Optimizing In-Home EV Charging: Real-Time Optimization with Time Series Transformers and Policy-based DRL

Shivendu Mishra [1,3*], Anurag Choubey [1,4], Harshit Dhankhar [2], Sri Vaibhav Devarasetty [1], Rajiv Misra [1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Patna, 801106, Bihar , India.
[2]Department of Mathematics, Indian Institute of Technology, Patna, 801106, Bihar, India.
[3]Department of Information Technology, Rajkiya Engineering College, Ambedkar Nagar, 224122, Uttar pradesh, India.
[4]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, 201310, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): shivendu_2021cs08@iitp.ac.in;
Contributing authors: anurag.pcs17@iitp.ac.in ;
harshit_2101mc20@iitp.ac.in; Vaibhavdevarasetty@gmail.com;
rajivm@iitp.ac.in;

Abstract

Managing EV charging is difficult due to limited battery capacity and unpredictable variables such as traffic, user behaviour, and fluctuating electricity prices. Researchers frequently use model-free approaches such as deep reinforcement learning (DRL) to address this. This paper presents a DRL-based Markov decision process (MDP) for optimising in-home EV charging. Forecasting, pricing, and scheduling are all part of EV charging management, with pricing and forecasts heavily influencing scheduling models. While LSTM and GRU are commonly used in AI-based forecasting, attention-based mechanisms like transformers are better at capturing long-term dependencies. This paper proposes a time series transformer-based network with DRL to improve in-home EV charging. Unlike previous works, which rely solely on the past 24 hours of price data. Unlike previous works, which rely solely on the past 24 hours of price data, our model investigates two additional time-frames: prices from the same hour over the

previous 24 days, and prices from the same hour on the same weekday over the previous 24 weeks, to automatically schedule actions that meet user needs while minimising charging costs. Extensive simulations on the above time-frames using three decision-making models—Deep Q-Networks (DQN), Deep Deterministic Policy Gradient (DDPG), and Proximal Policy Optimisation (PPO) (in both continuous and discrete forms)—integrated with Autoformer, Informer, and PatchTST-based feature extraction methods demonstrate the effectiveness of our transformer-based model. Our approach achieved full user satisfaction and significantly reduced EV charging costs compared to previous models. Our method lowers charging costs by 125.74 % in the continuous action space and 140.66 % in the discrete action space compared to previous studies.

Keywords: Deep reinforcement learning, Electric vehicles, EV charging and scheduling, Home EV Charging, Markov decision process, Proximal Policy Optimisation, Transformer.

1 Introduction

Electric vehicles (EVs) have become more popular as a green alternative to traditional gas-powered cars. But, if many EVs are charged simultaneously, it can create a high demand for electricity in one area. This can lead to problems like power losses, voltage swings, and grid overloads. To help with this, many utility companies have started using real-time power rates to encourage people to charge their EVs during off-peak times [1, 2]. Additionally, EVs can make money by using different modes such as vehicle-to-vehicle (V2V), mobile charging to vehicle (MC2V), wireless power transfer (WPT), and vehicle-to-grid (V2G). Optimizing the charging and discharging schedules of an EV can help reduce costs for the owner. However, managing these schedules is challenging because of unpredictable factors like traffic, when users arrive and leave, how much energy is used, commuting patterns, and changes in electricity prices [3–6].

Different methods have been used to manage EV charging and discharging, including various programming techniques, day-ahead scheduling, model-based strategies, and model-free methods like deep reinforcement learning (DRL). DRL has become especially popular recently, with many researchers successfully using it to solve specific EV charging problems. Managing EV charging and discharging involves three key elements: dynamic pricing, forecasting, and scheduling. How well these elements are managed greatly affects the overall performance. Forecasting models predict charge availability, electricity prices, driving patterns, and EV load demand, which help create effective schedules and pricing strategies. Many artificial intelligence-based models work as forecast models for EV charging and discharging. The most common ones use supervised learning techniques such as decision trees, k-nearest neighbours, random forests, linear regression, support vector machines, and methods based on gated recurrent units (GRUs), CNN, and RNN. Hybrid and ensemble approaches are also used. Among these, LSTM and GRU are the most widely used due to their ability to handle nonlinear and long-term dependencies [7].

However, attention-based mechanisms [8], unlike LSTM and GRU, can learn long-range dependencies in data sequences much better for several important reasons [9]. First, RNN-based networks like GRU and LSTM struggle with long-term dependencies and take much time to train. Second, feeding raw data directly into neural networks often adds noise to the training data. Third, model prediction and data denoising are treated as separate tasks, ignoring their correlation. Fourth, in GRU or LSTM models, the path lengths increase linearly as the distance between two points grows, whereas, in attention-based mechanisms, the path lengths stay the same no matter the distance. Lastly, attention-based mechanisms allow for more parallelization during training, which is especially useful for long sequences and dealing with memory constraints.

Our previous work [10] pioneered the use of an attention-based architecture for EV scheduling, building on the exploration of attention-based mechanisms and reinforcement learning (RL) in EV charging and discharging management. This work presents a DRL-based solution, framed as a Markov Decision Process (MDP), incorporating a novel multi-head attention-based model called "MHA-BiMGRU" to recognise patterns in historical electricity price data. This model employs deep reinforcement learning to make charging or discharging decisions based on incoming data on future electricity prices and charge status. The goal is to maximise in-home EV charging by accounting for fluctuating electricity prices and unpredictable commuting behaviour, ultimately reducing costs and increasing user satisfaction by leveraging price variations.

In the current work, unlike our previous approach that relied solely on the past 24 hours of price data to predict the current price, we have implemented three distinct cases, as outlined below:

- **Case-1 (C_1):** We used the electricity prices from the same hour over the past 24 days to predict the price for that same hour on the next day. For example, to predict the price at 10 AM, we utilized the prices recorded at 10 AM over the previous 24 days.
- **Case-2 (C_2):** We used the electricity prices from the same day at the same hour on a weekday. For example, to predict the price of Monday at 10 AM, we utilized the price records of the previous 24 Mondays at the same hour, i.e. 10 AM.
- **Case-3 (C_3):** We used the past 24 hours of price data to predict the current price.

Furthermore, to evaluate the performance of this work, we have used three decision models, DQN, DDPG, and PPO (continuous and discrete situations), along with autoformer, informer, and PatchTST-based feature extraction. The comparative analysis shows that the proposed model outperformed our previous models presented in [10] and the models used in related works [11, 12] in optimizing in-home EV charging scheduling management.

1.1 Motivation

As mentioned earlier, the significant advantages of attention-based mechanisms have opened up new opportunities for their effective integration into forecasting temporal information from electricity price sequences. Additionally, electricity prices depend on users' charging habits; for instance, the price at any given hour is closely related to the prices at the same hour on previous days or the same day at the same hours in past

weeks. This development has prompted our current study to explore the applicability of these mechanisms in predicting electric vehicle (EV) charging and discharging schedules. Our focus is optimizing in-home EV charging to minimize costs and enhance user satisfaction by leveraging price fluctuations. We have developed a novel neural network model incorporating a transformer-based architecture to achieve this. This model has proven highly effective and efficient in extracting pertinent information from electricity price sequences, enabling better predictions and optimizations in EV charging and discharging schedules.

1.2 Our contributions

This paper presents several key contributions, including:

i. The EV charging control model is structured as a Markov decision process (MDP), where the environment model accounts for dynamic energy prices and fluctuating charging demands.

ii. We have developed a transformer-based model that integrates Autoformer, Informer, and PatchTST architectures to extract historical energy price information across three distinct cases (C_1, C_2, and C_3). Specifically, C_1 utilizes prices from the same hour over the past 24 days, C_2 analyzes prices from the same hour on the weekday over the last 24 weeks, and C_3 leverages data from the previous 24 hours. This approach enables informed decision-making regarding real-time charging and discharging actions.

iii. To underscore the importance of feature extraction capabilities in our proposed approach and affirm the transformer-based model's efficacy, we conducted a comparative evaluation involving different variations of RNN-based feature extraction methods and recent attention-based models.

iv. Simulation results demonstrate that the proposed model outperforms recent feature extractors in terms of reduced charging costs and improved user satisfaction when evaluated with deep reinforcement learning benchmarks like DQN, DDPG, and PPO. Specifically, it achieves full user satisfaction and reduces charging costs by 125.74 % in the continuous space and 140.66 % in the discrete space, providing significant practical benefits for real-time EV charging management.

Table 1 displays our scheme's relevant notations and definitions.

1.3 Layout of the paper

The structure of this paper is organized as follows: In Section II, we provide a concise overview of prior research related to the EV charging scheduling problem. Section III transforms the single-EV charging and discharging scenario into an MDP and introduces the optimization objective of our proposed approach. Section IV provides a comprehensive overview of fundamental concepts, including deep reinforcement learning and transformer models. Section V details the proposed model. Section VI presents a performance analysis of the proposed model, including experimental details to illustrate its effectiveness. Lastly, in Section VII, we draw our conclusions.

Table 1: **Notations with their definitions**.

Notations	Definitions
RL, DRL, EV	Reinforcement learning, deep reinforcement learning, electric vehicle, respectively
$LSTM, JANET, MHA-BiMGRU$	Long short-term memory network, just another network, and multi-head attention-based bidirectional modified gated recurrent units, respectively
t, t^a, t^d	Current time, EV arrival time and departure time, respectively
s^t, s^{t+1}, a^t, r^t	Current state, next state, action, reward, respectively
$P^t, E^t, \Delta t$	Price at time t, EV state of energy at time t, the time interval between the departure time t^d and the current time t, respectively
S, R, A, T	State space, reward function, action space, and state transition function, respectively
P, E^{max}, E^{min}	The maximum charging/discharging action, the maximum and minimum capacity of EV, respectively
BSP, MDP, FAM	Bidirectional smart plug, Markov decision process, and feature analysis model, respectively
P^{mean}, α, β	Mean price of past 24 hours at time t and the real-valued coefficients, respectively
γ, CR^t, Q, K, V	Discount factor, cumulative reward, query, key, and value, respectively
$DNN, MGRU, DM,$	Deep neural network, modified GRU, and decision model, respectively
$MH, EL1, EL2, dsat$	Multi-Head, first encoder layer, second encoder layer, and degree of dissatisfaction, respectively
C_1, C_2, C_3	Case-1, Case-2, and Case-3, respectively

2 Related work

Researchers have used different programming techniques, like dynamic, non-linear, and linear programming, to optimize EV charging and discharging schedules [13–16]. However, these methods have some drawbacks. They can be time-consuming, may not scale well, and often require multiple tries to find the best solution. Because of the need for quick, real-time optimization to cut EV charging costs, relying only on these programming methods might not be practical [17]. To address this, various day-ahead scheduling methods have been suggested [18–24]. These methods aim to reduce the uncertainty in EV charging by using robust or stochastic optimization a day in advance. While day-ahead scheduling can help manage EV charging in somewhat uncertain situations, it is less suitable for real-time and large-scale EV charging problems with high demand and unpredictable prices.

Recently, model-free methods have made significant progress in complex decision-making applications. Furthermore, using reinforcement learning (RL), model-free methods can develop effective control policies without knowing the system beforehand, making them better than traditional model-based methods. In model-free methods,

the action-value function is a crucial part of evaluating the success of charging schedules. The main difference between these methods is how accurately they estimate the best action-value function [25–32].

The action-value function can be estimated using a Q-table-based approach when electricity prices and charging actions are discretized. However, this method has significant limitations. It can only handle a limited number of distinct states and actions, making it less effective in complex or continuously varying environments. The performance is also highly dependent on how states and actions are discretized. Poor discretization can lead to losing important information and reduced decision-making accuracy. To address these limitations, the authors of [28] used linear basis functions to approximate the action-value function. However, this method struggles with the non-linear nature of real-world electricity prices and commuting patterns. In [29], a non-linear kernel averaging regression operator was used to fit the action-value function accurately. Still, the performance depends heavily on the choice and parameters of the kernel function. Overall, these approximation methods still fall short in handling real-world situations effectively.

Researchers in [11, 12, 33–44] have used neural networks as universal approximators to estimate action-value matrices in reinforcement learning (RL). Deep neural networks have recently shown great potential in learning complex mappings from high-dimensional data. Many researchers have successfully applied deep reinforcement learning (DRL) to specific EV charging control problems, achieving excellent results. Recently, numerous studies [45–50] have also incorporated DRL techniques for various EV charging scenarios. Authors in [45] proposed a multi-agent deep deterministic policy gradient model for charging EVs at multiple stations and a time-of-use (TOU) dynamic pricing algorithm considering random EV arrivals and power outage constraints. Authors in [46] presented a deep reinforcement learning approach for fast-charging EVs at hubs, incorporating battery storage systems and home charging hubs. They addressed factors like random EV demand and power constraints as a mixed-integer linear programming problem. Authors in [47] proposed a household EV charging model at a charging station for grid filling and peak load shaving, utilizing a proximal policy optimization algorithm with limited knowledge of charging demand. Authors in [48] introduced a data-driven approach considering EV travel data and proposed a distributed proximal policy optimization (DPPO) algorithm with multiple actors. Authors in [49] offered a PV-powered self-sustainable household EV charging model constrained by local consumption profiles, employing an N-step DQN that considers historical smart-meter data to estimate local rewards. In [50], a multi-agent DNN model was proposed for creating a dynamic pricing scheme for off-peak charging, incorporating renewable energy generation and V2G and G2V provisioning. While these works explore different charging strategies at stations, there's a need for further research on home charging based on hourly TOU pricing data that enables both G2V and v2G to optimize EV charging costs and maximize discharging revenue.

EV charging management schemes [11, 12, 39] have utilized popular recurrent neural network (RNN)-based models as feature extractors. In [11], Long Short-Term Memory (LSTM) were introduced to extract meaningful features from price signals, combined with the Deep Q-Networks (DQN) learning algorithm for individual EV

charging scheduling problems. However, DQN is mainly suited for finite, discrete action spaces, posing a challenge for continuous charging rates. To address this, [39] proposed using the Deep Deterministic Policy Gradient (DDPG [51]) algorithm alongside an LSTM feature extractor, replacing the DQN. This adaptation enabled the management of continuous charging rates, which is crucial for realistic EV charging scenarios. Additionally, [12] suggested a deep reinforcement learning (DRL)-based approach for real-time optimization of EV charging management. This method combined "Just Another Network" (JANET) [52] and the DDPG algorithms. JANET was employed to extract and forecast valuable temporal information from electricity price sequences, enhancing the DDPG algorithm's performance. While these approaches leverage the strengths of DRL and advanced neural networks to make more effective real-time decisions in EV charging management, they suffer from drawbacks such as extended temporal dependencies and substantial training time expenditures.

To address the above challenge, our previous work [10] introduced an attention-based architecture for EV scheduling. We developed a DRL-based solution, structured as a Markov Decision Process (MDP), which incorporates a novel multi-head attention-based model named "MHA-BiMGRU" to detect patterns in historical electricity price data. Our innovative approach utilized the price data from the past 24 hours to forecast future electricity prices. However, upon further investigation, we found that future prices are more closely related to prices from the same hours on previous days or the same day at the same hour on weekdays. Therefore, in our current work, we restructured the data set based on the above investigation. We used a transformer model instead of the novel multi-head attention-based model as a feature extractor, which has proven more effective in experimental trials.

3 Problem formulation

We approach the real-time EV charging and discharging scheduling challenge from a driver/owner-centric viewpoint. In this scenario, the EV can connect to the grid at home and either draw power from it or feed power back into it. The bidirectional smart plug (BSP) is an intelligent charging device installed at the owner's home. Using our proposed transformer-based model, the BSP can manage hourly charging/discharging activities when the battery is connected. Our approach relies on accessing real-time battery state of charge (SOC) and energy price data across three distinct cases (C_1, C_2, and C_3). Specifically, C_1 leverages prices from the same hour over the past 24 days, C_2 uses prices from the same hour on the same weekday over the last 24 weeks, and C_3 using simply the past 24 hours of data. Additionally, on any given day, let's call it D, the EV's arrival time (AT) or plug-in time is denoted as t^a. The EV will depart (DT) at t^d on day $D+1$. The EV arrives at its residence on day D, and the episode concludes when the EV departs on day $D+1$. The main objective of our proposed method is to ensure that the EV charges during periods of low electricity prices and discharges during high electricity price intervals, ultimately leading to cost savings for the EV owner as well as maximizing user satisfaction for a maximum amount of energy level before departing home.

In the subsequent subsection, we formulate the real-time EV charging and discharging scheduling challenges as an MDP.

3.1 MDP formulation

The MDP (Markov Decision Process) provides a crucial mathematical framework for addressing sequential decision-making challenges [53]. In this context, the EV is considered as an agent that repeatedly observes the current state of the environment and responds accordingly. This work utilizes the four fundamental components (S, A, T, R) to formally define the MDP model for EV charging and discharging schedules. S represents the state space, encompassing all possible states. A denotes the action space containing feasible actions. T corresponds to the state transition function governing the system's evolution. R signifies the reward function, determining the immediate benefit or cost associated with each action taken.

The precise details of each element are expressed as follows:

i. **State:** The MDP's state at time t is represented as:

$$s^t = (E^t, \Delta t, P^{t-M}, \ldots, P^{t-1}) \tag{1}$$

where E^t signifies the State of Charge (SOC), representing the remaining battery energy in the EV. $\Delta t = t^d - t$ denotes the time difference between the current time t and the planned departure time t^d. A smaller Δt reflects a stronger user preference for immediate EV charging. $(P^{t-M}, \ldots, P^{t-1})$ corresponds to the electricity prices in the preceding M hours, where M equals 24 hours on previous days or the same day at the same hour on weekdays providing a comprehensive historical context of price trends.

ii. **Action:** The action a^t represents the power being charged or discharged given the state s^t. Let a^t be positive when the EV is charged and negative when it is discharged. The constraints on the charging and discharging power are defined as follows:

$$-P^m \leq a_t \leq P^m \tag{2}$$

where P^m denotes the maximum charging or discharging power the EV can perform at any given time step.

iii. **State transition function:** The state transition function can be expressed as follows:

$$T : s^t \times a^t \rightarrow s^{t+1} \tag{3}$$

In addition, to simulate the real-world scenario, we represent the dynamics of the EV battery as

$$E^{t+1} = E^t + a^t \tag{4}$$

iv. **Reward function:** The reinforcement the system receives right after action a^t is executed to move it from state s^t to s^{t+1}, known as the immediate reward, or r^t. In the next two cases, the reward is specified clearly:

(a) **When** $t \geq t^d$: In this context, the immediate reward is defined as:

$$r^t = -P^t.a^t - \alpha.(E^{max} - E^t) \tag{5}$$

(b) **When** $t^a < t < t^d$:

$$r^t = \begin{cases} -P^t.a^t - \alpha.(E^t - E^{max}), & E^t > E^{max} \\ -P^t.a^t - \alpha.(E^{min} - E^t), & E^t < E^{min} \\ -P^t.a^t + \beta.a^t.(P^{mean} - P^t), & E^{min} < E^t < E^{max} \end{cases} \tag{6}$$

In this context, E^{max} and E^{min} represent the maximum and minimum capacities of EVs, respectively. The timeframe when the EV is at home is denoted by $t^a < t < t^d$, with $t = t^d$ indicating the moment of departure. P^t represents the electricity price at time t, while P^{mean} is the mean price of the past 24 hours at time t. The terms α and β are real-valued coefficients. The charging cost at time step t is reflected in the reward as $P^t \cdot a_t$, which is positive when the EV charges. Conversely, it is negative if surplus energy is sold back to the grid using the BSP. This scenario is based on a net metering arrangement described in [11], where a bi-directional meter tracks electricity used for grid purchases and returns. Under this setup, the cost of buying electricity equals the revenue from selling it back to the grid.

Furthermore, in the reward structure, several penalty terms are used to ensure proper EV charging. The term $\alpha \cdot (E^{max} - E^t)$ penalizes an EV that leaves home without a full charge, while $\alpha \cdot (E^t - E^{max})$ prevents overcharging beyond the battery's maximum capacity. The term $\alpha \cdot (E^{min} - E^t)$ prevents undercharging. Additionally, the term $\beta \cdot a^t \cdot (P^{mean} - P^t)$ rewards charging when the price P^t is below the average price P^{mean} and rewards discharging when P^t is above P^{mean}. The coefficient β balances the cost of charging and the price-based rewards.

3.2 Optimization goal

The following are the optimization goals for the proposed EV charging/discharging model.

- **Minimize cumulative charging cost**: The description of cumulative charging costs is provided below:

$$\text{Cumulative\_cost } (T^1) = P^f \times E^{dif} + \sum_{i=t^a}^{t^d} P^t.a^t \tag{7}$$

  The components of the above cost analysis are as follows:
  - **Energy Difference (E^{dif})**: This denotes the difference in energy consumption, calculated as the maximum energy (E^{max}) minus the energy at time t^d (E^{t^d}).
  - P^f: This variable indicates the first price greater than zero after the time t^d.
  - **Dissatisfaction Cost (T^3)**: The dissatisfaction cost is defined as the product of the first price after t^d (P^f) and the energy difference (E^{dif}). It is also considered a measure of user satisfaction.
  - **Total Episode Cost (T^2)**: This cost is computed by summing the product of prices (P^t) and corresponding action values (a^t) over the time interval from t^a to t^d.

- **Maximize user satisfaction:** The proposed model aims to fulfil the EV user's desire for a specific amount of battery energy, denoted as E^{max}, before departing from home. Suppose the state of charge at departure, denoted as E^t, falls short of the desired battery energy level E^{max} at the departure time t^d. In that case, we quantify the degree of dissatisfaction experienced by the EV users as follows:

$$dsat = \alpha.(E^{max} - E^t) \tag{8}$$

Where α denotes the degree of dissatisfaction coefficient. According to Equation 8, the degree of dissatisfaction is directly proportional to the difference between $E^{\max}$ and E^t. In simple terms, the greater the difference between the desired battery energy level $E^{\max}$ and the actual state of charge E^t at the departure time t^d, the greater the dissatisfaction experienced by EV users.

4 Preliminaries

This section briefly overviews the core principles of deep reinforcement learning, including DQN, DDPG, and PPO models. It also introduces the basic concepts of transformer models.

4.1 DRL, DQN, and DDPG

Recent studies have underscored reinforcement learning (RL) as a promising strategy for addressing sequential decision-making challenges in tasks characterized by agent-environment interactions [54]. In RL, at each time step t, the agent observes the state of the environment, s^t, follows a predefined policy π, executes an action a^t, and receives an immediate reward r^t. Subsequently, the agent transitions to the next state, s^{t+1}. The main goal is to maximise cumulative reward, defined as:

$$Cr^t = \sum_{j=1}^{\infty} \gamma^{(j-t)} r_j \tag{9}$$

The discount factor, denoted by γ, reflects consideration for future potential rewards. The expected return Q for a specific state-action pair (s^t, a^t) is then estimated using reinforcement learning techniques, as follows:

$$\mathbb{Q}(s^t, a^t) = \mathbb{E}[Cr^t | s^t, a^t] \tag{10}$$

The $\mathbb{Q}$-learning algorithm updates the action-value function, represented as $\mathbb{Q}(s^t, a^t)$, iteratively utilizing the Bellman equation [55], which can be expressed as:

$$\mathbb{Q}_{j+1}(s, a) = \mathbb{E}\left[r^t + \gamma \max_{a^{t+1}} \mathbb{Q}_j(s^{t+1}, a^{t+1}) \,\middle|\, s^t = s, a^t = a \right] \tag{11}$$

Here, as the number of iterations $j \to \infty$ increases, the action-value function $\mathbb{Q}(s,a)$ will converge to the optimal action-value function $\mathbb{Q}^*(s,a)$. Next, a greedy strategy is used to determine optimal schedules:

$$a^* = \underset{a \in \mathbb{A}}{\operatorname{argmax}}\{\mathbb{Q}^*(s,a)\} \tag{12}$$

In the $\mathbb{Q}$-learning algorithm, the optimal action-value function, denoted as $\mathbb{Q}^*(s,a)$, is frequently represented by a lookup table, a common method in traditional reinforcement learning. However, estimating $\mathbb{Q}^*(s,a)$ using a lookup table becomes impractical when dealing with many states and actions, especially with high-dimensional inputs. To solve this problem, a deep neural network (DNN) is used as a function approximator, which resulted in the deep $\mathbb{Q}$-learning (DQN) technique. The integration of DNN and reinforcement learning creates Deep Reinforcement Learning (DRL), which effectively solves the "dimensional curse" [26]. DRL updates the DNN's parameters by minimising a loss function, which is expressed as follows:

$$\mathbb{L}(\Theta) = \left(r^t + \gamma \max_{a'} \mathbb{Q}(s^{t+1}, a'; \overline{\Theta}) - \mathbb{Q}(s^t, a^t; \Theta)\right)^2 \tag{13}$$

Here, Θ represents the Q-network parameters and $\overline{\Theta}$ represents the target Q-network parameters.

Due to the continuous and high-dimensional nature of electricity prices in our problem, DQN struggles with performance. DQN tries to create a state-action value for each action, which is difficult with a continuous action space. One common solution is to make the action space discrete, which isn't always ideal. The DDPG-based RL approach [51] handles continuous actions better. It uses neural networks with actors and critics: the actor-network generates actions to explore, and the critic network estimates the value of these actions.

4.2 Proximal Policy Optimization (PPO)

Reinforcement learning algorithms use methods from three categories: critic-only, actor-only, and actor-critic [56]. Critic-only or value-function-based techniques refine a deterministic policy by using the value function. Actor-only or policy gradient methods optimise the policy by updating parameters via gradient ascent without using any saved value function. Finally, Barto et al. [57] created actor-critic methods that use value functions and policy gradients to update parameters and refine policies. In DRL, the policy π is typically defined by a deep neural network, with weights expressing the parameter Θ. We use a practical policy gradient method, PPO [58], to create a real-time policy for our proposed MDP. PPO works as an on-policy DRL algorithm in the actor-critic framework. While retaining some of the advantages of TRPO [59], such as facilitating monotonic enhancement, PPO is easier to implement and has been experimentally shown to exhibit better sample efficiency.

The PPO algorithm utilizes a parameterized policy represented by the actor-network π_Θ, alongside a parameterized value function $\hat{V}$, aimed at minimizing training process variance. The parameterized policy undergoes updates by differentiating the policy loss as follows:

$$\mathbb{J}^{\text{policy}(\Theta)} = \mathbb{E}\left[\min(u_t(\Theta) \cdot A\tau, \text{clip}(\mu_t(\Theta), 1-\epsilon, 1+\epsilon) \cdot \hat{A}_t\right] \tag{14}$$

In this case, μ_t denotes the probability ratio and ϵ acts as a hyperparameter regulating the clip range:

$$\mu_t(\Theta) = \frac{\pi_\Theta(a^t|s^t)}{\pi_{\Theta_{\text{old}}}(a^t|s^t)} \tag{15}$$

Here, $\hat{A}_t$ denotes the estimation of the advantage function. The critic network, on the other hand, undergoes an update by minimizing the loss value [58]:

$$\mathbb{J}^{\text{critic}(\phi)} = \mathbb{E}\left[\min\left(V\phi(s^t) - V_\phi(s^t)\right)^2\right] \tag{16}$$

where $\hat{V}_t = \hat{A}_t + V_\phi(s^t)$.

4.3 Transformer

The Transformer [60] is a deep learning model used for tasks like translating languages, summarizing text, and analyzing meaning. Unlike traditional models called recurrent neural networks (RNNs), the Transformer doesn't process words one by one in a sequence. Instead, it uses layers of encoders and decoders built from self-attention and feed-forward neural networks, which allows it to work faster and more efficiently. The Transformer's structure is shown in Figure 1. Each layer in the Transformer has two parts: a self-attention layer and a feed-forward layer. The self-attention layer determines how each word in a sentence relates to every other word and gives them different levels of importance. The feed-forward layer then applies a non-linear transformation to the result from the self- attention layer.

To prevent problems like gradient vanishing or exploding, the Transformer uses residual connections and layer normalization after each sub-layer. This helps stabilise the training process and improves the model's performance. The decoder also has an encoder-decoder attention layer, which determines how each word in the decoder output relates to all words in the encoder output, assigning different importance levels to them. Since the Transformer doesn't use the RNN structure, it can't naturally understand the order of words in a sentence. To fix this, positional encoding is added to the input sequence, adding a vector representing each word's position relative to its word vector.

4.4 Autoformer, Informer, and PatchTST

This section describes transformer-based networks such as Autoformer, Informer, and PatchTST.

4.4.1 Autoformer

The Autoformer [61] is a type of Transformer model made for time-series forecasting. It improves efficiency and performance, especially for long-term forecasts and seasonal patterns. Autoformers use advanced attention mechanisms, like auto-correlation, to

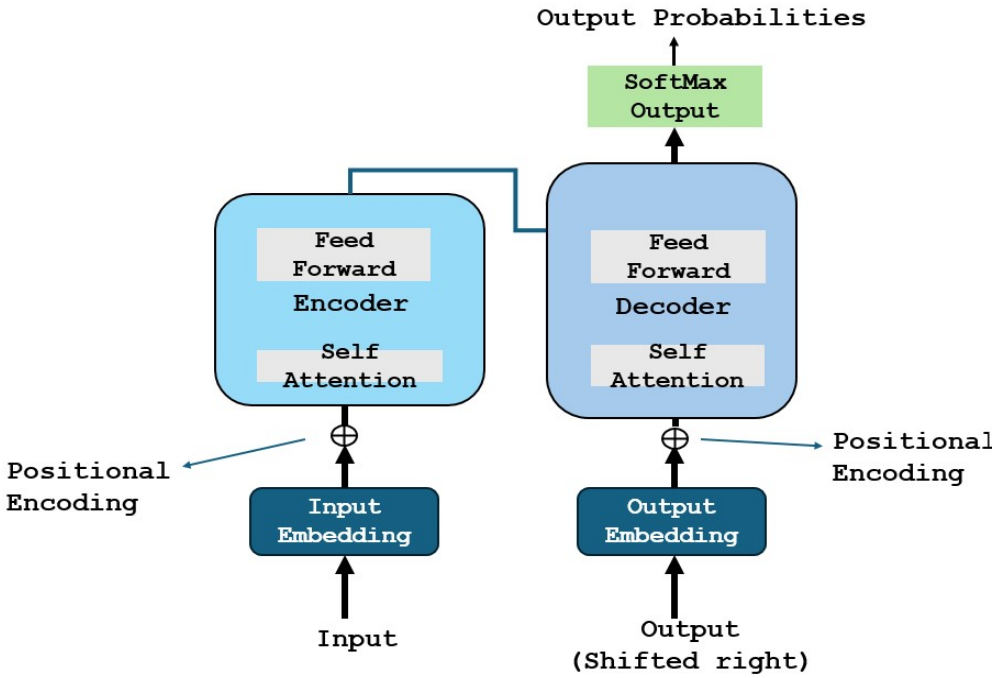


Fig. 1: Transformer Structure.

better capture short-term and long-term trends. They are designed to train quickly and with less computing power. They are also tailored for specific tasks like anomaly detection and other specialized uses. Autoformer improves traditional time series analysis by separating data into seasonal and trend parts. The series decomposition blocks in the encoder remove long-term trends, isolating seasonal patterns so the model can focus on these recurring patterns while ignoring noise. This way, the model captures the essential cycles in the time series.

The encoder-decoder auto-correlation mechanism is a key feature of Autoformer. This new approach replaces the standard self-attention used in traditional transformers, allowing the model to use period-based dependencies. By utilizing past seasonal information from the encoder, the auto-correlation mechanism improves the model's ability to predict future values based on historical patterns, enhancing overall performance. The decoder gradually integrates trend information from hidden variables provided by the encoder. This method ensures the model focuses on both short-term seasonal patterns and long-term trends. The decoder balances seasonal and trend data by progressively adding trend information.

4.4.2 Informer

The Informer model improves the Transformer model by solving several key issues:

- **Efficient Self-Attention**: It replaces traditional self-attention with ProbSparse Self-attention, which uses less time and memory [61, 62].

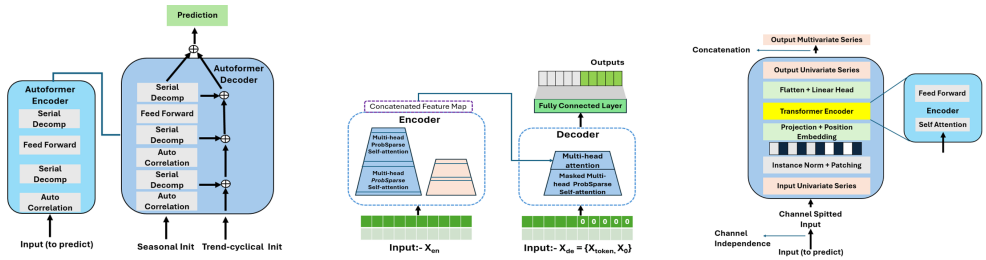
- **Downsampling**: It uses self-attention distilling to reduce the number of dimensions and network parameters, making it easier to handle long inputs.
- **Generative Decoder**: It has a generative decoder that can produce long outputs in one step, avoiding errors that build up over multiple steps

4.4.3 PatchTST

PatchTST [63] is a transformer model designed for time series forecasting, with two key features: patching and channel independence.

- **Patching:** This involves splitting the time series data into smaller sub-sequences (patches) that act as input tokens for the transformer.
- **Channel Independence:** Each channel (or variable) is treated separately as a univariate time series, but they all share the same transformer weights and embedding process.

The structure of the above variants of transformer networks is shown in figure 2, respectively.



(a) Autoformer (b) Informer (c) PatchTST

Fig. 2: Transformer variants

5 Propoed model

This section provides a comprehensive overview of our proposed framework for scheduling EV charging and discharging, including detailed explanations of the training process and the decision model used.

5.1 Framework

This paper optimizes EVs' suggested charging/discharging strategy using deep-reinforcement learning principles. We employ a transformer-based feature analysis model for pattern recognition of historical electricity price data. Subsequently, reinforcement learning (DQN, DDPG, and PPO) makes charging or discharging decisions by analyzing incoming features regarding future electricity prices and the state of charge (E^t). Figure 3 represents the proposed EV charging and discharging scheduling framework. The presented framework has two main parts. The first part uses a transformer-based model as input to predict future electricity prices based on the three

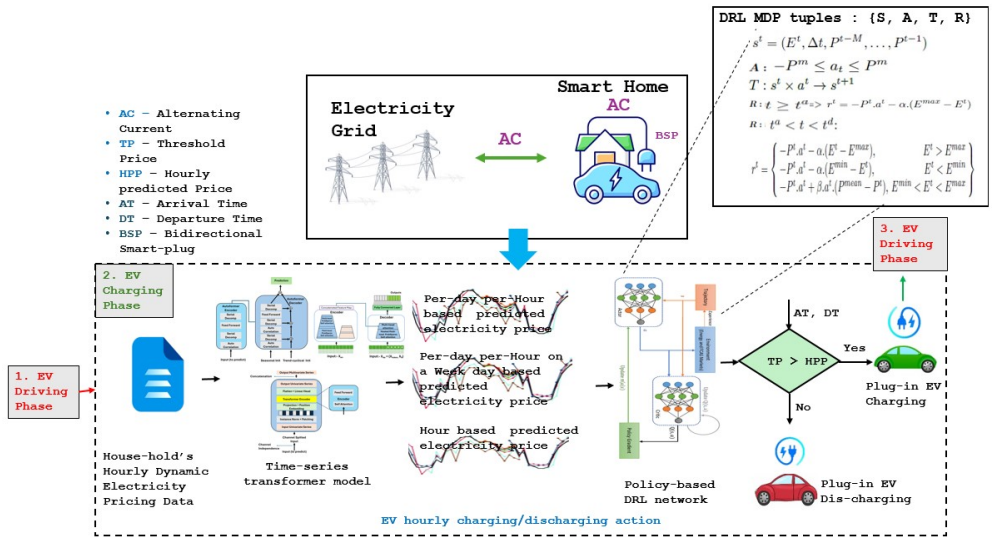


Fig. 3: Proposed EV charging and discharging scheduling framework.

distinct cases (C_1, C_2, and C_3). Then, in the second phase, reinforcement learning-based models (DQN, DDPG, and PPO) are used to suggest the action of charging or discharging at each hour during the EV plugged for charging at home.

5.2 Training process

The proposed FAM's training is carried out in a supervised manner. The training dataset comprises electricity prices from the initial 200 days of 2017 [64]. The training data is partitioned into input electricity prices and their corresponding target outputs for each training iteration. Throughout the training process, the FAM maps the input electricity prices and the desired output electricity prices. It iterative fine-tunes the parameters of the neural network to minimize the disparities between the electricity prices generated by the FAM and the target electricity prices.

5.3 Decision model

Scheduling the charging and discharging of electric vehicles (EVs) has become a prevalent practice in recent years. While Q-learning, a model-free reinforcement learning (RL) technique, has traditionally been used for optimizing policies by interacting with the environment, its effectiveness is limited in scenarios with constrained state and action spaces. Q-learning, introduced by Watkins in 1989 [65], has been widely applied in scheduling problems, including generating real-time charging and discharging schedules for EVs as proposed by Mhaisen et al. [66]. However, one of the significant challenges associated with Q-learning is the curse of dimensionality, which becomes particularly problematic when dealing with numerous states and actions, such as in the context of EV charging schedules. The conventional Q-learning approach, which

relies on a lookup table, is infeasible for complex problems, as highlighted by Mhaisen et al. [66] and Lee et al. [67].

Deep reinforcement learning (DRL), which combines reinforcement learning with deep neural networks, has been extensively adopted to address the curse of dimensionality. Deep-Q Networks (DQN), which integrate deep neural networks with Q-learning, have enhanced the applicability of RL in complex, high-dimensional environments [11, 68]. The DQN strategy, detailed in Algorithm 1, is effective in discrete action spaces but remains limited by its inability to explore continuous action spaces, which negatively impacts the efficiency of EV charging control algorithms. For tasks involving continuous state and action spaces, the Deep Deterministic Policy Gradient (DDPG) algorithm, discussed in Algorithm 2, has proven to be more suitable, as noted in recent studies [12, 39, 44]. Unlike DQN, DDPG is capable of handling the complexities associated with continuous action spaces, thereby improving the overall performance of the EV charging scheduling algorithm.

Furthermore, Proximal Policy Optimization (PPO) is another advanced reinforcement learning technique employed in this context. PPO is a policy gradient method designed to combine the data efficiency and robust performance of Trust Region Policy Optimization (TRPO) while relying only on first-order optimization techniques. The PPO approach to solving the proposed problem is detailed in Algorithm 3. By leveraging the strengths of DQN, DDPG, and PPO, a more comprehensive and effective solution for scheduling EV charging and discharging can be achieved.

6 Performance evaluation

In this section, we evaluate the performance of the proposed EV charging and discharging scheduling model. The experiments use Autoformer, Informer, and Patchtst-based feature extraction models across three distinct cases (C_1, C_2, and C_3). To demonstrate the effectiveness of the proposed approach, we employ three different reinforcement learning (RL) models: DQN, DDPG, and PPO (considering both continuous and discrete action spaces). The evaluation begins with an overview of the experimental setup and training results. Following this, we analyze the impact of the DQN, DDPG, and PPO decision models on the management and outcomes of EV charging and discharging processes.

6.1 Experimental circumstance

In this subsection, we provide a detailed discussion of the dataset, various parameters, and platforms used to implement the proposed work

6.1.1 Dataset and parameters

We validate our method using actual electricity price data from the COMED zone of PJM, USA, as reported in [64]. The dataset includes hourly retail prices reflecting wholesale market prices. We split the data into training (first 200 days of 2017) and testing (days 201 to 300) for evaluation. Following standard practices [11, 12, 38, 39], we assume predictable driving patterns for EV users, with routines like morning

Algorithm 1 Deep Q-Network (DQN).

Input: Price (P^t) (based on C_1, C_2, and C_3) , EV SOC (E^t), Δt, reward r^t
Output: DQN's parameter $\bar{\Theta}$
Begin by initializing the replay memory RM with a capacity of C.
Assign the Q-network with random weights $\bar{\Theta}$
Assign target Q-network with weights $\Theta^{'} = \bar{\Theta}$
Initialize exploration rate ϵ
for episode = 1 to 210,000 **do**
    Initialize state s^1
    for t = t^a to t^d **do**
        Obtain FAM output features following the reception of the preceding 24-units electricity price from s^t
        Combine these features with the battery SOC E^t
        Choose action a^t with ϵ-greedy policy
        Perform action a^t, then observe the resulting reward r^t and the subsequent state s^{t+1}.
        Save the transition (r^t, s^t, a^t, s^{t+1}) in the replay memory, denoted as RM.
        Sample a minibatch of transitions from RM
        Compute target Q-values:
            $Q(s^t, a^t) = r^t + \gamma \max_{a'} Q(s^{t+1}, a'; \Theta^{'})$
        Update Q-network using mean squared error loss:
            $\mathcal{L}(\bar{\Theta}) = \frac{1}{|B|} \sum_{(s,a,r,s') \in B} (Q(s, a; \bar{\Theta}) - Q(s, a))$
        Update target Q-network periodically:
            if $t \mod C == 0$: $\Theta^{'} = \bar{\Theta}$
        Decrease ϵ over time (exploration schedule)
    end for
end for

departures and evening returns. Based on [38], we model EV arrival and departure times using truncated normal distributions, with detailed EV commuting behaviour provided in Table 2.

Table 2: EV commuting behavior [38].

Parameter	Distribution	Threshold
Time of arrival	$t_a \sim \mathcal{N}(18, 1^2)$	$[15, 21]$
Time of departure	$t_d \sim \mathcal{N}(8, 1^2)$	$[6, 11]$
Energy remaining at time t_a	$SOC \sim \mathcal{N}(12, 2.4^2)$	$[4.8, 19.2]$

The arrival time is modelled with a mean of 18 and a standard deviation of 1 hour, constrained within the range $\left[15, 21\right]$. Departure time has a mean of 8, with a standard deviation of 1 hour, within the $\left[6, 11\right]$ range. Upon arriving home, the electric vehicle's (EV) battery energy is assumed to have a mean of 50 % and a standard deviation of 10 % of the battery's capacity. Our study focuses on the Nissan Leaf EV, which has a maximum battery capacity of 24 kWh. The permissible battery energy range is 2.4

Algorithm 2 Deep Deterministic Policy Gradient (DDPG).

Input: Price (P^t) (based on C_1, C_2, and C_3), EV SOC (E^t), Δt, reward r^t
Output: DDPG actor and critic parameters Θ, μ
Initiate the actor network $\mu(s;\Theta_\mu)$ and the critic network $Q(s,a;\Theta_Q)$ by assigning them random weights Θ_μ and Θ_Q.
Initiate the target actor network $\mu'(s;\Theta_{\mu'})$ and the target critic network $Q'(s,a;\Theta_{Q'})$ with weights by setting $\Theta_{\mu'}$ to Θ_μ and $\Theta_{Q'}$ to Θ_Q.
Commence by initializing the replay buffer, denoted as RB.
for episode = 1 to 210,000 **do**
    Obtain FAM output features following the reception of the preceding 24-units electricity price from s^t
    Combine these features with the battery SOC E^t
    Start by initializing a random process for action exploration.
    Observe the initial state, denoted as s^1.
    for t = t^a to t^d **do**
        Choose the action $a^t = \mu(s^t;\Theta_\mu) + \mathcal{N}^t$, where $\mathcal{N}^t$ represents the noise component.
        Perform action a^t, then observe the resulting reward r^t and the updated state s^{t+1}.
        Save the transition (r^t, a^t, s^t, s^{t+1}) in the replay buffer RB.
        Randomly select a minibatch of N transitions from the replay buffer RB.
        Update the critic network by minimizing its associated loss function:
            $\mathcal{L}(\Theta_Q) = \frac{1}{N}\sum_i (y^i - Q(s^i, a^i;\Theta_Q))^2$
        where $y^i = r^i + \gamma Q'(s^{i+1}, \mu'(s^{i+1};\Theta_{\mu'});\Theta_{Q'})$
        Update the actor network by employing sampled policy gradients:
            $\nabla_{\Theta_\mu} J(\Theta_\mu) \approx \frac{1}{N}\sum_i \nabla_a Q(s,a;\Theta_Q)|_{s=s^i, a=\mu(s^i)} \nabla_{\Theta_\mu}\mu(s;\Theta_\mu)|_{s^i}$
        Update the target networks as follows:
            $\Theta_{\mu'} \leftarrow \tau\Theta_\mu + (1-\tau)\Theta_{\mu'}$
            $\Theta_{Q'} \leftarrow \tau\Theta_Q + (1-\tau)\Theta_{Q'}$
    end for
end for

kWh (E^{min}) to 24 kWh (E^{max}), with a maximum charge or discharge rate of 6 kWh per hour. Thus, the action a_t can be selected from the range [-6, 6], where negative values represent discharging and positive values indicate charging. For the discrete action-based RL models (DQN and PPO), we assume the charger offers seven power levels: -6 kW, -4 kW, -2 kW, 0 kW, 2 kW, 4 kW, and 6 kW, for both charging and discharging.

6.1.2 Experimental platform

In our simulation experiments and analysis, we employed the Tyrone DIT400TR-48RL workstation, featuring 128 GB of RAM. This workstation has an NVIDIA Quadro RTX 5000 GPU card built on the Intel-C621 chipset. Additionally, our experimental

Algorithm 3 Training process of PPO

Input: Price (P^t) (based on C_1, C_2, and C_3), EV SOC (E^t), Δt, reward r^t
Output: PPO actor and critic parameters Θ, μ
Initiate the actor network $\mu(s;\Theta_\mu)$ and the critic network $Q(s,a;\Theta_Q)$ by assigning them random weights Θ_μ and Θ_Q.
Initiate the target actor network $\mu'(s;\Theta_{\mu'})$ and the target critic network $Q'(s,a;\Theta_{Q'})$ with weights by setting $\Theta_{\mu'}$ to Θ_μ and $\Theta_{Q'}$ to Θ_Q.
Commence by initializing the replay buffer, denoted as RM.
for episode = 1 to 210,000 **do**
    Obtain FAM output features following the reception of the preceding 24-units electricity price from s^t
    Combine these features with the battery SOC E^t
    Start by initializing a random process for action exploration.
    Observe the initial state, denoted as s^1.
    while $t^a \neq t^d$ **do**
        Use the $\hat{V}$ function to run π_θ in order to choose action a^t.
        Perform action a^t, then observe the resulting reward r^t and the updated state s^{t+1}.
        Save the transition (r^t, a^t, s^t, s^{t+1}) in the replay buffer RM.
        Select Sample batch $\phi = \{(s^t, a^t, r^t, s^{t+1})\}_{t=1}^{\#\phi}$ from RM
        Determine advantage estimates $\hat{A}_i = r_i + \gamma \hat{V}_\phi(s_i') - \hat{V}_\phi(s_i)$
        Apply gradient ascent to update the policy: $\theta \leftarrow \theta + \hat{\alpha} \nabla_\theta J^{\text{policy}}(\theta)$
        Compute value loss: $L(\phi) = \frac{1}{|\mathcal{D}|} \sum_i \left(\hat{V}_\phi(s_i) - (r_i + \gamma \hat{V}_\phi(s_i')) \right)^2$
        Revise the old PPO policy $\Theta_{\text{old}} \leftarrow \Theta$
        for $j = 1$ to N **do**
            Revise actor policy by policy gradient
            $\mu \leftarrow \mu - \hat{\alpha} \nabla_\mu \min \left[\rho(\pi(\theta), \pi_{\text{old}}(s)) \hat{A}_t, \min(\rho(\pi(\theta), \pi_{\text{old}}(s)), 1) \hat{A}_t \right]$
            Revise critic by:
            $\hat{V} \leftarrow \hat{V} + \beta \nabla_{\hat{V}} \frac{1}{2} [(\hat{V}(s^t) - \hat{r}^t)^2]$
        end for
        if every P steps **then**
            Reset $\overline{\Theta} = \Theta$
        end if
    end while
end for

environment is implemented in Python 3.8 with PyTorch 1.13.1+ and CUDA version 11.6.

6.2 *FAMs* train-test results

We conducted experiments with Autoformer, Informer, and PatchTST-based feature extraction models across three cases (C_1, C_2, and C_3). Using the parameters outlined in Table 3 and the previously described training and test datasets, we evaluated the

performance of these FAMs and their impact on the decision-making (DM) process in the following sections.

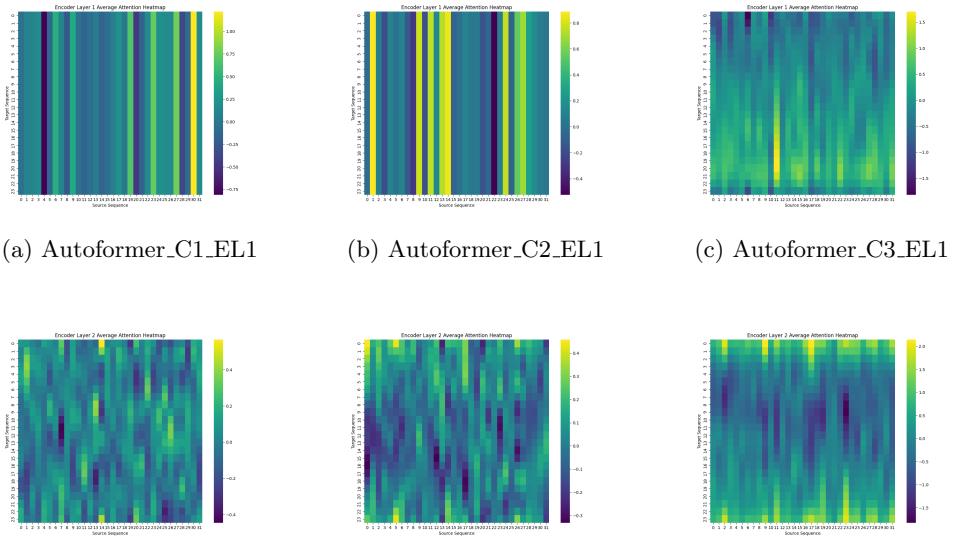
Table 3: List of parameters for all FAMs.

S.No	Parameters	Values
1	Training epoch	200
2	Hidden layer size	$[50, 50, 50, 50]$
3	Learning rate	0.0001
4	Fully connected layer	$[100,]$
5	Sequence length	24
6	Batch size	64

- **Training results:** To gain deeper insights into the effectiveness of feature extraction from raw electricity price data using Autoformer, Informer, and PatchTST, we present the average attention heatmaps for the encoder layers in Figures 4, 5, and 6. These figures illustrate the attention patterns across three cases (C1, C2, and C3) for each model. Figure 4 reveals that the average attention density in the first encoder layer is lower than in the second encoder layer. Notably, the Autoformer model in case 3 exhibits the highest overall attention density compared to cases 1 and 2. Similarly, Figures 5 and 6 show the average attention heatmaps for the Informer and PatchTST models, respectively. Both models follow the same trend as the Autoformer model, with case 3 displaying the highest overall attention density. This indicates that the transformer models in case 3 are more effective at extracting meaningful patterns from the data, which likely contributes to more accurate predictions that are also cleared with test loss results discussed in Table 4.
- **Comparative test loss:** To further showcase the effectiveness of Autoformer, Informer, and Patchtst-based feature extraction models across three cases (C_1, C_2, and C_3), we have visualized the comparative test losses in Figure 7 and presented these results in Table 4. The test loss outcomes unequivocally demonstrate that the innovative model PatchTST has the lowest test loss in each of the three cases. This superiority affirms its enhanced predictive capabilities.

Table 4: Comparative analysis of test loss for all FAMs.

Rank	**FAMs**	**Test loss (Rmse)**
rank-9	Autoformer_C_1	14.91
rank-8	Autoformer_C_2	7.30
rank-5	Autoformer_C_3	5.58
rank-7	Informer_C_1	6.21
rank-6	Informer_C_2	6.03
rank-4	Informer_C_3	4.87
rank-3	PatchTST_C_1	3.74
rank-1	PatchTST_C_2	3.68
rank-2	PatchTST_C_3	3.69



(a) Autoformer_C1_EL1 (b) Autoformer_C2_EL1 (c) Autoformer_C3_EL1

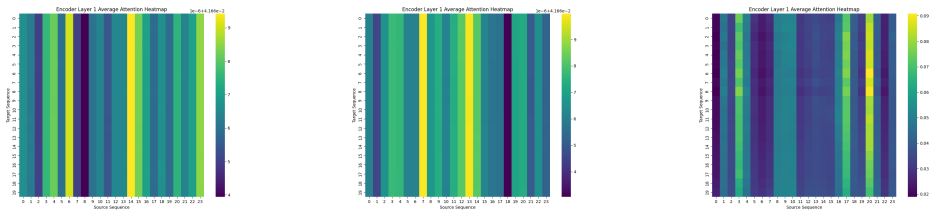
(d) Autoformer_C1_EL2 (e) Autoformer_C2_EL2 (f) Autoformer_C3_EL2

Fig. 4: The average attention Heatmaps for the encoder layer one (EL1) and two (EL2) of the Autoformer model.

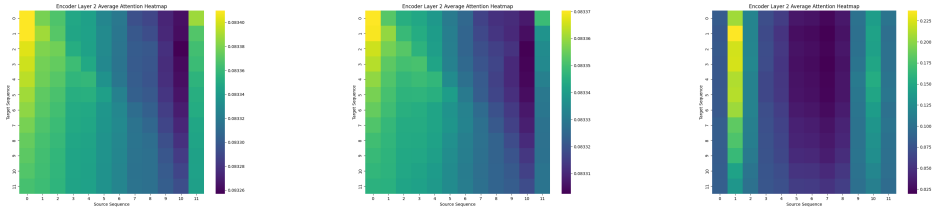
- **Comparative analysis of forecasting:** Finally, to illustrate the effectiveness of the suggested forecasting model, we present a visual comparison in Figure 8 that illustrates the forecasted electricity prices alongside the actual electricity prices for days 201 to 300 in 2017. As shown in Figure 8, the PatchTST and Informer models demonstrate a closer alignment with the actual price data compared to the Autoformer models. This closer fit underscores the superior forecasting performance of the proposed models on the test dataset

6.3 Decision models result

The proposed method utilizes FAM to identify recurring patterns in electricity price fluctuations. Then, a DRL algorithm is applied to optimize decision-making based on these features. Integrating deep learning and RL enhances robustness against uncertain electricity prices and varying EV owner behaviours. The proposed work implements DQN, DDPG, and PPO (continuous and discrete) for decision-making. In DDPG, the action a_t is selected from $\left[-6, 6\right]$, where negative values indicate discharging and positive values indicate charging. For DQN and PPO (discrete), seven power levels (-6 kW, -4 kW, -2 kW, 0 kW, 2 kW, 4 kW, 6 kW) are used for EV charging and discharging.



(a) Informer_C1_EL1 (b) Informer_C2_EL1 (c) Informer_C3_EL1



(d) Informer_C1_EL2 (e) Informer_C2_EL2 (f) Informer_C3_EL2

Fig. 5: The average attention Heatmaps for the encoder layer one (EL1) and two (EL2) of the Informer model.

6.3.1 Results with DQN

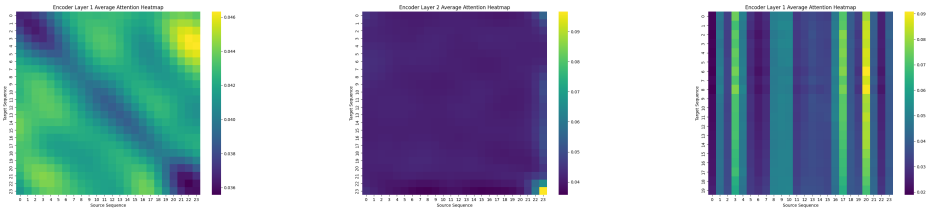
We trained the DQN model for 210,000 epochs to optimize EV charging and discharging schedules, with training conducted separately for each FAM. Each epoch starts when the EV arrives home and ends when it departs. All cases used the same parameters, as detailed in Table 5.

Table 5: Parameters for DQN.

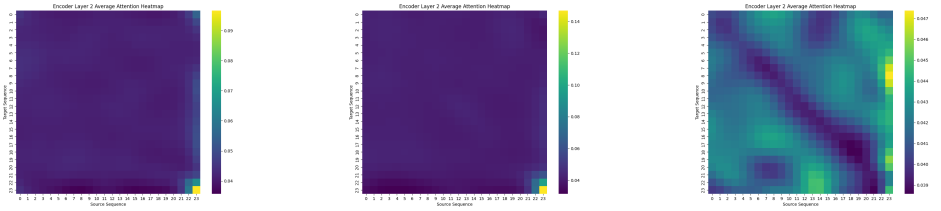
S.No	Parameters	Values
1	Discount factor gamma	0.95
2	Training epoch	2,10,000
3	Learning_rate	$3.5e-5$ to $e-5$
4	Batch_size	64
5	exploration_rate	0.1
6	Hidden fully connected Layer	$[400, 300, 300]$
7	buffer_size	1000000
8	τ	1

The performance results of the proposed model using DQN as the decision-making framework are outlined in the following paragraphs:

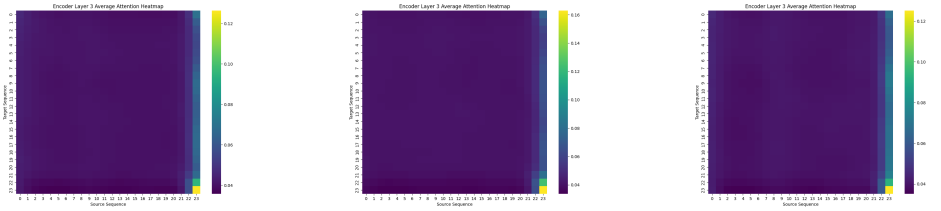
- **Comparative cost analysis using DQN:** To evaluate the effectiveness of all FAMs, we use the cumulative_cost metric defined in Equation 7. The cumulative_cost values for each FAM are listed in Table 6. Additionally, to visually



(a) PatchTST_C1_EL1 (b) PatchTST_C2_EL1 (c) PatchTST_C3_EL1



(d) PatchTST_C1_EL2 (e) PatchTST_C2_EL2 (f) PatchTST_C3_EL2



(g) PatchTST_C1_EL3 (h) PatchTST_C2_EL3 (i) PatchTST_C3_EL3

Fig. 6: The average attention Heatmaps for the encoder layer one (EL1) and two (EL2) of the PatchTST model.

demonstrate cost-effectiveness, we calculated the cumulative_cost reduction percentage for the proposed model compared to related our proposed work [10] (denoted as 'RW'). This calculation is based on Equation 17 and utilizes the computed cumulative_cost values.

$$\frac{Cumulative\_cost\ of\ model\ `RW' - Cumulative\_cost\ of\ proposed\ model}{Cumulative\_cost\ of\ model\ `RW'} \times 100 \tag{17}$$

The outcomes in Table 6 show that the proposed transformer-based approach consistently outperforms the related work [10]. Notably, our model, Autoformer_C_3, achieves the lowest cumulative cost across all models. A negative cumulative cost indicates financial gains for the driver/owner when using our

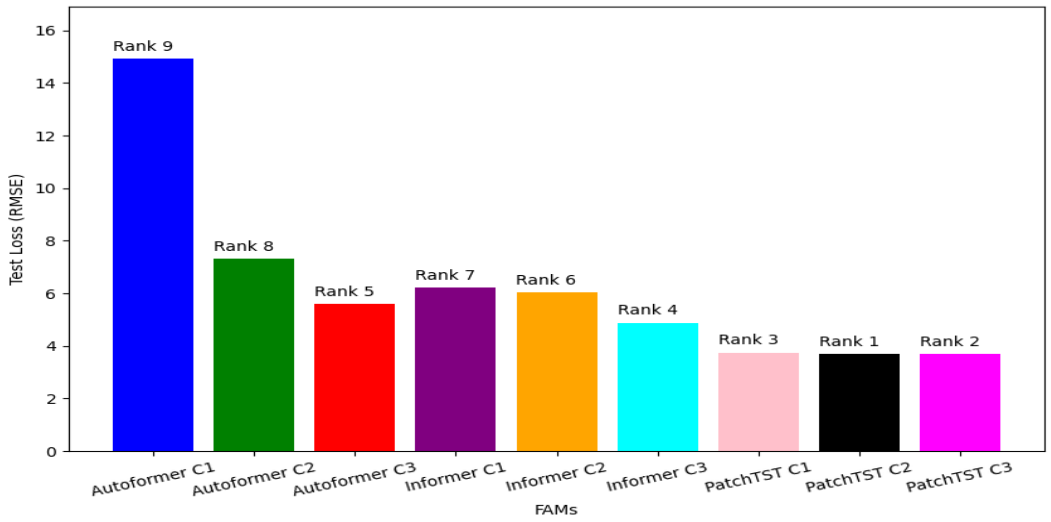


Fig. 7: Comparative test losses across all $FAMs$.

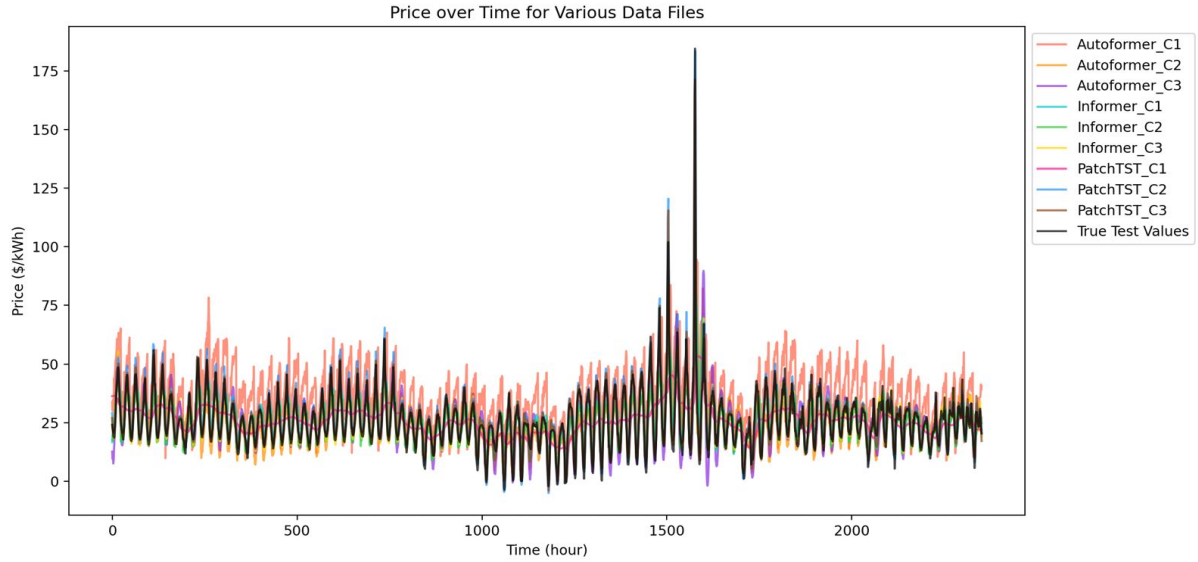
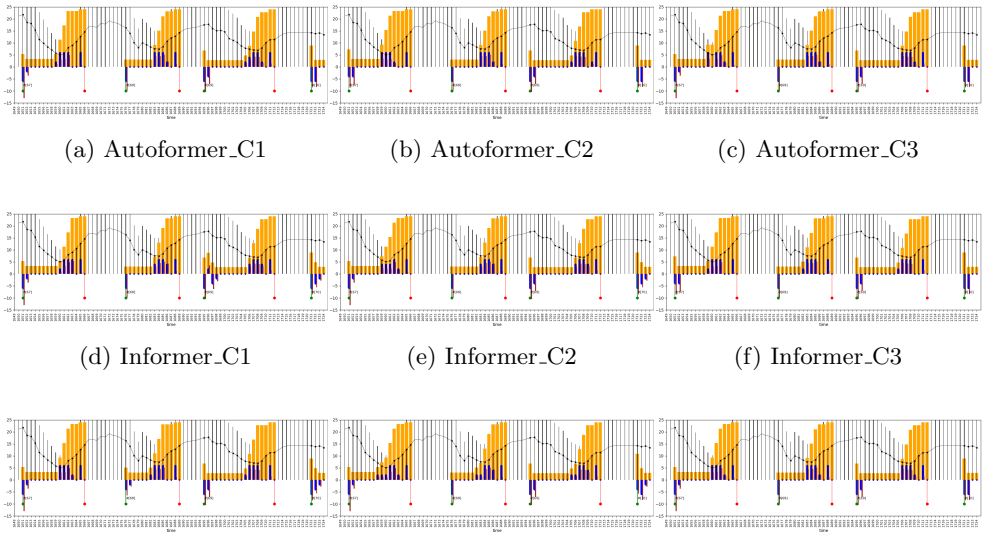


Fig. 8: Comparative analysis of forecasting to the actual price for 100 days test data.

innovative transformer-based method for scheduling charging and discharging. The table also highlights a significant trend: Autoformer models excel in Case-1, the Informer model performs best in Case-2, and PatchTST delivers the top results in Case-3. In contrast, the related work [10] yields the least favourable outcomes.

Moreover, the cumulative cost reduction percentage achieved by the proposed transformer-based approach, when compared to the related work *mgruA_bi_24_1* [10], is 125.18 %, 107.93 %, 136.44 %, 124.92 %, 107.52 %, 117.94 %, 117.11 %, 113.62 %, and 134.69 %, respectively, from top to bottom in Table 6. These results demonstrate that the innovative transformer-based model optimizes EV charging and discharging scheduling than the related work [10].

- **Charging/discharging patterns using DQN:** To validate the effectiveness of the proposed model, we present a detailed analysis of electricity prices alongside the corresponding charging and discharging patterns over three consecutive days (episodes 67, 68, and 69), as depicted in Figure 9



(a) Autoformer_C1 (b) Autoformer_C2 (c) Autoformer_C3

(d) Informer_C1 (e) Informer_C2 (f) Informer_C3

(g) PatchTST_C1 (h) PatchTST_C2 (i) PatchTST_C3

Fig. 9: Electricity price and charging/discharging behaviour over three consecutive days with DQN.

Table 6: Comparative analysis of cumulative_cost with DQN for each FAMs.

Rank	FAMs	Cumulative_cost	Cumulative_cost reduction %
rank-3	Autoformer_C_1	-1661.12	125.18
rank-8	Autoformer_C_2	-523.36	107.93
rank-1	Autoformer_C_3	-2403.53	136.44
rank-4	Informer_C_1	-1644.06	124.92
rank-9	Informer_C_2	-495.79	107.52
rank-5	Informer_C_3	-1183.16	117.94
rank-6	PatchTST_C_1	-1128.94	117.11
rank-7	PatchTST_C_2	-898.59	113.62
rank-2	PatchTST_C_3	-2288.69	134.69
rank-10	**mgruA_bi_24_1** [10]	**6596.83**	-

Figure 9 illustrates the charging and discharging behavior of $FAMs$ over a three-day period. The key elements depicted in the figure are as follows:

i. The interval between the green and red vertical lines represents the duration of a complete episode.
ii. The continuous black line, highlighted with dots, represents the fluctuations in electricity prices.
iii. The yellow bars indicate the state of charge after actions have been executed during specific hours.
iv. The blue vertical bars represent charging events, whereas the blue downward bars denote discharging events at different hours.
v. The red bars represent the associated costs incurred during specific hours.

The presentation above is intended to offer a comprehensive understanding of how the transformer-based FAM models manage EV charging and discharging schedules.

- **User satisfaction analysis with DQN:** To assess user satisfaction, we analyze dissatisfaction costs using the equation referenced in 7. Table 7 presents the dissatisfaction costs associated with different $FAMs$. The data in Table 7 indicate that the transformer-based FAM models achieve 100 % user satisfaction, outperforming existing approaches [10–12]. Specifically, for the transformer-based FAM, the cumulative_cost and Total episode cost are identical in all cases. This indicates that the EV owner leaves with a fully charged battery in all test episodes, resulting in 100 % user satisfaction. In contrast, existing works [10–12] show instances where users leave without a fully charged battery, resulting in dissatisfaction costs. Furthermore, a negative cumulative cost and total episode cost indicate that the EV owner or user benefits financially from the charging and discharging schedule.

Table 7: Comparative analysis of user dissatisfaction cost with DQN for each FAM.

FAMs	**Cumulative_cost** (T_1)	**Total Episode Cost**(T_2)	**Dissatisfaction Cost** ($T_1 - T_2$)
Autoformer_C1	-1661.12	-1661.12	0
Autoformer_C2	-523.36	-523.36	0
Autoformer_C3	-2403.53	-2403.53	0
Informer_C1	-1644.06	-1644.06	0
Informer_C2	-495.79	-495.79	0
Informer_C3	-1183.16	-1183.16	0
PtchTST_C1	-1128.94	-1128.94	0
PtchTST_C2	-898.59	-898.59	0
PtchTST_C3	-2288.69	-2288.69	0
lstm_uni_24_1 [11]	7574.93	-436.63	8011.56
janet_uni_24_1 [12]	8964.70	5244.53	3720.17
mgruA_bi_24_1 [10]	6596.83	4130.37	2466.46

6.3.2 Results with DDPG

We trained the DDPG model for 210,000 epochs to optimize EV charging and discharging schedules, with training conducted separately for each FAM. Each epoch starts when the EV arrives home and ends when it departs. All cases used the same parameters, as in Table 8.

Table 8: Parameters for DDPG.

S.No	Parameters	Values
1	Discount factor gamma	0.95
2	Training epoch	2,10,000
3	Learning_rate	$3.5e-5$ to $e-5$
4	Batch_size	64
5	Hidden fully connected Layer	[400, 300, 300]
6	buffer_size	1000000
7	τ	1
8	update episodes_rate	21000

The performance results of the proposed transformer-based model, using DDPG as the decision-making model, are detailed in the following paragraphs:

- **Comparative cost analysis using DDPG:** To evaluate the effectiveness of our innovative transformer-based feature extraction model with DDPG as the decision model, we employed the same calculation procedure discussed in the previous subsection and conducted a comparative cost analysis using DDPG. Compared to existing related models, the cumulative cost values for each FAM and the cumulative cost reduction percentage for the proposed transformer-based model are presented in Table 9.

Table 9: Comparative analysis of cumulative_cost with DDPG for each FAMs.

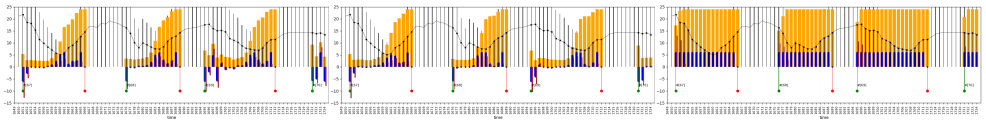
Rank	FAMs	Cumulative_cost	Cumulative_cost reduction %
rank-5	Autoformer_C1	1225.10	84.37
rank-1	Autoformer_C2	-846.98	110.81
rank-8	Autoformer_C3	39861.47	-408.58
rank-2	Informer_C_1	-627.47	108.01
rank-4	Informer_C_2	1104.47	85.91
rank-8	Informer_C_3	39861.47	-408.58
rank-3	PatchTST_C_1	532.10	93.21
rank-6	PatchTST_C_2	1609.29	79.47
rank-8	PatchTST_C_3	39861.47	-408.58
rank-7	mgruA_bi_24_1 [10]	7837.78	-

Table 9 shows that our novel transformer-based model, with *DDPG* as the decision model, consistently outperforms the related models in terms of cumulative cost (Except in Case-3). Notably, the Autoformer model in Case-2 has the lowest total cost (-846.98). The table also shows a significant trend: Autoformer excels in Case 2, while the Informer and PatchTST models excel in Case

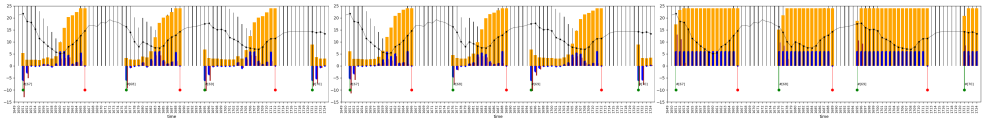
1. However, in Case 3, all transformer-based models perform poorly. In contrast, the related work [10] provides the least favourable overall results. Moreover, the cumulative cost reduction percentage achieved by the proposed transformer-based approach, when compared to the related work $mgruA\_bi\_24\_1$ [10], is 84.37 %, 110.81 %, -408.58 %, 108.01 %, 85.91 %, -408.58 %, 93.21 %, 79.47 %, and -408.58 %, respectively, from top to bottom in Table 9. These results demonstrate that the innovative transformer-based model optimizes EV charging and discharging scheduling than the related recent work [10].

Expanding on our previous discussion, it is evident that our proposed innovative transformer-based approach optimises the EV charging and discharging schedule with Case-1 and Case-2, even when DDPG serves as the decision model.

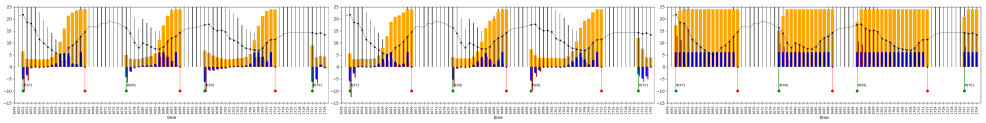
- **Charging/discharging patterns using DDPG:** To validate the effectiveness of the proposed model, we present a detailed analysis of electricity prices alongside the corresponding charging and discharging patterns over three consecutive days (episodes 67, 68, and 69), as depicted in Figure 10



(a) Autoformer_C1 (b) Autoformer_C2 (c) Autoformer_C3



(d) Informer_C1 (e) Informer_C2 (f) Informer_C3



(g) PatchTST_C1 (h) PatchTST_C2 (i) PatchTST_C3

Fig. 10: Electricity price and charging/discharging behaviour over three consecutive days with DDPG.

In Figure 10, we can observe the charging and discharging behaviour of $FAMs$ over three days. All the essential components depicted in the figure align with what we previously discussed in the context of the DQN case. This presentation provides a comprehensive overview of how the proposed transformer-based model effectively manages EV charging and discharging schedules with DDPG as the decision model. Similar to the results with DQN, the transformer-based model consistently outperforms its counterparts. It optimizes charging during low

electricity prices and discharging during high prices, ensuring the EV typically departs with a fully charged battery. As a result, the transformer-based approach maximizes user satisfaction when DDPG is employed as the decision model.

- **User satisfaction analysis with DDPG**: Similar to the approach with DQN as the decision model, we assess user satisfaction with DDPG by calculating dissatisfaction costs for different FAMs, as shown in Table 10. According to the data in Table 10, the transformer-based FAM models achieve 100 % user satisfaction, outperforming existing approaches [10–12].

Table 10: Comparative analysis of user dissatisfaction cost with DDPG for each FAMs.

FAMs	**Cumulative_cost** (T_1)	**Total Episode Cost**(T_2)	**Dissatisfaction Cost** ($T_1 - T_2$)
Autoformer_C1	1225.10	1225.10	0
Autoformer_C2	-846.98	-846.98	0
Autoformer_C3	39861.47	39861.47	0
Informer_C1	-627.47	-627.47	0
Informer_C2	1104.47	1104.47	0
Informer_C3	39861.47	39861.47	0
PatchTST_C1	532.10	532.10	0
PatchTST_C2	1609.29	1609.29	0
PatchTST_C3	39861.47	39861.47	0
lstm_uni_24_1 [11]	9222.69	4714.2	4508.49
janet_uni_24_1 [12]	12183.98	8175.2	4008.78
mgruA_bi_24_1 [10]	7837.78	6353.61	1484.17

6.3.3 Results with PPO

PPO can be applied to environments with either discrete or continuous action spaces. We trained the PPO model for both action types over 210,000 epochs to optimize EV charging and discharging schedules, with training conducted separately for each FAM. Each epoch begins when the EV arrives home and ends when it departs. The same parameters were used across all cases, as detailed in Table 11 for continuous action spaces and Table 12 for discrete action spaces.

Table 11: Parameters for PPO (Continuous).

S.No	**Parameters**	**Values**
1	Discount factor gamma	0.95
2	Training epoch	2,10,000
3	Learning_rate	$4.66e-5$ to $3e-5$
4	Batch_size	64
5	n_steps	256
6	n_epochs	30
7	clip_range	0.2

The performance results of the proposed transformer-based model, using PPO as the decision-making model, are detailed in the following paragraphs:

Table 12: Parameters for PPO (Discrete).

S.No	Parameters	Values
1	Discount factor gamma	0.95
2	Training epoch	2,10,000
3	Learning_rate	$4.66e-5$ to $3e-5$
4	Batch_size	64
5	n_steps	2048
6	n_epochs	25
7	clip_range	0.2

- **Comparative cost analysis using PPO:** Similar to the previous subsection, we assess the effectiveness of our innovative transformer-based feature extraction model combined with PPO in continuous and discrete action spaces as the decision-making model. Tables 13 and 14 compare the cumulative cost values for each FAM and the percentage reduction in cumulative costs achieved by the proposed transformer-based model with PPO against existing related models in continuous and discrete action spaces, respectively.

Table 13: Comparative analysis of cumulative_cost with PPO (continuous) for each FAMs.

Rank	FAMs	Cumulative_cost	Cumulative_cost reduction %
rank-1	Autoformer_C1	-2017.60	125.74
rank-4	Autoformer_C2	-865.78	111.05
rank-9	Autoformer_C3	647.80	91.74
rank-3	Informer_C_1	-980.05	112.50
rank-8	Informer_C_2	643.58	91.79
rank-6	Informer_C_3	-275.70	103.52
rank-7	PatchTST_C_1	102.33	98.69
rank-5	PatchTST_C_2	-668.03	108.52
rank-2	PatchTST_C_3	-1251.59	115.97
rank-10	mgruA_bi_24_1 [10]	7837.78	-

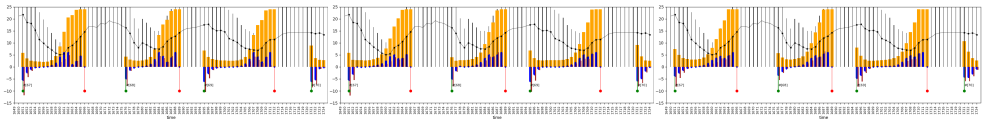
Table 14: Comparative analysis of cumulative_cost with PPO (Discrete) for each FAMs.

Rank	FAMs	Cumulative_cost	Cumulative_cost reduction %
rank-8	Autoformer_C1	-470.66	107.14
rank-9	Autoformer_C2	-182.10	102.76
rank-6	Autoformer_C3	-1116.95	116.93
rank-7	Informer_C_1	-810.71	112.29
rank-1	Informer_C_2	-2681.91	140.66
rank-5	Informer_C_3	-1118.52	116.96
rank-2	PatchTST_C_1	-1918.26	129.08
rank-4	PatchTST_C_2	-1422.28	121.56
rank-3	PatchTST_C_3	-1636.73	124.81
rank-10	mgruA_bi_24_1 [10]	6596.83	-

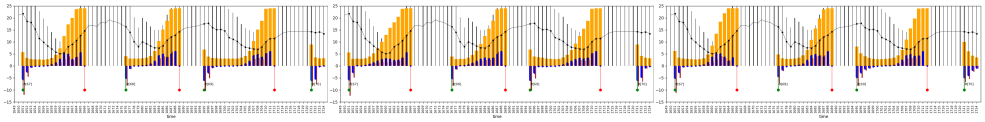
The results in Tables 13 and 14 demonstrate that our proposed transformer-based models, combined with the PPO decision model, consistently deliver the best performance in cumulative cost calculation across both continuous and discrete action space environments. Moreover, the proposed transformer-based approach achieves cumulative cost reduction percentages of 125.74 %, 111.05 %, 91.74 %, 112.50 %, 91.79 %, 103.52 %, 98.69 %, 108.52 %, and 115.97 %, compared to the related work $mgruA\_bi\_24\_1$ [10] in continuous action space. Furthermore, in the discrete action space environment, the reductions are 107.14 %, 102.76 %, 116.93 %, 112.29 %, 140.66 %, 116.96 %, 129.08 %, 121.56 %, and 124.81 %. These results demonstrate that the innovative transformer-based model optimizes EV charging and discharging scheduling than the recent related work $mgruA\_bi\_24\_1$ [10]. Moreover, Tables 13 and 14 reveal a notable trend: in the continuous action space, the Autoformer model in Case-1 achieves the lowest cumulative cost, while in the discrete action space, the Informer model in Case-2 records the lowest cumulative cost.

Expanding on our previous discussion, it is evident that our proposed innovative transformer-based approach optimises the EV charging and discharging schedule, even when PPO serves as the decision model.

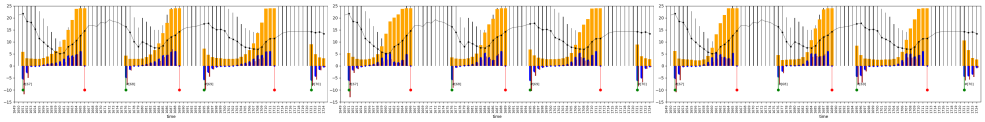
- **Charging/discharging patterns using PPO:** A detailed analysis of electricity prices alongside the corresponding charging and discharging patterns over three consecutive days (episodes 67, 68, and 69) using PPO (continuous) and PPO (discrete) is illustrated in Figures 11 and 12, respectively



(a) Autoformer_C1 (b) Autoformer_C2 (c) Autoformer_C3

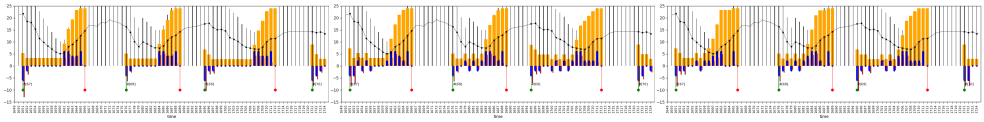


(d) Informer_C1 (e) Informer_C2 (f) Informer_C3

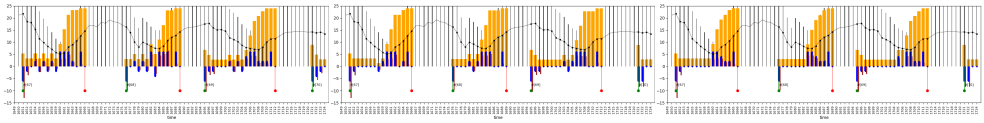


(g) PatchTST_C1 (h) PatchTST_C2 (i) PatchTST_C3

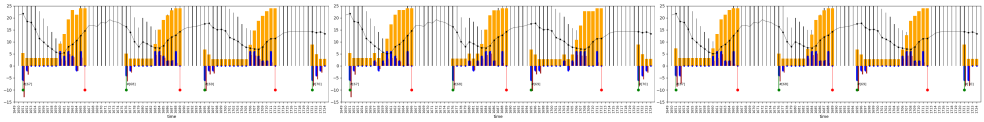
Fig. 11: Electricity price and charging/discharging behaviour over three consecutive days with PPO (continuous).



(a) Autoformer_C1 (b) Autoformer_C2 (c) Autoformer_C3



(d) Informer_C1 (e) Informer_C2 (f) Informer_C3



(g) PatchTST_C1 (h) PatchTST_C2 (i) PatchTST_C3

Fig. 12: Electricity price and charging/discharging behaviour over three consecutive days with PPO (Discrete).

Figures 11 and 12 show the charging and discharging behavior of $FAMs$ over three days. The transformer-based model, using PPO as the decision model, effectively manages EV schedules, similar to results with DQN and DDPG. It optimizes charging during low electricity prices and discharging during high prices, ensuring that the EV often departs fully charged and enhancing user satisfaction.

- **User satisfaction analysis with PPO**: Tables 15 and 16 demonstrate that the transformer-based FAM models achieve 100 % user satisfaction, outperforming existing approaches [10–12] in both continuous and discrete action space environments with PPO.

Table 15: Comparative analysis of user dissatisfaction cost with PPO (continuous) for each FAMs.

FAMs	**Cumulative_cost** (T_1)	**Total Episode Cost**(T_2)	**Dissatisfaction Cost** ($T_1 - T_2$)
Autoformer_C1	-2017.60	-2017.60	0
Autoformer_C2	-865.78	-865.78	0
Autoformer_C3	647.80	647.80	0
Informer_C_1	-980.05	-980.05	0
Informer_C_2	643.58	643.58	0
Informer_C_3	-275.70	-275.70	0
PatchTST_C_1	102.33	102.33	0
PatchTST_C_2	-668.03	-668.03	0
PatchTST_C_3	-1251.59	-1251.59	0
lstm_uni_24_1 [11]	9222.69	4714.2	4508.49
janet_uni_24_1 [12]	12183.98	8175.2	4008.78
mgruA_bi_24_1 [10]	7837.78	6353.61	1484.17

Table 16: Comparative analysis of user dissatisfaction cost with PPO (discrete) for each FAMs.

FAMs	**Cumulative_cost** (T_1)	**Total Episode Cost**(T_2)	**Dissatisfaction Cost** ($T_1 - T_2$)
Autoformer_C1	-182.10	-182.10	0
Autoformer_C2	-810.71	-810.71	0
Autoformer_C3	-1116.95	-1116.95	0
Informer_C_1	-2681.91	-2681.91	0
Informer_C_2	-2681.91	-2681.91	0
Informer_C_3	-1118.52	-1118.52	0
PatchTST_C_1	-1918.26	-1918.26	0
PatchTST_C_2	-1422.28	-1422.28	0
PatchTST_C_3	-1636.73	-1636.73	0
lstm_uni_24_1 [11]	7574.93	-436.63	8011.56
janet_uni_24_1 [12]	8964.70	5244.53	3720.17
mgruA_bi_24_1 [10]	6596.83	4130.37	2466.46

6.4 Discussion

We introduce a DRL-based charging strategy to maximize user charging requirements while minimizing associated expenses for EV owners. This strategy employs transformer-based FAMs, including Autoformer, Informer, and PatchTST, across three cases: C_1, C_2, and C_3.

In our previous work [10], we proposed a novel model, $MHA - BIMGRU$, an enhanced version of the GRU, which was used as the FAM to capture patterns in electricity price variations specifically in case C_3. This model combined the feature extraction capabilities of deep learning with the decision-making strengths of reinforcement learning, resulting in enhanced robustness against the uncertainties in electricity prices and EV owners' commuting behaviour. While the previous study validated the proposed approach using two well-known RL models, DQN and DDPG, the current work expands on this by validating the suggested approach for three cases with three well-known models. This includes applying the PPO RL model to continuous and discrete action space environments, as well as DQN and DDPG models.

In the proposed work, we conducted a comparative cumulative cost analysis using three reinforcement learning models, DQN, DDPG, and PPO, with the results visualized in Figure 13. Figure 13 illustrates the cumulative_cost of FAMs- Autoformer, informer, and PatchTST in three cases C_1, C_2, and C_3 hence the name categorized as Autoformer_c1, Autoformer_c2, Autoformer_c3, Informer_C1, Informer_C2, Informer_C3, PatchTST_C1, PatchTST_C2, and PatchTST_C3, respectively. In each case, each FAM has four results corresponding to the RL model: DQN, DDPG, PPO (Continuous), and PPO (Discrete).

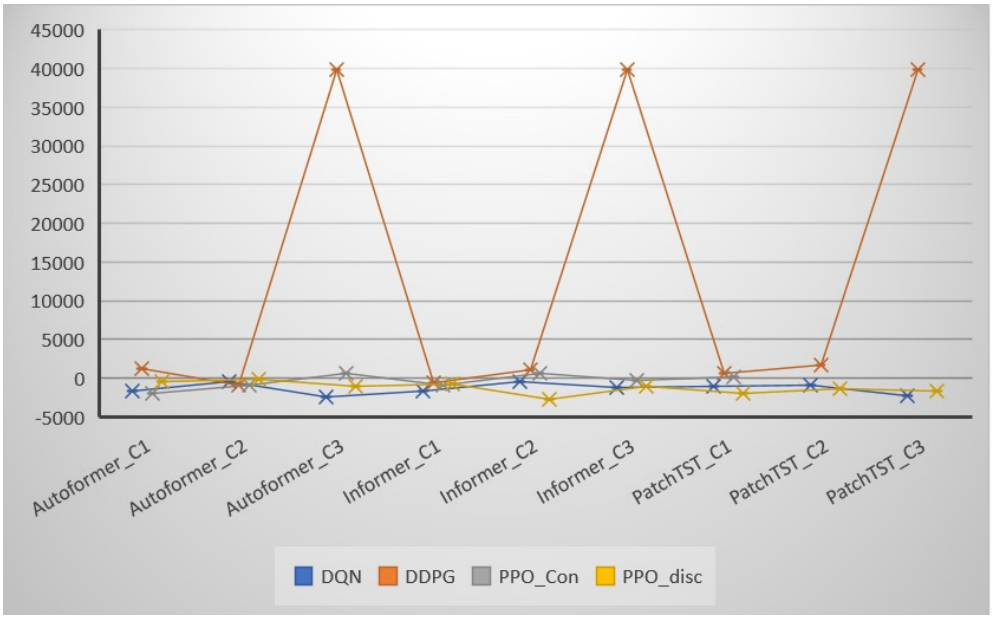


Fig. 13: Comparative cumulative_cost of each FAMs with DQN, DDPG, and PPO.

Notably, our FAM Autoformer_C1 achieves the best cumulative cost of -2017.60 when paired with the PPO (continuous) RL model. In contrast, Autoformer_C2 records a cumulative cost of -865.78, and Autoformer_C3 registers a cost of 647.80 with the same RL model. For the Informer model, Informer_C1 stands out with a cumulative cost of -1644.06 when integrated with DQN. Informer_C2 achieves the best cumulative cost of -2681.91 when paired with PPO (discrete), while Informer_C3 records a cost of -1118.52, also with PPO (discrete). For the PatchTST models, PatchTST_C1 achieves the lowest cumulative cost of -1918.26 with PPO (discrete), while PatchTST_C2 records a cumulative cost of -1422.28 with the same RL model. PatchTST_C3 delivers the best cumulative cost of -2288.69 when paired with DQN. Furthermore, in Case 3 (C3), the FAM model with DDPG exhibits the poorest performance compared to its results in Cases 1 and 2 under identical conditions, as well as overall. Notably, Informer_C2 surpasses all other models, achieving the best cumulative cost of -2681.91 with PPO (discrete). The negative cumulative cost indicates a net gain

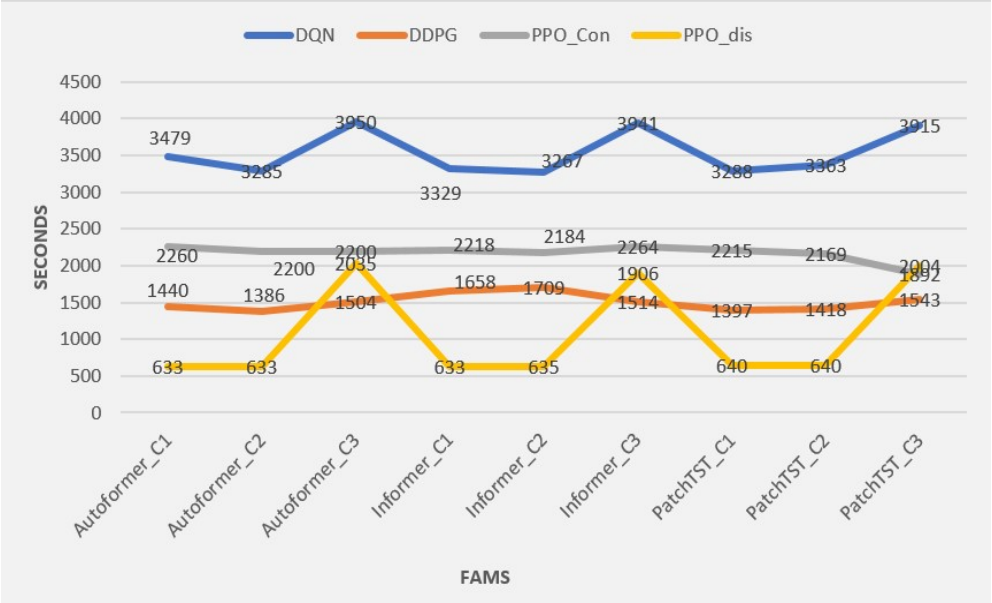


Fig. 14: Total time elapsed for each FAMs.

for the owner from the charging and discharging actions over 100 test episodes. Additionally, in the continuous action space, the Autoformer model in Case 1, paired with PPO, achieves the lowest cumulative cost of -2017.60 overall. In the discrete action space, the Informer model, combined with PPO in Case 2, records the lowest cost.

Thus, the preceding discussion highlights the exceptional versatility of the proposed transformer-based model, which excels in optimizing cumulative costs for discrete and continuous charging and discharging actions.

A comprehensive analysis of the running times for all these FAMS during our implementation with the decision models—DQN, DDPG, PPO (continuous), and PPO (discrete)—is presented in Figure 14. As depicted in the figure, all FAMs (except for Case-3) demonstrate the shortest running times when used with the PPO (discrete) model, while the longest running times are observed with the DQN model in all cases. Additionally, in the continuous action environment, all FAMs exhibit the shortest running times when paired with the DDPG model. Moreover, to assess the efficacy of our proposed model within the context of a reinforcement learning environment, we have incorporated three essential visual representations in Figures 15, and 16, and 17. These figures illustrate the progression of mean reward throughout the training epochs, the reward dynamics across test epochs, and the model-wise mean reward achieved in training.

The proposed model underwent training for 2,10,000 epochs to grasp the optimal EV charging/discharging actions. The evolution of mean reward across these epochs is depicted for each FAM in Figure 15. Notably, the mean reward exhibits a sharp increase from the beginning to the $40,000^{th}$ episode, followed by a gradual incline until the $210,000^{th}$ episode for all transformer-based FAMs under DQN. Upon

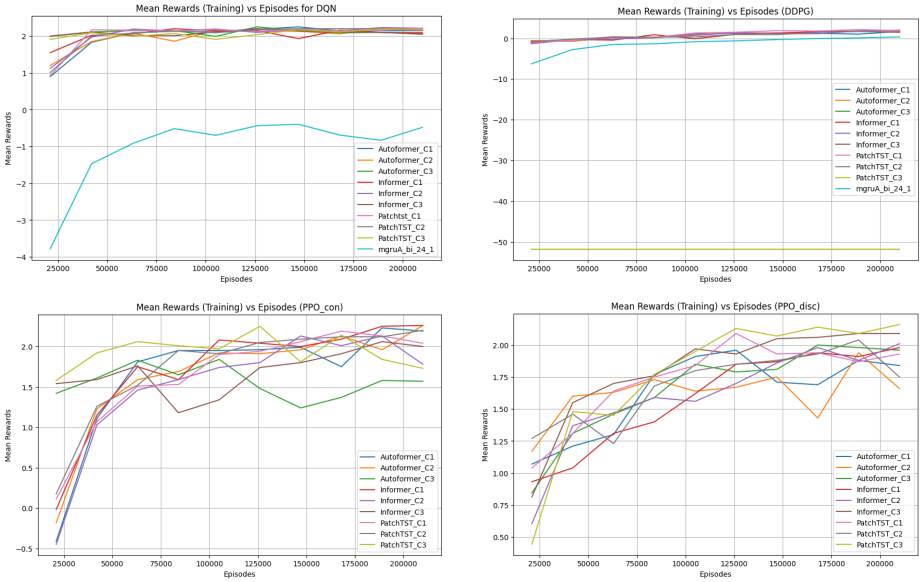


Fig. 15: Mean reward for each FAMs over training episodes.

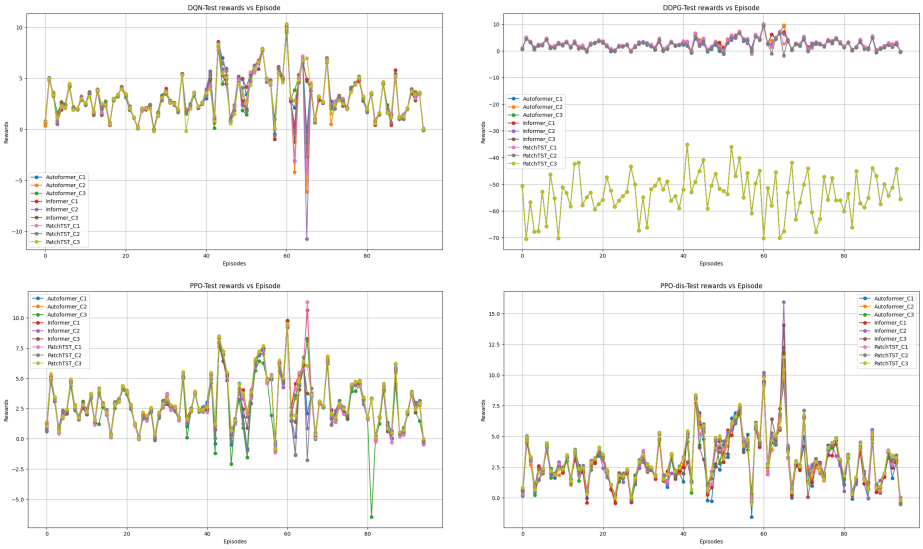


Fig. 16: Mean reward for each FAMs over Test episodes.

completion of training, the mean reward of the models, namely Autoformer_C1, Autoformer_C2, Autoformer_C3, Informer_C1, Informer_C2, Informer_C3, PatchTST_C1, PatchTST_C2, and PatchTST_C3, converges to 2.15, 2.19, 2.07, 2.09, 2.21, 2.05, 2.21,

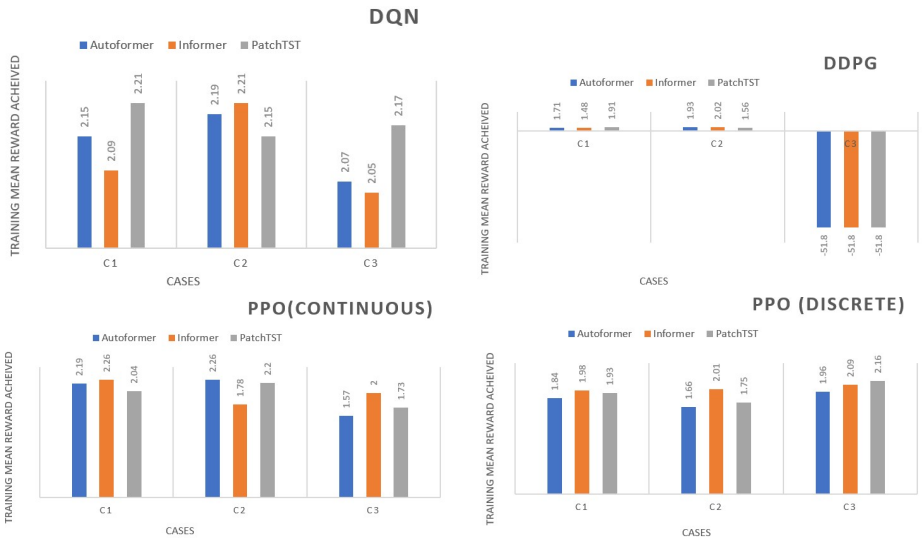


Fig. 17: Mean rewards achieved in training for four classes of FAMs.

2.15, and 2.17, respectively. Similarly, under DDPG, the convergence results are 1.71, 1.93, -51.8, 1.48, 2.02, -51.8, 1.91, 1.56, -51.8, and 0.352, respectively. Under PPO (continuous), mean reward convergence results are 2.19, 2.26, 1.57, 2.26, 1.78, 2.00, 2.04, 2.20, and 1.73, respectively, and PPO (discrete), mean reward convergence results are 1.84, 1.66, 1.96, 1.98, 2.01, 2.09, 1.93, 1.75, and 2.16. These results demonstrate that all transformer-based model variants successfully learned to increase the mean rewards, except in Case-3 with the DDPG model, where the FAM agent failed to learn effectively under the same settings as for the DDPG in Case-1 and Case-2. Moreover, compared to the best-related work, mgruA_bi_24_1, where the mean rewards were -0.479 with DQN and 0.352 with DDPG, the transformer-based models in this study demonstrated significantly higher mean rewards.

The evolution of rewards across test episodes for each FAM is visually presented in Figure 16. This figure also illustrates that, in combination with DQN, DDPG, and PPO, all the transformer-based model variants successfully learned a valid policy, achieving the highest mean rewards (except for Case-3 in DDPG). Additionally, Figure 17 provides a visual representation of the mean rewards obtained during training with DQN, DDPG, PPO (Continuous), and PPO (Discrete) as decision models across three transformer-based models—Autoformer, Informer, and PatchTST—under the three cases, C1, C2, and C3.

In reinforcement learning, a positive mean reward is widely acknowledged as a favourable outcome. The positive mean reward is a strong indicator that, on average, the RL agent accomplishes its objectives within the specified environment or task. Consequently, higher positive mean rewards are a reliable sign of more effective policies. Figure 17 visually illustrates our analysis of mean rewards during training when implementing DQN, DDPG, PPO (Continuous), and PPO (Discrete) in combination with transformer-based FAMS variants. Figure 17 shows that DQN, DDPG, and

PPO (Continuous) achieved higher mean rewards in either Case-1 or Case-2, while PPO (Discrete) performed better in Case-3. The highest mean reward in the discrete environment was 2.21, achieved by DQN in both Case-1 and Case-2. In the continuous action space, the highest mean reward was 2.26, achieved by PPO (Continuous) in Case-1 and Case-2. Overall, the proposed transformer-based model consistently outperformed in Case-1 and Case-2 compared to Case-3, highlighting its superior computational efficiency in those scenarios

The findings from the previous discussion highlight the effectiveness of the proposed transformer-based model in Case-1 and Case-2 compared to Case-3 and the related work [10–12].

7 Conclusion

This work demonstrates that integrating a transformer-based network with deep reinforcement learning (DRL) significantly enhances in-home EV charging optimization, meeting EV owner full charging requirements while minimizing charging costs. By employing advanced feature extraction models—Autoformer, Informer, and PatchTST—across three distinct cases (C_1, C_2, and C_3), we build upon our earlier approach, which only considered the past 24 hours of price data. Specifically, C_1 leverages prices from the same hour over the past 24 days, C_2 uses prices from the same hour on the same weekday over the last 24 weeks, and C_3 continues using the past 24 hours of data. Our previous study, which used the $MHA-BIMGRU$ model for Case C_3, demonstrated robustness against electricity price fluctuations through GRU-based FAMs and RL models like DQN and DDPG. In this work, we extend that approach by integrating broader time frames and evaluating multiple reinforcement learning models (DQN, DDPG, and PPO). Our comparative analysis shows significant reductions in cumulative costs, with the Autoformer model in C_1 excelling in the continuous action space and the Informer model in C_2 performing best in the discrete space. Overall, the proposed method achieves full user satisfaction and reduces charging costs by 125.74 % in the continuous space and 140.66 % in the discrete space, offering superior performance and practical benefits for real-time EV charging management.

Declarations

- Funding: This work did not receive financial support.
- Conflict of interest/Competing interests: The authors declare that there is no conflict of interest regarding the publication of this paper.
- Ethics approval: Not applicable
- Consent to participate: All the authors declare their consent to participate in this research article.
- Consent for publication: All the authors declare their consent for publication of the article on acceptance.
- Availability of data and materials: The data and materials are available within the manuscript.
- Code availability: Code will be provided on a need-to-know basis.

- Authors' contributions: The collaborative efforts of the authors were as follows: Shivendu Mishra contributed to the conceptualization methodology and was responsible for writing the - original Draft. Anurag Choubey played a key role in conceptualizing and meticulously revising the manuscript. Harshit Dhankhar and Sri Vaibhav Devarasetty primarily focused on designing and implementing part of the manuscript. Rajiv Misra provided valuable input through validation, supervision, and further refinement of the manuscript. Each author's commitment and involvement have been substantial, collectively assuming public responsibility for different facets of the content. Importantly, all authors have rigorously reviewed and approved the final manuscript.

References

[1] Ghosh, A.: Possibilities and challenges for the inclusion of the electric vehicle (ev) to reduce the carbon footprint in the transport sector: A review. Energies **13**(10), 2602 (2020)

[2] Zhang, J., Yan, J., Liu, Y., Zhang, H., Lv, G.: Daily electric vehicle charging load profiles considering demographics of vehicle users. Applied Energy **274**, 115063 (2020)

[3] Choubey, A., Sikarwar, A., Asoba, S., Misra, R.: Towards an ipfs-based highly scalable blockchain for pev charging and achieve near super-stability in a v2v environment. Cluster Computing, 1–42 (2024)

[4] Tan, J., Wang, L.: Real-time charging navigation of electric vehicles to fast charging stations: A hierarchical game approach. IEEE transactions on smart grid **8**(2), 846–856 (2015)

[5] Lee, W., Schober, R., Wong, V.W.: An analysis of price competition in heterogeneous electric vehicle charging stations. IEEE Transactions on Smart Grid **10**(4), 3990–4002 (2018)

[6] Silva, F.C., A. Ahmed, M., Martínez, J.M., Kim, Y.-C.: Design and implementation of a blockchain-based energy trading platform for electric vehicles in smart campus parking lots. Energies **12**(24), 4814 (2019)

[7] Chen, Q., Folly, K.A.: Application of artificial intelligence for ev charging and discharging scheduling and dynamic pricing: A review. Energies **16**(1), 146 (2022)

[8] Li, J., Wang, X., Tu, Z., Lyu, M.R.: On the diversity of multi-head attention. Neurocomputing **454**, 14–24 (2021)

[9] Reza, S., Ferreira, M.C., Machado, J.J.M., Tavares, J.M.R.: A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Systems with Applications **202**, 117275 (2022)

[10] Mishra, S., Choubey, A., Devarasetty, S.V., Sharma, N., Misra, R.: An innovative multi-head attention model with bimgru for real-time electric vehicle charging management through deep reinforcement learning. Cluster Computing, 1–31 (2024)

[11] Wan, Z., Li, H., He, H., Prokhorov, D.: Model-free real-time ev charging scheduling based on deep reinforcement learning. IEEE Transactions on Smart Grid **10**(5), 5246–5257 (2018)

[12] Li, S., Hu, W., Cao, D., Dragičević, T., Huang, Q., Chen, Z., Blaabjerg, F.: Electric vehicle charging management based on deep reinforcement learning. Journal of Modern Power Systems and Clean Energy **10**(3), 719–730 (2021)

[13] Iversen, E.B., Morales, J.M., Madsen, H.: Optimal charging of an electric vehicle using a markov decision process. Applied Energy **123**, 1–12 (2014)

[14] Hu, W., Su, C., Chen, Z., Bak-Jensen, B.: Optimal operation of plug-in electric vehicles in power systems with high wind power penetrations. IEEE Transactions on Sustainable Energy **4**(3), 577–585 (2013)

[15] Jin, C., Tang, J., Ghosh, P.: Optimizing electric vehicle charging: A customer's perspective. IEEE Transactions on vehicular technology **62**(7), 2919–2927 (2013)

[16] Ravey, A., Roche, R., Blunier, B., Miraoui, A.: Combined optimal sizing and energy management of hybrid electric vehicles. In: 2012 IEEE Transportation Electrification Conference and Expo (ITEC), pp. 1–6 (2012). IEEE

[17] Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., Blaabjerg, F.: Reinforcement learning and its applications in modern power and energy systems: A review. Journal of modern power systems and clean energy **8**(6), 1029–1042 (2020)

[18] Ortega-Vazquez, M.A.: Optimal scheduling of electric vehicle charging and vehicle-to-grid services at household level including battery degradation and price uncertainty. IET Generation, Transmission & Distribution **8**(6), 1007–1016 (2014)

[19] Zhao, J., Wan, C., Xu, Z., Wang, J.: Risk-based day-ahead scheduling of electric vehicle aggregator using information gap decision theory. IEEE Transactions on Smart Grid **8**(4), 1609–1618 (2015)

[20] Vayá, M.G., Andersson, G.: Optimal bidding strategy of a plug-in electric vehicle aggregator in day-ahead electricity markets under uncertainty. IEEE transactions on power systems **30**(5), 2375–2385 (2014)

[21] Sarker, M.R., Pandžić, H., Ortega-Vazquez, M.A.: Optimal operation and services scheduling for an electric vehicle battery swapping station. IEEE transactions on

power systems **30**(2), 901–910 (2014)

[22] Wu, D., Zeng, H., Lu, C., Boulet, B.: Two-stage energy management for office buildings with workplace ev charging and renewable energy. IEEE Transactions on Transportation Electrification **3**(1), 225–237 (2017)

[23] Guo, Y., Xiong, J., Xu, S., Su, W.: Two-stage economic operation of microgrid-like electric vehicle parking deck. IEEE Transactions on Smart Grid **7**(3), 1703–1712 (2015)

[24] Momber, I., Siddiqui, A., San Roman, T.G., Söder, L.: Risk averse scheduling by a pev aggregator under uncertainty. IEEE Transactions on Power Systems **30**(2), 882–891 (2014)

[25] Kim, S., Lim, H.: Reinforcement learning based energy management algorithm for smart energy buildings. Energies **11**(8), 2010 (2018)

[26] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., *et al.*: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)

[27] Wen, Z., O'Neill, D., Maei, H.: Optimal demand response using device-based reinforcement learning. IEEE Transactions on Smart Grid **6**(5), 2312–2324 (2015)

[28] Vandael, S., Claessens, B., Ernst, D., Holvoet, T., Deconinck, G.: Reinforcement learning of heuristic ev fleet charging in a day-ahead electricity market. IEEE Transactions on Smart Grid **6**(4), 1795–1805 (2015)

[29] Chiş, A., Lundén, J., Koivunen, V.: Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Transactions on Vehicular Technology **66**(5), 3674–3684 (2016)

[30] Bahrami, S., Wong, V.W., Huang, J.: An online learning algorithm for demand response in smart grid. IEEE Transactions on Smart Grid **9**(5), 4712–4725 (2017)

[31] Ruelens, F., Claessens, B.J., Vandael, S., De Schutter, B., Babuška, R., Belmans, R.: Residential demand response of thermostatically controlled loads using batch reinforcement learning. IEEE Transactions on Smart Grid **8**(5), 2149–2159 (2016)

[32] Shaarbaf, M.R., Ghayeni, M.: Identification of the best charging time of electric vehicles in fast charging stations connected to smart grid based on q-learning. In: 2018 Electrical Power Distribution Conference (EPDC), pp. 78–83 (2018). IEEE

[33] Chiş, A., Lundén, J., Koivunen, V.: Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Transactions on Vehicular Technology **66**(5), 3674–3684 (2016)

[34] Wan, Z., Li, H., He, H., Prokhorov, D.: A data-driven approach for real-time

residential ev charging management. In: 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5 (2018). IEEE

[35] Wan, Z., He, H.: Answernet: Learning to answer questions. IEEE Transactions on Big Data **5**(4), 540–549 (2018)

[36] Wan, Z., He, H., Tang, B.: A generative model for sparse hyperparameter determination. IEEE Transactions on Big Data **4**(1), 2–10 (2017)

[37] Wang, F., Gao, J., Li, M., Zhao, L.: Autonomous pev charging scheduling using dyna-q reinforcement learning. IEEE Transactions on Vehicular Technology **69**(11), 12609–12620 (2020)

[38] Li, H., Wan, Z., He, H.: Constrained ev charging scheduling based on safe deep reinforcement learning. IEEE Transactions on Smart Grid **11**(3), 2427–2439 (2019)

[39] Zhang, F., Yang, Q., An, D.: Cddpg: A deep-reinforcement-learning-based approach for electric vehicle charging control. IEEE Internet of Things Journal **8**(5), 3075–3087 (2020)

[40] Yan, L., Chen, X., Zhou, J., Chen, Y., Wen, J.: Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. IEEE Transactions on Smart Grid **12**(6), 5124–5134 (2021)

[41] Ye, Z., Gao, Y., Yu, N.: Learning to operate an electric vehicle charging station considering vehicle-grid integration. IEEE Transactions on Smart Grid **13**(4), 3038–3048 (2022)

[42] Jiang, Y., Ye, Q., Sun, B., Wu, Y., Tsang, D.H.: Data-driven coordinated charging for electric vehicles with continuous charging rates: A deep policy gradient approach. IEEE Internet of Things Journal **9**(14), 12395–12412 (2021)

[43] Cao, Y., Wang, H., Li, D., Zhang, G.: Smart online charging algorithm for electric vehicles via customized actor–critic learning. IEEE Internet of Things Journal **9**(1), 684–694 (2021)

[44] Chen, G., Shi, X.: A deep reinforcement learning-based charging scheduling approach with augmented lagrangian for electric vehicle. arXiv preprint arXiv:2209.09772 (2022)

[45] Hou, L., Li, Y., Yan, J., Wang, C., Wang, L., Wang, B.: Multi-agent reinforcement mechanism design for dynamic pricing-based demand response in charging network. International Journal of Electrical Power & Energy Systems **147**, 108843 (2023)

[46] Paudel, D., Das, T.K.: A deep reinforcement learning approach for power management of battery-assisted fast-charging ev hubs participating in day-ahead and real-time electricity markets. Energy, 129097 (2023)

[47] Qi, T., Ye, C., Zhao, Y., Li, L., Ding, Y.: Deep reinforcement learning based charging scheduling for household electric vehicles in active distribution network. Journal of Modern Power Systems and Clean Energy, 1–12 (2023) https://doi.org/10.35833/MPCE.2022.000456

[48] Zhang, J., Guan, Y., Che, L., Shahidehpour, M.: Ev charging command fast allocation approach based on deep reinforcement learning with safety modules. IEEE Transactions on Smart Grid, 1–1 (2023) https://doi.org/10.1109/TSG.2023.3281782

[49] Sykiotis, S., Menos-Aikateriniadis, C., Doulamis, A., Doulamis, N., Georgilakis, P.S.: A self-sustained ev charging framework with n-step deep reinforcement learning. Sustainable Energy, Grids and Networks **35**, 101124 (2023)

[50] Aljafari, B., Jeyaraj, P.R., Kathiresan, A.C., Thanikanti, S.B.: Electric vehicle optimum charging-discharging scheduling with dynamic pricing employing multi agent deep neural network. Computers and Electrical Engineering **105**, 108555 (2023)

[51] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)

[52] Jos, V., Lasenby, J.: The unreasonable effectiveness of the forget gate. Computer Science **2018**, 11–49 (2018)

[53] Song, H., Liu, C.-C., Lawarrée, J., Dahlgren, R.W.: Optimal electricity supply bidding by markov decision process. IEEE transactions on power systems **15**(2), 618–624 (2000)

[54] Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. mit press (2018)

[55] Bellman, R.: Dynamic programming. princeton university press, john wiley & sons (1958)

[56] Grondman, I., Busoniu, L., Lopes, G.A., Babuska, R.: A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews) **42**(6), 1291–1307 (2012)

[57] Barto, A.G., Sutton, R.S., Anderson, C.W.: Neuronlike adaptive elements that can solve difficult learning control problems. IEEE transactions on systems, man,

and cybernetics (5), 834–846 (1983)

[58] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[59] Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. In: International Conference on Machine Learning, pp. 1889–1897 (2015). PMLR

[60] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[61] Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems **34**, 22419–22430 (2021)

[62] Zhang, X., Yang, K., Zheng, L.: Transformer fault diagnosis method based on timesnet and informer. In: Actuators, vol. 13, p. 74 (2024). MDPI

[63] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers (2023). https://arxiv.org/abs/2211.14730

[64] PJM Zone COMED: Price Data Set: PJM Zone COMED. Accessed on July 3, 2023. https://www.engieresources.com/historical-data.

[65] Watkins, C.J., Dayan, P.: Q-learning. Machine learning **8**, 279–292 (1992)

[66] Mhaisen, N., Fetais, N., Massoud, A.: Real-time scheduling for electric vehicles charging/discharging using reinforcement learning. In: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), pp. 1–6 (2020). IEEE

[67] Lee, S., Choi, D.-H.: Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic system, energy storage system, and home appliances. Sensors **19**(18), 3937 (2019)

[68] Lee, J., Lee, E., Kim, J.: Electric vehicle charging and discharging algorithm based on reinforcement learning with data-driven approach in dynamic pricing scheme. Energies **13**(8), 1950 (2020)