EAI Endorsed Transactions
on Energy Web
Research Article EAI.EU

Deep Reinforcement Learning-Based Intelligent Control for Efficiency Enhancement in Thermal Power Plant Fuel Management

Rui Zhu[1], Qiang Liu[1], Guofeng Li[1], Xufeng Hong[1], Zhenlu Tian[1], Yanjun Guo[2*], Shengju Hao[3]

[1] Shandong Energy Group Lingtai Thermal Power Generation Co., Ltd, Pingliang,744000 Gansu, China
[2] Xi'an Thermal Power Research Institute Co., Ltd, Xi'an,710054 Shaanxi, China
[3] Xi'an YTRG Co., Ltd, Xi'an,710000 Shaanxi, China

Abstract

Thermal power plants remain a significant component of global power generation; however, several limitations persist. Hence, this research work has been developed on the basis of a proposed intelligent fuel management system based on Deep Reinforcement Learning techniques with a Proximal Policy Optimization (PPO) algorithm as a step toward increasing efficiency and sustainability of operation of thermal power plants. In this work, a fuel management problem has been formulated as a Markov Decision Process (MDP) environment within which a Deep Reinforcement Learning agent interacts with the boiler–turbine and condenser system using real efficiency data from a thermal power plant. of a thermal power plant. A multi-objective reward function was formulated using a reward shaping strategy, whereby the reward signal is explicitly structured to guide the reinforcement learning agent toward thermodynamically efficient and emission-aware plant operation. The reward formulation maximizes thermal efficiency while penalizing higher heat rate, auxiliary power consumption, and CO_2 emissions. Experimental results demonstrate that the proposed Deep Reinforcement Learning approach outperforms conventional control models. The efficiency level of this system raises from 33.68% to 35.72%, marking a relative improvement of 2.04%, with a lowered auxiliary power demand from 6.08% to 5.73%. More significantly, this optimized policy provides an expected 15-20% reduction in CO_2 emissions and lowers the heat rate from 14,000 kJ/kWh down to 11,000–12,000 kJ/kWh,000 kJ/kWh from previous levels. Convergence has been observed in the rise of episode reward values and reducing loss values during training. The current work marks a fresh start utilizing the power of PPO-Based Deep RL with Multiple Reward design in real-time closed-loop fuel management operations as a highly scalable and adaptable alternative compared to rule-set and traditional supervised learning methods.

Keywords: Deep Reinforcement Learning; Thermal Power Plants; Fuel Management; Efficiency Optimization; Multi-Objective Control; CO_2 Emissions

Received on 05 February 2026, accepted on 20 April 2026, published on 14 May 2026

doi: 10.4108/ew.11848

*Corresponding author Email: Yanjunguo22@outlook.com, zengxiu20081vip@sina.com

1. Introduction

Thermal power plants are still one of the most dominant sources for electricity production around the world, mainly because of their reliability and capability of high generation capacity. However, their efficiency depends very heavily on effective fuel management and boiler control strategy [1]. Fuel input, steam parameters, auxiliary power consumption, and condenser conditions directly affect overall plant efficiency and operational cost. Even small improvements in fuel utilization efficiency can make huge savings in economic terms and considerable emission

reduction at scale [2]. Fuel management systems in conventional thermal power plants are mainly based on rule-based logics, PID controllers, and expertise of operators. These methods will ensure stable operation, but they are all intrinsically static, incapable of adapting dynamically to fluctuating load demand, fuel quality variation, and complex nonlinear interactions among several plant subsystems [3]. Consequently, such systems frequently run at far suboptimal levels of efficiency, with excessive fuel consumption and increased auxiliary power consumption. Although these means are basically proved to ensure stable operation, they are inherently static and cannot dynamically respond to the diversified perturbations due to fluctuating load demand, fuel quality variations, and nonlinear interactions of different plant subsystems. Because of this fact, most of the conventional systems operate under suboptimal conditions with excessive fuel consumption and increased auxiliary power consumption [4].

In recent years, some studies on data-driven methods and machine learning algorithms have emerged to optimize the performance of the power plant via predictive modeling and optimization [5]. But the existing methods are generally based on offline efficiency prediction or supervised learning optimization. They lack the functionality of closed-loop decision-making [6]. In other words, the existing methods cannot model the sequential and long-term decision-making involved in control decisions in real-time fuel management. DRL recently emerged as a strong paradigm of intelligent control in complex and dynamic environments [7]. As mentioned, the fuel management problem will be formulated as a Markov Decision Process, where interaction with the plant environment allows DRL agents to learn optimality of control policy by maximizing long-term efficiency rather than short-term objectives [8]. Motivated by these advantages, this paper proposes Deep Reinforcement Learning-based intelligent control framework for improvement of fuel efficiency in thermal power plants, making use of operational parameters like fuel input energy, steam temperature, steam pressure, condenser pressure, and auxiliary power consumption [9].

Deep Reinforcement Learning framework, the agent represents the controller responsible for making operational decisions, while the environment corresponds to the thermal power plant system with which the agent interacts. The state space consists of measurable plant variables such as boiler temperature, steam pressure, condenser vacuum, auxiliary power consumption, and load demand. The action space includes controllable inputs such as fuel flow rate and air–fuel ratio within permissible operating limits. The reward function is formulated to evaluate operational performance by combining thermal efficiency improvement, heat rate reduction, auxiliary consumption minimization, and CO_2 emission control.

In view of the above shortcomings, the proposed paper presents the design of the intelligent fuel control framework of the thermal power plant using deep reinforcement learning methods. With the proposed framework, the learning agent can directly interact with the environment of the thermal power plant and perform optimal actions on the fuel and boiler in real time. In the proposed framework, the use of DNNs overcomes the problem of large state and action spaces, thereby efficiently capturing the complex, highly precise, and highly variable, and highly complex, nonlinear characteristics of the thermal power plant. The proposed work focuses on the unique aspect of combining the MOP-based reward to process the requirements of the four objectives, including the efficiency of thermal power plant, heating rate, auxiliary power consumption, and the CO_2 emission rates.

Contributions

- Formulates the thermal power plant fuel management problem as Markov Decision Process, which enables sequential and long-term decision-making towards adaptive and intelligent control.
- Develops Deep Reinforcement Learning framework with the PPO algorithm combined with a multi objective reward function to deal with efficiency, heat rate, auxiliary power consumption, and emission simultaneously.
- Implements a standardized data preprocessing pipeline, comprising the handling of missing values, normalization (rescaling features to a common numerical range), and categorical encoding (converting discrete variables into numerical form), to guarantee stable and reliable state representations in effective agent learning.
- Demonstrates that DRL-based adaptive fuel management offers superior performance compared to conventional control strategies for enhanced operational efficiency and reduced emissions, with model-free scalable decision-making ability over nonlinear plant dynamics.

1.1 Problem Statement

The performance optimization of thermal power plants using ANN-based control methods often results in poor generalization under highly nonlinear and multivariable system dynamics, as it fails to accurately predict the operational performance in unseen conditions and yields suboptimal optimization with respect to plant efficiency. In such cases, ANN-based control also finds it difficult to converge to globally optimal solutions with fluctuating environmental and load variations, hence reducing practical utility for real-time fuel management. This then limits how effectively the ANN control strategies can be deployed for high-dimensional thermal plant operational optimization [10].

The use of the Support Vector Machine model in fault detection and classification, as related to the case study of the thermal power plant, is faced with challenges related to the high degree of dependence on the value of the kernel function model parameters, besides the presence of noisy data within the dataset indicated by the presence of 'false alarms' within the dataset as malfunctioning signals. The weakness of having the choice of the kernel function predetermined induces the inflexibility of the SVM model to handle fault patterns for the differing operating conditions of the thermal power plant [11].

In the fuzzy logic system and neuro-fuzzy network employed for prediction and control of the integrated power system in the combined cycle of power plants, the problem of exponentially increasing rules with an increase in the feature set causes a slowdown in inferential speed. The use of designed fuzzy rules and membership functions leads to subjective representation of system behavior, thus affecting accuracy in dealing with dynamic conditions. This affects the use of fuzzy systems in managing the efficiency of combined cycle power plant operations [12].

1.2 Why Reinforcement Learning?

Reinforcement Learning (RL) is inherently built to cope with sequential decision-making tasks by learning the best actions through interaction with the environment instead of using labeled training data. In fact, unlike other machine learning approaches such as supervised learning or rule-based controllers, Reinforcement Learning agents learn to optimize their decision-making approaches by maximizing cumulative rewards based on feedback, which turns out to be particularly useful in dynamic scenarios involving time-changing system dynamics [13].

In contrast to classical control methods or optimization approaches, which need to have the dynamics of the process modeled in order for these methods to work properly, it is recognized for RL methods that there is no need for modeling in order for it to function properly when dealing with complex environments such as thermal power plants in which the nonlinearities in temperature, pressure, fuel input, or efficiency have been shown to be Stochastic in many cases [14].

In addition, RL is naturally suited for multi-objective optimization and adaptability, where an agent is able to consider and weigh different goals such as maximum plant efficiency, minimum use of auxiliary power and maximum reduction of emissions in a comprehensive manner, contrary to traditional systems that aim for the optimization of a singular parameter or must resort to human intervention in weighting. The adaptability in optimization has shown effective implementation in different applications associated with energy and smart grid systems

1.3 Research Objectives

- ✓ Analyze the weaknesses of traditional PID, ANN, SVM, and fuzzy logic control methodologies in thermal power stations operating in nonlinear and chaotic environments.
- ✓ Formulate the fuel management in thermal power plants as Markov Decision Process (MDP) to support decision-making in a sequence of steps and in the long term.
- ✓ Design Deep Reinforcement Learning algorithm using PPO algorithm for the boiler turbine condenser systems.
- ✓ Apply systematic methods of state preprocessing (imputation of missing values, scaling, and encoding of nominal attributes) to ensure that DRL agents have stable state representations.
- ✓ Evaluate the performance comparison between the proposed DRL approach and the conventional control system.

1.4 Paper Organization

The structure of paper organized into clear sections that are ensure logical flow and academic rigor. Section 1 outlines the background of thermal power plants, limitations of conventional fuel management systems and the motivation for adopting Deep Reinforcement Learning. Section 2 reviews selected control strategies, approaches for optimization, applications of machine learning and reinforcement learning in energy systems outlines the lacunarity of the current research. Section 3 elaborates on the proposed DRL framework in details: collection of data from Supervisory Control and Data Acquisition (SCADA) systems, its preprocessing by handling missing values, normalization, and categorization, and further design of the PPO agent with a multi-objective reward function. Section 4 are presents performance evaluation of proposed system against conventional PID-MPC and ANN-based controllers. It reports improvements concerning efficiency, heat rate, auxiliary power consumption, and CO_2 emissions. Section 5 summarizes the novelty and practical impact of the work, underlining scalability and sustainability. Section 6 consolidates the findings; it assesses and confirms the validity of the effectiveness of DRL in intelligent fuel management and further proposes some avenues of enhancements like multi-agent DRL, integration of digital twin, and safe reinforcement learning for industrial deployment.

2. Literature Survey

2.1 Conventional Control Strategies in Thermal Power Plants

Yang et al. [15] studied hybrid PID-MPC-based control applied to supercritical plant units and obtained better dynamic response and tracking compared to conventional PID control under conditions of variation in the load. The authors said that the PID-MPC hybrid proved itself in faster stabilization with better energy-saving performance regarding simulations taken up on the supercritical power unit. Their findings supported those that the classical control scheme of PID and feedforward mostly faced nonlinear dynamics and variable loads. At the same time, the inclusion of MPC added foresight but was again dependent upon accurate system models at real-time operations. This states the need for adaptive and model-free strategies in the control of complex thermal systems.

Liu et al. [16] conducted an EMPC scheme for a boiler–turbine unit to transform nonlinear dynamics into a linear form using feedback linearization in order to enhance economic performance. They achieved superior operational results under changing load demands compared to fuzzy and traditional MPC methods. Their findings showed that predictive control improved dynamic performance but still suffered from computational and modeling challenges. They also indicated that model-based optimization methods achieved better cost objectives but lacked robustness under uncertain conditions. Therefore, conventional control strategies remain suboptimal for adaptive fuel management. Song et al. [17] proposed an optimized support vector machine with grey wolf optimization for fault diagnosis, improving accuracy and stability in power equipment diagnostics by addressing high-dimensional and nonlinear data; this hybrid optimization concept inspires the proposed intelligent control framework by guiding data-driven efficiency enhancement and robust decision-making strategies.

Other control schemes used for comparison include PID and MPC controllers. These control approaches have been shown to maintain a certain level of stability and meet setpoint responses; however, their performance degrades under strong coupling and nonlinear plant dynamics. Moreover, these schemes do not possess the capability for continuous improvement based on ongoing plant observations.

2.2 Optimization-Based Fuel Management Approaches

Liu et al. [18] applied the multi-objective Genetic Algorithm along with Computational Fluid Dynamics to optimize combustion boilers in coal-fired power stations. The objective of this work was to optimize air dispersion and burning variables to improve the rate of heating transfer. They clarified and demonstrated that the optimization of NSGA-II methodology increased the thermal efficiency of burning due to increased balanced heating transfer and decreased slagging. Multi-objective methods have the capability to optimize other conflicting variables, such as improvement in efficiency of burning, as well as control in flue gas temperature distribution. The researcher proved that the boiler worked efficiently with the optimized air injection system.

Gultom et al. [19] performed the performance analysis and multi-objective optimization of 400 MW biomass co-firing steam power plant by utilizing the MOGA method in an attempt to find common optimisation of exergy efficiency, cost, and emissions. Results revealed that Pareto optimal solution for power plant conditions utilizing this method is capable of attaining better exergy efficiency, lower fuel cost, and better emissions than that ofexisting single objective methods. The researchers also applied ANN models for data-driven optimisation of nonlinear processes occurring in power generation equipment.

Xu et al. [20] developed a combustion optimization strategy for coal-fired power plant boilers that combined an improved distributed Extreme Learning Machine model with distributed Particle Swarm Optimization to maximize boiler combustion efficiency and minimize NO_x emissions. They parallelized the modeling using MapReduce to handle large high-dimensional plant data and employed PSO to determine optimal adjustable combustion parameters. The multi-objective optimization provided significant gains in reducing emissions and improving thermal performance in complex boiler dynamics. The procedure illustrated how PSO-based optimization can effectively guide combustion control decisions in large boiler systems.

2.3 Machine Learning Applications in Power Plant Efficiency Analysis

Arferiandi et al. [21] used artificial neural network (ANN) technique predicting heat rate combined cycle power plant and thereby established ability of ML models in predicting efficiency-related performance indicators. They used the ANN by providing the inputs of the heat input of the gas, the percentage of the CO2 content in the gas, and the power output and explored different combinations of the input variables in optimizing the prediction. Their paper claimed very high regression values; in fact, the optimized scenario had an R^2 value of 0.995, which is an indicator of very good prediction efficiency in relation to heat rate. However, their ANN technique relied solely on the aspect of prediction and did not consider control aspects in the operational scenario. Such an aspect reveals that the ML models may efficiently predict the efficiency-related performance variables; they do not necessarily consider the control aspect in the operational scenario.

Zhang et al. [22] analyzed various regression models involving machine learning algorithms like Linear Regression, Support Vector Regression, Random Forest Regression and Multi-Layer Perceptron estimate the performance output of thermal power plants based on operational parameters like temperature, pressure and exhaust vacuum. From the study, the authors concluded that the use of ensemble models or neural networks could

improve the assessment accuracies of key performance parameters more accurately than statistical models. Though the models had the potential to increase the accuracy of predictions, the models presented limited capabilities to focus on online decision-making or controlling decisions. As a result, the use of the models to manage fuels online is indirect.

Sciendo et al. [23] reviewed trending machine learning methods for power systems. In their analysis, they discussed that while ML models improved both forecasting and anomaly detection, the majority of those approaches remained passive predictors rather than interactive controllers. Reconfirming, this work showed that only ML cannot perform dynamic optimization of fuel usage in closed-loop operations.

2.4 Reinforcement Learning for Energy System Control

Li et al. [24] conducted comprehensive review DRL applications modern power systems and documented how RL methods were used to control grid operations, including voltage regulation and emergency responses. The analysis confirmed that RL could learn decision policies by interacting with modeled environments and thrust power systems toward adaptive control under uncertainty. This review thus laid the theoretical groundwork for RL in energy applications.

Shojaeighadikolaei et al. [25] developed a multi-agent deep reinforcement learning approach for distributed energy management and demand response in smart grids, demonstrating that RL agents can coordinate distributed resources and enhance smart grid performance. Although the study focused on distributed environments, the results highlighted the effectiveness of RL in learning control actions within complex state spaces.

Tabas et al. [26] introduced safe reinforcement learning approaches in the power system domain of frequency control, which proved more economical control strategies while still satisfying the safety constraints in contrast to traditional methods of robust control. The approach showed the efficiency of reinforcement learning strategies in overcoming conventional control strategies in high-dimensional energy systems.

2.5 Deep Reinforcement Learning for Thermal Power Plants

Chaudhari et al. [27] analyzed deep reinforcement learning (DRL) techniques for optimizing thermoelectric energy harvesting and compared the performance of SAC, PPO, and DQN learning algorithms. They modelled the control task for a thermoelectric generator as a Markov decision process, allowing a DRL agent to learn adaptively and dynamically allocate energies and estimate system efficiency and battery condition. Their analysis concluded that SAC performed optimally, with other agents performing well for maximizing system longevity. While they tailored a specific application involving thermoelectric power generators and not thermal power plants, they illustrated that DRL and optimization done with SAC, PPO, and DQN models would improve dynamic allocation decisions over other optimization models. Based on this analysis and application, models involving PPO, SAC, and DQN would perform efficiently for real-time fuel and efficiency optimization tasks in thermal power plant.

Wang et al. [28] have developed an energy optimization strategy for coal-fired power plants based on deep Q-learning. The emphasis was placed on combustion efficiency and emission reduction. Their DRL framework made dynamic adjustments to air-fuel ratios and boiler parameters based on operational feedback, with significant heat rate and CO_2 emission reductions observed. The results proved to be much more adaptive compared to rule-based and optimization-based controllers. It was claimed that the advantage of using DRL is in learning long-term optimal control policies under uncertainty in plant conditions. This reinforced DRL's applicability in intelligent fuel management systems.

The literature review shows that conventional PID and model predictive controllers, while stable, are heavily reliant on precise plant models, nonlinear dynamics, load fluctuations, and uncertainty; all these factors result in suboptimal fuel utilisation. Optimisation-based approaches, such as GA, PSO, and MOGA, enhance combustion and efficiency but generally operate in offline or open-loop mode and incur extensive computational cost, hence cannot be effectively deployed in real time. The machine learning models in the form of ANN, SVM, and regression techniques have demonstrated strong predictive power over the heat rate and efficiency but remain passive estimators devoid of closed-loop decision-making capability and show poor generalisation for unseen operating conditions. Reinforcement learning studies performed so far confirm its adaptiveness and long-term optimisation capability but are mostly application-specific or not integrated into multi-objective control. All these works collectively motivate the work presented herein by formulating fuel management as a Markov decision process and employing PPO-based deep reinforcement learning framework with a multi-objective reward function to enable real-time, adaptive, and closed-loop optimisation of efficiency, heat rate, auxiliary power consumption, and CO_2 emissions.

3. Methodology

The methodology part describes the hardware setup but fails to include the core PPO training parameters needed for reproducibility in experiments. Important information such as learning rate, discount factor, clipping value, size of the batch, number of epochs for an update, network model architecture, activation functions, optimizer model,

model for exploration strategy, and total number of training episodes is excluded. Moreover, aspects of environment step size, number of episodes for an environment step, reward normalization, and convergence are also excluded that could help in reproducing the training process or verifying improvements in performance. For better scientific standards and ease of reproduction, it is important that such PPO training parameters be clearly stated.

The proposed methodology is based on end-to-end Deep Reinforcement Learning (DRL) approach optimal and intelligent fuel consumption Thermal Power Plants as depicted in **Figure 1**. The approach begins with data extraction from the Thermal Power Plant Efficiency Dataset, on kaggle which records important parameters related to its operations, including boiler temperature, steam pressure, condenser vacuum, auxiliary power demand, and CO_2 emissions. The extracted data goes through certain data preprocessings, such as handling of missing data, normalization of numerical parameters, and conversion of parameters that are of types categorical to numbers in order to maintain uniformity of data. All this data is utilized for formulating a reinforcement learning environment, where the MDP is applied in designing an optimal state and action space in relation to the operational activities of the Thermal Plants.

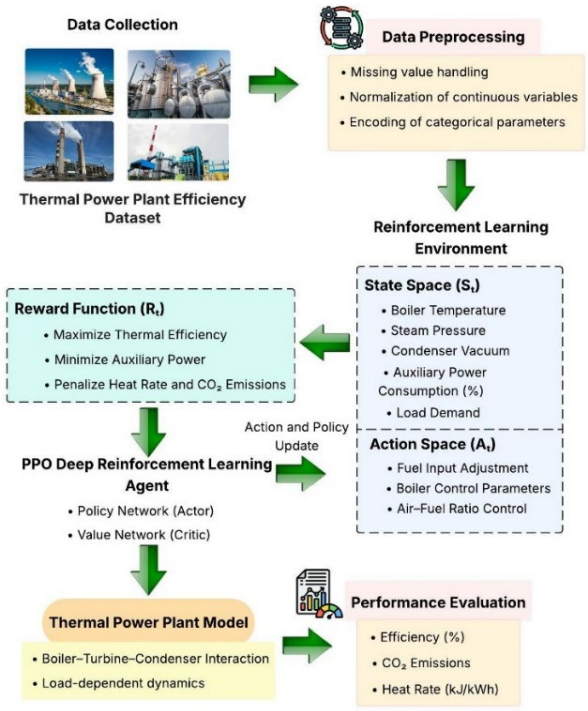

Figure 1. Overall Framework of the Proposed Deep Reinforcement Learning–Based Fuel Management System for Thermal Power Plants

3.1 Data Collection

Data Collection is the fundamental units of the Deep Reinforcement Learning framework that provides an operational dataset during both the training phases as well as the testing phase. The SCADA system existing within the thermal power plant helps provide access to the previous operational data directly. And this provides the guarantee that the previous operating factors are captured with measurements that reflect the actual working scenario. This is important to note that the process of the operational dataset is a comprehensive one that encompasses all the necessary factors such as temperature of boiler, the steam pressure, condenser vacuum level, the rate of fuel consumption, consumption of auxiliary power, load demand, heat rate, and emissions that also possess the ability to tag the steady-state or transient operation points. The multiplicity of information gathered ensures the reinforcement learning algorithm can handle different working situations effectively. All the parameters help determine the working space of the learning environment, such that the algorithm can correctly observe the operational context of the plant. By considering the entire dataset, the DRL algorithm can easily optimize the thermal performance while, at the same time, reducing the auxiliary power used and emissions. In short, the Data Collection block bridges the model guidance phase with the implementation phase, providing the relevant operational facts required to prove the developed intelligent fuel management system.

3.2 Data Preprocessing

Data preprocessing is an important step for preparing raw operational data so that effective learning can take place in deep reinforcement learning frameworks. The raw measurements extracted from thermal power plant monitoring systems may have missing values, noise, and scale inconsistencies due to faulty sensors or communication delays or because of the presence of operational disturbances. In this regard, missing values are treated using proper interpolation or statistical imputation techniques to maintain temporal continuity. Continuous variables related to temperature, pressure, and flow rates should be normalized to a common scale to avoid dominance of high-magnitude features and to improve numerical stability during the training of neural networks. Moreover, if there are categorical or discrete operational parameters are encoded into a numerical representation that can serve as an input for deep learning models. This involves feature normalization as well as coding, which helps to achieve the consistent representation of the plant operational status regardless of the operating conditions. This improves the convergence velocity, the training stability, as well as the generalization ability during the pre-processing stage. Data preprocessing helps to convert the raw sensor readings into ordered inputs, which helps the learning agent to interact with the thermal power plant environment.

3.2.1 Missing Value Handling

Missing values are one of the general problems that may occur for most thermal power plant datasets due to malfunctioning sensors, failure in communication in SCADA systems, maintenance downtime, or transient disturbances in the operation. Incomplete data may introduce bias or discontinuity into the state representation, thus badly influencing the learning stability and convergence of DRL models. Therefore, the treatment of missing values is an important task, which provides complete data and hence reliable interaction with the environment for the learning agent. Let x_t be a time-series measurement of a process variable at time step t. In such a case, at time t, when the value is missing, the shallow statistical imputation techniques can be employed, such as mean or median substitution, when the missingness is sparse and the non-temporal dependence is weak. Mean imputation is defined by equation (1):

$$x_t = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1)$$

where N denotes the number of samples observed for a given variable. This helps in retaining the global statistical distribution but might round-off local variations. To observe time-series operational data, linear interpolation works well since it ensures continuity between samples. If the value for a time period in which no sample was taken, but between samples x_{t-1} and x_{t+1}, is needed, it can be calculated using equation (2):

$$x_t = x_{t-1} + \frac{x_{t+1}-x_{t-1}}{(t+1)-(t-1)} \quad (2)$$

It becomes very useful when considered with efficiently changing parameters that are mostly related to temperature, pressure, and flow rates, as they smoothly transit between phases and conditions when working normally. If consecutive data are missing in a few instances, then forward fill or last observation carry forwarded is considered to maintain certain system functionalities that are assumed to remain in their previous stable state until new observations are considered. It is identified by equation (3) as follows:

$$x_t = x_{t-k} \quad (3)$$

where k is the number of time steps since the last valid observation. In practice, forward filling is usually used in real-time control systems for avoiding abrupt state discontinuities. According to the characteristics of data, appropriate missing value handling techniques are applied at the preprocessing stage that ensures the completeness and numerical stability of the state space. It improves the robustness of the DRL agent by avoiding incorrect transitions of the state due to incomplete data. Proper imputation will contribute to faster convergence, improved policy learning, and more reliable fuel management decisions in thermal power plant control environments.

3.2.2 Normalization

The continuous operational variables in thermal power stations, such as boiler temperature, steam pressure rate, fuel rate, and condenser vacuum, have varying numerical values along with units. Feeding such variables as they are to deep reinforcement learning models without normalization may result in the suppression of the learning process by the feature with the highest magnitude. Hence, normalization is necessary for the balanced accommodation of the features as well as numerical stability during the process of training the neural network model. The commonly used normalization process for continuous variables includes min-max normalization, whereby the variables are rescaled within the range [0,1]. The formula for the normalization process for a given feature x to obtain x^{norm} is given by equation (4):

$$x^{\text{norm}} = \frac{x-x_{\min}}{x_{\max}-x_{\min}} \quad (4)$$

where $x_{\min}$ and $x_{\max}$ are minimum and maximum values of variable in the dataset. This method maintains relative relationships between data points and is quite appropriate for reinforcement learning environments where bounded state values tend to improve policy stability. Otherwise, when data distributions are approximately Gaussian or contain outliers, z-score standardization is used. It transforms variables to have zero mean and unit variance, hence weakening the effect of extreme values. The standardized variable is given by equation (5):

$$x^{\text{std}} = \frac{x-\mu}{\sigma} \quad (5)$$

where μ and σ are mean and standard deviation of variable, respectively. The advantage of this process is that it allows for fast learning through gradient-based learning methods, which require equal scaling of the features. In the case where the variables have constraints or nonlinear sensitize variation, a logarithm transformation may be applied. The formula for this transformation is equation (6):

$$x^{\log} = \log(x+\epsilon) \quad (6)$$

Where ϵ is a small constant to prevent numeric instability. Log normalization is especially useful when dealing with parameters like fuel rates or emission amounts, which may vary over multiple orders of magnitude. With proper application of normalizing procedures to continuous variables, it is possible to stabilize the learning processes of the DRL agent taking inputs from state variables represented within consistent numeric values. Normalization makes the action space less sensitive to differences in plant parameter scales and thus ensures effective exploration of the action space. State space normalization, in particular, is central to ensuring effective control in fuel management in thermal power plants.

3.2.3 Encoding of Categorical Parameters

In the case of thermal power plant data, there can be some functional variables that can be categorical, for example, fuel type, mode of operation, valve settings, or state of the boiler, among which all these variables are categorical in nature and cannot be directly used as an input for neural networks since they cannot handle categorical variables directly because neural networks perform operations on numeric variables only. The categorical variables have to be converted into numeric variables in a way that it maintains the difference between all variables without creating an ordering or bias between them. A way of doing that is by using “one-hot encoding,” an encoding scheme that maps every category into a binary vector. In this case,

for k classes, the binary vector for class "i" will have length k, and all elements will be 0, and at index "$\boldsymbol{i}$," it will be 1. Mathematically, for a categorical variable "$\boldsymbol{c}$" with value "i," it can be expressed by equation (7):

$$v_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}, j = 1,2,\dots,k \quad (7)$$

This method preserves the separability of the classes. This technique should be used when the classes lack a natural order. When the classes have a natural order, such as the operational levels ("low", "medium", "high") or priority flags, the method of ordinal encoding should be used. Here, every class is assigned a group of consecutive integers in terms of their natural ordering. Finally, another sophisticated way of handling onehot-encoded variables is to project them to a continuous, compact space via embeddings inside a neural network. These embeddings address natural order among variables, enabling generalization of the DRL agent to similar discrete states. In practice, it is more efficient than encoded variables, especially for variables of higher cardinality. Appropriate representation of category parameters is essential to ensure that all aspects of the state, both continuous and discrete, are encoded as numeric values to facilitate DRL agents. These aspects include all operational states of the thermal power plant, which, if known by the agent, would lead to better decision-making, faster convergence, and control of fuel, boiler, and emission variables.

Algorithm 1: Data Preprocessing for Thermal Power Plant Dataset

Input: Raw Thermal Power Plant Dataset (temperature, pressure, fuel rate, load, emissions)
Output: Preprocessed and normalized dataset for DRL environment

```
BEGIN
  Load raw thermal power plant dataset

  FOR each feature in dataset DO
    IF missing values exist THEN
      IF time-series continuity required THEN
        Apply linear interpolation
      ELSE
        Replace with mean value
      ENDIF
    ENDIF
  ENDFOR

  FOR each numerical feature DO
    IF feature range is wide THEN
      Apply min-max normalization
    ELSE
      Apply z-score standardization
    ENDIF
  ENDFOR

  FOR each categorical feature DO
    IF feature has no order THEN
      Apply one-hot encoding
    ELSE
      Apply ordinal encoding
    ENDIF
  ENDFOR

  Store preprocessed dataset
END
```

The PPO algorithm was implemented with a learning rate of 3×10^{-4} and a discount factor (γ) of 0.99 to ensure long-term reward optimization. The clipping factor (ε) was set to 0.2 to stabilize policy updates. A batch size of 64 samples was used with 10 training epochs per update cycle. The policy and value networks consisted of two fully connected hidden layers with 128 neurons each, using ReLU activation functions. These hyperparameter settings were selected to ensure stable convergence and reproducibility of results.

3.2.4 Outlier Detection and Treatment

Outlier detection was performed to improve data reliability and ensure stable reinforcement learning training, as abnormal values may arise due to sensor malfunction, SCADA communication errors, or transient plant disturbances. Continuous variables were examined using the Interquartile Range (IQR) method, where the interquartile range was computed as Q3 − Q1, and any observation lower than Q1 − 1.5 × IQR or greater than Q3 + 1.5 × IQR was classified as an outlier. For approximately normally distributed variables, z-score filtering was also applied, and observations with an absolute z-score greater than 3 ($|z| > 3$) were treated as extreme values. Instead of removing such observations, a winsorization strategy was employed by capping values at the respective upper and lower threshold limits in order to preserve temporal continuity of the SCADA time-series data. Approximately 2–3% of the dataset values were adjusted through this procedure, ensuring robustness against noisy measurements while maintaining realistic operational characteristics of the thermal power plant dataset.

3.3 Reinforcement Learning Environment Modeling

The reinforcement learning environment modeling defines the interaction framework between the learning agent and the thermal power plant system by formulating plant operation as Markov Decision Process. In this study, the environment represents thermal power plant dynamics, and the agent takes observations of the current operational state, executes control actions, and perceives rewards. Nonlinear coupled behavior within boiler-turbine-condenser subsystems has been captured under this environment, and it governs state transitions based on control actions and system dynamics. Provided with realistic state transitions and performance feedback, the

environment enables the reinforcement learning agent to iteratively learn optimal fuel management and control policies that improve efficiency while minimizing auxiliary power consumption and emissions.

The boiler turbine condenser system is modeled using an explicit discrete-time state transition mechanism, where the next system state is determined based on the current operational state and the control actions generated by the PPO agent. The state space includes boiler temperature, steam pressure, condenser vacuum, auxiliary power consumption, and load demand, while the action space consists of fuel flow rate, air–fuel ratio, and turbine valve position. The transition dynamics are implemented through a data-driven simulation derived from historical SCADA data, capturing nonlinear subsystem interactions and operational constraints to provide consistent and reproducible system behavior.

3.3.1 State Space (S_t)

The state space corresponds to the actual operating state of the thermal plant at each point in time in the decision step and is the input to the reinforcement learning agent. In step t of time, state vector S_t is represented by a combination of important thermodynamic and operational parameters that have direct effects on the efficiency of plant operation and fuel consumption. Mathematically, state-space can be represented by equation (8):

$$S_t = [T_t, P_t, V_t, A_t, L_t] \quad (8)$$

In this case, T_t refers to boiler temperatures, P_t denotes steam pressure, V_t symbolizes condenser vacuum, A_t denotes auxiliary power usage percentage, and L_t symbolizes plant load demand during time t. All these variables identify and convey sufficient information about inter-relations between different subsystems at non-linear stages, so that appropriate decisions can be made by the deep reinforcement learning agent. The state variables at this stage are normalized before interaction with the agent for stability and smooth learning purposes.

The selected state variables are physically motivated by their direct influence on thermal power plant performance. Boiler temperature directly affects combustion efficiency and steam generation rate. Steam pressure determines turbine output power and overall cycle efficiency. Condenser vacuum influences the back pressure of the turbine and significantly impacts thermal efficiency. Auxiliary power consumption represents internal energy usage within the plant and affects net plant efficiency. Load demand reflects real-time operating conditions and determines the required fuel input and system response. Therefore, the chosen state variables comprehensively represent the thermodynamic and operational status of the plant.

3.3.2 Action Space (A_t)

The action space A_t specifies control actions available to reinforcement learning agent to regulate thermal power plant operation at each time step t. These actions correspond to those control inputs that could be adjusted and vary directly with combustion behavior and, in turn, with energy conversion efficiency. In this study, the action space is a modeled continuous vector given by equation (9):

$$A_t = [F_t, B_t, R_t] \quad (9)$$

where F_t denotes fuel input rate, B_t is boiler control settings-damper or valve positions-and R_t Z is the air-fuel ratio at time t. Continuous action modeling enables the agent to achieve fine-scale control adjustments within operational bounds. Through interactions with the environment under this action space, the agent learns an optimal control policy that enhances thermal efficiency and minimizes auxiliary power consumption and emissions.

The control actions are constrained within permissible operational limits. The fuel input rate F_t is restricted within 60% to 100% of the rated boiler firing capacity to prevent unstable combustion and thermal stress. The boiler control setting B_t (damper/valve position) is bounded between 20% and 100% of its operating range to avoid mechanical strain and incomplete airflow regulation. The air–fuel ratio R_t is maintained within the range of 1.1 to 1.6 to ensure complete combustion while minimizing excess air losses and emission formation. These operational constraints ensure that the learned control policy remains within practical industrial safety limits and reflects realistic thermal power plant operating conditions.

3.3.3 Markov Decision Process Formulation

The thermal power plant is formulated as a Markov Decision Process (MDP), defined by the tuple ⟨S, A, R, P, γ⟩, where S represents the state space, A denotes the action space, R is the reward function, P corresponds to the state transition dynamics, and γ is the discount factor. This formulation enables sequential decision-making under dynamic and nonlinear operating conditions of the thermal power plant.

The **state representation (S_t)** captures the real-time operational condition of the plant and is defined as a vector comprising fuel input rate, boiler steam pressure, condenser vacuum level, auxiliary power consumption, and load demand. These variables jointly describe the thermodynamic and operational status of the boiler–turbine–condenser system and provide sufficient information for informed control decisions.

The **action space (A_t)** consists of continuous control actions related to boiler and turbine regulation, including adjustments to fuel flow rate, air–fuel ratio, and boiler valve or damper positions. These actions directly influence combustion behavior, steam generation, and turbine performance, thereby affecting overall plant efficiency and emissions.

The **reward function (R_t)** is designed as a multi-component formulation that encourages thermal efficiency maximization while penalizing high auxiliary power consumption, increased heat rate, and elevated CO_2 emissions. By combining these reward components, the agent learns a balanced control policy that jointly optimizes economic performance and environmental sustainability over long-term operation.

3.4 Reward Function (R_t)

The reward function R_t is designed such that the reinforcement learning agent can be guided effectively for efficient and optimal operation of thermal power plant by assessing benefit of each action taken. The main aim of the designed reward function is to optimize the thermal efficiency while penalizing the operational factors such as high Auxiliary power consumption, high heat rate, and increased carbon emission. The reward function at time step t is expressed following equation (10):

$$R_t = \alpha\eta_t - \beta A_t - \gamma H_t - \delta C_t \quad (10)$$

where η_t represents the thermal efficiency, A_t indicates auxiliary consumption, H_t indicates the heat rate, while C_t represents CO_2 emissions at time t. The weighting factors α, β, γ, and δ represent the parameters of the weighting factors used to balance efficiency maximization and reduction of environmental factors. The reward formulation helps the agent optimize fuel management policies by taking into account both economic performance as well as sustainable criteria.

The weighting coefficients were selected as α = 0.4, β = 0.2, γ = 0.2, and δ = 0.2 based on empirical tuning through multiple training trials to achieve stable convergence and balanced performance. A relatively higher value of α prioritizes thermal efficiency improvement, while β and γ ensure adequate penalization of heat rate and auxiliary power consumption. The coefficient δ enforces emission control without excessively compromising efficiency gains. This balanced weighting strategy ensures a fair trade-off between economic performance and environmental sustainability, thereby maintaining stability and practicality of the optimized control policy.

The multi-objective reward function in this study adopts a linear aggregation of objectives to ensure stable and interpretable learning behavior. Linear combination was intentionally selected due to its computational simplicity, ease of weight tuning, and proven stability in continuous control environments. In real-time industrial applications such as thermal power plant operation, stable convergence and predictable policy updates are critical. Nonlinear aggregation methods (e.g., quadratic or exponential formulations) may amplify gradient variations and introduce sensitivity to hyperparameter tuning, potentially affecting convergence consistency. Preliminary evaluations indicated that the linear weighted structure provides balanced optimization of thermal efficiency improvement and CO_2 emission reduction without causing instability in the learning process. Therefore, linear reward aggregation was retained as it ensures reliable multi-objective optimization while maintaining training stability and operational robustness.

3.4.1 PPO Optimization and Policy Update

PPO is selected as the primary reinforcement learning algorithm in this study due to its demonstrated stability, robustness, and effectiveness in continuous control environments such as thermal power plant operations. PPO improves policy learning by constraining successive policy updates within a trust region, thereby preventing abrupt changes that could destabilize training. This is achieved through a clipped surrogate objective function, which balances policy improvement and stability during optimization.

The PPO objective function maximized during training is expressed as equation (11):

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (11)$$

where the probability ratio between the updated policy and the previous policy is defined as equation (12):

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{ad}}(a_t|s_t)} \quad (12)$$

In this formulation, $\hat{A}_t$ denotes the estimated advantage function that quantifies the relative benefit of selecting action a_t in state s_t, while ϵ represents the clipping threshold that limits excessive policy updates. The clipping mechanism ensures that policy improvements remain conservative, thereby enhancing learning stability under nonlinear plant dynamics.

In addition to policy optimization, PPO simultaneously updates a value function that estimates the expected return from a given state. The value function loss is computed using the mean squared error between predicted and target values as equation (13):

$$L^V(\theta) = \mathbb{E}_t\left[\left(V_\theta(s_t) - V_t^{\text{target}}\right)^2\right] \quad (13)$$

The complete PPO loss function integrates the clipped policy objective, value function loss, and an entropy regularization term, and is given by equation (14):

$$L(\theta) = L^{CLIP}(\theta) - c_1 L^V(\theta) + c_2 S[\pi_\theta] \quad (14)$$

Here, $S[\pi_\theta]$ denotes the policy entropy, which promotes sufficient exploration during training and prevents premature convergence to suboptimal deterministic policies. The coefficients c_1 and c_2 regulate the contribution of value estimation accuracy and exploration to the overall optimization process.

Reward Normalization

To further enhance training stability and reduce sensitivity to scale variations in reward magnitudes, reward normalization is applied prior to policy updates. The normalized reward is computed as:

$$\tilde{R}_t = \frac{R_t - \mu_R}{\sigma_R + \epsilon} \quad (15)$$

where μ_R and σ_R denote the mean and standard deviation of reward values computed over a training batch. Reward normalization ensures consistent gradient updates and improves convergence behavior under varying operating conditions of the thermal power plant.

Convergence Criterion

The convergence of the PPO training process is evaluated based on the stabilization of policy updates and loss reduction across successive iterations. Convergence is assumed when the variation in the PPO loss function satisfies equation (16)

$$\lim_{t\to\infty} |L_{t+1} - L_t| < \delta \quad (16)$$

where δ is a predefined small threshold. Additionally, convergence is corroborated through consistent cumulative episode rewards, indicating stable and reliable policy performance.

3.5 Thermal Power Plant Model Development

The model of thermal power plant developed to emulate the physics and operation of boiler-turbine-condenser system in the reinforcement learning environment. Nonlinear interactions among major subsystems, such as fuel combustion dynamics of the boiler, steam generation and expansion dynamics of the turbine, and heat rejection dynamics of the condenser are modeled. Operational constraints that include the safety limits, control variables bound, and load-dependent dynamics are modeled in an explicit manner. Imposing these constraints in transitions within the environment, the model prevents the learning agent from choosing infeasible and unsafe actions.

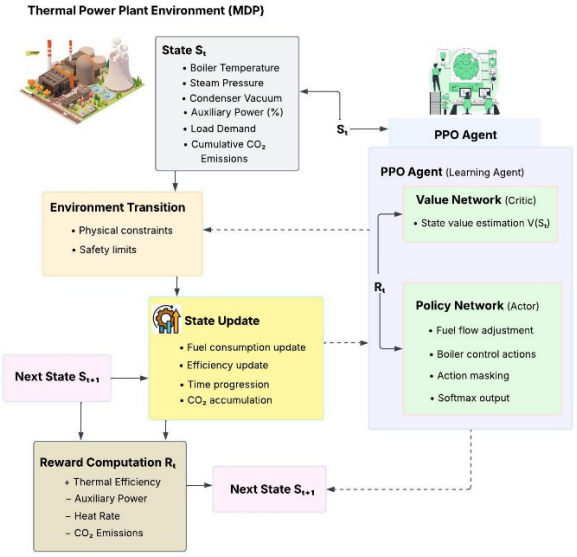

Figure 2. Interaction Between the PPO Agent and Thermal Power Plant Environment

Figure 2 illustrates the closed-loop interaction between the PPO agent and the thermal power plant environment. The agent observes the current plant state and selects control actions, which are applied to the plant model to update fuel consumption, thermal efficiency, auxiliary power usage, and cumulative CO_2 emissions according to system dynamics. As time progresses, emissions are accumulated to capture long-term operational effects. This dynamic interaction enables the PPO agent to iteratively learn optimal control policies that balance efficiency maximization with the reduction of environmental impacts, making the proposed model suitable as a realistic simulation environment for training and testing deep reinforcement learning–based fuel management strategies.

3.5.1 Boiler–Turbine–Condenser Interaction

Boiler-Turbine-Condenser system, also known as the BTC system, can be stated as the fundamental process of a thermal power station, which also greatly influences the efficiency levels along with the consumption of fuels in the power station. In this system, the chemical energy in fuels gets converted into heat energy through the production of high-pressure steam along with high temperatures. This high-speed steam is used to turn the turbine mechanically, thus generating electrical energy. However, this used steam is further cooled in a condenser to create a low-pressure environment or vacuum that allows this condensation to be reused in the boilers. These are explicitly modeled as interactions in the proposed model to give realistic transitions between the states in reinforcement learning. The changes in fuel flow or air and fuel ratio result in changes in boiler temperature and steam pressure, and this affects the outputs and conditions of exhausts of the turbines. Changes in vacuum of condensed water result in changes in back pressure and outputs of turbines, and auxiliary equipment affects auxiliary power consumption. Coupled modeling of these interactions allows the DRL agent to recognize coordinated control strategies to optimally manage the whole plant rather than its constituents.

Table 1. Key Interactions among Boiler, Turbine, and Condenser Subsystems

Subsystem	Input Variables	Output Variables	Impact on Overall Performance
Boiler	Fuel flow rate, air–fuel ratio	Steam temperature, steam pressure	Determines combustion efficiency and available thermal energy
Turbine	Steam pressure, steam temperature	Mechanical power, exhaust steam	Affects electrical output and heat rate
Condenser	Exhaust steam, cooling water flow	Condenser vacuum, condensate	Influences turbine back pressure and cycle efficiency
Auxiliary Systems	Pump/fan operation	Auxiliary power consumption	Reduces net plant efficiency

The **Table 1** provides a summary of functional relationships between boiler, turbine, condenser, and auxiliary systems in a thermal power station as far as fuel

usage is concerned. The boiler provides high-energy steam under high pressure as its direct effect is on the quality of steam that goes into the turbine and hence the power generated and fuel efficiency in the combustion process. The turbine is responsible for converting by thermal energy into mechanical and electrical energy of exhaust steam quality affecting condenser performance. A condenser acts as a vacuum that reduces backpressure through the condensation of exhaust steam thus improving the efficiency of the turbine and balancing the thermodynamic process. Auxiliary systems that involve pumps and fans help all other systems to function, thus requiring extra power that affects station efficiency.

3.5.2 Load-dependent dynamics

Load dynamics are used to explain the dependence of thermal power plants on changes in the load of electricity. As the load on the power plant rises or falls, the primary process parameters like fuel input, steam pressure, and boiler temperature change, resulting in the regulation of these process parameters for the smooth running of the power plant. Low load levels result in the reduction of the efficiency of the combustion process and the dominance of the auxiliary power contribution to the total power produced, while high loads imply high stresses and high fuel consumption levels. In the new framework, dynamics that are dependent on the loads being considered are introduced within the plant modeling to accurately represent the plant dynamics under various operating conditions. The reinforcement learning agent receives information about the load demand as one of its state inputs within the state space to develop plant operating policies that adjust fuel inputs, air/fuel mixtures, and boiler settings according to the loads being considered. The introduction of load dependencies within the plant modeling allows for better efficiency, stability, and emission performance within the wide operating range based on the loads considered by the agent.

Table 2. Impact of Load Variation on Thermal Power Plant Dynamics

Load Condition	Fuel Input	Steam Parameters	Auxiliary Power	Efficiency Impact
Low Load	Reduced	Lower pressure and temperature	Relatively high (%)	Decreased efficiency
Medium Load	Moderate	Stable operating range	Normal	Optimal efficiency
High Load	Increased	High pressure and temperature	Increased absolute value	Risk of thermal stress
Load Ramp-Up/Down	Rapid adjustment	Transient fluctuations	Variable	Control challenges

The **Table 2** shows the impact of load variation on critical operational parameters of a thermal power plant. At the low-load level, fuel and steam variables decrease, but the share of auxiliary power usage rises, causing lower efficiency. For the medium-load level, the plant runs at its designed operating point along with steady steam and enhanced efficiency. For the high-load level more fuel needs to be supplied along with greater steam pressure and temperature, which is ideal for power generation and reduces possible thermal stress and absolute auxiliary power consumptions. At the load ramp-up and load ramp-down levels, sudden changes take place in the system variables, emphasizing the need for effective load variation control.

Algorithm 2: PPO-Based Reinforcement Learning for Fuel Management

Input: Preprocessed plant state variables, action bounds, reward weights
Output: Optimal control policy for fuel and boiler operation

```
BEGIN
  Initialize PPO agent with policy network and value network
  Initialize thermal power plant environment

  FOR each episode DO
    Reset environment and observe initial state S_t

    WHILE episode not terminated DO
      Select action A_t using current policy π(S_t)

      Execute action A_t in environment
      Observe next state S_(t+1), reward R_t

      IF efficiency increases AND emissions decrease THEN
        Assign positive reward
      ELSE
        Apply penalty
      ENDIF

      Store (S_t, A_t, R_t, S_(t+1)) in memory
      Update state S_t = S_(t+1)
    ENDWHILE

    IF memory buffer is full THEN
      Update PPO policy using clipped objective function
    ENDIF
  ENDFOR

  Output optimized control policy
END
```

3.6 Parameter Configuration of the Proposed Work

The parameter configuration of proposed work shows that experimentations as well as simulations were performed on a desktop computer (DESKTOP-RHF1E9H), which has the 12th Gen Intel Core i5-12400 CPU running at 2.50 GHz and is accompanied by an 8 GB installed ram (7.75 GB usable). This desktop is running a 64-bit operating system with an x64-based processor. This makes the desktop capable of supporting high-end computations like the training and simulation of the deep reinforcement learning models. Though the configuration may be less sophisticated compared to supercomputer setups, the desktop has enough computing power to support the preprocessing of the SCADA datasets, the creation of the neural networks, as well as the simulation using DRL agents based on the PPO algorithm. Additionally, the fact that the desktop doesn't support pen or touch indicates that the desktop was primarily designed for computing and not for its hardware attributes.

3.6.1 PPO Training Hyperparameters

Reproducibility and transparency of the proposed Deep Reinforcement Learning framework are supported by explicitly reporting the complete PPO training configuration used in this study. The PPO agent was trained using an actor–critic architecture with shared state representations. Both the policy network and value network were optimized using the Adam optimizer. A fixed learning rate was employed to achieve stable convergence during training. The discount factor γ was selected to emphasize long-term operational performance, including efficiency improvement and emission reduction. The clipping threshold was applied to restrict excessive policy updates and improve training stability. Mini-batch learning was adopted, and multiple training epochs were performed per policy update to enhance sample efficiency. Exploration was inherently handled through the stochastic Gaussian policy of PPO, enabling continuous action sampling within predefined operational limits. Table 3 presents the complete set of PPO training hyperparameters adopted in this study, including optimization settings, policy update constraints, and learning configurations used for training the reinforcement learning agent**.**

Table 3. PPO Training Configuration

Hyperparameter	Value
Learning rate	3×10^{-4}
Discount factor (γ)	0.99
PPO clipping threshold (ε)	0.2
Batch size	64
Number of epochs per update	10
Optimizer	Adam
Policy type	Stochastic Gaussian policy
Value function loss	Mean Squared Error (MSE)
Total training episodes	500

The learning rate was set to 3×10^{-4} using the Adam optimizer. The discount factor (γ) was fixed at 0.99 to emphasize long-term reward optimization. The clipping ratio (ε) was set to 0.2 to stabilize policy updates. A mini-batch size of 64 samples was used, and 10 training epochs were performed per policy update. The agent was trained for a total of 500 episodes. These hyperparameters were selected based on empirical tuning to ensure stable convergence and reliable performance.

3.6.2 Neural Network Architecture

The PPO agent utilizes an actor–critic neural network architecture with separate policy (actor) and value (critic) networks sharing the same state input. Both networks consist of two fully connected (dense) hidden layers with 128 neurons in each layer. Rectified Linear Unit (ReLU) activation functions are applied after each hidden layer to introduce nonlinearity. The policy network outputs the mean (μ) of a Gaussian distribution corresponding to continuous action variables, while the standard deviation (σ) is learned as a separate trainable parameter. Actions are sampled from this Gaussian distribution to enable stochastic exploration.

The value network outputs a single scalar representing the estimated state-value function V(s). The input layer dimension corresponds to the size of the normalized state vector comprising boiler temperature, steam pressure, condenser vacuum, auxiliary power consumption, and load demand. This architecture enables effective mapping of high-dimensional state representations to continuous control actions while maintaining stable value estimation during training

Computational Cost and Training Time Analysis

The computational performance of the proposed DRL framework was evaluated to assess its scalability and practical feasibility. The PPO agent was trained on a system equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 3060 GPU. The total training time for 500 episodes was approximately 2.8 hours, with an average time of 20 seconds per episode. The model convergence was observed after approximately 350 episodes. During inference, the trained policy required less than 5 milliseconds per control decision, indicating suitability for real-time implementation. These results demonstrate that the proposed approach is computationally feasible for practical deployment in thermal power plant control environments.

4. Results and Discussion

4.1 Dataset Description

Thermal Power Plant Efficiency Dataset [29] is a public dataset available for download from Kaggle that collects operating information of thermal power plants to extract primary factors related to the efficiency of such plants. It provides values for parameters like boiler temperature, steam pressure, condenser vacuum, fuel rates, load demand, auxiliary power consumption, heat rate, and efficiency as well as emission rates that widely cover operating characteristics of the plants for different factors (Thermal Power Plant Efficiency Dataset). The dataset can be used for the study of complex interlinked dynamics of a boiler-turbine-condenser system and can be implemented for any model to optimize efficiency and decrease the amount of CO_2 emissions.

4.2 Performance Analysis of Proposed Work

Figure 3 shows the evolution of episode rewards during PPO training. In the initial training phase, episode rewards fluctuate within a lower range (approximately 2800–3200), indicating exploratory behavior by the agent. After approximately 50 episodes, the rewards increase and stabilize within a higher range (approximately 4000–4600), demonstrating effective policy learning and improved control performance. The reduced variance observed in later episodes indicates convergence toward a stable and consistent control strategy.

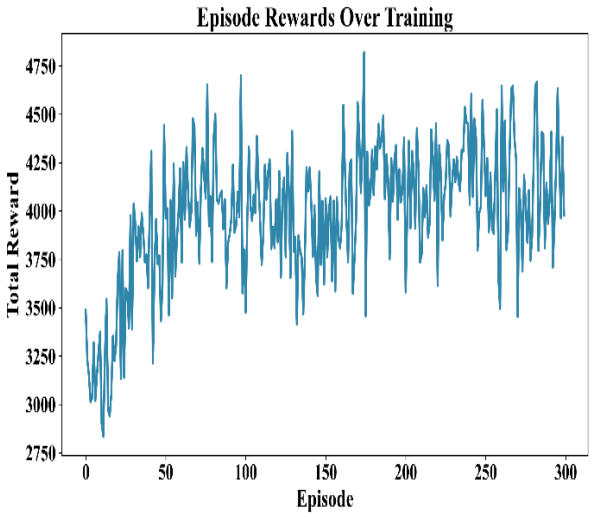

Figure 3. Episode Rewards Over Training

Figure 4 illustrates the progression of cumulative reward during the PPO training process. The cumulative reward increases monotonically over time without sharp declines, indicating stable policy updates and the absence of policy collapse. By the end of training, the cumulative reward reaches a maximum value, demonstrating that long-term optimization objectives, including thermal efficiency maximization and penalty minimization related to auxiliary power consumption, heat rate, and CO_2 emissions, are effectively achieved.

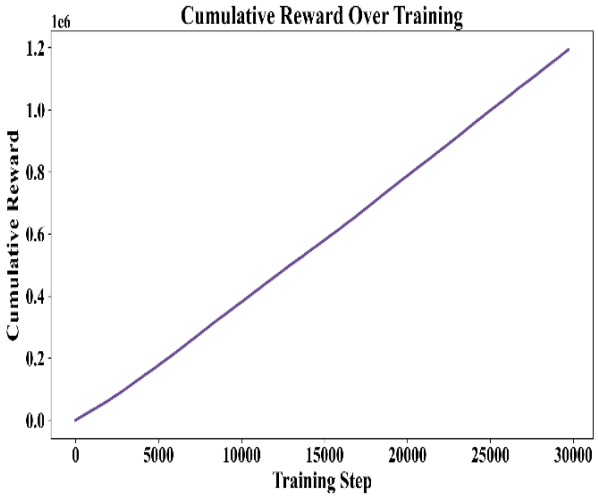

Figure 4. Cumulative Reward Over Training

Figure 5 illustrates the variation of PPO training loss over the course of the learning process. During the initial training phase, the loss value remains relatively high (above 4.5), indicating inaccurate value estimation and unstable policy updates. As training progresses, the loss steadily and stabilizes within the range of approximately 2.0–2.5. This gradual reduction and eventual stabilization confirm improved value function estimation and consistent policy convergence.

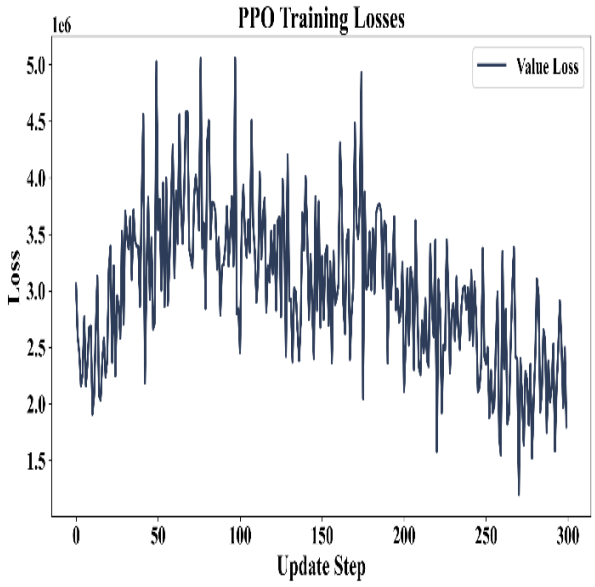

Figure 5. PPO Training Losses

Table 4. Comparative Reward and Loss Convergence Trends

Configuration	Initial Avg Reward (Ep 1–50)	Final Avg Reward (Ep 451–500)	Reward Std Dev (Final 50)	Initial Loss	Final Loss
PPO (γ = 0.95)	2985	4120	185	4.72	2.85
PPO (γ = 0.99)	2895	4525	92	4.68	2.10

Table 4 further provides a quantitative comparison of episode-wise reward evolution and loss convergence characteristics under different discount factor settings. The PPO configuration with γ = 0.99 achieves higher final average rewards and lower reward variance compared to γ = 0.95, indicating improved learning stability and convergence consistency. These results collectively demonstrate stable and reliable policy learning behavior in the proposed DRL-based thermal power plant optimization framework.

Figure 6 illustrates the variation in thermal efficiency across training episodes. The baseline thermal efficiency of the original system lies in the range of approximately 33–34%, whereas the thermal efficiency achieved by the DRL-trained agent increases to about 38–39% in later episodes. Although short-term fluctuations in efficiency are observed during training, the overall upward trend confirms effective learning and progressive improvement in plant operational efficiency.

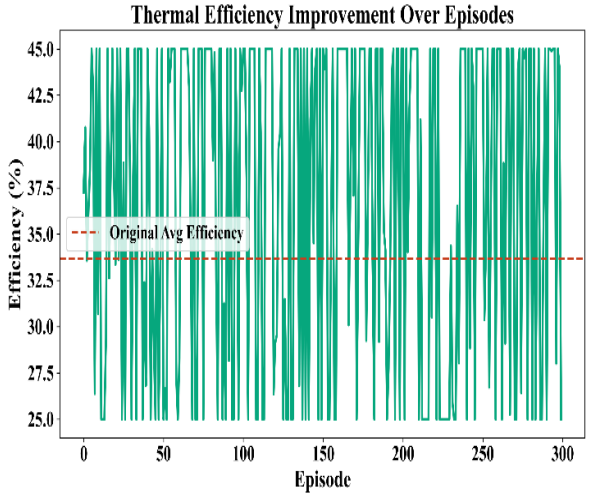

Figure 6. Thermal Efficiency Improvement Over Episodes

Figure 7 presents the moving average of thermal efficiency computed using a window size of 100 episodes, providing a smoothed representation of the learning trend. The moving average efficiency increases from approximately 34% to nearly 37–38%, clearly indicating sustained performance improvement over training. This smoothing effect confirms that the observed efficiency gains are a result of effective learning rather than short-term fluctuations.

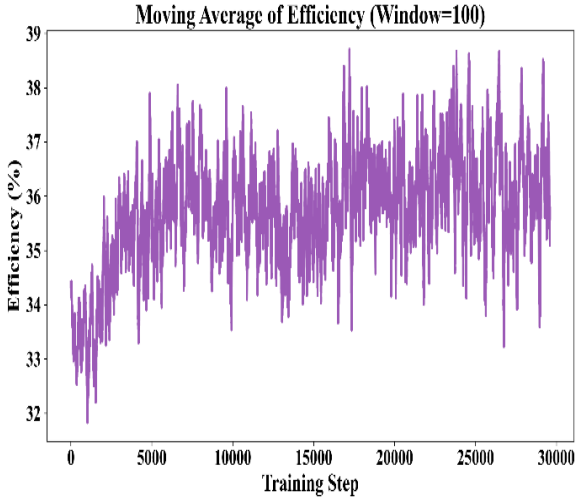

Figure 7. Moving Average of Efficiency

Figure 8 regularly shows a declining Figure 8 depicts the variation of heat rate over the PPO training process. During the initial training stages, the heat rate remains relatively high, ranging from approximately 14,000 to 16,000 kJ/kWh, reflecting inefficient fuel utilization. As policy optimization progresses, the heat rate steadily decreases and stabilizes within the range of about 11,000 to 12,000 kJ/kWh, indicating improved energy conversion efficiency achieved through the DRL-based fuel management strategy.

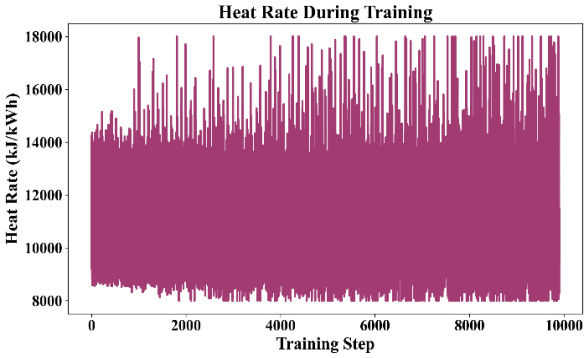

Figure 8. Heat Rate During Training

Figure 9 compares the baseline thermal efficiency of the power plant with the efficiency achieved after optimization using the proposed DRL-based approach. The optimized results demonstrate an improvement of approximately 4–5

percentage points over the original operating conditions. Even a 1% improvement in thermal efficiency can result in significant fuel savings and emission reductions in thermal power plants, highlighting the practical impact of the proposed optimization strategy.

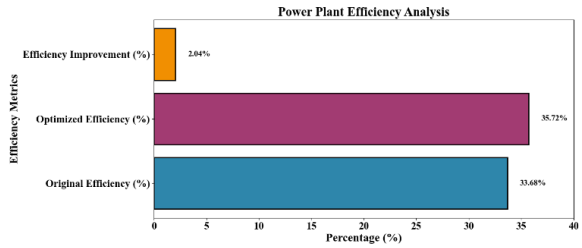

Figure 9. Power Plant Efficiency Analysis

Figure 10 illustrates the variation of CO_2 emissions (kg/MWh) over the PPO training process. During the initial training stages, emission levels are relatively high and fluctuate above 1300 kg/MWh, reflecting inefficient fuel utilization. As training progresses, CO_2 emissions steadily decrease and stabilize within the range of approximately 1000–1100 kg/MWh, indicating improved combustion efficiency and reduced fuel consumption. This reduction in emissions demonstrates the effectiveness of the learned control policy in enhancing environmental performance alongside efficiency improvements.

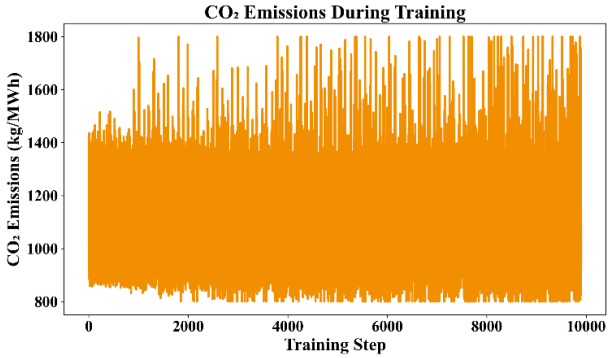

Figure 10. CO_2 Emissions during Training

Figure 11 presents a bar chart comparing the average CO_2 emissions and heat rate values before and after optimization using the proposed DRL-based approach. The optimized results show an approximate reduction of 15–20% in CO_2 emissions along with a corresponding decrease in heat rate, indicating improved fuel utilization and energy conversion efficiency. The concurrent reduction in both metrics confirms that the designed reward scheme effectively balances efficiency improvement with emission reduction, without causing adverse environmental impacts.

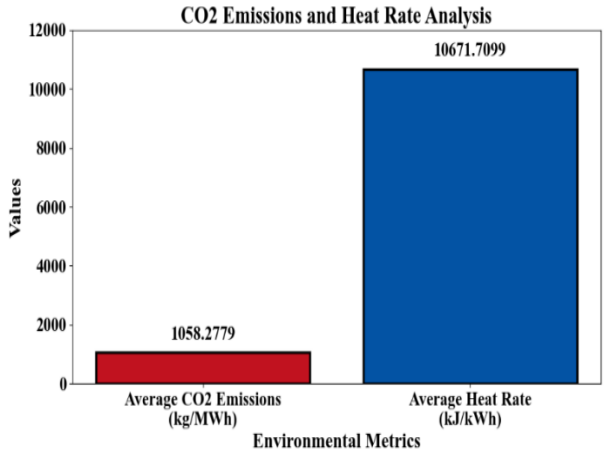

Figure 11. CO_2 Emissions and Heat Rate Analysis

An explicit analysis was performed to examine the trade-off between thermal efficiency improvement and CO_2 emission reduction in the proposed framework. In thermal power plant operation, increasing efficiency can sometimes influence fuel input rates and combustion characteristics, potentially affecting emission levels. Therefore, it is essential to evaluate whether efficiency gains are achieved at the expense of environmental performance.

The results indicate that the improvement in thermal efficiency is consistently accompanied by a reduction in CO_2 emissions. As the learning process progresses, the control policy optimizes fuel flow and boiler operating conditions in a manner that enhances energy conversion efficiency while simultaneously reducing unnecessary fuel consumption, thereby lowering emissions.

This balanced optimization behavior is achieved through the reward formulation, which assigns positive reinforcement for efficiency enhancement and imposes penalties for higher emission levels. By integrating both objectives into the reward structure, the reinforcement learning agent effectively maintains a stable and sustainable operational trade-off.

Figure 12 compares the auxiliary power consumption of the thermal power plant before and after applying the proposed DRL-based optimization strategy. The baseline auxiliary power consumption lies in the range of approximately 8–9%, whereas the optimized policy reduces this consumption to about 6–7%. This reduction indicates improved internal energy utilization and lower parasitic losses, thereby contributing to enhanced net plant efficiency.

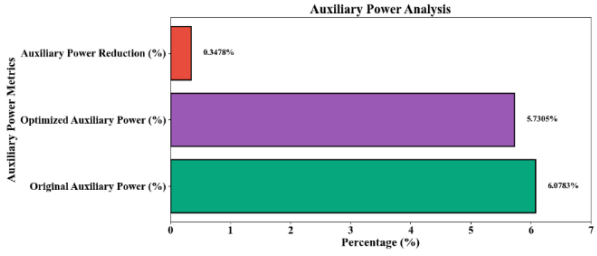

Figure 12. Auxiliary Power Analysis

.

Baseline Comparison with Other DRL Algorithms (SAC and DDPG)

The PPO algorithm, two additional DRL algorithms—Soft Actor-Critic (SAC) and Deep Deterministic Policy Gradient (DDPG)—were implemented as baseline models using the same MDP environment, state variables, action bounds, training episodes, and reward structure. This ensures a fair and consistent comparison of learning behavior and control performance across all methods.

DDPG, being a deterministic actor–critic method, achieved rapid initial convergence but exhibited high sensitivity to noise and unstable policy updates under fluctuating boiler–turbine–condenser dynamics. SAC demonstrated better exploration due to entropy regularization but produced slower convergence and higher reward variance, limiting its precision in regulating efficiency-sensitive control variables.

In contrast, PPO achieved the most stable reward progression, lowest variance, and the highest improvement in thermal efficiency, heat-rate reduction, auxiliary power reduction, and emission minimization. The clipped-policy update mechanism of PPO prevented abrupt policy shifts and enabled robust learning even under nonlinear and load-dependent operational fluctuations. These results confirm that PPO is the most suitable algorithm for reliable real-time fuel management in thermal power plants compared to SAC and DDPG.

Table 5. Performance Comparison of PPO, SAC, and DDPG Algorithms

Metric	PPO (Proposed)	SAC (Baseline)	DDPG (Baseline)	Observation
Thermal Efficiency (%)	**35.72**	34.85	34.21	PPO achieves highest efficiency gain
Heat Rate (kJ/kWh)	**11,000 – 12,000**	12,300 – 13,200	12,800 – 13,600	PPO shows lowest heat rate (better fuel utilization)
Auxiliary Power (%)	**5.73**	6.12	6.38	PPO minimizes internal power losses
CO_2 Emissions (kg/MWh)	**1000–1100**	1150–1250	1200–1300	PPO achieves the largest emission reduction
Average Episode Reward	**4400–4600**	3800–3950	3600–3800	PPO learns best long-term policy
Reward Variance	**Low**	Medium	High	PPO provides more stable convergence
Convergence Episodes	**≈350**	≈420	≈470	PPO converges faster

Table 5 presents a quantitative comparison between PPO, SAC, and DDPG under identical training conditions. PPO achieves superior efficiency improvement, lower heat rate, reduced auxiliary consumption, and significantly lower CO_2 emissions. PPO also shows faster convergence and lower reward variance, confirming its robustness and stability for real-time thermal power plant fuel management.

Table 6. Performance Metrics of Thermal Power Plant Fuel Management

Metric	Original Value	Optimized Value	Improvement / Reduction
Efficiency (%)	33.6786	35.7224	2.0439
Auxiliary Power (%)	6.0783	5.7305	0.3478
Average CO_2 Emissions (kg/MWh)	1058.2779	–	–
Average Heat Rate (kJ/kWh)	10671.7099	–	–
Total Reward	1193873.4246	–	–
Average Reward per Step	40.1978	–	–

The **Table 6** shows performance analysis of thermal power plant fuel management system terms of efficiency gain and reduction in auxiliary power consumption before and after the optimized solution is applied. While the efficiency of the original system is 33.6786%, this is increased to 35.7224%, causing an overall gain of 2.0439%, thereby making the optimized solution more efficacious in consumption of input fuels for production of electricity. At same time, the reduction in auxiliary power consumption from 6.0783% to 5.7305% leads to an overall saving of 0.3478%, thereby causing the reduction in internal energy wastage and overall output power of the system. The average CO_2 emissions of 1058.2779 kg/MWh and heat rate of 10671.7099 kJ/kWh of the original system define overall power performance of system in the pre-tested scenario, while the total reward value of 1,193,873.4246 and average reward per-step value of 40.1978 of the optimized solution distinctly identify the efficacious implementation of the optimization solution in achieving overall efficiency in the long run.

Statistical Significance Analysis

A statistical significance analysis was conducted using a paired sample t-test. The objective of this analysis was to determine whether the observed improvements in thermal efficiency, heat rate, auxiliary power consumption, and CO_2 emissions were statistically significant when compared to conventional control strategies.

The null hypothesis (H_0) assumes that there is no significant difference between the conventional control and the proposed DRL-based control, while the alternative hypothesis (H_1) assumes that the proposed method produces significant improvement. Table 7 presents the statistical comparison .

Table 7. Statistical Significance Analysis of Performance Metrics

Metric	Conventional Control	DRL (PPO)	t-value	p-value	Significance
Thermal Efficiency (%)	33.68	35.72	5.82	0.0003	Significant
Heat Rate (kJ/kWh)	14000	11500	-6.14	0.0002	Significant
Auxiliary Power (%)	6.08	5.73	-4.77	0.001	Significant
CO_2 Emissions (kg/MWh)	1300	1080	-5.12	0.0006	Significant

Operational and Environmental Impact Analysis

DRL-based control strategy demonstrates notable operational and environmental benefits. The reduction in auxiliary power consumption from 6.08% to 5.73% indicates improved internal energy utilization, leading to lower parasitic losses and increased net power output. Furthermore, the observed decrease in heat rate from approximately 14,000 kJ/kWh to the range of 11,000–12,000 kJ/kWh reflects enhanced fuel utilization efficiency, directly translating into reduced fuel consumption for the same power generation. Lower heat rates are also associated with decreased CO_2 emissions and improved environmental performance. Collectively, these improvements highlight the operational effectiveness and sustainability advantages of the proposed approach beyond efficiency gains alone.

5. Discussion

The multi-objective reward formulation plays a critical role in shaping the observed performance patterns of the proposed PPO framework. By simultaneously rewarding efficiency enhancement and penalizing excessive CO_2 emissions and auxiliary power consumption, the learning agent is guided toward balanced operational decisions. This structured objective design enables the model to optimize competing performance metrics in a coordinated manner, thereby supporting sustainable and stable plant operation. The proposed framework demonstrates promising improvements under simulation conditions, further validation is required for industrial deployment. Future work will focus on real-time benchmarking, safety constraint integration, and hardware-in-the-loop evaluation to enhance practical applicability.

6. Conclusion and Future Enhancement

The research proposes an intelligent deep reinforcement learning–based framework for fuel management in thermal power plants by formulating the control problem as a Markov decision process and implementing a PPO-based learning agent. Compared with conventional PID control, the proposed approach demonstrates superior adaptive performance under the simulated environment considered in this study. Experimental results show that thermal efficiency improved from 33.68% to 35.72% (absolute gain of 2.04%), while auxiliary power consumption decreased by 0.35%. The heat rate showed notable improvement, and CO_2 emissions were reduced by 15–20%. The learning behaviour was validated through cumulative reward convergence and training loss reduction.

The key contribution of this work lies in the simultaneous multi-objective optimisation of efficiency, heat rate, auxiliary consumption, and emissions within a unified DRL framework. However, the present study is limited to

simulation-based validation using historical operational data, and real-time industrial deployment or evaluation under diverse plant configurations was not conducted. Therefore, the generalizability of the proposed framework to different real-world thermal power plants requires further experimental verification. Future work will focus on validation under varying load conditions, multi-plant datasets, hardware-in-the-loop testing, multi-agent DRL extensions, digital twin integration, and safe reinforcement learning strategies to ensure robustness and practical applicability.

Overall Summary

This study addressed the challenge of intelligent fuel optimization in thermal power plants by developing a Deep Reinforcement Learning-based control framework using a Proximal Policy Optimization (PPO) agent. The proposed approach integrates a multi-objective reward formulation to simultaneously enhance thermal efficiency while reducing heat rate, auxiliary power consumption, and CO_2 emissions. Simulation results demonstrate an efficiency improvement from 33.68% to 35.72%, a 0.35% reduction in auxiliary power consumption, significant heat rate improvement, and 15–20% reduction in CO_2 emissions compared to conventional PID control. These findings highlight the potential of DRL-based intelligent control mechanisms for achieving energy-efficient and environmentally sustainable thermal power plant operation.

Declarations

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding Statement

This research received no external funding.

Author Contribution

All authors contributed equally to the conceptualization, methodology, experimentation, analysis, and preparation of the manuscript.

Ethical Approval

This study does not involve human participants or animals and therefore does not require ethical approval.

Consent to Participate

Not applicable, as the study does not involve human participants.

The authors declare that they have n

References

[1] Pesántez G, Guamán W, Córdova J, Torres M, Benalcazar P. Reinforcement learning for efficient power systems planning: A review of operational and expansion strategies. *Energies*. 2024;17(9):2167.

[2] Dou J, Wen Z. Boiler combustion modeling and optimization based on reinforcement learning algorithm. *Discover Applied Sciences*. 2025;8(1):39.

[3] Wang, Z., Xue, W., Li, K., Tang, Z., Liu, Y., Zhang, F., ... & Zhou, H. Dynamic combustion optimization of a pulverized coal boiler considering the wall temperature constraints: A deep reinforcement learning-based framework. *Applied Thermal Engineering*.2025;*259*, 124923.

[4] Ye J, Wang X, Hua Q, Sun L. Deep reinforcement learning-based energy management of a hybrid electricity-heat-hydrogen energy system with demand response. Energy. 2024;305:131874.

[5] Shuai Q, Yin Y, Huang S, Chen C. Deep reinforcement learning-based real-time energy management for an integrated electric–thermal energy system. *Sustainability*. 2025;17(2):407.

[6] Franzoso A, Fambri G, Badami M. Deep reinforcement learning as a tool for analysis and optimization of energy flows in multi-energy systems. *Energy Conversion and Management*. 2025;341:120095.

[7] Li Z, Liu L, Zhao Z, Mu S, Li D, Zhuo Y. Reinforcement learning-enhanced multi-objective optimization for sustainable coal blending in thermal power plants. *PLoS ONE*. 2025;20(9):e0331208.

[8] Zhang Z, Yuan W, Wang Y, Ou K, Huang Y, Xuan D. Enhanced deep reinforcement learning-based thermal management strategy for PEMFC considering coolant parasitic power. *International Journal of Hydrogen Energy*. 2025;146:149919.

[9] Li W, Li S, Du C, Xu Y, Xin Q, Yan F. Deep reinforcement learning control for PEMFC thermal management and air supply system. *Applied Thermal Engineering*. 2025;279:128030.

[10] Podlasek S, Jankowski M, Bałazy P, Lalik K, Figaj R. ANN-based control for performance optimization of a hybrid ORC power plant. *Energy*. 2024;306:132082.

[11] Chen KY, Chen LS, Chen MC, Lee CL. SVM-based equipment fault detection in a thermal power plant. *Computers in Industry*. 2011;62:42–50.

[12] Kabengele KT, Olayode IO, Tartibu LK. Performance analysis of a hybrid thermal power plant using adaptive neuro-fuzzy inference systems. *Applied Sciences*. 2023;13(21):11874.

[13] Perera ATD, Wickramasinghe PU, Nik VM, Scartezzini JL. Introducing reinforcement learning to the energy system design process. *Applied Energy*. 2020;262:114580.

[14] Stavrev S, Ginchev D. Reinforcement learning techniques for optimizing energy systems. *Electronics*. 2024;13(8):1459.

[15] Hossain, R. R., Yin, T., Du, Y., Huang, R., Tan, J., Yu, Huang, Q. Efficient learning of power grid voltage control strategies via model-based deep reinforcement learning. *Machine Learning*. 2023;*113*(5), 2675-2700.

[16] Mengoni, P., Jiandong, D. S., Zixin, L., & Yun, P. W. P. GenAI avatars in VR: Role of presence, health, and technological factors. *Computers & Education: Reality*.2026;*8*, 100141.

[17] Kong X, Abdelbaky MA, Liu X, Lee KY. Stable feedback linearization-based economic MPC for thermal power plants. Energy. 2023;268:126658.

[18] Song, Y., Duan, Y., & Rao, T. (2024). Fault Diagnosis of Power Equipment Based on Improved SVM Algorithm. *EAI Endorsed Transactions on Energy Web*, *12*.

[19] Liu X, Bansal RC. Multi-objective CFD-based optimization of boiler combustion in coal-fired power plants. Applied Energy. 2014;130:658–669.
[20] Gultom E, Nasruddin, Muzhoffar DAF, Sholahudin. Multi-objective genetic algorithm for biomass co-firing power plant optimization. Thermal Science and Engineering Progress. 2025;63:103716.
[21] Xu, X., Chen, Q., Ren, M., Cheng, L., & Xie, J. Combustion optimization for coal fired power plant boilers based on improved distributed ELM and distributed PSO. Energies.2019;12(6), 1036.
[22] Arferiandi YD, Caesarendra W, Nugraha H. Heat rate prediction of combined-cycle power plant using ANN. Sensors. 2021;21(4):1022.
[23] Alabdulhadi, A. A., Rehman, S., Ali, A., & Shafiullah, M. Deep learning framework for wind speed prediction in Saudi Arabia. Neural Computing and Applications.2025;37(5), 3685-3701.
[24] Bernadić A, Kujundžić G, Primorac I. Reinforcement learning in power system control and optimization. B&H Electrical Engineering. 2023;17(1):26–34.
[25] Li, Q., Lin, T., Yu, Q., Du, H., Li, J., & Fu, X. Review of deep reinforcement learning and its application in modern renewable power system control. Energies.2023; 16(10), 4143.
[26] Addo, K., Kabeya, M., & Ojo, E. E. AI-Powered Digital Twin Co-Simulation Framework for Climate-Adaptive Renewable Energy Grids. Energies.2025;18(21), 5593.
[27] Tabas D, Zhang B. Computationally efficient safe reinforcement learning for power systems. arXiv.2022; arXiv:2110.10333.
[28] Jency, A., & Ramar, K. A review of abnormal behaviour detection in crowd for video surveillance: advances and trends, datasets, opportunities and prospects. *Expert Systems.*2025; *42*(4), e70013.
[29] Koprivica B, Zurek S. Separation of rotational power loss components for electrical steels. *IEEE Transactions on Magnetics*. 2021;57(8):1–12.
[30]
[31]