E_{BSS,t} = E_{dis,t} - E_{cha,t}$ with $E_{cha,t}E_{dis,t} = 0$. The BSS can charge solely from surplus PV energy and discharge no more than the residual demand. The exchange of energy with the national grid is determined via (2), which is the energy balance equation:

$$E_{D,t} - E_{PV,t} - E_{BSS,t} - E_{GEx,t} = 0, \tag{2}$$

where $E_{GEx,t}$ is the energy exchanged with the grid, $E_{D,t}$ is the electric demand and $E_{PV,t}$ is the energy generated by the PV station. Two further quantities are defined: grid energy purchase ($E_{GP,t}$) and surplus sale ($E_{GS,t}$). If $E_{GEx,t} \geq 0$, then $E_{GP,t} = E_{GEx,t}$ and $E_{GS,t} = 0$; otherwise, $E_{GP,t} = 0$ and $E_{GS,t} = |E_{GEx,t}|$.

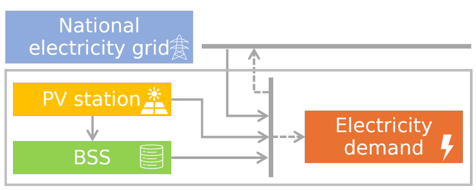


Fig. 1: Scheme of the energy system.

III. Problem Formulation

An energy management system is introduced to optimize the BSS schedule to minimize costs and constraint violations operating on an hourly time step. At each step, the system receives information on SOC, demand, PV generation, and single national price (SNP). The latter is the wholesale reference price for electricity exchanged with the national grid. Then, the system determines the BSS charging or discharging action. PV self-consumption and grid exchange amounts are subsequently calculated to ensure the energy balance. A schematic of this problem is shown in Fig. 2.

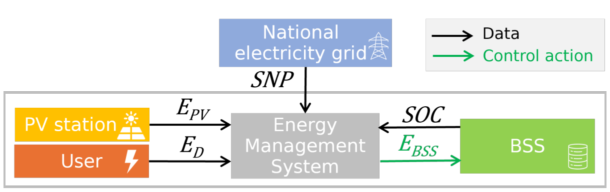


Fig. 2: Diagram of the energy management problem.

A. Objective function

The objective function is the Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) [48]:

$$EBITDA = P - C, \tag{3}$$

where P is the revenue from selling the surplus electric energy to the grid and C is the cost of the electric energy purchased from the national grid. They are defined below:

$$P = E_{GS}(SNP + spread_S), \tag{4}$$
$$C = E_{GP}(SNP + spread_P), \tag{5}$$

where E_{GP} and E_{GS} are derived from E_{GEx} (see section II), SNP is the single national price, $spread_P = 0.080$ €/kWh and $spread_S = 0.020$ €/kWh are the fixed purchasing and selling price, respectively. Costs related to the operation and maintenance of the BSS were not considered.

B. Constraints

At each time step t, the system is subject to the following set of constraints to ensure that it operates within physical system boundaries:

$$SOC^{min} \leq SOC_t \leq SOC^{max} \tag{6}$$
$$E_{GEx,t} - E_{D,t} + E_{PV,t} - E_{cha,t} + E_{dis,t} = 0 \tag{7}$$

In (6), SOC_t is the current SOC of the BSS, SOC^{min} and SOC^{max} define the operation range of the SOC. Equation (7) is the balance of energy introduced in (2). In addition, $E_{dis,t}$ and $E_{cha,t}$ are subject to the following constraints:

$$E_{cha,t}E_{dis,t} = 0 \tag{8}$$
$$E_{cha,t} \leq E^{max}_{cha,t} \tag{9}$$
$$E_{cha,t} \leq \min(0, E_{D,t} - E_{PV,t}) \tag{10}$$
$$E_{dis,t} \leq E^{max}_{dis,t} \tag{11}$$
$$E_{dis,t} \leq \max(0, E_{D,t} - E_{PV,t}) \tag{12}$$

Here, $E^{max}_{cha,t}$ and $E^{max}_{dis,t}$ represent the maximum amount of energy that can be charged or discharged at each time step, respectively:

$$E^{max}_{cha,t} = \frac{SOC^{max} - SOC_t}{\eta_{cha}}, \tag{13}$$
$$E^{max}_{dis,t} = (SOC_t - SOC^{min})\eta_{dis}. \tag{14}$$

C. Markov Decision Process formulation

RL problems are commonly modelled as Markov Decision Processes (MDPs), a discrete-time stochastic framework for sequential decision-making. An MDP comprises a state space S, an action space A, transition dynamics $T: S \times A \to \Delta(S)$, and a reward function $r = S \times A \times S \to \mathbb{R}$ with discount factor γ. Central to the definition of an MDP is the Markov property, which assumes that the probability distribution of the next state is dependent solely on the current state-action pair. Nevertheless, full state observability is rare in real-world environments [49]: such problems are typically modelled as partially observable MDPs (POMDPs) [50]. Formally, a POMDP is described by the 6-tuple (S, A, T, r, Ω, O), where the agent receives observations $o \in \Omega$ drawn from a distribution, $o \sim O(s)$, rather than having direct access to the true state. The following sections define the main elements within this framework.

1) State: In the BSS scheduling problem presented in this work, the agent must account for two stochastic exogenous variables that evolve independently of its actions: the aggregate electric demand $E_{D,t}$ of the user and the on-site PV generation $E_{PV,t}$. Neither variable is directly controllable and their future realizations are unknown at decision time. This uncertainty requires formulating the task as a POMDP. Specifically, at each time step t, the latent state of the environment $s_t \in S$ is defined as:

$$s_t = (h_t, SOC_t, \mathbf{SNP}_t^{(H)}, E_{D,t}, E_{PV,t}). \tag{15}$$

Here, h_t represents the hour of the day, and $\mathbf{SNP}_t^{(H)} = (SNP_t, ..., SNP_{t+H-1})$ is the SNP for the next $H = 12$ hours. The SNP is known 12 hours in advance. Because $E_{D,t}$ and $E_{PV,t}$ are not observable, they are replaced with their respective H-step-ahead forecasts $\hat{\mathbf{E}}_{D,t}^{(H)} = (\hat{E}_{D,t|t}, ..., \hat{E}_{D,t+H-1|t})$ and $\hat{\mathbf{E}}_{PV,t}^{(H)} = (\hat{E}_{PV,t|t}, ..., \hat{E}_{PV,t+H-1|t})$, where $\hat{E}_{D,t+k|t}$ and $\hat{E}_{PV,t+k|t}$ are the forecasts made at time t of $E_{D,t+k}$ and $E_{PV,t+k}$, respectively. The observable surrogate state supplied to the RL agent, therefore, becomes

$$\tilde{s}_t = (h_t, SOC_t, \mathbf{SNP}_t^{(H)}, \hat{\mathbf{E}}_{D,t}^{(H)}, \hat{\mathbf{E}}_{PV,t}^{(H)}). \tag{16}$$

2) Action: The action $a_t \in A$ denotes the BSS energy exchange at time t, i.e. $a_t = E_{BSS,t}$. If $a_t > 0$ the BSS discharges: $E_{dis,t} = E_{BSS,t}$, $E_{cha,t} = 0$. If $a_t < 0$ it charges: $E_{cha,t} = -E_{BSS,t}$, $E_{dis,t} = 0$. The SOC of the BSS is then updated based on (1).

3) Reward: The EBITDA objective function in (3) includes stochastic PV generation and electric demand, which lead to a high-variance learning signal. To isolate the economic benefit of the BSS and stabilize training, the following reward is defined:

$$r_t = k_r(AC_t + P_t) \tag{17}$$

where P_t is revenue from selling surplus energy in (4) and $AC_t = E_{dis,t}(SNP_t + spread_P)$ is the cost saved by self-consuming BSS discharge. Omitting the grid-purchase term C_t ensures an always positive, low-variance signal that accelerates and stabilizes policy optimisation. The coefficient $k_r = 10^4$ leaves the underlying MDP unchanged while scaling rewards to $\mathcal{O}(10^0 - 10^2)$, which empirically speeds up convergence .

IV. Proposed safe BSS scheduling framework

This section outlines the components of the proposed safe BSS scheduling framework, designed with a focus on safety constraints. It first describes the PV generation and electric demand forecaster, followed by a custom feature extractor that incorporates this forecaster to exploit diverse time-series data. Next, the main RL algorithm used (i.e., PPO) is explained, and the safety paradigms implemented are summarized. Lastly, a description of the overall workflow of the training and testing processes is provided.

A. PV generation and electric demand forecaster

Accurate H-step-ahead forecasts of E_{PV} and E_D strongly impact the decision-making performance of the battery-storage scheduling agent. State-of-the-art multi-step time-series forecasters range from classical statistical models to deep-learning architectures. For PV and demand forecasting, however, encoder-decoder Long Short-Term Memory (LSTM-ED) networks have shown a trade-off between accuracy and computational cost [51], [52]. Therefore, a sequence-to-sequence LSTM-ED network that maps historical PV generation and electric demand to their respective H-hour-ahead forecasts was pre-trained. Fig. 3 summarizes the architecture of the LSTM-ED model implemented.

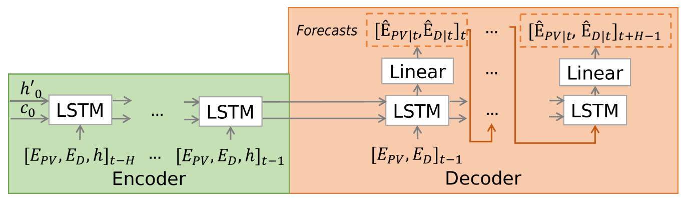


Fig. 3: Architecture of the LSTM-ED model. A compact notation is adopted, where the temporal indices for all input variables are specified outside the brackets to avoid repetition.

B. Feature extractor

The surrogate state $\tilde{s}_t$ mixes instantaneous variables (e.g., SOC_t), deterministic signals (h_t, $\mathbf{SNP}_t^{(H)}$), and stochastic multi-step forecasts $(\hat{\mathbf{E}}_{D,t}^{(H)}, \hat{\mathbf{E}}_{PV,t}^{(H)})$ output by the pre-trained LSTM-ED. Therefore, the agent must simultaneously take into account the present operational state and longer-term dependencies. In order to reduce the dimensionality of the state while separately handling present and future components of the state, a custom feature extractor is designed. Fig. 4a illustrates the two-headed feature extractor that transforms the observable surrogate state $\tilde{s}_t$ into the compact representation $\boldsymbol{\phi}_t \in \mathbb{R}^{d_\phi}$ processed by the RL agent. In particular, the present-state head aggregates the state of variables at the decision instant t:

$$\mathbf{p}_t = \left[h_t,\ SOC_t,\ SNP_t,\ \hat{E}_{PV,t|t},\ \hat{E}_{D,t|t}\right]^\top \in \mathbb{R}^5. \tag{18}$$

This vector is passed unchanged to the final concatenation node. The lookahead-state head is composed of a single-layer LSTM encoder (LSTM-LH) with hidden size d_z which receives the forecasts of PV generation and electric demand, the known SNP and hour variables across the next $H-1$ steps. It then outputs a fixed-length context vector $\mathbf{z}_t \in \mathbb{R}^{d_z}$, capturing longer-term dependencies. Finally, the two branches are concatenated and forwarded to the shared backbone of the RL agent:

$$\boldsymbol{\phi}_t = \left[\mathbf{p}_t^\top,\ \mathbf{z}_t^\top\right]^\top \in \mathbb{R}^{d_\phi}, \qquad d_\phi = 5 + d_z. \tag{19}$$

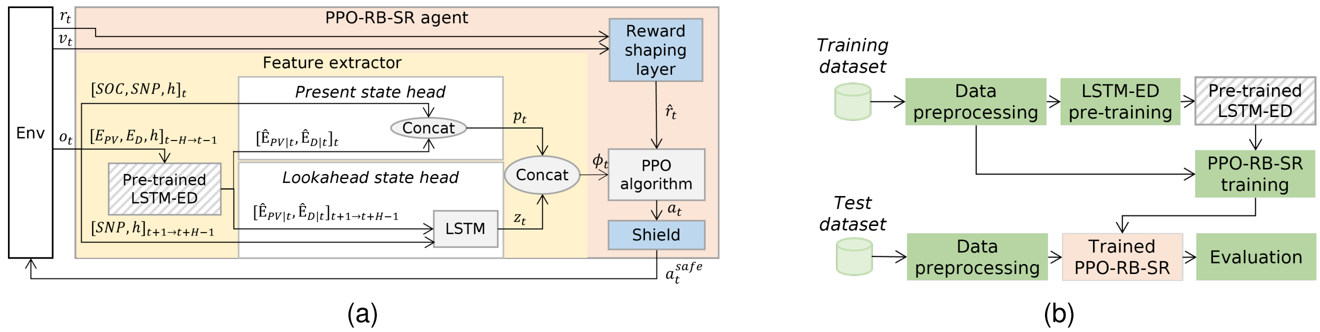


Fig. 4: (a) Schematic representation of the PPO-RB-SR agent, integrating the feature extractor and the dual safety mechanism. A compact notation is adopted, where the temporal indices for all input variables are specified collectively outside the brackets to avoid repetition, and time ranges are expressed as $t_0 \rightarrow t_1$. The block denoted as *Env* represents the environment summarized in Fig. 2. (b) High-level algorithmic flowchart illustrating the sequential stages of data preprocessing, training, and evaluation.

C. Proximal policy optimization algorithm

The RL model selected in this work is Proximal Policy Optimization (PPO). PPO is a state-of-the-art first-order, on-policy deep RL algorithm with an actor-critic architecture. It was proposed in [53] as an improvement over standard policy gradient methods. It serves as a core component for numerous approaches to safe RL. The PPO loss in [53] is computed as

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_{\substack{s\sim d^\pi \\ a\sim\pi}} [\mathcal{L}_{\text{CLIP}}(\theta) + c_{\text{vf}}\,\mathcal{L}_{\text{vf}}(\theta)], \tag{20}$$

where π is the current policy with parameters θ, d^π is the discounted future state distribution, $\mathcal{L}_{\text{vf}} = (V_\theta(s_t) - V_t^{targ})^2$ is the squared-error value-function loss with weighting coefficient c_{vf}, and $\mathcal{L}_{\text{CLIP}}(\theta)$ is the clipped surrogate objective, defined in [53] as

$$\mathcal{L}_{\text{CLIP}}(\theta) = \mathbb{E}_{\substack{s\sim d^\pi \\ a\sim\pi}}\Big[\min\big\{\rho(\theta)\,A^\pi,\ \text{clip}\big(\rho(\theta),\,1-\varepsilon,\,1+\varepsilon\big)\,A^\pi\big\}\Big] \tag{21}$$

where $\rho(\theta)$ is the probability ratio between the new and the old policy, A^π is the generalised advantage estimate (GAE), and $\varepsilon \in (0, 1)$.

In this study, the default flattening extractor of PPO is replaced with the custom feature extractor (Section IV-B) as a shared layer preceding both actor and critic networks to process time-series data.

D. Safety strategies

To ensure safe and reliable operation of the BSS within the microgrid, two safety paradigms are implemented:

1) Shielding: A shielding mechanism based on predicted PV generation and demand is incorporated, adopting an *action replacement* approach: it substitutes any unsafe action a_t with a default safe action a_t^{safe}, ensuring safety with minimal intervention.

2) Reward Shaping: To guide the RL agent towards safe and effective behavior, reward shaping is employed. The original reward function is modified by incorporating a penalty term for actions that violate system constraints and an additive bonus term to incentivize safe actions:

$$\hat{r}_t = r_t - p\mathbb{I}_{\{\sum_i v_t^i \geq 1\}} + b\mathbb{I}_{\{\sum_i v_t^i = 0\}}, \tag{22}$$

where $v_t^i = \mathbb{I}_{\{\text{violation of the } i\text{-th constraint at time } t\}}$ are violation indicators. The scalar weights are $p = 0$ and $b = 0.01$. They are fixed constants experimentally tuned to balance reward maximization with safety compliance.

As illustrated in Fig. 4a, the two safety strategies, namely, reward bonus (RB) and shielding via replacement (SR), are integrated within the RL architecture as distinct layers that interface between the PPO backbone and the environment. The resulting agent is denoted as PPO-RB-SR. At each time step t, the feature extractor maps the observation o_t received from the environment to a context vector $\boldsymbol{\phi}_t$, which is the input for the PPO algorithm to produce a candidate action a_t. The shielding block intercepts a_t and evaluates constraint violations based on o_t and the predicted PV generation and demand. If a potential violation is detected, the action is replaced with a fixed safe action a_t^{safe} for execution in the environment. The environment then returns a reward r_t and a violation indicator vector v_t. Finally, the reward shaping layer adjusts r_t based on v_t to produce the augmented reward $\hat{r}_t$, which is used to update the weights of the PPO model.

E. Algorithmic workflow

The overall algorithmic workflow of the proposed framework, encompassing both the training and testing phases, is depicted in Fig. 4b. The training phase begins with the preprocessing of the training dataset, which is used to pre-train the LSTM-ED forecaster to provide the multi-step-ahead forecasts required for the state representation. The pre-trained LSTM-ED is frozen and integrated into the feature extractor. During the training loop, the agent learns the optimal scheduling policy by interacting with the environment through the shielding and reward-shaping mechanisms. Finally, the performance of the trained agent is validated using a test dataset. This ensures that the agent's ability to optimize BSS scheduling while maintaining safety generalizes to unseen PV and demand profiles.

V. Case study

This section introduces the case study, which concerns the microgrid of a transport hub located in Italy, the experimental setup of the proposed approach, the baseline methods for benchmarking, and the metrics used in this work.

A. Dataset and pre-processing

The collected dataset comprises generation data from the on-site PV station, electric demand data from the transport hub facilities, and the SNP data from January 1, 2023, to December 31, 2023, aggregated hourly. Daily 12-hour-ahead SNP values are publicly available in Italy and were gathered throughout the specified period. The PV station has a nominal power equal to 22 MW. The BSS has a nominal capacity SOC^{max} of 10 MWh ($SOC^{min} = 0$ MWh), a charging and discharging efficiency (η_{cha} and η_{dis}, respectively) of 0.935, and requires one hour for a complete charge or discharge cycle. The electric demand of the transport hub ranges from 3937 kWh to 13379 kWh, and the SNP ranges from 0.001 €/kWh to 0.870 €/kWh.

A training set has been extracted, selecting one day for every month of the year (288 data points). Then, the whole dataset (8760 data points) has been used for testing to evaluate yearly metrics. The data sets have been scaled in the interval $[-1, 1]$ with min-max scaling to improve training stability and convergence.

B. Experimental setup

The PV generation and electric demand forecaster LSTM-ED has been implemented via the Pytorch framework [54]. It comprises an encoder and a decoder, each containing two LSTM layers; the decoder includes a fully connected output layer. The LSTM hidden-state dimension is 32. The training data have been randomly partitioned into 80% and 20% subsets, used for its training and validation, respectively. Training employs mini-batches of 32 samples and the Adam optimizer with a learning rate equal to $7.74 \cdot 10^{-3}$ and early stopping. The model has been trained for 103 epochs.

The proposed PPO agent (PPO-RB-SR) features a hybrid safety approach: it combines the reward-shaping bonus technique (with $p = 0$ and $b = 0.01$) and action replacement shielding ($a_t^{safe} = 0$). The hyperparameters p and b are fixed constants and have been tuned to identify the configuration that optimally balances operational efficiency with constraint satisfaction. The PPO implementation employed is the discrete version of PPO available in the Stable Baselines3 library [55], with the addition of the custom feature extractor in place of the default flattening extractor. Its hyperparameters are summarized in Table II. The action space comprises 98 discrete values ranging from -10000 kW to 9350 kW in 200 kW increments; the upper limit is truncated at 9350 kW, which is the maximum feasible discharge after accounting for efficiency. Although the BSS uses a continuous power range, a discrete action space was chosen to improve RL model performance and convergence speed [56]. This approach is standard for energy storage arbitrage, allowing for more robust policies [57].

C. Baseline methods

To validate the economic return and constraint compliance of the proposed BSS scheduling approach, the performance of PPO-RB-SR is benchmarked against different state-of-the-art methods. They include multiple safe RL methods and two MPC controllers.

TABLE II: Main hyperparameters of PPO

Hyperparameter	PPO	Hyperparameter	PPO
No. environments	4	Advantage norm.	Yes
Actor/critic layers	2	c_{vf}	0.5
Actor/critic units	128	Gradient clipping	0.5
Learning rate	$3 \cdot 10^{-4}$	LSTM-LH layers	1
Steps	288	LSTM-LH units (d_z)	16
Batch size	64	λ	0.95
Epochs	10	ϵ	0.2
γ	0.99		

Five PPO agents have been trained, each employing a different safe RL strategy and the same PPO hyperparameters detailed in Table II. These benchmarks are categorized as follows:

- *Shielding*: Two variants have been evaluated. First, a PPO agent with action projection shielding (PPO-SP), mapping a_t to the nearest action a_t^{safe} within the safe action space via constraint-based clipping. Second, a PPO agent with action replacement shielding (PPO-SR), where invalid actions were replaced with the fixed action $a_t^{safe} = 0$.
- *Reward shaping*: Two variants have been evaluated. First, a PPO agent featuring the penalty approach with $p = 0.01$ and $b = 0$ (PPO-RP). Second, a PPO agent featuring the bonus approach with $p = 0$ and $b = 0.01$ (PPO-RB).
- *Trust region method*: A Penalized Proximal Policy Optimization (P3O) [43] model has been created by customizing the PPO algorithm. P3O integrates safety constraints into the PPO objective via a ReLU-based penalty. In this work, the adaptive update of the penalty coefficient $\kappa_i \leftarrow \min(\rho\kappa_i, \kappa_{\max})$ is executed only when the empirical discounted cost J_{C_i} exceeds the limit d_i, and the critic includes an additional cost-value head. P3O hyperparameters are $\kappa_0 = 0.1$, $\rho = 0.1$, $\kappa_{max} = 10$, and $d_i = 0$.

Additionally, performance has been benchmarked against two MPC controllers. MPC is widely used in industrial processes, building automation [58], and microgrid power management [16]. It optimizes control trajectories by minimizing a cost function subject to constraints, executing only the first action at each time step. In this work, a 1-hour time step and a 12-hour horizon are considered. The state variable is the SOC in (1), controllable inputs are $E_{cha,t}$ and $E_{dis,t}$, and auxiliary variables are $E_{BSS,t}$ and $E_{GEx,t}$. The disturbances are SNP_t and forecasts $(\hat{\mathbf{E}}_{D,t}^{(H)}, \hat{\mathbf{E}}_{PV,t}^{(H)})$, with $H = 12$. The objective function in (3) is formulated as a cost minimization. Two MPC variants have been evaluated:

- A deterministic MPC (MPC-D), which uses perfect realizations $(\mathbf{E}_{D,t}^{(H)}, \mathbf{E}_{PV,t}^{(H)})$ to establish an upper-bound performance benchmark, since the problem is convex.
- A stochastic MPC (MPC-S), which uses forecasts $(\hat{\mathbf{E}}_{D,t}^{(H)}, \hat{\mathbf{E}}_{PV,t}^{(H)})$, with $H = 12$, to manage uncertainty and enforce operational constraints in a realistic setting.

They are developed using the discrete implementation available in the do-mpc library [59] with an Interior Point Optimizer (IPOPT) solver [60]. Its tolerance is set to 10^{-7} and its maximum number of iterations to 50000 via hyperparameter tuning.

All the experiments have been conducted using the same training and testing sets. To reflect realistic operational practices incorporating multiple safety layers (e.g., protection devices and emergency shutdowns), actions leading to constraint violations are replaced with a fixed action ($a_t = 0$) during testing. This ensures a fair assessment of agent performance within a broader, realistic safety context.

D. Metrics

The primary economic metric is EBITDA (see (3)). Safety is evaluated using two main violation metrics. The overall violation rate (VR) measures the percentage of time steps with any violation during the entire test simulation. A second metric, VRA, computes the violation rate only for time steps where $a_t \neq 0$ (VRA), providing a more granular measure of constraint compliance during active operation. The trade-off between EBITDA and VRA can be summarized with the trade-off index I:

$$I = \frac{|\text{EBITDA}| - |\text{EB}_{\min}|}{|\text{EB}_{\max}| - |\text{EB}_{\min}|} - \frac{\text{VRA} - \text{VRA}_{\min}}{\text{VRA}_{\max} - \text{VRA}_{\min}}, \quad (23)$$

where $\text{EB}_{\min} = 0$, $\text{EB}_{\max}$ is the EBITDA of MPC-D, $\text{VRA}_{\min} = 0$, and $\text{VRA}_{\max} = 100$.

VI. Results and Discussion

A. PV generation and electric demand forecasting

The performance of the LSTM-ED model is benchmarked against two state-of-the-art baselines, a gated recurrent unit encoder-decoder (GRU-ED) and a one-dimensional convolutional neural network (CNN-1D), using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2) metrics. The GRU-ED consists of two layers with 32 hidden units, while the CNN-1D employs 64 channels and a kernel size of 7; both were trained with specific learning rates ($4.23 \cdot 10^{-3}$ and $2.33 \cdot 10^{-3}$, respectively) and a batch size of 32. To ensure a rigorous comparison, all models were trained and evaluated under identical conditions.

As shown in Table III, for E_{PV}, the LSTM-ED emerges as the best performing model. In the case of E_D, all models exhibit good predictive performance ($R^2 = 0.92$), with GRU-ED providing marginal improvement. Overall, the LSTM-ED maintains the most consistent performance across both variables.

TABLE III: Performance comparison of LSTM-ED, GRU-ED, and CNN-1D

Variable	Algorithm	MAE (kWh)	RMSE (kWh)	R^2
E_{PV}	LSTM-ED	**847.78**	**1720.46**	**0.87**
	GRU-ED	876.00	1742.52	0.86
	CNN-1D	949.65	1815.14	0.85
E_D	LSTM-ED	421.59	559.59	**0.92**
	GRU-ED	**420.85**	**541.83**	**0.92**
	CNN-1D	428.73	554.75	**0.92**

B. BSS scheduling

Fig. 5 shows the mean episode reward and episode violation count of each RL strategy during training. The rewards are not directly comparable because different bonus or penalty magnitudes shift the overall reward scale; however, it is possible to observe that all the models converge to a stable policy with a low violation count, with PPO-RB, PPO-RP, and PPO-RB-SR exhibiting the best safety performance during training.

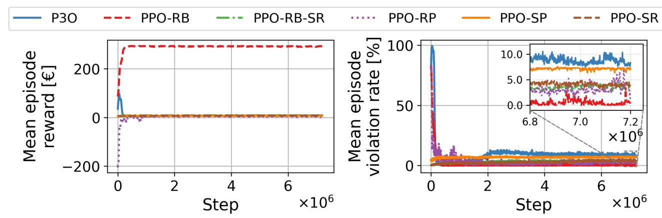


Fig. 5: Mean episode reward (left) and episode violation count (right) of each RL strategy as functions of training step, with an inset providing a zoomed view of late-stage behavior.

The following sections describe the comparative analysis of the performance of each safety strategy over the full test year of operation based on economic and safety metrics.

1) Economic and safety performance analysis: Table IV presents a comparative analysis of the economic and safety performance of each control strategy, revealing a clear trade-off between profitability and operational safety. MPC-D exhibits the theoretical best performance, achieving an EBITDA of -€16,917,608 with a perfect operational record, i.e., zero constraint violations; for this reason, it is omitted from the study of violation types and frequencies in the following sections. Introducing stochasticity with MPC-S results in a marginal decrease in economic performance (-0.79%) and an associated small VR of 4.50% overall. Its VRA, however, is considerably high, leading to a trade-off index of 0.76. The P3O strategy exhibits a significantly higher VR, a very high VRA, and the worst economic performance overall, indicating it is a less robust approach. Reward shaping strategies achieve better safety outcomes than the MPC. Shielding approaches, specifically PPO-SR, further enhance safety: PPO-SR displays the lowest overall VR, equal to 1.26%, and a VRA of 10.33%. The proposed hybrid strategy, PPO-RB-SR, achieves a competitive trade-off with $I = 0.95$: it minimizes the VR and exhibits the lowest VRA among all tested strategies (6.47%), while providing the second highest EBITDA among RL strategies, after PPO-SP.

TABLE IV: Total annual EBITDA, violation rates, and trade-off index for each strategy

Type	Strategy	EBITDA (€)	VR (%)	VRA (%)	I
MPC	*MPC-D*	*-16,917,608*	*0.00*	*0.00*	*1.00*
	MPC-S	**-17,051,322**	4.50	25.16	0.76
Trust region	P3O	-17,229,296	16.96	79.33	0.23
Reward shaping	PPO-RB	-17,217,605	3.47	10.49	0.91
	PPO-RP	-17,216,126	3.18	16.45	0.85
Shield	PPO-SP	-17,104,787	5.20	31.39	0.70
	PPO-SR	-17,151,938	**1.26**	10.33	0.91
Hybrid	PPO-RB-SR	-17,133,910	1.55	**6.47**	**0.95**

Of the seven strategies, two models were chosen for a quarterly analysis: the MPC-S controller for its peak EBITDA, and the PPO-RB-SR agent for its optimal balance of economic return and constraint satisfaction. Table V summarizes quarterly EBITDA and VRA for both models. PPO-RB-SR consistently outperforms the MPC-S benchmark, maintaining a lower VRA across all quarters. The most notable improvement occurs in Q4, where the RL policy achieves an 86.81% relative VRA reduction over the MPC-S. Both controllers achieve nearly identical EBITDA (under 1% difference), proving that PPO-RB-SR's enhanced constraint satisfaction does not sacrifice economic performance. For both, the third quarter yields the lowest EBITDA because of increased summer electricity demand. Overall, the PPO-RB-SR policy is flexible and generalizes well across seasonal scenarios.

TABLE V: Quarterly performance of PPO-RB-SR and MPC-S in terms of EBITDA and VRA. Q1-Q4 is the notation for the four trimesters of the year 2023

Metric	Quarter	PPO-RB-SR	MPC-S
EBITDA (€)	Q1	-3,262,979	-3,245,080
	Q2	-2,615,495	-2,593,966
	Q3	-7,611,531	-7,587,165
	Q4	-3,634,301	-3,615,506
VRA (%)	Q1	2.76	23.49
	Q2	4.20	15.78
	Q3	12.62	30.04
	Q4	5.07	38.43

2) Analysis of constraint violation types and magnitudes: Fig. 6a illustrates the relative frequency of constraint violation types across the implemented safety strategies. The four main violation types are *Charge exceeding PV surplus* (see (10)), *Discharge exceeding net load* (see (12)), *Maximum SOC exceeded* (see right-hand side of (6)), and *Minimum SOC exceeded* (see left-hand side of (6)). The MPC-S and PPO strategies with shielding, including PPO-RB-SR, predominantly display the *Charge exceeding PV surplus* violation type, which occurs with a relative frequency over 80%. This outcome reflects difficulties in anticipating and reacting to uncertain exogenous factors, particularly PV generation, emphasizing the importance of enhancing PV forecasting accuracy. The lack of *Maximum SOC exceeded* or *Minimum SOC exceeded* violations is explained by the design of these strategies, which explicitly restrict actions within SOC limits. In contrast, the reward-shaping agents frequently incur *Minimum SOC exceeded* violations (83.15% with PPO-RP and 47.65% with PPO-RB). This behavior is likely rooted in the temporal credit assignment problem inherent to RL: the agents may fail to associate unsafe actions (aggressive discharging) with delayed negative consequences (SOC depletion). Moreover, their tendency to cause *Charge exceeding PV surplus* violations (43.62% with PPO-RB and 15.38% with PPO-RP) suggests limited robustness against PV forecasting uncertainty. Finally, P3O primarily causes *Discharge exceeding net demand* violations (92.78% relative frequency), suggesting that its penalty-based learning process makes the agent prioritize economic performance at the expense of compliance to safety limits.

Fig. 6b presents a comparison of the violation magnitudes through box-and-whisker plots. The PPO-RB strategy stands out with its tightly clustered distribution and minimal outlier occurrence: this shows that this strategy tends to be very conservative. Conversely, P30 and PPO-RP are characterized by a significantly high median and interquartile range: this suggests that these strategies may be more prone to occasional but substantial constraint violations. Analysis of the control action sequences shows that P3O and PPO-RP prioritize aggressive discharging when high electricity prices coincide with high demand and PV generation. This behavior leads to severe constraint violations. Conversely, most other strategies (e.g., PPO-RB-SR) maintain a lower median and tighter IQR by consistently dampening aggressive actions that approach safety thresholds, regardless of the potential reward. Fig. 7 illustrates an example of this behaviour during a challenging day (July 29, 2023): during intervals of volatile SNP and high demand, the PPO-RP agent (Fig. 7a) executes an aggressive discharging sequence to attempt to minimize costs. This behavior causes breaches in the lower bound of the SOC, leading to the large magnitude violations seen in the previous analysis. In contrast, the PPO-RB-SR agent (Fig. 7b) executes a conservative charging sequence, allowing for an effective discharge event at 21:00, when SNP peaks.

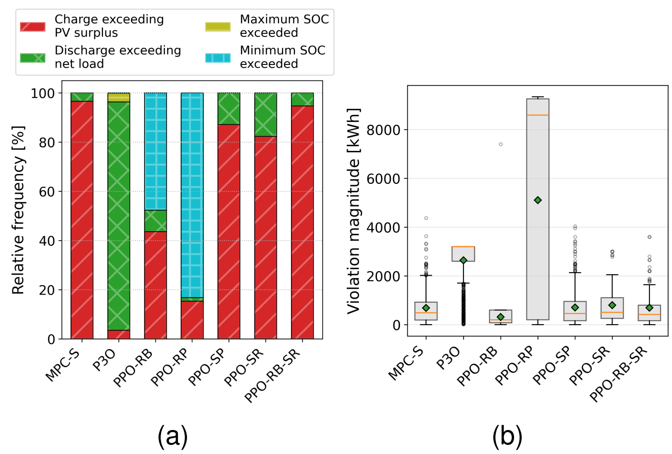
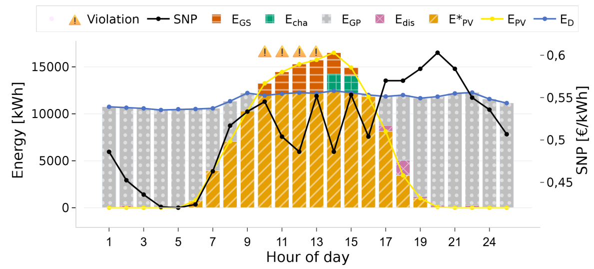
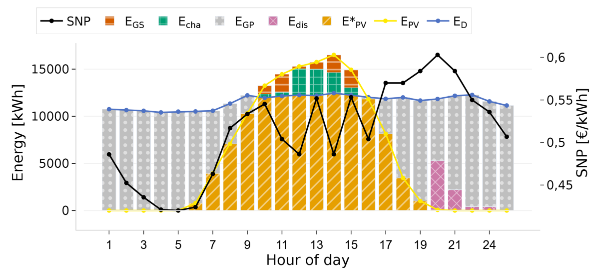


Fig. 6: (a) Relative frequency of constraint violation types for each strategy. (b) Box-and-whisker plot of absolute violation magnitudes across strategies. Boxes represent the inter-quartile range (IQR, 25th-75th percentiles); horizontal bars denote medians, diamonds denote means, whiskers extend to $1.5{\cdot}$IQR, and circles highlight outliers.

3) Analysis of the impact of LSTM-ED accuracy: To investigate the impact of LSTM-ED accuracy on PPO agent training, we conducted an experiment where the proposed agent was trained using real measurements of PV generation and electric demand in place of the LSTM-ED forecasts. Fig. 8 illustrates the learning curves for both the original PPO-RB-SR agent and this newly trained agent. While both agents achieve satisfactory performance, the agent trained with real data consistently exhibits a slightly higher mean episode reward and lower constraint violation rate. This highlights the sensitivity of PPO training to the quality of the input forecasts



(a)



(b)

Fig. 7: Daily energy profiles and SNP for a challenging day (July 29, 2023): (a) PPO-RP, (b) PPO-RB-SR. E^*_{PV} represents self-consumed PV generation. The warning symbol indicates a *Minimum SOC exceeded* violation.

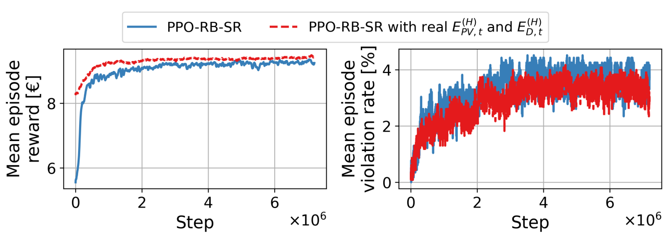


Fig. 8: Mean episode reward (left) and episode violation count (right) as functions of training step for PPO-RB-SR (blue) and PPO-RB-SR using real $E^{(H)}_{PV,t}$ and $E^{(H)}_{D,t}$ in place of LSTM-ED predictions (red).

and underscores the importance of accurate forecasting for optimal BSS scheduling.

Considering all agents at test time, Fig. 9 shows a heat-map of the per-hour violation frequency for each strategy. It is possible to observe that 76.02% of the total violations across all models (2,235 out of 2,940) occur between 10:00 and 15:00, aligning with peak PV output and, thus, suggesting a temporal dependence driven by forecasting errors. This temporal clustering is most pronounced in the P3O strategy, which accounts for 43.36% of the recorded violations in this time. PPO-RB-SR accounts for only 5.64% in this interval and 4.52% of the total violations interval: even if the shield fails due to high prediction errors, reward shaping encourages a proactive, conservative behavior, making the agent less likely to select actions that trigger violations. The temporal dependence is further supported by Table VI, which reports the Pearson correlation coefficients between VRs and the hourly MAE of forecasts: RL-based strategies tend to be sensitive to PV forecasting errors: P3O has the highest correlation, which reflects their midday violation peaks in Fig. 9. Correlations with demand forecasting errors are lower overall. Thus, violation risk often stems from the "perception and prediction layer", but the "control and decision layer" mitigates it. As shown in Fig. 7, the conservative policy learned via the dual safety mechanism maintains a sufficient buffer to prevent violations. These results further highlight the critical role of accurate PV generation forecasting for safe battery control.

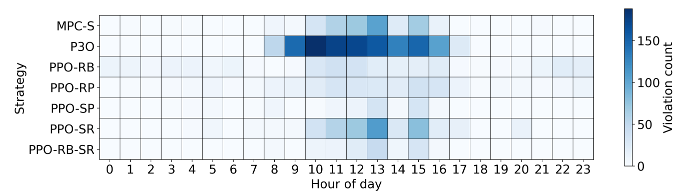


Fig. 9: Heat-map of the per-hour violation frequency.

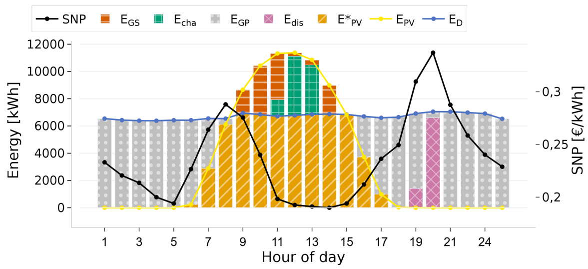
TABLE VI: Pearson correlation coefficients between the VR of each strategy and the hourly MAE of PV and demand predictions

Correlation	Hourly MAE of PV forecasts	Hourly MAE of demand forecasts
Hourly VR of MPC-S	0.68	0.31
Hourly VR of P3O	0.92	0.52
Hourly VR of PPO-RB	0.45	0.24
Hourly VR of PPO-RP	0.83	0.42
Hourly VR of PPO-SP	0.52	0.26
Hourly VR of PPO-SR	0.68	0.27
Hourly VR of PPO-RB-SR	0.59	0.24

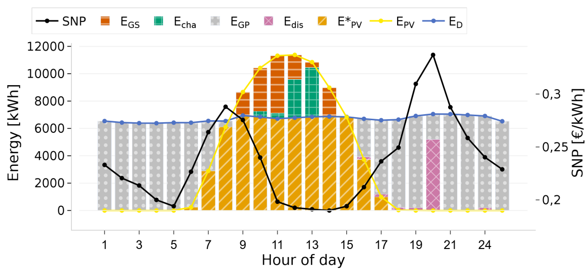
4) Daily energy profiles: Fig. 10 illustrates daily energy profiles for a representative day (March 21, 2023). Only the two best models, i.e. PPO-RB-SR and MPC-S, are considered. BSS charging and discharging patterns differ across strategies. Under MPC-S (a), charging occurs predominantly in midday hours when PV generation peaks and SNP is low, while discharging is concentrated during the evening price peak. PPO-RB-SR (b) exhibits a midday charging trend similar to MPC-S but with a larger safety margin, combined with moderate discharging during the evening. This indicates a more distributed approach, which allows the agent to minimize violations especially during charging, as previously discussed.

VII. Conclusion and Future Work

In prior safe RL studies for BSS scheduling, two research gaps emerge: the lack of a systematic comparison of performance and safety compliance with traditional control methods, and the lack of explicit and efficient handling of uncertainty and temporal dependencies in partially observable environments. The proposed approach addresses both by introducing a unified framework for BSS scheduling. Central to the proposed approach is a custom RL feature extractor designed to address the partially-observable nature of the problem and capture heterogeneous temporal dependencies. A PPO-RB-SR agent is built based on the proposed architecture and trained with a dual safety mechanism, comprising reward shaping to steer the



(a)



(b)

Fig. 10: Daily energy profiles and SNP for a representative day (March 21, 2023): (a) MPC-S, (b) PPO-RB-SR. E^*_{PV} represents self-consumed PV generation.

policy toward feasible actions, and shielding to block unsafe actions. Its performance is benchmarked against multiple state-of the art safe RL and MPC strategies.

The results show that PPO-RB-SR achieves the best overall balance between economic performance and safety metrics, with an EBITDA of –€17,133,910 and a VRA of 6.47%: it provides a 74.28% decrease in VRA compared to the MPC-S baseline, while decreasing the EBITDA by less than 1%. The quarterly performance analysis highlights that PPO-RB-SR consistently outperforms the MPC-S benchmark, maintaining a lower VRA and achieving an EBITDA that is less than 1% lower across all quarters. All the reward shaping and shielding RL strategies substantially improve safety compliance relative to MPC-S thanks to the inherent flexibility provided by the exploration and exploitation trade-off. However, the prevalence of discharging violations for RL agents highlights challenges in temporal credit assignment, where short-term rewards can lead to delayed safety breaches. Violation timing patterns concentrate around midday PV peaks, aligning with periods of high PV forecast error. Even though PPO-RB-SR accounts for only 5.64% violations in this time interval (and 4.52% of the total violations), this finding reinforces the necessity of accurate PV forecasting for safe BSS operation and suggests that further integration of probabilistic forecasts could enhance safety.

Overall, the comparative analysis shows that the hybrid PPO-RB-SR strategy can improve safety relative to state-of-the-art methods under uncertainty, effectively balancing profitability and operational safety. The novelty of this work lies in jointly addressing partial observability, heterogeneous temporal dependencies, and multi-metric safety evaluation in a unified framework for BSS scheduling.

Future work will enhance the proposed method in multiple directions. First, to address potential scalability limitations posed by a discrete action space when dealing with high-dimensional, multi-asset scenarios or finer temporal scales, the policy architecture will be modified to handle a continuous action space. Second, to enhance decision-making under uncertainty, the deterministic point forecasts from the pre-trained LSTM-ED model will be replaced with probabilistic predictions. Furthermore, the long-term reliability and robustness of the approach will be validated by expanding the current dataset to include multi-year data and extreme weather events. Finally, battery degradation metrics will be integrated to ensure that control actions remain compliant with hardware longevity and safety requirements.

Acknowledgments

Special thanks to Pellerey Valeria, Porro Gioele, Chiurato Simona, and Morciano Davide Bruno of TRIGENIA Srl. for their support of this work.

References

[1] D. M. López González and J. Garcia Rendon, "Opportunities and challenges of mainstreaming distributed energy resources towards the transition to more efficient and resilient energy markets," *Renewable and Sustainable Energy Reviews*, vol. 157, p. 112018, 2022.

[2] Y. Yang, S. Xia, P. Huang, and J. Qian, "Energy transition: Connotations, mechanisms and effects," *Energy Strategy Reviews*, vol. 52, p. 101320, 2024.

[3] S. Nyamathulla and C. Dhanamjayulu, "A review of battery energy storage systems and advanced battery management system for different applications: Challenges and recommendations," *Journal of Energy Storage*, vol. 86, p. 111179, 2024.

[4] C. Zhao, P. B. Andersen, C. Træholt, and S. Hashemi, "Grid-connected battery energy storage system: a review on application and integration," *Renewable and Sustainable Energy Reviews*, vol. 182, p. 113400, 2023.

[5] G. Ghione, V. Randazzo, E. Pasero, and M. Badami, *Comparison of LSTM-Based Models for the 6-Hour-Ahead and Day-Ahead Forecasting of Photovoltaic Power Generation*. Singapore: Springer Nature Singapore, 2025, pp. 45–54.

[6] G. Ghione, M. A. Judge, M. Badami, E. Pasero, V. Franzitta, and G. Cirrincione, "A comparison of univariate methods for day-ahead short-term load forecasting," in *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 2024, pp. 662–667.

[7] M. M. Rana, M. Uddin, M. R. Sarkar, G. Shafiullah, H. Mo, and M. Atef, "A review on hybrid photovoltaic – battery energy storage system: Current status, challenges, and future directions," *Journal of Energy Storage*, vol. 51, p. 104597, 2022.

[8] C. Cienfuegos, P. M. Rodrigo, I. Cienfuegos, and A. Diaz-Ponce, "Comparative analysis of battery energy storage systems' operation strategies for peak shaving in industries with or without installed photovoltaic capacity," *Renewable Energy Focus*, vol. 49, p. 100574, 2024.

[9] J. Salpakari and P. Lund, "Optimal and rule-based control strategies for energy flexibility in buildings with pv," *Applied Energy*, vol. 161, pp. 425–436, 2016.

[10] F. Y. Melhem, O. Grunder, Z. Hammoudan, and N. Moubayed, "Energy management in electrical smart grid environment using robust optimization algorithm," *IEEE Transactions on Industry Applications*, vol. 54, no. 3, pp. 2714–2726, 2018.

[11] D. Cho and J. Valenzuela, "Optimization of residential off-grid pv-battery systems," *Solar Energy*, vol. 208, pp. 766–777, 2020.

[12] A. C. Real, G. P. Luz, J. Sousa, M. Brito, and S. Vieira, "Optimization of a photovoltaic-battery system using deep reinforcement learning and load forecasting," *Energy and AI*, vol. 16, p. 100347, 2024.

[13] S. Zhang and Y. Tang, "Optimal schedule of grid-connected residential pv generation systems with battery storages under time-of-use and step tariffs," *Journal of Energy Storage*, vol. 23, pp. 175–182, 2019.

[14] Z. Hu, Y. Gao, L. Sun, M. Mae, and T. Imaizumi, "Improved robust model predictive control for residential building air conditioning and photovoltaic power generation with battery energy storage system under weather forecast uncertainty," *Applied Energy*, vol. 371, p. 123652, 2024.

[15] S. Teleke, M. E. Baran, S. Bhattacharya, and A. Q. Huang, "Optimal control of battery energy storage for wind farm dispatching," *IEEE Transactions on Energy Conversion*, vol. 25, no. 3, pp. 787–794, 2010.

[16] C. J. Jena, P. K. Ray, G. Panda, and S. Nandkeolyar, "Power quality enhancement and power management of pv-hess based grid-tied microgrid using model predictive control approach," *IEEE Transactions on Industry Applications*, vol. 61, no. 2, pp. 3424–3434, 2025.

[17] H. Song *et al.*, "Smart optimization in battery energy storage systems: An overview," *Energy and AI*, vol. 17, p. 100378, 2024.

[18] Y. Yang, S. Bremner, C. Menictas, and M. Kay, "Modelling and optimal energy management for battery energy storage systems in renewable energy systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 167, p. 112671, 2022.

[19] R. Subramanya, S. A. Sierla, and V. Vyatkin, "Exploiting battery storages with reinforcement learning: A review for energy professionals," *IEEE Access*, vol. 10, pp. 54 484–54 506, 2022.

[20] G. Ghione, V. Randazzo, E. Pasero, and M. Badami, "Optimal cogeneration scheduling: A comparison of genetic and pomdp-based deep reinforcement learning approaches," *IEEE Access*, vol. 13, pp. 128 562–128 581, 2025.

[21] Q. Meng, S. Hussain, F. Luo, Z. Wang, and X. Jin, "An online reinforcement learning-based energy management strategy for microgrids with centralized control," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 1501–1510, 2025.

[22] H. Hou *et al.*, "Interval multi-objective optimization for low-carbon building energy management system upon deep reinforcement learning," *IEEE Transactions on Industry Applications*, vol. 61, no. 2, pp. 2193–2202, 2025.

[23] F. Sanchez Gorostiza and F. M. Gonzalez-Longatt, "Deep reinforcement learning-based controller for soc management of multi-electrical energy storage system," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5039–5050, 2020.

[24] D. Domínguez-Barbero, J. García-González, M. A. Sanz-Bobi, and A. García-Cerrada, "Energy management of a microgrid considering nonlinear losses in batteries through deep reinforcement learning," *Applied Energy*, vol. 368, p. 123435, 2024.

[25] E. Hosseini *et al.*, "Reinforcement learning-based energy management system for lithium-ion battery storage in multilevel microgrid," *Journal of Energy Storage*, vol. 109, p. 115114, 2025.

[26] C. Guo, X. Wang, Y. Zheng, and F. Zhang, "Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning," *Energy*, vol. 238, p. 121873, 2022.

[27] M. Alonso, H. Amaris, D. Martin, and A. de la Escalera, "Proximal policy optimization for energy management of electric vehicles and pv storage units," *Energies*, vol. 16, no. 15, 2023.

[28] X. Fang, P. Hong, S. He, Y. Zhang, and D. Tan, "Multi-layer energy management and strategy learning for microgrids: A proximal policy optimization approach," *Energies*, vol. 17, no. 16, 2024.

[29] K. Zhou, K. Zhou, and S. Yang, "Reinforcement learning-based scheduling strategy for energy storage in microgrid," *Journal of Energy Storage*, vol. 51, p. 104379, 2022.

[30] L. Brunke *et al.*, "Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. Volume 5, 2022, pp. 411–444, May 2022.

[31] V.-H. Bui, S. Das, A. Hussain, G. V. Hollweg, and W. Su, "A Critical Review of Safe Reinforcement Learning Techniques in Smart Grid Applications," Sep. 2024, arXiv:2409.16256 [eess].

[32] S. Gu *et al.*, "A Review of Safe Reinforcement Learning: Methods, Theories, and Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 216–11 235, Dec. 2024.

[33] T. Su, T. Wu, J. Zhao, A. Scaglione, and L. Xie, "A review of safe reinforcement learning methods for modern power systems," *Proceedings of the IEEE*, vol. 113, no. 3, pp. 213–255, 2025.

[34] S. Gu, A. Kshirsagar, Y. Du, G. Chen, J. Peters, and A. Knoll, "A human-centered safe robot reinforcement learning framework with interactive behaviors," *Frontiers in Neurorobotics*, vol. Volume 17 - 2023, 2023.

[35] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[36] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 2656–2667, 2023.

[37] A. Ajagekar and F. You, "Deep reinforcement learning based unit commitment scheduling under load and wind power uncertainty," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 2, pp. 803–812, 2023.

[38] G. Dalal, K. Dvijotham, M. Vecerík, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *CoRR*, vol. abs/1801.08757, 2018.

[39] Z. Yi, X. Wang, C. Yang, C. Yang, M. Niu, and W. Yin, "Real-time sequential security-constrained optimal power flow: A hybrid knowledge-data-driven reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1664–1680, 2024.

[40] X. Wang, P. Wang, R. Huang, X. Zhu, J. Arroyo, and N. Li, "Safe deep reinforcement learning for building energy management," *Applied Energy*, vol. 377, p. 124328, 2025.

[41] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Applied Energy*, vol. 309, p. 118403, 2022.

[42] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 15 338–15 349.

[43] L. Zhang *et al.*, "Penalized Proximal Policy Optimization for Safe Reinforcement Learning," Jun. 2022, arXiv:2205.11814 [cs].

[44] Y. Ye, H. Wang, P. Chen, Y. Tang, and G. Strbac, "Safe deep reinforcement learning for microgrid energy management in distribution networks with leveraged spatial–temporal perception," *IEEE Transactions on Smart Grid*, vol. 14, no. 5, pp. 3759–3775, 2023.

[45] M. Tarle, M. Larsson, G. Ingeström, L. Nordström, and M. Björkman, "Safe reinforcement learning to improve facts setpoint control in presence of model errors," *IEEE Transactions on Industry Applications*, pp. 1–10, 2025.

[46] G. Hao, Y. Li, Y. Li, K. Guang, and Z. Zeng, "Safe reinforcement learning for active distribution networks reconfiguration considering uncertainty," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 1757–1769, 2025.

[47] G. Ghione, V. Randazzo, M. Badami, and E. Pasero, "Optimizing battery storage systems in energy microgrids: A reinforcement learning approach comparing multiple reward functions," in *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, 2024, pp. 642–646.

[48] J. Bouwens, T. De Kok, and A. Verriest, "The prevalence and validity of EBITDA as a performance measure," *Comptabilité-Contrôle-Audit*, vol. 25, no. 1, pp. 55–105, 2019.

[49] K. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174–205, 1965.

[50] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[51] E. Ekinci, "A comparative study of lstm-ed architectures in forecasting day-ahead solar photovoltaic energy using weather data," *Computing*, vol. 106, no. 5, pp. 1611–1632, 2024.

[52] G. La Tona, M. Luna, and M. Di Piazza, "Day-ahead forecasting of residential electric power consumption for energy management using long short-term memory encoder–decoder model," *Mathematics and Computers in Simulation*, vol. 224, pp. 63–75, 2024.

[53] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[54] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," 2019.

[55] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, "Stable baselines3," https://github.com/DLR-RM/stable-baselines3, 2019.

[56] Y. Tang and S. Agrawal, "Discretizing continuous action space for on-policy optimization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5981–5988, Apr. 2020.

[57] J. Jeong, T.-Y. Ku, and W.-K. Park, "Exploring the preference for discrete over continuous reinforcement learning in energy storage arbitrage," *Energies*, vol. 17, no. 23, 2024.

[58] M. Schwenzer, M. Ay, T. Bergs, and D. Abel, "Review on model predictive control: An engineering perspective," *The International Journal of Advanced Manufacturing Technology*, vol. 117, no. 5, pp. 1327–1349, 2021.

[59] F. Fiedler *et al.*, "do-mpc: Towards fair nonlinear and robust model predictive control," *Control Engineering Practice*, vol. 140, p. 105676, 2023.

[60] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, pp. 25–57, 2006.