**Bulletin of Electrical Engineering and Informatics**
Vol. 15, No. 4, August 2026, pp. 3183~3196
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.11211


# Proximal policy optimization with self adaptive penalty function for vehicular resource allocation

**Irshad Khan[1], Neetha Papanna Umalakshmi[2], Somshekhar Durgaiah[3], Vijetha Acharya[4], Suman Joseph[5], Varshini Totliganahalli Rajanna[6]**
[1]Presidency School of Artificial Intelligence and Advanced Computing, Presidency University, Bengaluru, India
[2]Department of Computer Science and Engineering, BMS Institute of Technology and Management, Bengaluru, India
[3]Department of Computer Science and Engineering, Harsh Institute of Technology, Bengaluru, India
[4]Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, India
[5]Department of Computer Science and Engineering, SJC Institute of Technology, Chikkabalapura, India
[6]Department of Computer Science and Engineering, Cambridge Institute of Technology North Campus, Bengaluru, India

**Article Info**

***Article history:***

Received Aug 14, 2025
Revised May 6, 2026
Accepted Jun 19, 2026

***Keywords:***

Proximal policy optimization
Resource allocation
Self-adaptive penalty function
Vehicle-to-everything
Vehicular communications

**ABSTRACT**

The vehicle-to-everything (V2X) is a significant technology that improves road safety, travel experience and entertainment services. Resource allocation (RA) in vehicular networks defines the strategic distribution of communication resources like power, time slots and bandwidth among vehicles and infrastructures to provide effective data transmission. However, RA faces challenges in managing limited transmission resources because of network delays and high reception time by frequent changes in network topology and varying user demands through high mobility of vehicles. Therefore, this research proposes a proximal policy optimization with self adaptive penalty function (PPO-SAPF) based RA for vehicular communications. The PPO-SAPF optimizes the RA in dynamic vehicular networks by adjusting the coefficient matrices, which ensures better adaptability to network topology and user demands. The SAPF provides fine-tuning of policy updates, maintaining a better balance between exploration and exploitation, thereby enhancing performance under different network conditions. The PPO-SAPF achieves a less inter-packet reception time of 119 ms for 16 vehicle-to-vehicle (V2V) links in case 3 compared to context-aware RA (CARA). These results demonstrate that the proposed PP-SAPF is suitable for real-time deployment in intelligent transportation systems (ITS) and autonomous vehicles where low latency, reliable connectivity, and adaptive resource management is significant.



***Corresponding Author:***

Irshad Khan
Presidency School of Artificial Intelligence and Advanced Computing, Presidency University
Bengaluru, Karnataka 560064, India
Email: research.irshad@gmail.com

## 1. INTRODUCTION

Autonomous driving is recognised as a key solution for increasing road safety, traffic efficiency, thereby providing environmental sustainability. The modern autonomous vehicles are prepared with advanced sensors like radar, cameras and advanced driver assistance systems (ADAS) [1]. As a significant module of intelligent transportation systems (ITS), vehicular networks have gained more attention in recent years due to their potential to improve road safety, support various infotainment requirements and optimise traffic flow [2]. The urban transportation system includes traffic congestion, which leads to time loss and increased safety concerns, thereby resulting in casualties. To address this issue, the vehicle-to-vehicle (V2V) communication

technology has been employed, enabling autonomous driving and intelligent vehicle systems [3]. The V2V and vehicle-to-infrastructure (V2I) communications are combined in the vehicular network to enhance the performance of ITS [4]. The need for direct data exchange among vehicles has led to the incorporation of vehicle-to-everything (V2X) communication into long-term evolution (LTE) and its advancement in 5G new radio (NR) standards [5]. Specifically, non-safety-critical applications such as entertainment services and traffic efficiency are required to transmit lot of data frequently to base station (BS) that are handled by V2I links due to the high-capacity nature [6]-[8]. Therefore, it is significant to manage radio resources to ensure the quality of service (QoS) for vehicles in V2V network [9]. Recent studies have been conducted to manage radio resources in V2X communications that aim to enhance the total throughput, mitigate interference and reduce latency [10]. Resource sharing has emerged as a significant approach to enhance resource utilisation by allowing multiple users to reuse the same spectrum band. However, severe interference among users degrades the transmission rate that resulting in less QoS [11]. A single type of road side unit (RSU) or BS are insufficient to support V2X services when communicating with vehicle terminals [12].

Recently, the traditional V2X resource optimisation methods such as location-based, clustering-based and mode-based resource allocation (RA) in vehicular networks, have been analyzed extensively [13]. To further improve transmission performance, the RA is extended to joint optimization where the spectrum scheduling and power allocation are considered together [14]. A mixed-integer nonlinear programming issue is formulated to enhance sum ergodic ability of V2I links while ensuring a particular probability of delayed outages in V2V links [15], [16]. One possible solution is to enable the sensing functionality of the V2I system, enabling simultaneous communication and vehicle tracking within an incorporated platform [17]. Among the various challenges, the RA in high dynamic V2X communications is specifically significant, which has seen a significant rise from both academia and industry [18]. Deep reinforcement learning (DRL) has been implemented in V2V networks to manage challenges in large-scale data and decision making under ambiguity [19]. This section provides various RA techniques used in the existing research and their limitations for vehicular communications. Liu and Deng [20] developed a multi-agent DRL with attention mechanism (MADRL-AM) for RA in vehicular communication. Initially, two critic networks were applied to estimate the global and local reward functions, which achieve joint optimization of power and spectrum allocations. It balances the individual agent interests with collective advantages, thereby reaching the lower latency communication of V2V links. The AM was helpful to concentrate on critical information which adjust RA dynamically. However, the MADRL-AM has computational overhead that enhances the latency. Although various RA has developed for vehicular networks, existing approaches faces limitations such as high computational complexity, inefficient interference management, dependence on node cooperation, and higher latency in highly dynamic environments. Jin *et al*. [21] developed a DRL based two-dimensional RA for V2I communications. The RA agent in BS allocated a QoS into a two-dimensional resource block to every vehicle to enhance sum of achievable data quantity (ADQ). This results in vehicle allocation data that contains time-frequency location, TTI and bandwidth that was a two-dimensional RA solution. However, it does not address the interference, thereby leading to an impact on the performance in vehicular networks.

Liu *et al*. [22] introduced a joint double deep q-network (DDQN)-based centralized spectrum allocation and distributed power control for V2X communication. It decouples the channel matching and transmission power selection to balance the resource utilization and user experience satisfaction levels. The graph theory-based centralized spectrum scheme was introduced to minimise the multiple interference among V2V and V2I links, which enhances the system's ability of vehicle networks. However, the graph theory-based approach has computational complexity that leads to challenges in large-scale networks. Zhang and Wang [23] suggested a context-aware RA (CARA) for V2V communications in cellular V2X networks. Initially, the inter-packet reception was developed to present delay among dual successful and consecutive reception packets at the receiver end. Then, the application-specific function was developed with the utility based on vehicle safety context and packet reception performance. The game model guided every node to cooperate and obtain efficiency and fairness in channel RA. However, it depends on cooperation among nodes that leads to impracticality in dynamic vehicular environments. Chai and Ren [24] introduced a multi-objective evolutionary optimization and power control for spectrum allocation in vehicular communications. Initially, the network capacity of V2I links as optimization objective that enhances network transmission rate when providing reliability of V2V links. The power consumption efficiency was taken to manage interference in the network and generate optimal energy efficiency. The evolutionary algorithm requires a high training time which affects the adaptability in vehicular networks. Most of the existing approaches considered spectrum allocation and power control separately, the conventional PPO-based approaches rely on fixed penalty mechanism which does not adapt to change network conditions.

To address these issues, this research proposes a proximal policy optimization with self-adaptive penalty function (PPO-SAPF) for RA in vehicular communications. The proposed PPO-SAPF introduces a self-adaptive penalty mechanism which dynamically regulates the policy updates, optimize spectrum along

with power allocation thereby enhancing convergence ability, reducing inter-packet reception time and enhance transmission reliability in dynamic V2X environments. The contributions are as follows:

- PPO-SAPF optimizes the RA in dynamic vehicular networks through effectively adjusting the coefficient matrices, thereby ensuring enhanced adaptability to network topology changes and user demands.
- The SAPF enables the model to fine-tune the policy updates, maintaining a balance among exploration and exploitation, thereby enhancing performance in various network conditions.
- By using PPO-SAPF, RA is effective which resulting in reducing latency, enhanced transmission reliability and network connectivity in V2X communications.

This research paper is arranged as follows: section 2 delivers a description of the research method, section 3 explains results and discussion and the conclusion of this research is given in section 4.

## 2. RESEARCH METHOD

This research proposed a PPO-SAPF for RA in vehicular networks. The PPO-SAPF process the dynamic environment data and determines the possible states, actions and rewards for better decision-making. The PPO algorithm is applied to optimise the policy for RA and the SAPF is used to adaptively optimize and dynamically adjust the penalty values to balance exploration and exploitation. The RA decision output provides optimal spectrum and power allocation and finally this performance is measured by using the sum capacity performance of V2I links, Inter-packet reception time, the successful transmission ratio of V2V link and average V2V link rate. Figure 1 shows the block diagram for RA in vehicular communication. It denotes how the network state information from vehicles is processed by the PPO agent while SAPF controls policy updates to ensure stable learning. The resultant spectrum and power allocation decision are estimated by communication performance metrics that are given as rewards to enhance the allocation process.

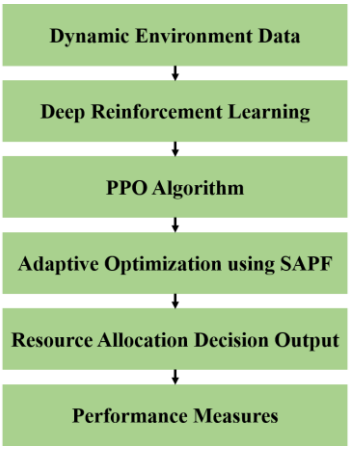


Figure 1. Block diagram of the proposed PPO-SAPF method for RA in vehicular communications

### 2.1. System model

The network contains macro base station (MBS) and $K$ vehicles, where a group of vehicles is denoted as $K$. The agent deployed in MBS dynamically assigns resource blocks to vehicles according to their received power and current occupation state of resources. The objective of agent is to enhance a sum of ADQ while considering the presence or absence QOS constraint. The resource space operates over a time duration of $T$ and its occurrence range is $B$. The numerology defines a transmission time interval (TTI) of a resource block.

When numerology of vehicle $k \in K$ is $\mu_k \in M = \{0, 1, \ldots, \mu^{max}\}$, TTI of vehicle $k$ is $T_k = 2^{-\mu_k} T^{max}$, here $T^{max}$ is a TTI for $\mu = 0$. The vehicle $k$ bandwidth is $B_k \in B = \{B^1, B^2, \ldots, B^b\}$. Here, resource space is divided into resources. The time duration is $T^{min}$ and bandwidth is $B^{min}$. The $T^{min}$ is a TTI for $\mu^{max}$ and $B^{min}$ is a minimum bandwidth. The resource element $i^{th}$ time and $j^{th}$ frequency is named as $(i, j)$ th element. The TTI and bandwidth for vehicle $k$ is separated through $T^{min}$ and $B^{min}$ as $T_k^e$ and $B_k^e$ correspondingly. The $T_k^e$ and $B_k^e$ indicates the number of resource elements are required to express the $T_k^e$ and $B_k^e$ correspondingly. The MBS assigns the resource block vertically or horizontally and assigns repeated resource to every vehicle in space. The interference from other vehicles on vehicle $k$ is denoted as (1):

$$I_k = \frac{1}{T_k} \sum_{k' \neq k} \sum_{i,j} \rho k, ij \rho k', ij \frac{P_{k'} \hat{h}_{k'}}{B_{k'}} B^{min} T^{min} \quad (1)$$

Where, $\rho k$ is a transmit power, $B_{k'}$ is a bandwidth allocated to vehicle $k'$, $B^{min}$ is a minimum gain factor to normalize channel interference, $T^{min}$ is a minimum time slot duration, $\rho k'$ is a transmit power of vehicle $k'$, $\hat{h}_{k'}$ is a channel power improvement on interference from vehicle $k'$, and $\rho k, ij \in \{0,1\}$ denotes whether vehicle $k$ uses $(i,j)$th elements. Consider that, interference energy in every resource element is estimated, summed and then divided by $T_k$ to obtain interference power. The signal interference noise ratio (SINR) of the vehicle $k$ is given by (2), while ADQ for vehicle $k$ is defined in (3):

$$\gamma k = \frac{P_k h_k}{N_0 B_k + I_k} \tag{2}$$

$$D_k = B_k T_k \, log_2(1 + \gamma k) \tag{3}$$

Where, $h_k$ is a channel power gain, $N_0$ is a noise spectral density. Consider that, instead of capacity, $D_k$ is used that is adequate metric if the $T_k$ of every user is similar as in the past RA. However, $D_k$ is affected through expanded $T_k$ that are not reflected if we utilize the capacity instead of $D_k$. The detailed system model serves as a foundation of defining optimization problem to achieve effective and reliable RA in vehicular communications.

### 2.2. Problem formation

Based on the defined system model, the RA challenge in vehicular networks are formulated as an optimization problem to enhance performance while satisfying QoS requirements for all vehicles. The optimisation task involves allocating resource blocks through repetitive resource elements to vehicles, for enhancing sum of $D_k$ when ensuring the QoS constraint for every vehicle. The optimization problem is denoted in (4):

$$\begin{aligned} &\max_{(\rho k,ij)} \textstyle\sum_{k\in K} D_K \\ &s.t. \textstyle\sum_{i=i_s}^{i_s+T_k^e-1} \rho k, ij = T_k^e, j \in \left\{j_{k,s}, j_{k,s+1}, \dots, j_{k,f}\right\}, \forall k \in K \\ &\textstyle\sum_{j=j_s}^{j_s+B_k^e-1} \rho k, ij = B_k^e, i \in \left\{i_{k,s}, i_{k,s+1}, \dots, i_{k,f}\right\}, \forall k \in K \\ &D_K > D_{min,} \, \forall k \in K, \end{aligned} \tag{4}$$

Where, $D_{min}$ is a QoS constraints, $i_{k,s}$ and $i_{k,f}$ are initial and ending location of an allocated source block for vehicle $k$ in time correspondingly. The $j_{k,s}$ and $j_{k,f}$ are the initial and ending locations of allocated source block for a vehicle $k$ in frequency correspondingly. The initial two restraints mean source blocks assigned to every vehicle are constant and final is QoS constraint.

### 2.3. Proximal policy optimization with self-adaptive penalty function

The traditional convex optimization cannot handle the sum rate optimization process in wireless networks because it changes often. The PPO based DRL [25] helps the proposed scheme solve beamforming vector and coefficient matrix optimization challenges for sum rate. The coefficient matrices are properly adjusted along with the symbol rate output. Training with PPO leads to improved network performance and enables the wireless system to operate more reliably under varying traffic loads. Figure 2 shows the framework of PPO-SAPF algorithm. The actor-critic network provides policy actions and values evaluation based on the network state while the SAPF adjusts the penalty term to control the policy updates. This framework provides effective convergence and robustness under dynamic vehicle network conditions.

The actor network receives all system inputs at the time step and creates the output variable $s_t^d$. Due to the large differences between channel information and rate feedback elements of state input data, z-score normalization normalizes the inputs when collecting all episode data for system. The actor network delivers outcomes that include both the standard deviation $\sigma(s_t^d)$ and average vector $\mu(s_t^d)$ of its policy $\vartheta_\varphi$. This phrase illustrates the Gaussian distribution structure that the policy matches as (5). The agent chooses the starting action $a_t^0$ by sampling from action distribution $\vartheta_\phi(a_t^d|s_t^d)$ while this distribution is linked to matrix coefficients and symbol rates. The integration of an alternative goal function within PPO algorithms ensures both continuous policy advancement and delivery of reliable performance. Generally, the function is expressed in (6).

$$\vartheta_\phi(a_t^d|s_t^d) = \frac{1}{\sigma(s_t^d)\sqrt{2\pi}} exp\left(-\frac{(a_t^d - \mu(s_t^d))^2}{2\sigma(s_t^d)^2}\right) \tag{5}$$

$$G(\phi)^{CPI} = \hat{\mathbb{E}}_t\left[\frac{\vartheta(a_t^d|s_t^d)\hat{B}_t{}^{\vartheta_{\phi_{past}}}}{\vartheta_{\phi_{past}}(a_t^d|s_t^d)}\right] = \hat{\mathbb{E}}_t\left[c_t(\vartheta)\hat{B}_t{}^{\vartheta_{\phi_{past}}}\right] \tag{6}$$

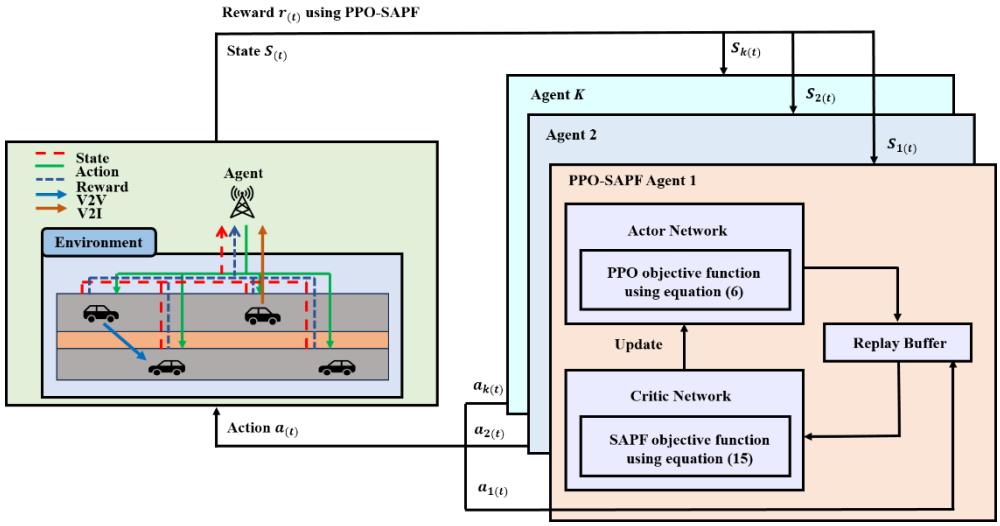


Figure 2. Internal learning structure of the PPO-SAPF algorithm under dynamic vehicle network conditions

Where $CPI$ is a conservative policy operation. The probability ratio function is represented through $c_t(\phi)$ while the calculated mean value of limited cases is $\hat{\mathbb{E}}(\cdot)$. The ratio $\frac{\vartheta(a_t^d|s_t^d)}{\vartheta_{\phi_{past}}(a_t^d|s_t^d)}\hat{B}_t^{\vartheta_{\phi_{past}}}$ computes the extent by which present actions excel or lose to typical policy moves in specific temporal contexts at time $t$. It uses the following equation to approximate the advantage function of the prior policy as (7). The evaluation function follows represented by $\hat{Q}_t^{\vartheta_{\phi_{past}}}$ is presented in (8), after redefining it through the action value $\hat{U}_t^{\vartheta_{past}}(s_t;\phi)$ is presented in (9).

$$\hat{B}_t^{\vartheta_{\phi_{past}}}(s_t^d, a_t^d:\phi) = \hat{Q}_t^{\vartheta_{\phi_{past}}}(s_t^d, a_t;\phi) - \hat{U}_t^{\vartheta_{past}}(s_t;\phi) \tag{7}$$

$$\hat{B}_t^{\vartheta_{\phi^{past}}}(s_t^d, a_t^d:\phi) = \hat{\mathbb{E}}_{a_t\sim\vartheta_{\phi_{past}}, s_t^d\sim p}\left[\sum_{t'=t}^{\prime}\gamma^{t'-t}C_{t'}\right] \tag{8}$$

$$\hat{U}_t^{\vartheta_{past}}(s_t;\phi) = \hat{\mathbb{E}}_{a_t}\sim p\left[\sum_{t'=t}^{\prime}\gamma^{t'-t}C_{t'}\right] \tag{9}$$

The absence of limitations on policy alteration between $\vartheta_\phi$ and $\vartheta_{\phi_{past}}$ leads a decrease in performance during computations when maximizing $G^{CPI}$. The problem is solved through a SAPF $G^{SAPF}(\phi)$ is defined through (10). The SAPF is presented as $g\left(\varepsilon, \hat{B}_t^{\vartheta\phi past}\right)$ which is applied to limit the policy adjustments compared to past policies as (11). The actor-network parameters are updated as in (12). During training the agent collects environmental data that the system uses to create $\hat{\mathbb{E}}_t[\cdot]$ gets calculated from what the agent learns during its environment interactions. Next, Monte Carlo sampling generates random values to evaluate the expectation operator $\hat{\mathbb{E}}$ as given in (13):

$$G^{SAPF}(\phi) = \hat{\mathbb{E}}_t\left[min\ c_t(\phi)\hat{B}_t^{\vartheta\phi past}, g\left(\varepsilon, \hat{B}_t^{\vartheta\phi past}\right)\right] \tag{10}$$

$$g\left(\varepsilon, \hat{B}_t^{\vartheta\phi past}\right) = \begin{cases}(1+\varepsilon)\hat{B}_t^{\vartheta\phi past} if \hat{B}_t^{\vartheta\phi past} \geq 0\\ (1-\varepsilon)\hat{B}_t^{\vartheta\phi past} if \hat{B}_t^{\vartheta\phi past} < 0\end{cases} \tag{11}$$

$$\phi = arg\max_{\phi}\ \hat{\mathbb{E}}_t\left[min\ c_t(\phi)\hat{B}_t^{\vartheta\phi past}, g\left(\varepsilon, \hat{B}_t^{\vartheta\phi past}\right)\right] \tag{12}$$

$$\varsigma = arg\min_{\varsigma}\ \hat{\mathbb{E}}_t\left[(U_\varsigma - \upsilon)^2, (\ clip(U_\varsigma, U_{\varsigma_{past}} - e, U_{\phi_{past}} + e) - \upsilon\ )^2\right. \tag{13}$$

No training begins until the network values. At start-up, the model parameters $\phi$ $\phi past$ and $\varsigma$ receive defined values. In SAPF, the optimizer employs $f(X)$ it directly as its penalty term. SAPF tends to less effective when the objective function value is small. In this case, the SAPF uses a set number (integer) to improve the output

of the objective function $f(X)$. The SAPF formulation is given in (14) to (16). The power consumption efficiency is estimated based on the (17).

$$SAPF = f(X) + \int \times (\sum_{i=1}^{n} g_i(X) + \sum_{i=1}^{m} h_i(X)) \quad (14)$$

$$Minimize\ f(X) = f(x_1, x_2, x_3 \ldots, x_N) \quad (15)$$

$$\begin{aligned} g_i(X) &\leq 0, I = 1,2, \ldots, n \\ h_i(X) &= 0, I = 1,2, \ldots, m \end{aligned} \quad (16)$$

$$Power\ consumption\ efficiency = \frac{P_{max}^{d} - \sum_{j \in J} y_j P_j^d}{P_{max}^{d}} \quad (17)$$

Where, $f(X)$ is objective function, $g_i(X)$ and $h_i(X)$ are imbalance and balance constraints. The agent is continuously trained until it convergence, that occurs when it obtains the specified performance level or reaches the maximum number of repetitions. Where, $j, P$ and $d$ denotes the V2V link index, total transmission power and resource block index respectively. The $P_j^d$ is an allocated power for $j$th V2V link in resource block $d$, the $P_{max}^{d}$ is a maximum allowable power over resource block. The $y_j \in \{0, 1\}$ presents power control with binary indicator, if the transmit power of $j^{th}$ DUE is allocated when value of $y_j = 1$ otherwise 0. This research utilizes SAPF that dynamically adjusts the penalty term in PPO based on policy behavior. The SAPF integrates feedback-driven adjustment strategy and threshold-based recalibration when the penalty becomes ineffective thereby ensuring continuous policy improvements in higher mobility and congested scenarios. This adaptability and robustness to environment variability is not addressed in the existing PPO-based V3X scheduling works. Moreover, this work integrates spectrum and power allocation while many existing PPO based methods handle separately. The PPO-SAPF improves resource utilization efficiency through integrating SAPF into policy gradient updates thereby resultant in enhanced convergence, and reduced inter-packed reception time. The step-by-step implementation of this optimization process is outlined in Pseudocode 1, which details the environment observation, policy updates, and adaptive penalty adjustments.

Pseudocode 1. Pseudocode for PPO-SAPF

```
Initialize PPO-SAPF parameters such as policy parameters ϕ, value function parameters ϑ,
clipping parameter ε, penalty coefficient γ,
Initialize environment with vehicles and MBS,
Set state space S and action space A for vehicles,
Start
        while not convergence do:
                Observe state s_t,
                Compute action a_t using policy as in (5),
                Perform action a_t in the environment,
                Observe reward r_t and next state s_t+1,
                Compute the advantage function using (7),
                Compute policy loss using PPO objective function as (6),
                Compute value loss using the evaluation function as (8),
                Compute SAPF penalty using (10),
                Compute SAPF range using (11),
                Compute SAPF enhanced PPO objective function using (15),
                Update policy parameters ϕ of PPO using gradient ascent as in (12),
                if SAPF penalty is ineffective:
                        Adjust the penalty term using (14),
                end if
                Store experience (s_t, a_t, r_t, s_t+1),
                 if training condition met:
                                Perform Monte Carlo estimation using (13),
                        Optimize policy and value function,
                end if
        end while
return optimized PPO-SAPF model
end
```

## 3. RESULTS AND DISCUSSION

The PPO-SAPF is simulated based on MATLAB R2020b with Intel i7 processor, Windows 10 OS and, 16 GB RAM. The performance metrics such as sum capacity performance of V2I links (Mbps), inter-packet reception time (ms), successful transmission ratio of V2V link, average V2V link rate (Mb/s), and

power consumption efficiency are estimated for RA in vehicular networks. Table 1 signifies the simulation parameters considered for this research.

Table 1. Simulation parameter

| Parameters | Values |
|---|---|
| Carrier frequency | 2 GHz |
| Channel bandwidth | 4 MHz |
| Vehicle antenna height | 1.5 m |
| Vehicle antenna gain | 3 dBi |
| Vehicle transmit power | 23 dBm |
| BS antenna gain | 25 dBi |
| BS antenna height | 8 m |
| Path-loss model | Log-distance path loss |
| Random seed | 42 |
| Mobility model | Random waypoint |

### 3.1. Performance analysis

This section describes the performance of proposed PPO-SAPF with various metrics such as sum capacity performance of V2I links, inter-packet reception time, successful V2V transmission ratio, average V2V link rate, and power consumption efficiency. The results are compared with existing methods such as RL, DRL, and PPO to illustrates the enhancements achieved by adaptive RA, interference management and policy optimization. Figures 3 to 7 provides the effectiveness of PPO-SAPF under various vehicular network condition.

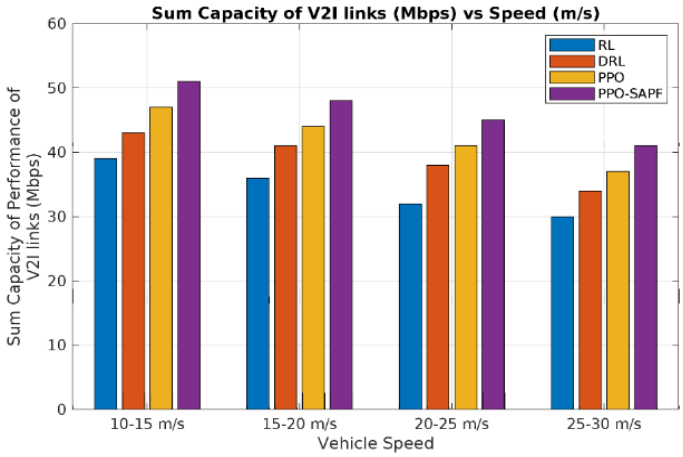


Figure 3. Sum capacity performance of V2I links (Mbps) under varying vehicle speeds

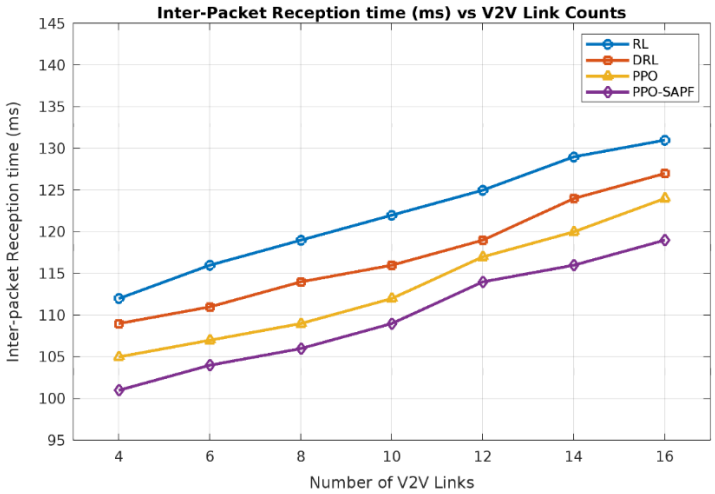


Figure 4. Inter-packet reception time (ms) for different number of V2V links

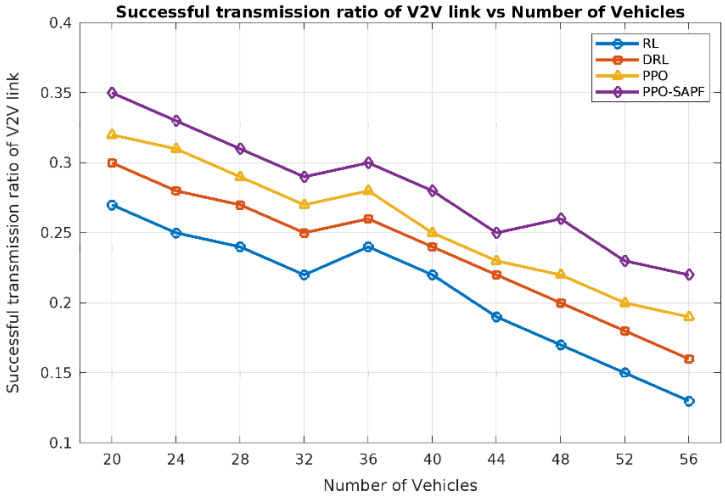


Figure 5. Successful transmission ratio for different number of vehicles

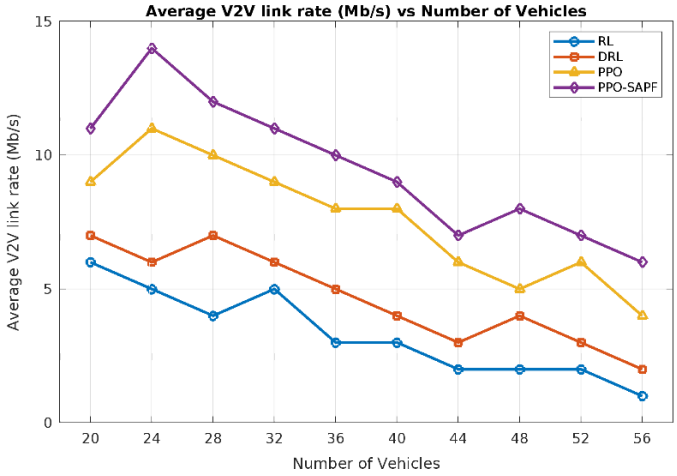


Figure 6. Average V2V link rate (Mb/s) for different number of vehicles

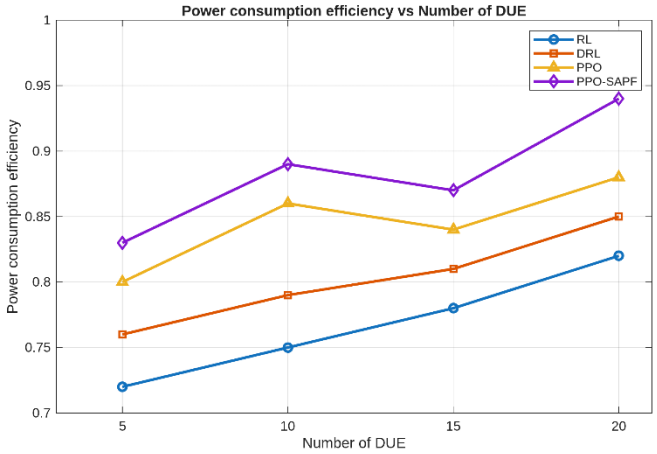


Figure 7. Power consumption efficiency for different number of DUE

In Figure 3, the performance of proposed PPO-SAPF is given for sum capacity performance of V2I links. The existing methods such as RL, DRL, and PPO are compared with the proposed PPO-SAPF. Compared to these existing methods, the PPO-SAPF achieves sum capacity of 51 Mbps, 48 Mbps, 45 Mbps, and 41 Mbps

for the vehicle speed such as [10, 15] m/s, [15, 20] m/s, [20, 25] m/s, and [25, 30] m/s. It achieves better sum capacity performance because it dynamically adjusts the reward constraints to optimize the RA. The proposed PPO-SAPF achieves higher capability compared to RL, DRL, and conventional PPO due to its adaptive policy updates and effective interference-based RA. This enhancement shows its ability to maintain the higher throughput in high-mobility environments.

In Figure 4, performance of the proposed PPO-SAPF is given for inter-packet reception time (ms). The existing methods such as RL, DRL, and PPO are compared with the proposed PPO-SAPF. Compared to these existing methods, the PPO-SAPF achieves inter-packet reception time of 101 ms, 104 ms, 106 ms, 109 ms, 114 ms, 116 ms, and 119 ms for a number of V2V links such as 4, 6, 8, 10, 12, 14, and 16. It achieves better inter-packet reception time by effectively managing RA and dynamically adjusting the communication parameters to reduce the delay. The PPO-SAPF achieves less delay across all scenarios by optimizing spectrum and power allocation dynamically thereby ensuring its suitability of the proposed method for latency-based vehicular applications.

In Figure 5, the performance of the PPO-SAPF is given for the successful transmission ratio of V2V link. The existing methods, such as RL, DRL, and PPO are compared with the proposed PPO-SAPF. Compared to these existing methods, the PPO-SAPF achieves successful transmission ratio of V2V link of 0.35, 0.33, 0.31, 0.29, 0.30, 0.28, 0.25, 0.26, 0.23, and 0.22 for number of vehicles such as 20, 24, 28, 32, 36, 40, 44, 48, 52, and 56. It achieved a better successful transmission ratio of V2V link through effectively optimizing the spectrum and power allocation when dynamically adjusting penalty constraints to reduce the interference. From 32 to 36 vehicles minor increase occurs due to adaptive adjustment of the penalty function in PPO-SAPF that temporarily enhances the spectrum reuse and minimizes the collisions. From 44 to 48 vehicles, the increase results from reinforcement-based learning, which adapts to congestion through reallocating spectrum and power effectively, thereby improving success rate and link coordination. The PPO-SAPF maintains a higher success ratio by reducing interference with its adaptive penalty control and effective spectrum reuse. The lesser performance variation reflects the learning agent's adaptation to the network congestion.

In Figure 6, performance of the PPO-SAPF is given for the average V2V link rate (Mb/s). The existing methods such as RL, DRL, and PPO are compared with the proposed PPO-SAPF. Compared to these existing methods, the PPO-SAPF attains average V2V link rates of 11 Mb/s, 13 Mb/s, 10 Mb/s, 9 Mb/s, 8 Mb/s, 6 Mb/s, 7 Mb/s, 6 Mb/s, 5 Mb/s, and 4 Mb/s for the number of vehicles such as 20, 24, 28, 32, 36, 40, 44, 48, 52, and 56. It achieved a better average V2V link rate by dynamically optimizing spectrum and power allocation, thereby ensuring efficient resource utilization when reducing interference. The dip in the figure occurs as the number of vehicles increases, which leads to higher channel congestion and enhanced interference, thereby reducing the average V2V link rate. The increase observed in certain regions, especially in PPO-SAPF due to adaptive RA and scheduling that optimises the communication effectively in particular traffic conditions. From 20 to 24 vehicles, the increase in V2V sum rate is due to active V2V links being formed where the available resources are effectively allocated among lower to moderate network loads, thereby enhancing throughput. From 44 to 48 vehicles, the rise occurs as the learning agent optimises scheduling, transmission time and managing interference, thereby enhancing link rates. The proposed PPO-SAPF achieves high link rates compared to baseline models by jointly optimizing spectrum and power allocation. The performance degradation at high vehicle densities is due to enhanced interference and channel contention.

In Figure 7, the performance of the proposed PPO-SAPF is given for power consumption efficiency. The existing methods such as RL, DRL, and PPO, are compared with the proposed PPO-SAPF. Initially, from 5 to 10 arrears, electricity consumption efficiency improves due to better RA and load balance, optimization of electricity use.

However, a slight dip is observed between 10 and 15 device-to-device (D2D) user equipment (DUE) due to intervention and suboptimal power distribution as the network size increases. After 15 DUE, the efficiency increases again, possibly adaptive power management techniques or reinforcement learning-based adaptation strategies that reduce interference and increase power allocation. Between the approaches, the PPO-SAPF displays better performance than RL, DRL, and PPO, consistently maintains high power consumption efficiency. This benefit is probably attributed to improved policy adaptation and adaptive scheduling of PPO-SAPF which effectively balances power consumption and intervention management, making it more efficient in dynamic network conditions. PPO-SAPF outperforms existing methods by balancing transmission power and interference control with adaptive learning thereby leading to enhanced energy efficiency in dynamic networks.

### 3.2. Comparative analysis

To evaluate the effectiveness and robustness of PPO-SAPF, a comparative analysis is performed over existing approaches. Each case evaluates specific scenarios, including sum ADQ, minimum ADQ, sum capacity under varying speeds, V2V transmission success probability, and power consumption efficiency. The carrier frequency, vehicle transmit power and vehicle antenna height are considered as simulation parameters and its values are presented in Table 2.

Table 2. Simulation parameters for various cases

| Parameters | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| Carrier frequency | 2 GHz | 2 GHz | 2 GHz | 2 GHz |
| Vehicle transmit power | 23 dBm | 23 dBm | 23 dBm | NA |
| Vehicle antenna height | 5 m | 1.5 m | 1.5 m | 1.5 m |

Figures 8 to 11 present the comparative analysis of PPO-SAPF with existing methods such as QoS20 [21], DDQN [22], and CARA [23]. Case 1 is for QoS20 [21], case 2 is for DDQN [22], case 3 is for CARA [23], and case 4 is for multi-objective evolutionary algorithm [24]. The results in Figures 8 to 11 show that PPO-SAPF achieves better performance, particularly in dynamic environments. The model strength remains in its adaptive penalty mechanism and DRL-based decision making that assist in optimizing spectrum reuse, minimize interference and power allocation under various network conditions.

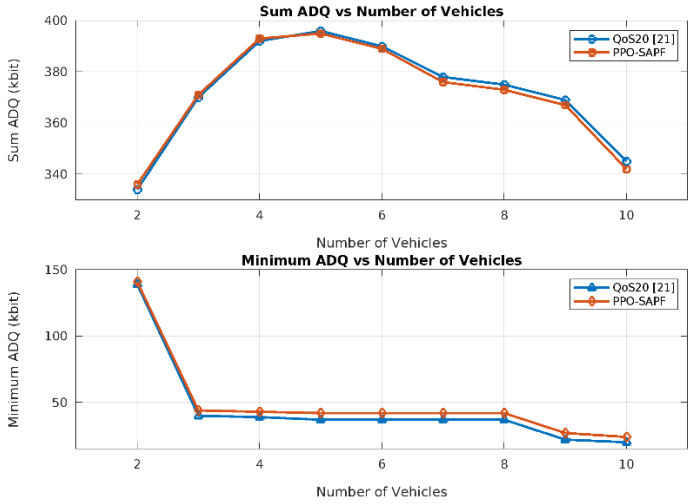


Figure 8. Comparison of PPO-SAPF for different number of vehicles in Case 1

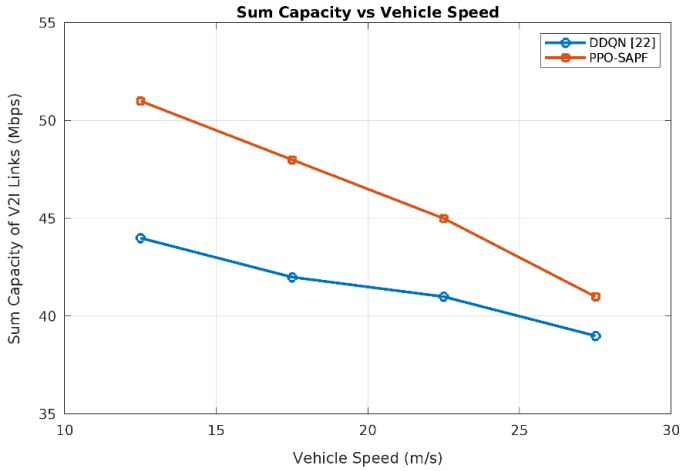


Figure 9. Comparison of PPO-SAPF for different number of vehicles in Case 2

Figure 8 compares the PPO-SAPF with existing methods under Case 1 simulation where the results shows enhanced performance in terms of ADQ thereby showing the robustness of the proposed approach under the QoS scenario. Figure 9 shows the performance comparison under Case 2 where PPO-SAPF outperforms DDQN based approaches because its reduced computational complexity and penalty regularization thereby enhancing scalability in vehicular networks. Figure 10 shows the comparative results for Case 3 where the proposed PPO-SAPF achieves less inter-packet reception time and high reliability than CARA thereby denoting its effectiveness in non-cooperative and high dynamic vehicular networks. Figure 11 compares the

PPO-SAPF with multi-objective evolutionary approach where PPO-SAPF achieves better performance with lower training time and better adaptability thereby making it suitable for V2X communication.

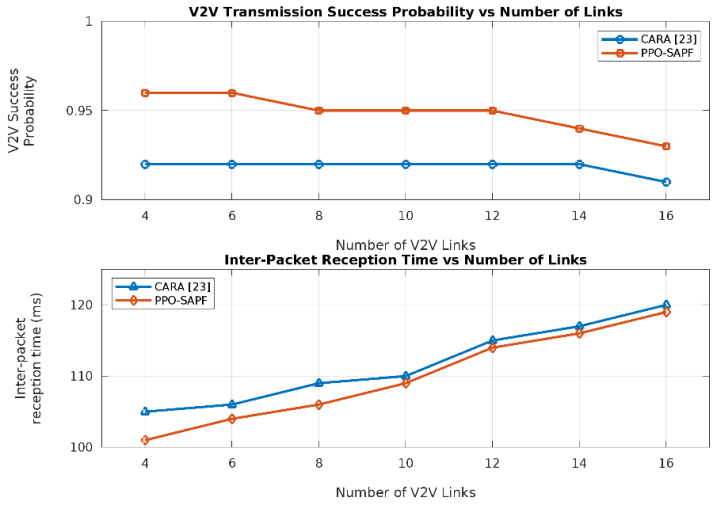


Figure 10. Comparison of PPO-SAPF for different number of vehicles in Case 3

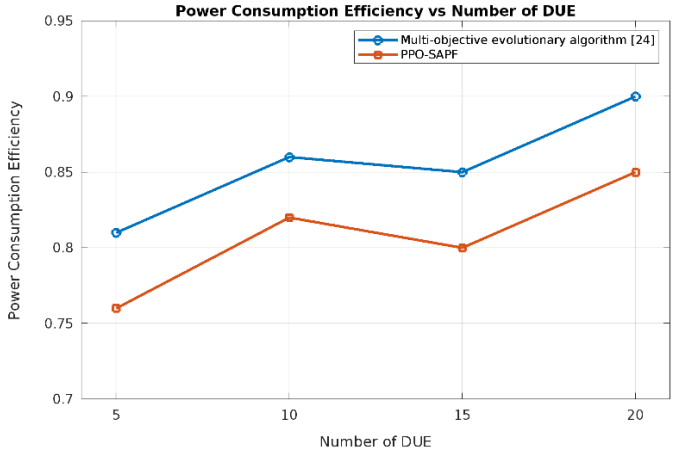


Figure 11. Comparison of PPO-SAPF for different number of vehicles in Case 4

### 3.3. Discussion

The advantages of PPO-SAPF and the limitations of existing research are explained in this section. The existing method has limitations, like QoS20 [21] does not address the interference, thereby leading to an impact on the performance in vehicular networks. In DDQN [22], the graph theory-based approach has computational complexity that leads to a challenge in large-scale networks. CARA [23] depends on cooperation among nodes that leads to impracticality in dynamic vehicular environments. The evolutionary algorithm [24] requires a high training time that affects the adaptability in vehicular networks. MADRL-AM [20] has computational overhead that enhances the latency. The proposed PPO-SAPF overcomes these limitations by dynamically adjusting its allocation parameters with changing network topology and user demands. The SAPF enables fine-tuned policy updates that ensure a better balance between exploration and exploitation, thereby leading to stable and effective vehicular communications. Furthermore, it enhances transmission efficiency, minimizes inter-packet reception time and network delays, and makes it better for real-time vehicular applications like autonomous driving, ITS and V2X communications. Compared to baseline methods, the PPO-SAPF outperforms standard PPO by its adaptive penalty to accelerate convergence and outperforms DQN and DDPG through providing better stability to handle continuous action spaces in high dynamic vehicular environments. The attained results denote that the proposed PPO-SAPF is suitable for real-time smart city and autonomous driving environments by higher vehicle density and changing network conditions. By adaptively optimizing spectrum and power allocation, the PPO-SAPF supports scalable V2X deployments. In autonomous

driving system, the safety applications required end-to-end latency within 100-120 ms that is ensured through achieved inter-packet reception time of 119 ms for 16 V2V links. Furthermore, the enhanced V2I capability, V2V transmission reliability and power efficiency allows effective traffic management, energy-aware ITS.

## 4. CONCLUSION

This research proposes a PPO-SAPF based RA framework for vehicular communications. PPO optimises RA dynamically by adjusting the coefficient matrices and ensuring better adaptability to changing network topologies and user demands. The SAPF fine-tunes the policy updates, handling optimal balance among exploration and exploitation, thereby enhancing network performance under various conditions. It significantly minimizes network delays and enhances overall communication performance which makes it better for high-mobility environments. Subsequently, PPO-SAPF reduces the inter-packet reception time and provides better capacity utilization for V2X communications. The simulation result shows the PPO-SAPF in various vehicular communication scenarios. The PPO-SAPF achieves an inter-packet reception time of 119 ms for 16 V2V links, which is better than existing approaches. This development makes the PPO-SAPF provide a better solution for ITS, real-time vehicular applications and autonomous driving where low-latency and effective communications are required. In future, other optimisation algorithms with multi-agent scenarios will be considered to further enhance the RA performance for vehicular communications.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

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

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Irshad Khan | ✓ | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | | | | |
| Neetha Papanna | | ✓ | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Umalakshmi Somshekhar Durgaiah | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| Vijetha Acharya | | | | | ✓ | | | | | ✓ | | | | |
| Suman Joseph | | | | | ✓ | | ✓ | | ✓ | ✓ | | ✓ | | ✓ |
| Varshini Totliganahalli Rajanna | ✓ | ✓ | | ✓ | | ✓ | | | ✓ | | | ✓ | ✓ | |

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

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

## REFERENCES

[1] I. Lee and D. K. Kim, "Decentralized Multi-Agent DQN-Based Resource Allocation for Heterogeneous Traffic in V2X Communications," *IEEE Access*, vol. 12, pp. 3070–3084, 2024, doi: 10.1109/ACCESS.2023.3349350.
[2] D. Yuan, D. Hu, and X. Chen, "Resource Allocation in C-V2X Mode 3 Based on the Exchanged Preference Profiles," *Electron.*, vol. 12, no. 5, pp. 1-17, Feb. 2023, doi: 10.3390/electronics12051071.
[3] J. Zhao *et al.*, "Boosting vehicular connectivity through resource allocation algorithm based on Heterogeneous Agent Proximal Policy Optimization," *Veh. Commun.*, vol. 50, Dec. 2024, doi: 10.1016/j.vehcom.2024.100856.

[4] B. Ji *et al.*, "Optimization of Resource Allocation for V2X Security Communication Based on Multi-Agent Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 74, no. 2, pp. 1849–1861, Feb. 2025, doi: 10.1109/TVT.2023.3340424.
[5] A. M. A. Ibrahim, Z. Chen, Y. Wang, H. A. Eljailany, and A. A. Ipaye, "Optimizing V2X Communication: Spectrum Resource Allocation and Power Control Strategies for Next-Generation Wireless Technologies," *Appl. Sci.*, vol. 14, no. 2, pp. 1-18, Jan. 2024, doi: 10.3390/app14020531.
[6] G. Chai, W. Wu, Q. Yang, and F. R. Yu, "Data-Driven Resource Allocation and Group Formation for Platoon in V2X Networks with CSI Uncertainty," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7117–7132, Dec. 2023, doi: 10.1109/TCOMM.2023.3311455.
[7] M. Azizi, F. Zeinali, M. Robat, and S. Shokrollahi, "Efficient AoI-Aware Resource Management in VLC-V2X Networks via Multi-Agent RL Mechanism," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 14009–14014, Sep. 2024, doi: 10.1109/TVT.2024.3392738.
[8] D. Han and J. So, "Energy-Efficient Resource Allocation Based on Deep Q-Network in V2V Communications," *Sensors*, vol. 23, no. 3, pp. 1-15, Jan. 2023, doi: 10.3390/s23031295.
[9] Y. Xu, K. Zhu, H. Xu, and J. Ji, "Deep Reinforcement Learning for Multi-Objective Resource Allocation in Multi-Platoon Cooperative Vehicular Networks," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 9, pp. 6185–6198, Sep. 2023, doi: 10.1109/TWC.2023.3240425.
[10] J. S. Alrubaye and B. S. Ghahfarokhi, "Geo-Based Resource Allocation for Joint Clustered V2I and V2V Communications in Cellular Networks," *IEEE Access*, vol. 11, pp. 82601–82612, 2023, doi: 10.1109/ACCESS.2023.3300294.
[11] Y. Xu, L. Zheng, X. Wu, Y. Tang, W. Liu, and D. Sun, "Joint Resource Allocation for UAV-Assisted V2X Communication With Mean Field Multi-Agent Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 1209–1223, Jan. 2025, doi: 10.1109/TVT.2024.3466116.
[12] S. H. An and K. H. Chang, "Resource Management for Collaborative 5G-NR-V2X RSUs to Enhance V2I/N Link Reliability," Sensors, vol. 23, no. 8, pp. 1-18, Apr. 2023, doi: 10.3390/s23083989.
[13] Z. Zhong and Z. Peng, "Joint Resource Allocation for V2X Sensing and Communication Based on MADDPG," *IEEE Access*, vol. 13, pp. 12764–12776, 2025, doi: 10.1109/ACCESS.2025.3527049.
[14] Y. Xu, X. Wu, Y. Tang, J. Shang, L. Zheng, and L. Zhao, "Joint Resource Allocation for V2X Communications with Multi-Type Mean-Field Reinforcement Learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 10, pp. 17462–17476, Oct. 2025, doi: 10.1109/TITS.2024.3484764.
[15] C. Xu, S. Wang, P. Song, K. Li, and T. Song, "Intelligent Resource Allocation for V2V Communication with Spectrum–Energy Efficiency Maximization," *Sensors*, vol. 23, no. 15, pp. 1-16, Jul. 2023, doi: 10.3390/s23156796.
[16] M. Huang, Z. Shen, and G. Zhang, "Joint Spectrum Sharing and V2V/V2I Task Offloading for Vehicular Edge Computing Networks Based on Coalition Formation Game," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 11918–11934, Sep. 2024, doi: 10.1109/TITS.2024.3371096.
[17] Y. Guo, W. Qin, Y. Xu, Y. Gu, C. Yin, and B. Xia, "Predictive Beam Tracking and Power Allocation with Cooperative Sensing for V2I Communication," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 6048–6063, 2024, doi: 10.1109/OJCOMS.2024.3416297.
[18] B. Zhao, J. Liu, B. Mao, and S. Li, "Optimal Resource Allocation for Random Multiple Access Oriented SCMA-V2X Networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10921–10932, Aug. 2023, doi: 10.1109/TVT.2023.3262274.
[19] I. Khan and M. S. Haladappa, "Risk-Aware Multi-Agent Advantage Actor-Critic Based Resource Allocation for C-V2X Communication in Cellular Networks," *Proc. Eng. Technol. Innov.*, vol. 29, pp. 47–60, Feb. 2025, doi: 10.46604/peti.2024.14136.
[20] Z. Liu and Y. Deng, "Resource allocation strategy for vehicular communication networks based on multi-agent deep reinforcement learning," *Veh. Commun.*, vol. 53, Jun. 2025, doi: 10.1016/j.vehcom.2025.100895.
[21] H. Jin, J. Seo, J. Park, and S. C. Kim, "A Deep Reinforcement Learning-Based Two-Dimensional Resource Allocation Technique for V2I Communications," *IEEE Access*, vol. 11, pp. 78867–78878, 2023, doi: 10.1109/ACCESS.2023.3298953.
[22] P. Liu, H. Cui, and N. Zhang, "DDQN-Based Centralized Spectrum Allocation and Distributed Power Control for V2X Communications," *IEEE Trans. Veh. Technol.*, vol. 74, no. 3, pp. 4408–4418, Mar. 2025, doi: 10.1109/TVT.2024.3493137.
[23] F. Zhang and G. Wang, "Context-aware resource allocation for vehicle-to-vehicle communications in cellular-V2X networks," *Ad Hoc Networks*, vol. 163, Oct. 2024, doi: 10.1016/j.adhoc.2024.103582.
[24] Z. Chai and J. Ren, "Multi-Objective Evolutionary Optimization for Spectrum Allocation and Power Control in Vehicular Communications," *Wirel. Pers. Commun.*, vol. 129, no. 4, pp. 2767–2789, Apr. 2023, doi: 10.1007/s11277-023-10257-y.
[25] S. Chaudhary, I. Budhiraja, R. Chaudhary, S. Garg, B. J. Choi, and M. Alrashoud, "Proximal Policy Optimization based sum rate maximization scheme for STAR-RIS-assisted vehicular networks underlaying UAV," *Alexandria Eng. J.*, vol. 118, pp. 700–710, Apr. 2025, doi: 10.1016/j.aej.2024.12.115.

## BIOGRAPHIES OF AUTHORS

**Irshad Khan** is currently working as an Assistant Professor in the Presidency School of Artificial Intelligence and Advanced Computing, Presidency University, Bengaluru, India. He earned his Bachelor Degree in Computer Science and Engineering from Visvesvaraya Technological University and pursued a Master's degree in Computer Networks from the same university. Currently he is pursuing his Ph.D. degree from Bangalore University in the field of AI-based Resource Allocation and Optimization Strategies for C-V2X Networks. He expertise spans diverse domains, including computer networks, V2X communication, intelligent transportation systems, artificial intelligence, machine learning, neural networks, and autonomous vehicles. He has authored 10 Scopus indexed publications, 1 Web of Science (WoS) indexed journal paper, and 2 book chapters published by Taylor & Francis and the Springer Lecture Notes in Networks and Systems (LNNS) series. He has an H-index of 3, reflecting the impact of his research contributions. He is also an active Life Member of the Indian Society for Technical Education (ISTE) and the Computer Society of India (CSI). He continues to contribute to teaching, research, and innovation in emerging areas of artificial intelligence and advanced computing. He can be contacted at email: research.irshad@gmail.com.

**Dr. Neetha Papanna Umalakshmi** is an accomplished researcher and academician specializing in deep learning applications for Alzheimer's Disease detection and prediction. Currently serving as an Assistant Professor at BMS Institute of Technology and Management, she holds a Ph.D. in Computer Science and Engineering from UVCE, Bengaluru. Her expertise spans AI, machine learning, data analysis, neural networks, blockchain, and cloud computing. She has led several innovative research projects, authored multiple publications in reputed journals and conferences, and earned recognitions including the Best Poster Award from IISc. She has one book and 2 patents in her name. She has actively contributed as a session chair, reviewer, and speaker in various academic events. With strong technical skills in Python, LaTeX, and C, and a record of impactful workshops, and certifications, she is well-versed in tackling challenges like class imbalance and multi-modality in disease classification. Her career reflects a deep commitment to research, teaching, and technological innovation in healthcare. She can be contacted at email: neethapu@bmsit.in.

**Somshekhar Durgaiah** is a part-time research scholar at the Department of Computer Science and Engineering, Visvesvaraya Technological University, India. He earned his Bachelor's degree in Computer Science and Engineering from Visvesvaraya Technological University and completed his Master's degree in Computer Networks from Bangalore University, Bengaluru, India. He expertise spans diverse domains, including blockchain, computer networks, V2X communication, machine learning, neural networks, reinforcement learning, and autonomous vehicles. His dedication to cutting-edge research is evident through the presentation of two IEEE conference papers, underscoring his passion for advancing technological understanding and contributing to the future of computer science and intelligent transportation systems. He can be contacted at email: somshekharsoma1997@gmail.com.

**Vijetha Acharya** earned her bachelor's degree in computer science and engineering from Srinivas Institute of Technology under Visvesvaraya Technological University and pursued a master's degree in computer science and engineering from St. Joseph Engineering College under Visvesvaraya Technological University. She expertise spans diverse domains, including computer networks, cloud computing, artificial intelligence, machine learning, and neural network. She can be contacted at email: vijetha0623@gmail.com.

**Suman Joseph** received his B.E. and M.Tech. degree in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi. He is pursuing his Ph.D. at Bangalore University and his current research focuses on image processing and deep learning. He can be contacted at email: sumansj1992@gmail.com.

**Varshini Totliganahalli Rajanna** is a research scholar at REVA University Department of Computer Science and Engineering, working as Assistant professor at Cambridge Institute of Technology North Campus, she earned her bachelor's degree and Master's degree in Computer science and Engineering from Bangalore University. She expertise spans diverse domains, including machine learning, artificial intelligence, neural networks, digital forensics, and cyber security. Her passion for understanding and contributing technologies to the future of computer science. She can be contacted at email: varshinivandana94@gmail.com.