International Journal of Power Electronics and Drive System (IJPEDS)
Vol. 17, No. 4, December 2026, pp. 2912~2923
ISSN: 2088-8694, DOI: 10.11591/ijpeds.v17.i4.pp2912-2923

Baseline stability-oriented proximal policy optimization for energy storage in microgrids

Citra Dewi[1], Ricky Maulana[1], Randi Purnama Putra[2]
[1]Department of Electrical Engineering, Universitas Negeri Padang, Padang, Indonesia
[2]Department of Mechanical Engineering, Universitas Negeri Padang, Padang, Indonesia

Article Info

Article history:

Received Dec 1, 2025
Revised Sep 1, 2026
Accepted Sep 13, 2026

Keywords:

Energy storage system
Proximal policy optimization
Reinforcement learning
Rule-based benchmark
Stability-oriented dispatch
State-of-charge regulation

ABSTRACT

The increasing penetration of renewable energy in microgrids introduces power variability and supply-demand imbalance, making energy storage system (ESS) control important for stability-oriented operation. This study develops a reproducible baseline stability-oriented ESS control framework using proximal policy optimization (PPO) in a simplified PV, load, and ESS microgrid environment. The control problem is formulated as a continuous-action Markov decision process, where the agent observes load demand, PV generation, state of charge, net imbalance, and previous ESS command to generate bounded charge/discharge actions. The reward function penalizes residual power imbalance, SOC deviation, and abrupt command variation. The framework is implemented using MATLAB/Simulink and a function-based reinforcement learning environment, and a rule-based controller is included as a benchmark. Training results show that PPO improves episode reward from approximately -72.63 to an average reward of about -2.51 after 300 episodes. During evaluation, the simplified voltage-response indicator remained within 0.9985–1.0016 p.u, keeps SOC within 49.4830–50.0000%, and generates bounded ESS commands. Compared with the rule-based benchmark, PPO demonstrates feasible and bounded operation, although the rule-based controller achieves lower average residual imbalance in the simplified deterministic profile. These results establish a transparent baseline for future stability-oriented DRL-based ESS control studies.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Citra Dewi
Department of Electrical Engineering, Universitas Negeri Padang
Padang, Indonesia
Email: citradewi@ft.unp.ac.id

1. INTRODUCTION

Microgrids are increasingly deployed to integrate renewable energy resources (RES), such as photovoltaic (PV) and wind generation, into local power systems. However, the stochastic and non-dispatchable characteristics of RES introduce uncertainty in generation scheduling and may create supply-demand imbalance under variable load conditions [1], [2]. In microgrids with high renewable penetration, these fluctuations may affect voltage regulation and overall operational stability if flexibility resources are not properly coordinated [3]. Energy storage systems (ESS) are therefore widely used to buffer renewable variability, improve operational flexibility, and support stability-oriented microgrid operation. Previous studies have shown that ESS and converter-based control strategies can enhance microgrid operation under PV generation and hybrid AC/DC configurations [4], [5]. In practical applications, ESS dispatch is commonly handled using rule-based logic, proportional-integral-derivative (PID) control, and model predictive control (MPC). Rule-based and hybrid intelligent strategies are attractive because of their simple implementation, but they often require careful tuning and may not adapt well to changing PV-load patterns [6], [7]. MPC-based

strategies can explicitly handle prediction and optimization objectives, but their performance depends on model accuracy, forecast quality, and computational complexity [8]-[10].

Deep reinforcement learning (DRL) has emerged as a model-free control alternative that learns control policies from interaction data. It has been increasingly explored for microgrid energy management, demand flexibility, economic dispatch, and PV-battery operation under uncertain operating conditions [11]-[13]. Comparative studies have also shown that reinforcement learning can provide an alternative to model-based control for PV-battery systems, although its performance depends strongly on the training environment and reward formulation [14], [15]. In addition, DRL has been applied to distribution-network resilience and microgrid formation, indicating its potential beyond purely economic scheduling problems [16]. For ESS and microgrid control, recent DRL studies have investigated learning-based strategies for hybrid energy storage integration, autonomous energy management, renewable stochastic operation, and frequency-constrained energy balancing [17]-[20]. These studies indicate that learning-based controllers can support flexible microgrid operation, but they also show that controller performance is sensitive to the design of the reward function, state representation, and evaluation scenario.

Despite these developments, several open issues remain in learning-based ESS control for microgrids. Recent reviews emphasize that reproducibility, robustness, and deployment-oriented evaluation are still important challenges in reinforcement learning-based microgrid control [21]. More recent scheduling and multi-agent studies have also reported promising results for ESS coordination in simulated microgrid environments [22], [23]. However, broader multi-objective reinforcement learning studies suggest that transparent benchmarking and application-specific evaluation remain necessary for interpreting learning-based control performance [24]. In particular, many reported proximal policy optimization (PPO)-based or DRL-based microgrid studies primarily emphasize economic objectives, such as cost minimization, peak shaving, market-oriented scheduling, or general energy management. Stability-oriented ESS control for real-time power balancing, voltage-response assessment, safe state of charge (SOC) regulation, and smooth charge/discharge behavior is still less explicitly discussed. Moreover, some DRL studies are implemented using highly abstracted environments, which can limit reproducibility and make it difficult to connect the learned policy with continuous-time simulation workflows commonly used in power electronics and microgrid studies.

To address these gaps, this study develops a baseline stability-oriented PPO framework for ESS charge/discharge control in a simplified PV-load-ESS microgrid. This work does not propose a new PPO algorithm; instead, it provides a transparent and reproducible baseline for evaluating stability-focused ESS dispatch. The PPO agent observes PV generation, load demand, SOC, and instantaneous power imbalance to generate bounded charge/discharge commands. The reward function is designed to reduce residual power imbalance, maintain SOC within a safe operating region, and discourage abrupt control changes. The main contributions are: i) Formulation of a PPO-based DRL controller for stability-oriented ESS management; ii) Development of a reward structure that considers power-balance regulation, SOC safety, and smooth dispatch; iii) Presentation of a reproducible MATLAB/Simulink simulation framework; iv) Quantitative assessment using representative PV-load operating conditions; and v) Explicit discussion of the baseline model limitations, including simplified converter representation, absence of hardware validation, and the need for broader benchmarking and robustness analysis in future work.

2. METHOD

This study develops a baseline stability-oriented control framework for ESS charge/discharge management in a simplified microgrid using PPO. The methodology consists of six main stages: i) development of a MATLAB/Simulink signal-level microgrid testbed, ii) modeling of ESS SOC dynamics, iii) formulation of the reinforcement learning environment, iv) design of the stability-oriented reward function, v) PPO agent training and evaluation, and vi) comparison with a rule-based benchmark controller. The general workflow of the proposed baseline PPO-based ESS controller is illustrated in Figure 1. The process begins by initializing the simplified PV-load-ESS microgrid model, defining the reinforcement learning environment, and setting the PPO controller parameters. During training, the agent observes the current operating state, computes an ESS charge/discharge action, applies the action to update the SOC response, and receives a reward based on residual power imbalance, SOC deviation, and action variation. The PPO policy parameters are updated iteratively until the training performance becomes relatively stable. The trained policy is then evaluated and compared with a rule-based benchmark controller.

The proposed framework is implemented using MATLAB/Simulink and the reinforcement learning toolbox. The microgrid testbed consists of a grid-voltage input, simplified load-dynamics block, PV generation profile, load demand profile, ESS SOC block, delay block, controller block, and output logging blocks for voltage response, ESS command, SOC, and reward. The Simulink implementation of the testbed is shown in Figure 2. The model is intentionally designed as a reproducible baseline environment to evaluate the feasibility of learning-based ESS dispatch under PV-load variations. Therefore, detailed distribution-

network power-flow constraints, switching-level converter dynamics, semiconductor switching losses, thermal behavior, battery aging, and hardware-in-the-loop implementation are not included in this baseline version.

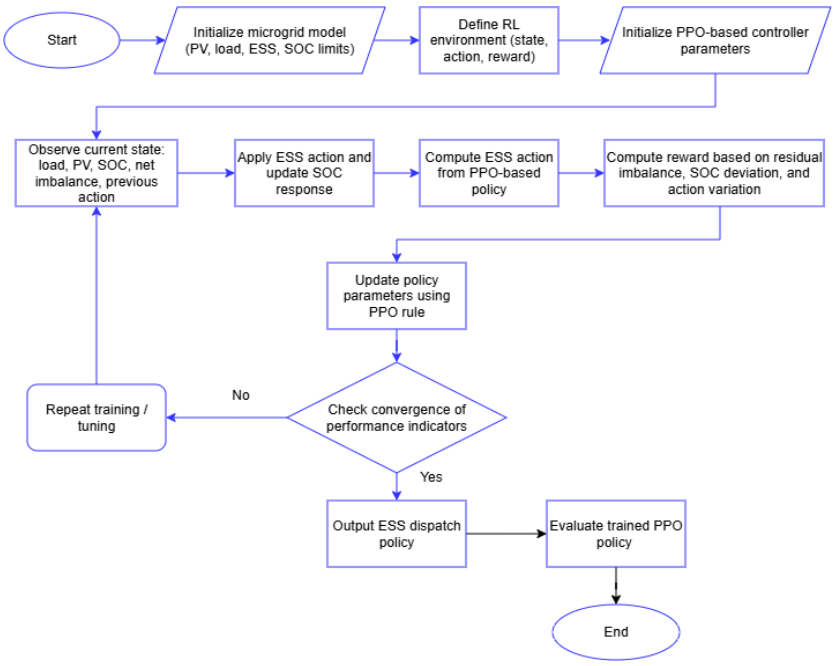

Figure 1. Workflow of the proposed baseline PPO-based ESS control framework in the simplified PV-load-ESS microgrid model

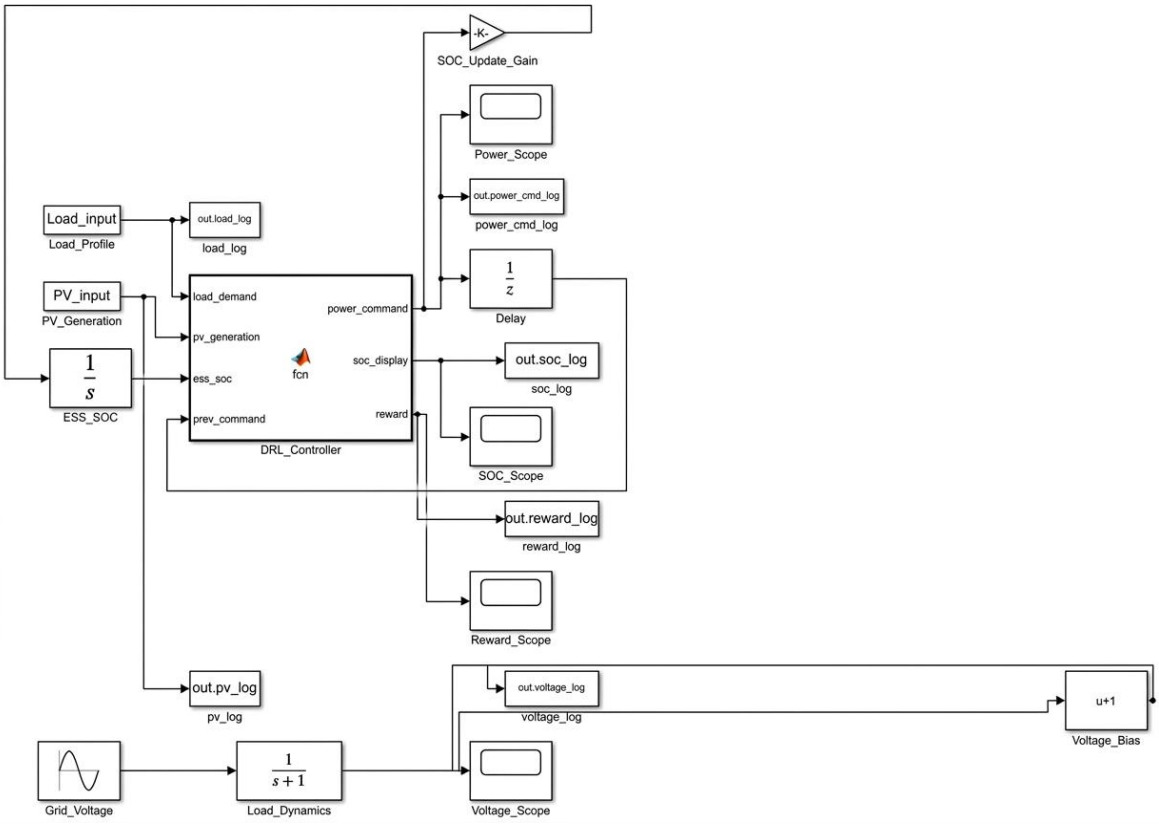

Figure 2. Simulink implementation of the simplified PV-load-ESS microgrid testbed with ESS controller, SOC dynamics, and output logging blocks

2.1. Simulink-based microgrid testbed

The simplified microgrid model consists of PV generation, load demand, voltage dynamic response, and battery-based ESS. The PV generation and load demand profiles are defined as normalized time-series input signals and supplied to the simulation environment. The normalized PV and load profiles used in the evaluation window are summarized in Table 1.

The load-side dynamic response is represented using a first-order transfer function, as in (1).

$$G(s) = \frac{1}{(s+1)} \tag{1}$$

The output of this block is shifted around the nominal operating voltage of 1 p.u. and used as a voltage-response indicator in the simplified testbed. Because this work adopts a signal-level microgrid model, the voltage response should be interpreted as a dynamic response indicator rather than a detailed distribution network bus voltage obtained from power-flow calculation. The simulation uses a fixed-step ODE4 solver with a sample time of 0.1 s. The evaluation window is 10 s, resulting in 101 simulation steps per episode. The main simulation settings are listed in Table 2.

Table 1. Normalized PV generation and load demand profiles used in the simulation

Time (s)	PV generation (p.u.)	Load demand (p.u.)
0	0.20	0.60
1	0.35	0.70
2	0.70	0.80
4	0.90	0.75
6	0.60	0.90
8	0.30	0.65
10	0.20	0.70

Table 2. Simulation configuration

Parameter	Value	Parameter	Value
Simulation platform	MATLAB/Simulink	Initial SOC	50%
Toolbox	Reinforcement Learning Toolbox	SOC lower limit	20%
Solver	Fixed-step ODE4	SOC upper limit	90%
Sample time	0.1 s	Action range	[-1,1] p.u.
Evaluation window	10 s	Positive action	ESS charging
Number of evaluation steps	101	Negative action	ESS discharging

2.2. ESS state-of-charge model

The ESS is represented using a simplified SOC dynamic model. The SOC is updated based on the charging or discharging command generated by the controller. In general, the ESS energy balance can be expressed as (2).

$$SOC_{t+1} = SOC_t + \frac{\eta_c P_{ch,t} - P_{dis,t}/\eta_d}{E_{rated}} \Delta t \tag{2}$$

Where SOC_t is the state of charge at time step t, $P_{ch,t}$ and $P_{dis,t}$ are the charging and discharging power, η_c and η_d are the charging and discharging efficiencies, E_{rated} is the rated ESS capacity, and Δt is the simulation time step. In the implemented baseline testbed, the normalized ESS command is used to update SOC as (3).

$$SOC_{t+1} = SOC_t + k_{soc} a_t \tag{3}$$

Where a_t is the normalized ESS action and $k_{soc} = 0.025$ is the SOC update gain used in the simulation environment. The SOC is constrained between 20% and 90% to prevent overcharging and deep discharging. The initial SOC is set to 50%, allowing both charging and discharging actions during the evaluation period.

2.3. Reinforcement learning environment formulation

The ESS control problem is formulated as a continuous-action Markov decision process (MDP), following the standard reinforcement learning representation in which an agent interacts with an environment through states, actions, and rewards [25]. At each time step, the agent observes the operating state of the microgrid and generates a normalized ESS charge/discharge command. The observation vector is defined as (4).

$$s_t = [P_{load}(t), P_{PV}(t), SOC_{norm}(t), P_{net}(t), a_{t-1}] \tag{4}$$

Where $P_{load}(t)$ is the normalized load demand, $P_{PV}(t)$ is the normalized PV generation, $SOC_{norm}(t)$ is the normalized SOC, $P_{net}(t)$ is the net power imbalance before ESS action, and a_{t-1} is the previous ESS command. The net imbalance before ESS action is calculated as (5).

$$P_{net}(t) = P_{load}(t) - P_{PV}(t) \tag{5}$$

The action space is defined as a continuous scalar, as in (6).

$$a_t \in [-1,1] \tag{6}$$

Where $a_t > 0$ represents charging, $a_t < 0$ represents discharging, and $a_t = 0$ represents idle operation. The residual power imbalance after ESS action is calculated using the adopted sign convention:

$$P_{res}(t) = P_{load}(t) - P_{PV}(t) + a_t \tag{7}$$

Because positive a_t represents charging, it increases the net demand. Conversely, negative a_t represents discharging and therefore reduces the net demand. This sign convention is used consistently in the reward function and performance evaluation.

2.4. Stability-oriented reward function

The reward function is designed to encourage three control objectives: reducing residual power imbalance, maintaining SOC near a safe operating point, and avoiding abrupt ESS command changes. The reward at each time step is formulated as (8).

$$R_t = -\alpha \mid P_{res}(t) \mid -\beta \frac{|SOC(t) - SOC_{ref}|}{50} - \gamma \mid a_t - a_{t-1} \mid \tag{8}$$

Where $P_{res}(t)$ is the residual power imbalance after ESS action, SOC_{ref} is the reference SOC, and $\mid a_t - a_{t-1} \mid$ is the absolute action variation. The reward weights α, β, and γ determine the relative importance of power-balance regulation, SOC regulation, and control smoothness.

The final reward-weight configuration was selected from preliminary tuning to obtain bounded ESS commands, safe SOC behavior, and stable learning. A more aggressive setting with a larger power-imbalance penalty was also tested during development, but it produced more aggressive ESS commands and larger action variation. Therefore, the final configuration adopts a more balanced reward setting, as shown in Table 3.

Table 3. Reward function configuration

Reward component	Symbol	Value	Purpose
Residual power imbalance penalty	α	1.00	Reduce PV-load mismatch after ESS action
SOC deviation penalty	β	0.05	Maintain SOC near reference level
Action variation penalty	γ	0.10	Avoid abrupt ESS command changes
SOC reference	SOC_{ref}	50%	Preferred operating point

2.5. PPO agent design and training configuration

PPO is selected because it is suitable for continuous-action control problems and can limit excessive policy updates through a clipped surrogate objective [26]. This characteristic is relevant for ESS control because the agent must learn a charge/discharge policy without producing highly unstable control commands. The PPO agent uses an actor-critic structure. The actor network generates the ESS command, while the critic network estimates the state value. Both networks use two fully connected hidden layers with 128 neurons per layer and rectified linear unit (ReLU) activation functions. The actor output uses a continuous Gaussian policy, with a mean output constrained by a hyperbolic tangent activation to remain within the normalized action range.

The PPO agent is trained offline using a MATLAB function environment constructed to reproduce the same PV-load profile, SOC dynamics, action convention, and reward formulation used in the Simulink testbed. This approach allows repeatable PPO training while preserving consistency with the simplified signal-level simulation model. The training process records the episode reward and moving average reward to evaluate learning convergence. The main PPO hyperparameters are listed in Table 4.

Table 4. PPO training hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Learning rate	3×10^{-4}	Actor hidden layers	2 fully connected layers
Clipping parameter	0.2	Critic hidden layers	2 fully connected layers
Discount factor	0.99	Neurons per hidden layer	128
Experience horizon	2048	Activation function	ReLU
Mini-batch size	256	Maximum episodes	300
Update epochs	10	Maximum steps per episode	101
Entropy coefficient	0.01		

2.6. Rule-based benchmark controller

A rule-based ESS controller is implemented as a reference benchmark. The purpose of this benchmark is to provide a simple deterministic comparison for evaluating whether the PPO controller produces feasible and physically meaningful ESS actions. The rule-based controller uses the sign of the net power imbalance and SOC limits to determine the ESS command. If load demand exceeds PV generation and SOC is above the lower limit, the ESS discharges. If PV generation exceeds load demand and SOC is below the upper limit, the ESS charges. Otherwise, the ESS remains idle. The controller also includes a command-rate limitation to avoid abrupt changes. The rule-based command is expressed as (9).

$$a_t^{RB} = \begin{cases} -\min\left(|\, P_{net}(t)\,|, 1\right), & P_{net}(t) > 0,\ SOC > SOC_{min} \\ \min\left(|\, P_{net}(t)\,|, 1\right), & P_{net}(t) < 0,\ SOC < SOC_{max} \\ 0, & otherwise \end{cases} \quad (9)$$

A command-rate limiter is applied as (10).

$$|\, a_t - a_{t-1}\,| \leq \Delta a_{max} \quad (10)$$

Where $\Delta a_{max} = 0.15$. The same SOC limits, PV-load profile, sign convention, and evaluation window are used for both the PPO controller and the rule-based benchmark.

2.7. Evaluation metrics

The trained PPO controller and rule-based benchmark are evaluated using the same PV-load profile and simulation settings. The performance assessment focuses on five indicators: voltage response, ESS command, SOC trajectory, reward behavior, and residual power imbalance. The SOC range is calculated as in (11).

$$SOC_{range} = SOC_{max} - SOC_{min} \quad (11)$$

The mean absolute residual power imbalance is calculated as (12).

$$MAE_{res} = \frac{1}{N}\sum_{t=1}^{N} |\, P_{res}(t)\,| \quad (12)$$

The command smoothness index is calculated as (13).

$$J_{smooth} = \frac{1}{N-1}\sum_{t=2}^{N} |\, a_t - a_{t-1}\,| \quad (13)$$

Reward statistics are also reported using minimum, maximum, mean, standard deviation, and mean absolute step change. For comparison purposes, each signal is summarized using (14).

$$x_{min}, x_{max}, \bar{x}, \sigma_x, \text{ and } \frac{1}{N-1}\sum_{t=2}^{N} |\, x_t - x_{t-1}\,| \quad (14)$$

These indicators are used to determine whether the learned controller produces bounded, safe, and physically meaningful ESS dispatch. The evaluation does not claim full-scale voltage or frequency stability validation because the present testbed does not include detailed network constraints, converter switching dynamics, or electromechanical frequency models. Instead, the results are interpreted as a baseline assessment of stability-oriented ESS dispatch behavior in a simplified reproducible simulation environment.

3. RESULTS AND DISCUSSION

3.1. PPO training performance

The PPO agent was trained offline using the reinforcement learning environment described in section 2. The training process was conducted for 300 episodes, with 101 steps per episode. Figure 3 shows the episode reward and average reward during PPO training. At the beginning of training, the episode reward

was highly negative because the agent initially generated exploratory ESS charge/discharge actions. As the training progressed, the reward gradually improved, indicating that the agent learned to reduce residual power imbalance, maintain SOC near the desired operating region, and avoid excessive control variation. The episode reward was approximately -72.63 at the start of training, while the moving-average reward reached approximately -2.51 by episode 300. This result confirms that the PPO agent was able to learn a feasible control policy in the simplified PV, load, and ESS environment. However, the training curve should be interpreted as evidence of learning convergence in the proposed baseline testbed, not as a direct indication that PPO universally outperforms deterministic control methods.

3.2. Evaluation profile

The trained PPO controller and the rule-based benchmark were evaluated using the same normalized PV generation and load demand profiles. The evaluation profile is shown in Figure 4. The profile includes both load-deficit and PV-surplus conditions, allowing the ESS controller to perform discharging and charging actions during the 10 s evaluation window. At the beginning of the simulation, the load demand is higher than PV generation, requiring the ESS to discharge. Around the middle of the evaluation window, PV generation becomes closer to or higher than the load demand, allowing the ESS to reduce discharging or perform charging. Toward the end of the window, the load demand again exceeds PV generation, requiring additional ESS support.

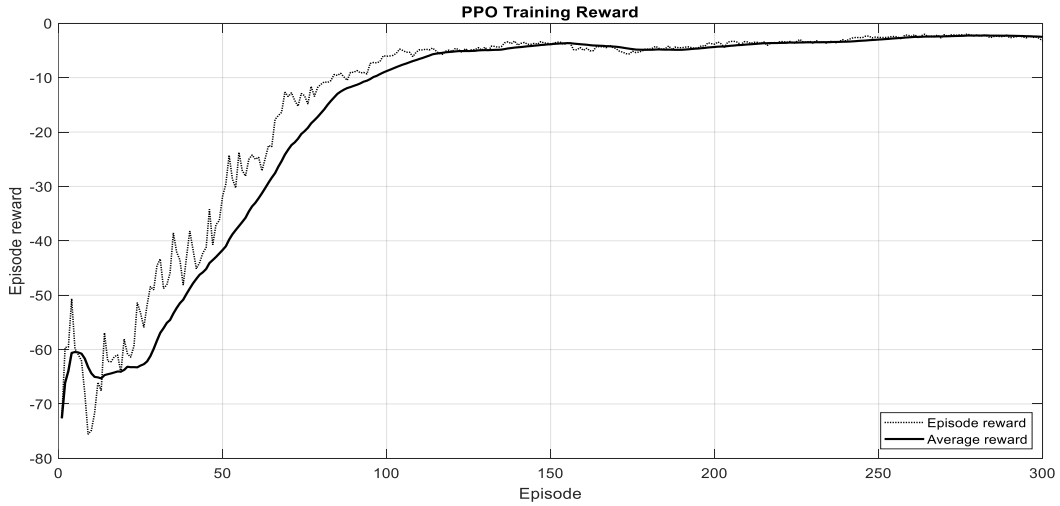

Figure 3. PPO training reward over 300 episodes

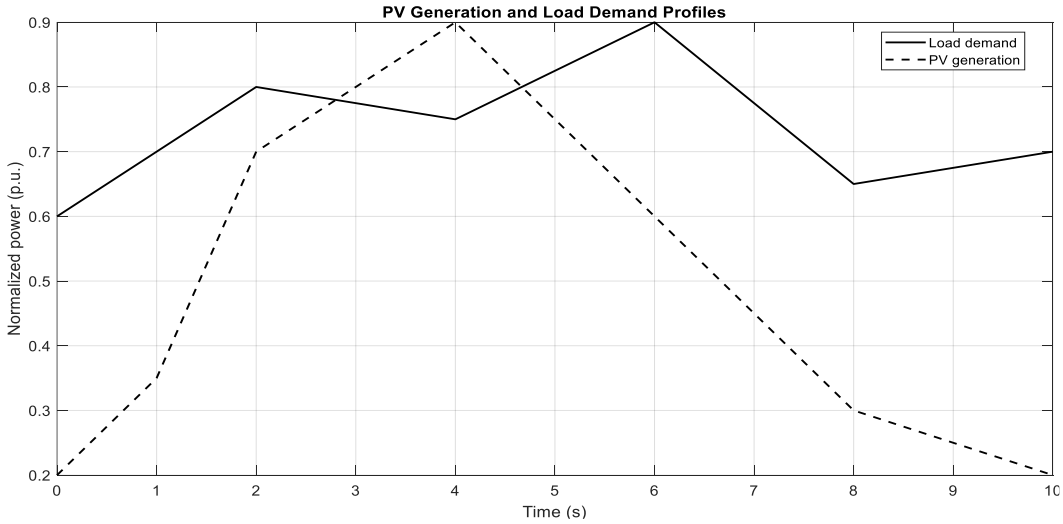

Figure 4. PV generation and load demand profiles

3.3. Dynamic response of rule-based and PPO controllers

Figure 5 compares the dynamic responses of the rule-based and PPO controllers under the same PV-load profile. The comparison includes voltage response, ESS command, SOC trajectory, and residual power imbalance. As shown in Figure 5(a), both controllers maintain the voltage response close to the nominal value of 1 p.u. This indicates that the simplified voltage-response indicator remains close to the nominal operating point during the evaluation window. Figure 5(b) shows that both controllers generate bounded ESS commands within the normalized action range. The rule-based controller produces smoother command transitions, while the PPO controller shows more fluctuation due to the learned stochastic policy behavior. Figure 5(c) shows that the SOC remains within a safe operating region for both controllers. The SOC variation is small, indicating that neither controller causes excessive charging or deep discharging during the evaluation period. Figure 5(d) shows the residual power imbalance after ESS action. The rule-based controller achieves lower residual imbalance under this deterministic profile, while the PPO controller still produces feasible and bounded ESS dispatch.

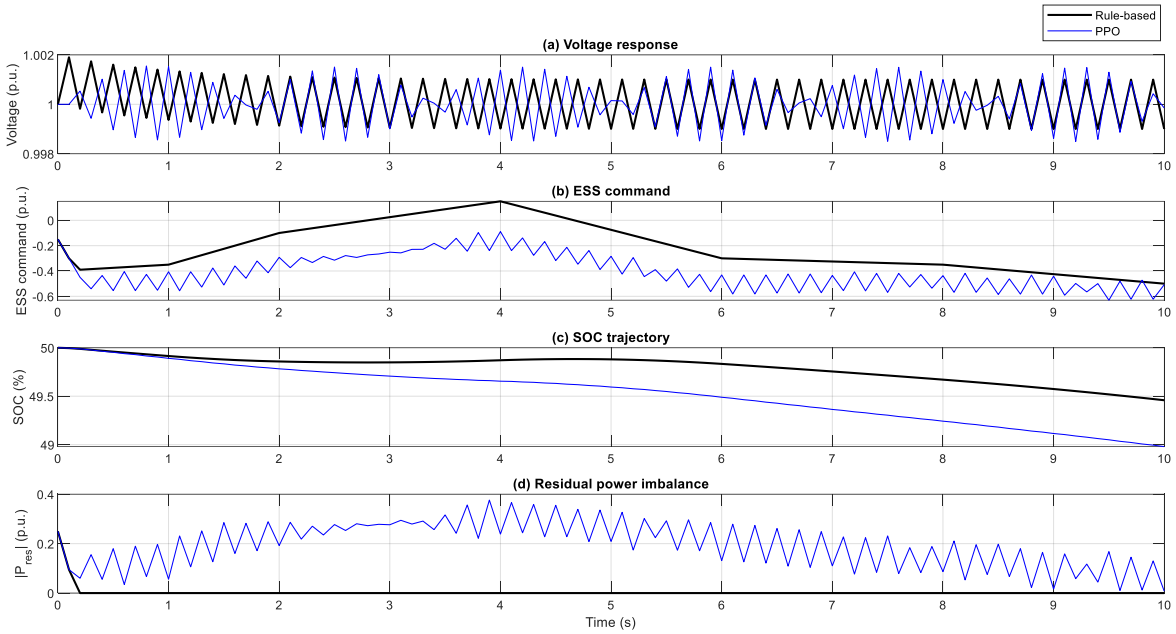

Figure 5. Comparative dynamic responses of the rule-based and PPO controllers under the same PV-load profile: (a) voltage response, (b) ESS command, (c) SOC trajectory, and (d) residual power imbalance

3.4. Quantitative performance comparison

The statistical performance of the rule-based baseline controller is summarized in Table 5. The voltage response remains close to 1 p.u., with a mean value of 1.0001 p.u. and a standard deviation of 0.0010. The ESS command is bounded between -0.5000 and 0.1500 p.u. The SOC remains within a narrow range of 49.4590–50.0000%, confirming safe ESS operation. The mean residual power imbalance is 0.0034 p.u., indicating that the rule-based controller can effectively compensate for the simplified PV-load mismatch.

The PPO-based ESS control performance is summarized in Table 6. The PPO controller maintains the voltage response within 0.9985–1.0016 p.u. and keeps SOC within 49.4830–50.0000%. The ESS command remains bounded between -0.5225 and 0.1376 p.u. These results show that PPO successfully learns a feasible charge/discharge policy without violating the SOC safety range. However, the mean residual power imbalance of PPO is 0.0275 p.u., which is higher than that of the rule-based controller.

Table 5. Rule-based baseline performance

Signal	Min	Max	Mean	Std	Mean $\lvert\Delta\rvert$
Voltage	0.9990	1.0019	1.0001	0.0010	0.0020
ESS command	-0.5000	0.1500	-0.2189	0.1862	0.0143
SOC	49.4590	50.0000	49.7940	0.1335	0.0061
Reward	-0.2650	-0.0004	-0.0052	0.0283	0.0027
Power imbalance	0.0000	0.2500	0.0034	0.0265	0.0025

Table 6. PPO-based ESS control performance

Signal	Min	Max	Mean	Std	Mean |Δ|
Voltage	0.9985	1.0016	1.0000	0.0011	0.0019
ESS command	-0.5225	0.1376	-0.2089	0.1761	0.0386
SOC	49.4830	50.0000	49.7960	0.1271	0.0058
Reward	-0.2650	-0.0022	-0.0317	0.0350	0.0217
Power imbalance	0.0001	0.2500	0.0275	0.0331	0.0216

Table 7 provides a direct comparison between the rule-based and PPO controllers. Both methods maintain voltage close to 1 p.u. and keep SOC within the safe operating region. The PPO controller produces a slightly smaller SOC range, indicating slightly lower SOC excursion. However, the rule-based controller achieves lower average residual power imbalance and smoother ESS command variation.

The comparison shows that PPO achieved the main feasibility objectives: bounded control action, safe SOC operation, and voltage response close to the nominal value. Nevertheless, in this simplified deterministic test profile, the rule-based controller remained highly competitive and outperformed PPO in terms of average residual power imbalance and command smoothness. This result is reasonable because the rule-based controller directly uses the sign of the net imbalance to determine the ESS command, while PPO learns the policy through stochastic exploration and reward feedback.

Table 7. Quantitative comparison between rule-based and PPO controllers

Metric	Rule-based	PPO	Interpretation
Voltage range (p.u.)	0.9990–1.0019	0.9985–1.0016	Both responses remained close to 1 p.u.
Voltage standard deviation	0.0010	0.0011	Comparable
SOC range (%)	0.5410	0.5170	PPO slightly lower SOC excursion
Mean ESS command (p.u.)	-0.2189	-0.2089	Comparable discharge tendency
Mean |ΔESS command|	0.0143	0.0386	Rule-based smoother
Mean residual power imbalance (p.u.)	0.0034	0.0275	Rule-based lower
Maximum residual power imbalance (p.u.)	0.2500	0.2500	Equal peak value
Mean reward	-0.0052	-0.0317	Rule-based lower penalty

3.5. Discussion of stability-oriented behavior and limitations

The results confirm that the proposed PPO framework can serve as a reproducible baseline for stability-oriented ESS dispatch in a simplified PV, load, and ESS microgrid environment. The PPO agent learned to produce bounded charge/discharge commands and maintained the SOC within a narrow and safe range. This supports the main objective of the study, which is to establish a baseline learning-based ESS control framework. The comparison with the rule-based controller provides an important practical insight. In a simple deterministic operating profile with direct access to PV-load imbalance, a rule-based controller can perform very well because its logic directly compensates the mismatch between PV generation and load demand. Therefore, PPO does not necessarily outperform rule-based control in all metrics under this simplified condition. PPO may merit further evaluation in environments with greater uncertainty, nonlinear dynamics, delayed responses, and multi-objective constraints.

The PPO result should therefore be interpreted as a feasibility demonstration of a stability-oriented learning framework. The simulated voltage-response indicator remained near the nominal value, kept SOC within 49.4830–50.0000%, and produced bounded ESS commands within the normalized action limits. At the same time, the higher mean residual power imbalance and higher action variation compared with the rule-based controller indicate that additional training refinements, richer disturbance scenarios, and broader benchmarking are required before claiming clear performance advantages.

This finding also clarifies the limitations of the baseline model. The current testbed does not include detailed converter switching dynamics, distribution-network constraints, battery aging, or electromechanical frequency dynamics. Therefore, the voltage response reported in this study should be interpreted as a simplified dynamic response indicator, not as full network-level voltage stability validation. Future work should incorporate more detailed PV-storage dynamic models and voltage predictive control structures [27]. The framework should also be extended to include converter and inverter representations, multi-ESS coordination, economic and degradation objectives, and broader benchmarking against conventional and learning-based controllers such as MPC, PID, deep deterministic policy gradient (DDPG), twin delayed deep deterministic policy gradient (TD3), or soft actor-critic (SAC). In addition, hardware implementation or hardware-in-the-loop validation should be considered to assess the real-time feasibility of the proposed controller under more realistic operating conditions [28].

4. CONCLUSION

This study presented a reproducible baseline stability-oriented PPO framework for ESS charge/discharge management in a simplified PV-load-ESS microgrid environment. The ESS control problem was formulated as a continuous-action MDP, where the agent observes load demand, PV generation, SOC, net imbalance, and previous ESS command to generate bounded charging or discharging actions. The training results showed that the PPO agent was able to learn a feasible dispatch policy, as indicated by the improvement of episode reward during training. During evaluation, the PPO controller maintained the voltage response close to 1 p.u., kept SOC within approximately 49.4830–50.0000%, and produced bounded ESS commands within the normalized action limits.

Compared with the rule-based benchmark, the PPO controller demonstrated feasible and bounded operation, but it did not outperform the rule-based strategy in all metrics. The rule-based controller achieved lower mean residual power imbalance and smoother ESS command variation under the simplified deterministic PV-load profile. This finding indicates that PPO should not be claimed as universally superior in simplified operating conditions, but it remains promising for more complex, uncertain, nonlinear, and multi-objective microgrid environments.

The main contribution of this study is the development of a transparent and reproducible baseline framework for stability-oriented learning-based ESS control. The study contributes to the field by providing a clear simulation structure, reward formulation, benchmark comparison, and quantitative evaluation procedure. Future work should extend the model by including detailed converter and inverter dynamics, frequency-response assessment, battery aging and economic objectives, broader disturbance scenarios, broader controller benchmarking, and hardware-in-the-loop validation.

FUNDING INFORMATION

This research was funded by the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Universitas Negeri Padang, under the internal research grant contract No. 2045/UN.35.15/LT/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Citra Dewi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Ricky Maulana	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Randi Purnama Putra	✓	✓		✓		✓			✓	✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Simulation data and MATLAB/Simulink models supporting this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] S. E. Eyimaya and N. Altin, "Review of energy management systems in microgrids," *Applied Sciences*, vol. 14, no. 3, Feb. 2024, doi: 10.3390/app14031249.

[2] D. Ahmed, M. Ebeed, A. Ali, A. S. Alghamdi, and S. Kamel, "Multi-objective energy management of a micro-grid considering stochastic nature of load and renewable energy resources," *Electronics*, vol. 10, no. 4, Feb. 2021, doi: 10.3390/electronics10040403.

[3] M. Ebeed *et al.*, "Optimal energy planning of multi-microgrids at stochastic nature of load demand and renewable energy resources using a modified Capuchin Search Algorithm," *Neural Computing and Applications*, vol. 35, no. 24, pp. 17645–17670, Aug. 2023, doi: 10.1007/s00521-023-08623-9.
[4] D. T. Antalem, M. V, and A. Bhattacharya, "Decentralized control of islanding/grid-connected hybrid DC/AC microgrid using interlinking converters," *Science and Technology for Energy Transition*, vol. 77, Dec. 2022, doi: 10.2516/stet/2022019.
[5] D. Emara, M. Ezzat, A. Y. Abdelaziz, K. Mahmoud, M. Lehtonen, and M. M. F. Darwish, "Novel control strategy for enhancing microgrid operation connected to photovoltaic generation and energy storage systems," *Electronics*, vol. 10, no. 11, May 2021, doi: 10.3390/electronics10111261.
[6] C. P. Agupugo, H. M. Kehinde, and H. N. N. Manuel, "Optimization of microgrid operations using renewable energy sources," *Engineering Science & Technology Journal*, vol. 5, no. 7, pp. 2379–2401, Jul. 2024, doi: 10.51594/estj.v5i7.1360.
[7] O. Akbulut, M. Cavus, M. Cengiz, A. Allahham, D. Giaouris, and M. Forshaw, "Hybrid intelligent control system for adaptive microgrid optimization: integration of rule-based control and deep learning techniques," *Energies*, vol. 17, no. 10, May 2024, doi: 10.3390/en17102260.
[8] J. Guo *et al.*, "Low-carbon robust predictive dispatch strategy of photovoltaic microgrids in industrial parks," *Frontiers in Energy Research*, vol. 10, Jul. 2022, doi: 10.3389/fenrg.2022.900503.
[9] Z. Wang, Y. Wang, J. Wu, and K. Zhang, "Fuzzy model predictive control for frequency regulation of temporary microgrids during load restoration," *IET Renewable Power Generation*, vol. 18, no. 7, pp. 1247–1259, May 2024, doi: 10.1049/rpg2.12842.
[10] J. Hu, P. Yan, and G. Tan, "A two-layer optimal scheduling method for microgrids based on adaptive stochastic model predictive control," *Measurement Science and Technology*, vol. 36, no. 2, Feb. 2025, doi: 10.1088/1361-6501/ada39b.
[11] R. Yan and Y. Xu, "A data-driven method for microgrid bidding optimization in electricity market," *Energy Conversion and Economics*, vol. 4, no. 4, pp. 276–286, Aug. 2023, doi: 10.1049/enc2.12093.
[12] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, Mar. 2021, doi: 10.1016/j.segan.2020.100413.
[13] F.-J. Lin, C.-F. Chang, Y.-C. Huang, and T.-M. Su, "A deep reinforcement learning method for economic power dispatch of microgrid in OPAL-RT environment," *Technologies*, vol. 11, no. 4, Jul. 2023, doi: 10.3390/technologies11040096.
[14] S. Zhan, Y. Lei, and A. Chong, "Comparing model predictive control and reinforcement learning for the optimal operation of building-PV-battery systems," *E3S Web of Conferences*, vol. 396, Jun. 2023, doi: 10.1051/e3sconf/202339604018.
[15] G. Muriithi and S. Chowdhury, "Optimal energy management of a grid-tied solar PV-battery microgrid: a reinforcement learning approach," *Energies*, vol. 14, no. 9, May 2021, doi: 10.3390/en14092700.
[16] Y. Huang, G. Li, C. Chen, Y. Bian, T. Qian, and Z. Bie, "Resilient distribution networks by microgrid formation using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4918–4930, Nov. 2022, doi: 10.1109/TSG.2022.3179593.
[17] K. Kumar, S. Kwon, and S. Bae, "Deep reinforcement learning-based control strategy for integration of a hybrid energy storage system in microgrids," *Journal of Energy Storage*, vol. 108, Feb. 2025, doi: 10.1016/j.est.2024.114936.
[18] I. Ioannou, S. Javaid, Y. Tan, and V. Vassiliou, "Autonomous reinforcement learning for intelligent and sustainable autonomous microgrid energy management," *Electronics*, vol. 14, no. 13, Jul. 2025, doi: 10.3390/electronics14132691.
[19] A. B. Esan and H. Shareef, "Augmented deep reinforcement learning for the energy management of microgrids considering renewable stochastic parameters," *Engineering Applications of Artificial Intelligence*, vol. 162, Dec. 2025, doi: 10.1016/j.engappai.2025.112785.
[20] K. Parhizkar, B. Yousefi, M. Rezvani, and A. N. Shirazi, "Coordinated deep policy learning for frequency-Constrained energy management for second-to-second energy balancing of microgrids," *Engineering Applications of Artificial Intelligence*, vol. 158, p. 111479, Oct. 2025, doi: 10.1016/j.engappai.2025.111479.
[21] M. S. Nkambule, A. N. Hasan, and T. Shongwe, "A review of intelligent control strategies for energy management systems in microgrids," *Energy Conversion and Management: X*, vol. 28, Oct. 2025, doi: 10.1016/j.ecmx.2025.101323.
[22] S.-W. Jung, Y.-Y. An, B. Suh, Y. Park, J. Kim, and K.-I. Kim, "Multi-agent deep reinforcement learning for scheduling of energy storage system in microgrids," *Mathematics*, vol. 13, no. 12, Jun. 2025, doi: 10.3390/math13121999.
[23] C. P. Agupugo, M. F. C. Tochukwu, K. A. Ogunmoye, A. S. Mosha, and F. Sabbih, "Review of smart microgrid platform integrating AI and deep reinforcement learning for sustainable energy management," *International Journal of Future Engineering Innovations*, vol. 2, no. 3, pp. 01–17, 2025, doi: 10.54660/IJFEI.2025.2.3.01-17.
[24] T. N. Y. Huynh, A. T. Nguyen, Y. Ahn, B. L. Oo, and B. T. H. Lim, "Multi objectives reinforcement learning for smart buildings: A systematic review of algorithms, applications and future perspectives," *Energy and Buildings*, vol. 345, Oct. 2025, doi: 10.1016/j.enbuild.2025.116045.
[25] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and P. Joelle, "An introduction to deep reinforcement learning," *Foundations and Trends in Machine Learning*, vol. 11, no. 3–4, pp. 219–354, Dec. 2018, doi: 10.1561/2200000071.
[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *Arxiv*, 2017, doi: 10.48550/arXiv.1707.06347.
[27] H. Zhao *et al.*, "Research on hybrid logic dynamic model and voltage predictive control of photovoltaic storage system," *Energies*, vol. 17, no. 17, Aug. 2024, doi: 10.3390/en17174285.
[28] H. M. Hussein, S. M. S. H. Rafin, M. S. Abdelrahman, and O. A. Mohammed, "Hardware implementation of a resilient energy management system for networked microgrids," *World Electric Vehicle Journal*, vol. 15, no. 5, May 2024, doi: 10.3390/wevj15050209.

BIOGRAPHIES OF AUTHORS



Citra Dewi received her Bachelor of Education degree in electrical engineering education from Universitas Negeri Padang, Indonesia, in 2011, and her Master of Engineering degree in electrical engineering from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2014. She is currently an assistant professor in the Department of Electrical Engineering at Universitas Negeri Padang and is pursuing a doctoral degree in electrical engineering at Universitas Andalas, Indonesia. Her research interests include power system analysis, renewable energy integration, and intelligent control in modern power networks. She can be contacted at email: citradewi@ft.unp.ac.id.



Ricky Maulana received his Bachelor of Engineering degree in electrical engineering from Universitas Andalas, Padang, Indonesia, in 2012, and his Master of Engineering degree in electrical engineering from Institut Teknologi Bandung, Indonesia, in 2015. He is currently an assistant professor in the Department of Electrical Engineering at Universitas Negeri Padang, Indonesia, and is pursuing a doctoral degree in electrical engineering at Universitas Andalas. His research interests include renewable energy, power system analysis, and power system computing. He can be contacted at email: ricky.maulana@ft.unp.ac.id.



Randi Purnama Putra received his Bachelor of Science degree in automotive engineering education from Universitas Negeri Padang, Indonesia, in 2011, his Master of Engineering degree in mechanical engineering from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2013, and his Ph.D. degree in mechanical engineering from the same institution in 2020. He is currently a lecturer in the Department of Mechanical Engineering at Universitas Negeri Padang, Indonesia. He can be contacted at email: randipurnama@ft.unp.ac.id.