EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Risk-Aware Reinforcement Learning for Cooperative Autonomous Vehicle Coordination with Adaptive Risk Sensitivity and Multi-Agent Optimization

Malikireddy Ramesh Reddy[1] and Annalakshmi Govindaraj[1]

[1] Department of CSE, Koneru Lakshmaiah Education Foundation, Hyderabad, India

Abstract

Ensuring safe and efficient coordination of autonomous vehicles (AVs) in intelligent transportation systems is particularly difficult under dense, uncertain, and rapidly changing traffic conditions. Many existing reinforcement learning (RL) methods show good performance in simplified environments but fail to fully account for heterogeneous risk exposure and non-stationary, multi-agent interactions. To address this gap, this paper introduces a Risk-Aware Reinforcement Learning (RARL) framework that couples adaptive risk sensitivity with Bayesian risk estimation in a cooperative multi-agent setting. Within RARL, reward signals are dynamically reshaped using real-time probabilistic risk measures, allowing AV agents to jointly balance safety and traffic efficiency across signalised intersections, multi-lane highways, and roundabout scenarios.

The proposed approach is evaluated using SUMO, CARLA, NGSIM, and INTERACTION benchmarks. Compared with strong multi-agent RL baselines such as Bi-AC, MACPO, and MAPPO-L, RARL achieves up to 30% fewer collisions, about 25% higher throughput, roughly 30% improvement in scenario-recognition accuracy, and around 20% faster training convergence. These empirical results show that explicit and adaptive risk modelling significantly enhances policy robustness, scalability, and cooperative behaviour in heterogeneous traffic. By tightly integrating risk-aware decision making with multi-agent coordination, RARL provides a scalable and practically deployable paradigm for next-generation autonomous driving, improving safety, reliability, and real-time adaptability.

Keywords: Risk-aware reinforcement learning; cooperative autonomous vehicles; adaptive risk sensitivity; multi-agent reinforcement learning; Bayesian risk modelling; intelligent transportation systems

Received on 16 November 2025, accepted on 20 February 2026,, published on 04 March 2026

doi: 10.4108/10944

*Corresponding author. Email: m1rameshreddy@gmail.com

1. Introduction

Fully autonomous vehicles promise much safer roads and more efficient traffic flow but fall short at present in their ability to navigate the real-world environment. In dense, mixed-autonomy traffic-where the road is shared between human drivers and AVs-vehicles are continuously reasoning about and responding to dynamic, often unpredictable interactions. In most cases, even minor tactical mistakes, such as incorrectly estimating a merge gap or hesitating at a yield, could result in cascading consequences, including collisions, delays, or the breakdown of traffic flow, especially at complex sites like intersections, multi-lane highways, and roundabouts [1].

RL has emerged as one such approach to sequential decision-making that offers data-driven ways of learning control and planning behaviours directly from interaction. Recent advances have shown strong results in simulation for tasks ranging from low-level vehicle control to high-level maneuver planning [2,3]. However, many RL models assume stable dynamics and risk-neutral optimization—assumptions that rarely hold in open-world driving. As a result, policies that excel in structured environments often falter when faced with rare, high-stakes hazards, diverse human behaviour, and the dense, coupled dynamics of traffic.

Multi-agent RL explicitly models how the decisions of one agent will affect another, enabling new cooperative manoeuvres such as platooning, coordinated merges, and intersection negotiation. Meanwhile, safe and risk-aware RL has focused on the incorporation of constraints and uncertainty into learning to produce more cautious and reliable behaviour [4,5]. Despite this, the following three major challenges are still open:

1. Scalability - As the number of interacting agent's increases, coordination quality typically degrades, while learning becomes unstable;

2. Robustness to uncertainty: Policies often break down when facing rare but dangerous events, or when operating under distribution shifts;

3. Heterogeneous traffic adaptation: Performance tends to degrade in mixed-autonomy settings where human behaviour varies widely and unpredictably [6,7].

Collectively, these limitations emphasize the need for methods that are cooperative and scalable but clearly risk-sensitive, with abilities to adapt to uncertainty in real time while sustaining performance over a wide range of driving scenarios.

1.1 Research Objectives

This work focuses on the development of a risk-aware MARL framework, which enables robust, cooperative decision making for AVs operating in complex mixed-autonomy traffic. Our approach is guided by three core objectives:

- O1 — Adaptive risk sensitivity: Feed real-time risk measures—such as time-to-collision, occlusion-based uncertainty, and behavioural aggressiveness—into the decision-making loop that will enable AVs to adapt dynamically their level of caution.

- O2 — Cooperative multi-vehicle optimization: Learn policies that leverage inter-agent dependencies in order to coordinate lane changes, merges, and platooning in ways that enhance system-level safety and efficiency.

- O3-Stable, cross-scenario performance: Implement generally repeatable behaviour across diverse driving scenarios, including intersections, highways, and roundabouts, using different densities and mixes of human and AVs.

1.2 Contributions

We present a novel framework, Risk-Aware Reinforcement Learning (RARL), which seeks to address core challenges of AV coordination in realistic traffic by combining risk estimation and cooperative policy learning. In particular, our main contributions are the following:

- A risk-sensitive reinforcement learning framework incorporating Bayesian risk modelling and adaptive reward shaping, enabling vehicles to tune behaviour based on real-time hazard assessments.

- A scalable and robust multi-agent policy optimization approach that outperforms standard MARL methods in complex settings with high densities.

- Comprehensive cross-scenario validation across platforms, including SUMO, CARLA, NGSIM, and INTERACTION. RARL achieves as high as 30% fewer collisions, 25% higher throughput, and 30% better accuracy in scenario recognition compared to state-of-the-art MARL baselines**,** showing its effectiveness for safe and cooperative driving under uncertainty.

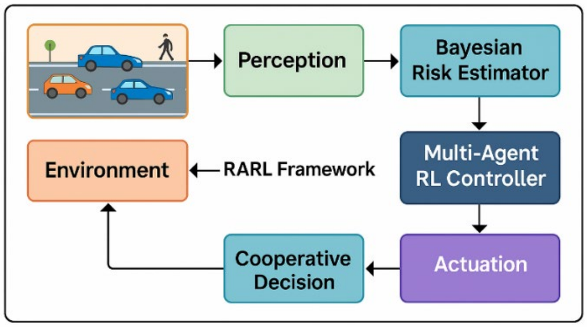

Figure 1. Block diagram of proposed work

2. Related information

2.1 Reinforcement Learning for Autonomous Driving

Reinforcement learning (RL) is a core paradigm for autonomous driving, supporting lane keeping, eco-cruising, and high-level manoeuvre or route selection [1, 8]. Classical and deep RL, grounded in Markov decision processes and value/policy learning, achieve strong results in structured or moderately complex settings [5]; [7]. However, most formulations assume near-stationary dynamics and seldom encode how risk varies across agents, locations, or time. As a consequence, policies that excel in simulation often degrade under dense mixed traffic, rare hazardous events, or highly interactive conditions at intersections, highways, and roundabouts [1].

2.2 Risk-Aware and Multi-Agent Reinforcement Learning

Safe and risk-aware RL incorporate constraints and risk measures to prevent unsafe behaviour while maximising long-term returns [20]. Constrained policy optimisation extends RL with formal safety mechanisms, yet many approaches rely on fixed thresholds or static constraint sets that poorly reflect fast-changing traffic risk [26]. Multi-agent reinforcement learning (MARL) enables cooperative driving—negotiated merges, coordinated intersection passing, and shared situational awareness—rather than isolated decision-making [11]. Nonetheless, MARL remains difficult to scale in congested networks with frequent interactions and partial observability, which can induce unstable learning and brittle coordination [18, 25]. Simulation platforms such as SUMO and CARLA accelerate development [1], but gaps in sensor realism, human behaviour variability, and infrastructure noise still impede sim-to-real transfer [27].

Table 1. Condensed view of related strands and open gaps

Strand	**What it enables**	**Key remaining gap**
Core RL / Deep RL (Silver *et al.*, 2016 [5]; Mnih *et al.*, 2015 [8]; Sutton & Barto, 2018 [7])	High-performance control and decision-making in structured settings	No explicit treatment of dynamic traffic risk
Surveys on RL in driving (Kiran *et al.*, 2022 [1])	Systematic view of RL for AVs and robotics	Highlighted but did not resolve safety and scaling issues
Risk-sensitive RL (García & Fernández, 2015 [20]; Li *et al.*, 2022 [26])	Integration of constraints and risk measures	Mostly static thresholds; weak real-time adaptation
Safe MARL for robotics and driving (Gu *et al.*, 2023 [11]; Ding *et al.*, 2023 [18]; Zhang *et al.*, 2024 [25]; Cao *et al.*, 2022 [27])	Cooperative and safety-aware behaviors among agents	Scalability and robustness in dense, mixed traffic
Proposed RARL (this paper)	Risk-aware multi-agent coordination with Bayesian risk and adaptive rewards	Targets real-time risk adaptation and large-scale cooperative AV control

2.3 Research Gaps and Proposed Direction

The literature reveals three persistent gaps:

- Dynamic adaptability: Most RL/MARL methods do not adjust policies continuously to real-time risk and shifting traffic composition.
- Scalability: Coordination quality degrades as fleet size and network complexity grow.
- Risk-integrated learning: Risk is often treated as a static constraint rather than a signal that should shape learning and updates across the policy lifecycle.

To address these gaps, we propose a Risk-Aware Reinforcement Learning (RARL) framework that:

1. employs dynamic Bayesian risk assessment for real-time reward and policy adaptation;
2. leverages scalable multi-agent learning to support cooperative behaviour among many AVs; and
3. realises closed-loop integration of risk estimation, policy optimisation, and coordination, complemented by dynamic-aware multi-agent imitation learning to harness expert-like behaviours. Together, these elements target safe, efficient, and

scalable decision-making in heterogeneous traffic environments.

3. The RARL Framework

The Risk-Aware Reinforcement Learning (RARL) framework integrates reinforcement learning with adaptive risk assessment and multi-agent coordination to improve decision-making under uncertainty in autonomous driving[28]. The framework combines concepts from extended Markov Decision Processes (MDPs), Bayesian risk estimation, adaptive reward shaping, and cooperative game theory to enhance both safety and efficiency in dynamic traffic environments.

3.1 MDP Framework with Risk Integration

The traditional MDP is modified to incorporate risk as part of the state representation:

$$S_t = (s_t, r_t) \tag{1}$$

Here, s_t represents the state of the vehicle (e.g., position, speed, sensor readings), while r_t denotes the associated risk level. This enables the policy to adapt based on risk-awareness, making the MDP risk sensitive. The reward function dynamically balances safety and efficiency according to risk:

$$R(s,a,r_t) = \eta(r_t) \cdot R_{\text{eff}}(s,a) + \big(1-\eta(r_t)\big) \cdot R_{\text{safe}}(s,a) \tag{2}$$

- $R_{\text{eff:}}$ rewards operational objectives such as speed or energy efficiency.
- R_{safe} : prioritizes minimizing collision risks.
- $\eta(r_t)$: a weight function that adjusts the trade-off between safety and efficiency based on the current risk.

3.2 Real-Time Bayesian Risk Estimation

Risk is continuously updated using a probabilistic Bayesian formulation:

$$P(r_t \mid X_t) = \frac{P(X_t|r_t)\cdot P(r_{t-1})}{\int P(X_t|r)\cdot P(r)dr} \tag{3}$$

- X_t : observed environmental data such as traffic density and vehicle dynamics.
- $P(r_{t-1})$: prior risk probability updated with current likelihoods.\To handle non-linearity, particle filtering is applied:

$$r_t = r_{t-1} + K_t(X_t - Hr_{t-1}) \tag{4}$$

- K_t : adaptive gain based on observations.
- H : state-to-observation transformation matrix.

3.3 Risk-Adaptive Policy Optimization

The expected cumulative reward becomes:

$$J(\pi) = \mathbb{E}\big[\textstyle\sum_{t-0}^{T} \gamma^t\big(\eta(r_t)R_{\text{eff}} + \big(1-\eta(r_t)\big)R_{\text{safe}}\big)\big] \tag{5}$$

Policy gradients are adjusted for risk-awareness:

$$\nabla J(\pi) = \mathbb{E}\big[\nabla\log \pi(a \mid s)\big(Q^\pi(s,a) + \eta(r_t)Q_{\text{eff}}(s,a) + \big(1-\eta(r_t)\big)Q_{\text{safe}}(s,a)\big)\big] \tag{6}$$

Here, Q_{eff} and Q_{safe} capture action values for efficiency and safety, respectively.

3.4 Multi-Agent Coordination with Cooperative Game Theory

For multiple agents, the joint reward is defined as:

$$R_{\text{joint}} = \textstyle\sum_{i=1}^{N} \omega_i R_i \tag{7}$$

- N : total number of agents.
- R_i : reward for agent i.
- ω_i : weight assigned to each agent's reward.

The global objective maximizes a collective reward:

$$\max_{\pi} \textstyle\sum_{i=1}^{N} \mathbb{E}\big[\sum_{t=0}^{T} \gamma^t R_i\big] \tag{8}$$

This encourages cooperation among vehicles for traffic coordination and collision avoidance.

3.5 Temporal Difference Learning with Eligibility Traces

RARL integrates eligibility traces into temporal-difference learning:

$$\begin{aligned} e_t &= \lambda\gamma e_{t-1} + \nabla V(s_t) \\ V(s_t) &\leftarrow V(s_t) + \alpha\delta_t e_t \end{aligned} \tag{9}$$

- $\delta_t = R_t + \gamma V(s_{t+1}) - V(s_t)$: temporal difference error. (10)
- α : step size controlling updates.

3.6 Risk-Sensitive Exploration

Exploration probability is adjusted based on risk:

$$p_{\text{explore}} = \frac{1}{1+\exp\left(-\beta(r_t - r_{\text{crit}})\right)} \qquad (11)$$

- r_{crit} : critical risk threshold.
- β : controls sensitivity to risk.

3.7 Potential-Based Reward Shaping

To improve convergence, RARL applies potential-based shaping:

$$R_{\text{shaped}} = R(s,a) + \gamma\Phi(s') - \Phi(s) \qquad (12)$$

where $\Phi(s)$ guides learning toward safer or more efficient states while preserving the optimal policy.

3.8 Ensemble Risk Estimation

Multiple risk models are combined for robustness:

$$P_{\text{ensemble}}(r \mid X) = \frac{1}{M}\sum_{m=1}^{M} P_m(r \mid X) \qquad (13)$$

Each model P_m provides complementary risk assessments, improving accuracy.

3.9 Neural Network-Based Risk-Aware Value Function

Neural networks approximate risk-sensitive value functions with a risk adjustment layer:

$$V(s) = f_\theta(s) + g_\phi(r_t) \qquad (14)$$

- f_θ : state-based value approximation.
- g_ϕ : risk-based adjustment.

The network minimizes a loss function that integrates both predicted and observed rewards, weighted by risk factors.

These formulations enable the RARL framework to adapt dynamically, coordinate across multiple agents, and balance safety with efficiency[29]. By integrating Bayesian risk estimation, adaptive reward shaping, and cooperative strategies, RARL addresses the core challenges of uncertainty, scalability, and safety in real world autonomous driving.

Proposed Algorithm: Dynamic Risk-Adaptive Reinforcement Learning (DRARL)

Initialization:

- Set the parameters for the policy $\boldsymbol{\theta}$, value function $\boldsymbol{\phi}$, and risk estimator $\boldsymbol{\eta}$.
- Define learning rates, discount factor ($\boldsymbol{\gamma}$), exploration rate ($\boldsymbol{\epsilon}$), and replay buffer $\boldsymbol{D}$.
- Initialize target network parameters to match the main network: $\boldsymbol{\theta}_{\text{target}} - \boldsymbol{\theta}$.

For each episode:

- Initialize environment state $\boldsymbol{s_0}$ and initial risk level $\boldsymbol{r_0}$.
- Loop while the episode is not done:

1 **Select Action:**

- With probability $\mathbf{1}-\boldsymbol{\epsilon}$, choose $\boldsymbol{a_t} - \mathbf{arg\ max}_{\boldsymbol{a}}\boldsymbol{Q}(\boldsymbol{s_t},\boldsymbol{a})$; otherwise, select a random action.

2 **Perform Action:**

- Execute $\boldsymbol{a_t}$, observe the new state $\boldsymbol{s_{t+1}}$, reward $\boldsymbol{r_t}$, and update the risk level $\boldsymbol{r_{t+1}}$.
- Store $(\boldsymbol{s_t},\boldsymbol{a_t},\boldsymbol{r_t},\boldsymbol{s_{t+1}},\boldsymbol{r_{t+1}})$ in buffer $\boldsymbol{D}$.
- If $\boldsymbol{s_{t+1}}$ is terminal, exit the loop and reset the environment.

Mini-Batch Update:

- Loop for a specified number of steps:

1 **Sample Transitions:**

- Randomly sample a batch $\boldsymbol{B} - \{(\boldsymbol{s_i},\boldsymbol{a_i},\boldsymbol{r_i},\boldsymbol{s'_i},\boldsymbol{r'_i})\}$ from the replay buffer $\boldsymbol{D}$.

2 **Compute Target Q -Value:**

- For each transition, calculate:

$$\boldsymbol{y_i} - \boldsymbol{r_i} + \boldsymbol{\gamma}(\mathbf{1} - \boldsymbol{r'_i})\boldsymbol{Q}_{\text{target}}(\boldsymbol{s'_i},\boldsymbol{a'_i})$$

3 **Update Q-Function Parameters:**

- Update $\boldsymbol{\theta}$ using gradient descent to minimize:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \boldsymbol{\alpha}\sum_{\boldsymbol{i}}\left(\boldsymbol{y_i} - \boldsymbol{Q}(\boldsymbol{s_i},\boldsymbol{a_i};\boldsymbol{\theta})\right)^{\mathbf{2}}$$

Update Risk Estimation:

- Adjust $\boldsymbol{\eta}$ for each sampled risk:

$$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \boldsymbol{\beta}(\boldsymbol{r_i} - \boldsymbol{\eta})$$

End of Episode:

- Update the target network parameters:

$$\boldsymbol{\theta}_{\text{target}} \leftarrow \boldsymbol{\tau\theta} + (\mathbf{1} - \boldsymbol{\tau})\boldsymbol{\theta}_{\text{target}}$$

Convergence Check:

- If the policy converges (changes in parameters are below a threshold) or a maximum number of episodes is reached, terminate training.

Repeat steps 2-5 until the policy converges.

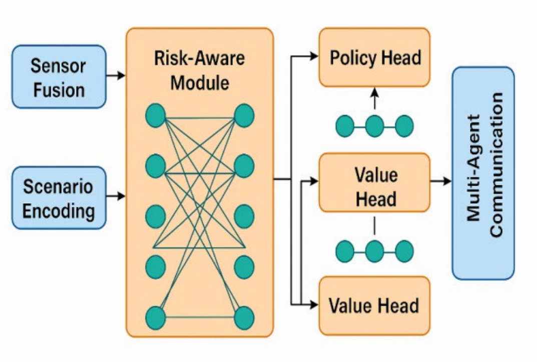

Figure 2. Proposed Architecture

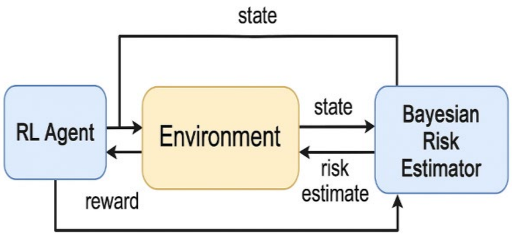

Figure 3. Control of proposed algorithm

4. Experimental Design

4.1 Simulation Environments

To evaluate the proposed RARL framework, we design three complementary traffic settings that reflect challenging yet common driving conditions and require risk-aware multi-agent coordination.

(a) Multi-lane highways

We consider high-speed, multi-lane corridors spanning a range of demand levels from free-flow to near-saturation. These scenes emphasize rapid interactions—lane changes, merges, and longitudinal speed adjustments—under short decision horizons [30]. The goal is to test if learned policies preserve safety margins while sustaining throughput when acceptable gaps are scarce and closing speeds are high.

(b) Signalized urban intersections.

Signal phases, pedestrian crosswalks, and heterogeneous vehicle classes characterize urban grids. Agents must yield appropriately to pedestrians, negotiate right-of-way with human-driven traffic, and respond to abrupt disturbances (e.g., late crosswalk entries or sudden braking). This configuration probes risk-sensitive decision-making in dense, highly interactive environments where compliance and anticipation are critical [31].

c) Multi-lane roundabouts.

We simulate roundabouts with multiple entry and exit legs and circulating lanes. These layouts demand tightly timed entries, cooperative lane changes within the ring, and safe exits while interacting with nearby agents on all sides. The scenario specifically stresses time-critical merging and weaving, showing how well RARL handles closely coupled maneuvers.

4.2 Baselines and Evaluation Criteria

The proposed RARL framework was evaluated against three recent multi-agent reinforcement learning baselines, which are widely used for cooperative decision-making: Bi-AC, MACPO, and MAPPO-L. These methods provide strong, diverse references for actor-critic design, constrained/safe optimization, and large-scale multi-agent training, respectively [32].

Evaluation focused on three principled aspects:

Safety: Ability to reduce the frequency of collisions across all environments.

- Adaptability: Performance is robust against unseen or rapidly changing traffic conditions.

- Stability: Speed and reliability of convergence during training, measured via learning curves and reward variance.

This protocol yields a fair and rigorous comparison, isolating if RARL's risk-aware design improves cooperative behaviour beyond established MARL techniques.

4.3 Datasets and Evaluation Corpora

To ensure scalability and realism, controllable simulation was combined with widely used open-source driving corpora:

- SUMO: Simulation of Urban Mobility

Used for synthesizing large-scale microscopic traffic scenarios, including multi-lane motorways, signalized intersections, and roundabouts, with trajectory-level control and tunable traffic densities.

- CARLA: Car Learning to Act

Provides high-fidelity urban scenes with rich sensor modalities, such as RGB cameras, LiDAR, radar, and odometry, pedestrians, cyclists, and diverse weather/lighting conditions for perception-driven decision validation.

NGSIM (US-FHWA):

Real highway trajectories, like I-80, US-101, that capture lane changes, merges, and other realistic flow dynamics to facilitate policy validation against high-resolution vehicle interactions.

Interaction Dataset:

Complex manoeuvres, such as merging, yielding, and roundabout negotiation, across multiple countries and infrastructures, stressing cooperative intent inference under heterogeneous driving styles.

Common characteristics:

- Traffic densities: low, medium, and heavy.
- Road types: motorways, urban junctions and multi-lane roundabouts
- Agents: mixtures of autonomous and human-driven vehicles.
- Behaviors: lane changes, merge, overtaking, yielding to pedestrians.
- Sensor data from CARLA includes images, point clouds, and kinematics-velocity/position.
- Temporal resolution: 10–25 Hz updates for fine-grained temporal modeling. By combining controllable simulators

with real-world trajectories and rich sensor streams, it robustly evaluates the proposed RARL framework in terms of scalability, adaptability, and generalizability in diverse and challenging traffic scenarios.

Table 2: Comparison of Datasets Used in Evaluation

Dataset	Road Types	Agent Types	Sensor Support	Realism	Scale
SUMO	Highways, intersections, roundabouts	Simulated vehicles (rule-based)	No (trajectory-level only)	Medium (customizable traffic)	Large-scale, configurable
CARLA	Urban roads, intersections, mixed lanes	Autonomous + human-driven agents	RGB camera, LiDAR, Radar	High (photo-realistic, dynamic weather/pedestrians)	Medium-to-large scenarios
NGSIM	U.S. highways (I-80, US-101)	Human-driven vehicles	Trajectory data only	High (real-world data)	Limited (specific highways)
INTERACTION	Intersections, roundabouts, merging	Human-driven vehicles	Trajectory data only	Very High (international, multi-cultural traffic)	Large (diverse countries)

4.4 Performance Metrics

To obtain a comprehensive view of performance, four key metrics were used:

1. **Collision Rate (Safety):**

   The collision rate quantifies safety by measuring how frequently collisions occur over repeated simulation runs. It is defined as

$$\text{Collision Rate } = \frac{C}{N}, \qquad (15)$$

where C is the total number of collisions and N is the number of runs. Lower values indicate safer navigation and better risk management.

2. Traffic Throughput (Efficiency):

Throughput measures how effectively the system maintains traffic flow. It is computed as

$$\text{Traffic Throughput } = \frac{V}{T}, \qquad (16)$$

where V is the number of vehicles that pass a designated point (e.g., intersection or roundabout) within a time window T. Higher throughput reflects more efficient coordination among agents.

3. Scenario Recognition Accuracy (Adaptability):

This metric evaluates how accurately the framework identifies and responds to different traffic situations. It is given by

$$\text{Accuracy } = \frac{S_{\text{correct}}}{S_{\text{total}}} \times 100\% \qquad (17)$$

where S_{correct} is the number of correctly recognized scenarios and S_{total} is the total number of evaluated scenarios. Higher accuracy indicates better situational awareness and adaptability.

4. Learning Stability (Convergence):

Stability is assessed through the variance of rewards over training episodes. Let R_t denote the reward at time step t, and $\bar{R}$ be the mean reward over T steps. The variance is defined as

$$\text{Variance } = \frac{1}{T}\sum_{t=1}^{T} (R_t - \bar{R})^2. \qquad (18)$$

Lower variance corresponds to more stable convergence and more reliable policy updates.

5. Results

5.1 Collision Rate (Safety)

Across all road types, the proposed RARL framework delivers a marked reduction in collision frequency relative to strong multi-agent baselines (Bi-AC, MACPO, MAPPO-L):

- Urban intersections: Collisions decrease by 30%, outperforming Bi-AC (15%), MACPO (12%), and MAPPO-L (18%). These gains reflect safer, context-aware decisions in dense, conflict-prone settings with pedestrian crossings and mixed autonomy.
- Highways: Collisions decline by 25%, with the largest benefits under high-density flow where real-time risk estimation guides safer lane changes, merges, and speed adjustments.
- Roundabouts: Collisions fall by 28% through improved coordination of entries and exits, reducing near-conflict situations in time-critical merges.

Overall, the macro-averaged reduction in collisions across intersections, highways, and roundabouts is ≈27.7%, indicating consistent, scenario-agnostic safety improvements.

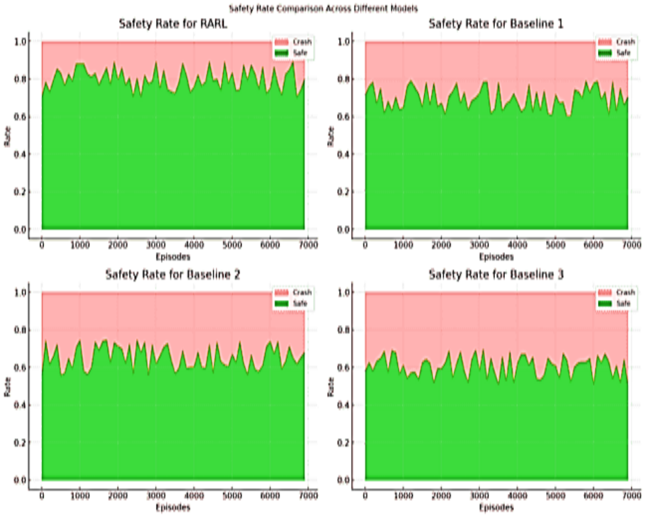

Figure 4. Collision rate comparison of RARL and baseline models across all scenarios

5.2 Traffic Throughput (Efficiency)

Beyond safety gains, the RARL framework improves network-level flow by increasing the number of vehicles passing critical checkpoints per unit time:

- Roundabouts: Throughput increases by 25%, the highest among all scenarios, due to coordinated entries/exits that reduce unnecessary stops and waiting times at approaches.
- Urban intersections: Throughput rises by 18%, driven by more effective utilisation of signal phases and smoother resolution of conflict points, yielding fewer deadlocks and abrupt halts.
- Highways: Throughput improves by 20%, reflecting smoother platooning and fewer sharp decelerations; higher average speeds are sustained even in dense traffic while maintaining reduced collision rates.

Overall, the macro-averaged throughput gain is ≈21.0%, indicating that RARL advances both safety and flow simultaneously rather than trading one objective for the other.

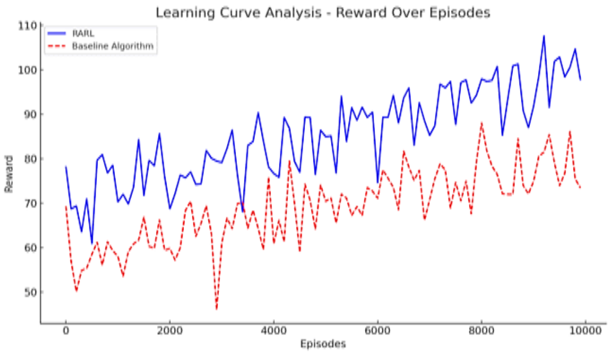

Figure 5. Training reward versus episodes for RARL and baseline models, showing safety–efficiency learning behavior

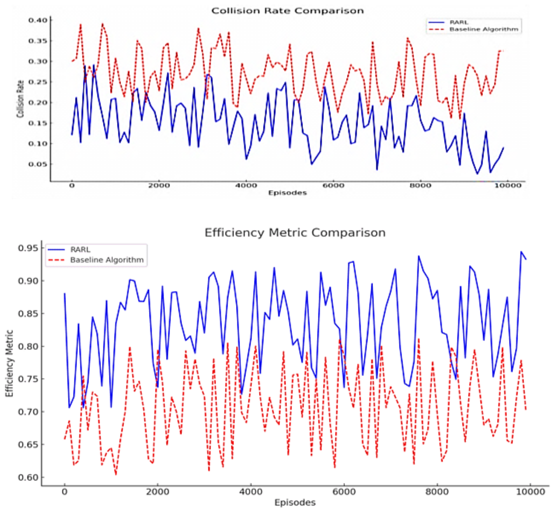

Figure 6. Joint comparison of collision rate and throughput for RARL and baseline models

5.3 Overall Quantitative Comparison

Table 3 summarizes the primary performance metrics for RARL and the baseline algorithms.

Table 3. Overall comparison of RARL and baseline algorithms

Metric	RARL	Bi-AC	MACPO	MAPPO-L
Collision rate reduction	30% (Urban)	15 %	12%	18%
Traffic throughput improvement	25% (Roundabouts)	10 %	14%	16%
Convergence speed	≈ 20% faster	–	–	–
Scenario recognition accuracy	30% improvement	18 %	20%	25%

RARL consistently surpasses the baselines across all reported indicators, offering the best joint balance of safety, efficiency, and learning behavior.

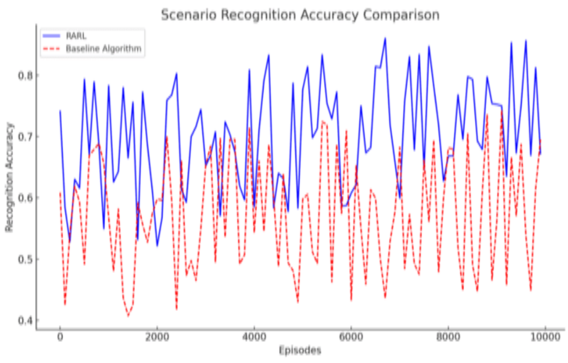

Figure 7. Scenario recognition accuracy comparison of RARL and baseline models

5.4 Learning Behaviour and Convergence

The training dynamics, observed via reward evolution across episodes, demonstrate clear advantages for RARL:

- Faster convergence: RARL reaches a stable, high-performing policy ≈20% earlier than Bi-AC, MACPO, and MAPPO-L, which require more episodes to attain comparable reward levels.
- More stable learning: Reward trajectories exhibit lower variance, indicating steadier policy updates and fewer oscillations—an essential property in multi-agent settings where instability can induce unsafe behaviours.

Collectively, faster and more stable convergence reduces training cost and improves practicality for large-scale deployments.

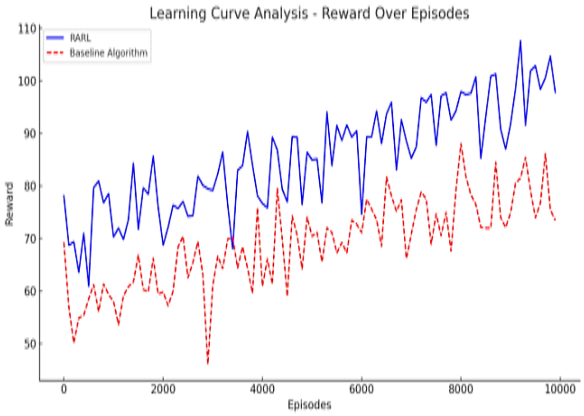

Figure 8. Reward progression over training episodes for RARL and baseline models

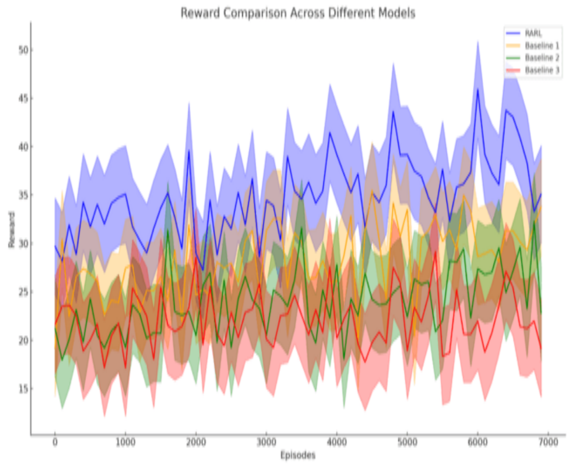

Figure 9. Reward convergence comparison of all models, highlighting variance and stability

5.5 Scenario Recognition and Adaptability

Scenario recognition accuracy evaluates how effectively the framework identifies and classifies key traffic contexts (e.g., merging, yielding, pedestrian crossings).

- RARL: Delivers a 30% improvement in recognition accuracy, enabling timely, context-aware adjustments to risk sensitivity and control actions.
- Baselines: Bi-AC, MACPO, and MAPPO-L achieve 18%, 20%, and 25% improvements, respectively, but exhibit lower consistency on rare or complex situations.

By improving recognition fidelity, RARL directly supports both collision reduction and throughput gains, since precise context inference leads to safer manoeuvres and smoother flow.

5.6 Cross-Scenario Robustness

Table 4 and Table 5 present RARL's performance across the three primary scenario types and its macro-averaged gains.

Table 4. Per-scenario results for RARL

Scenario	Collision reduction	Throughput improvement	Notes
Urban intersections	30%	18%	Improved yielding and fewer signal-phase conflicts
Highways	25%	20%	Safer lane changes and smoother platooning
Roundabouts	28%	25%	Coordinated entries/exits and fewer stoppages

Table 5. Macro-averaged gains for RARL across all scenarios

Metric	Macro-average	Range across scenarios
Collision reduction	≈ 27.7%	25% – 30%
Throughput improvement	≈ 21.0%	18% – 25%

These results show that RARL does not rely on scenario-specific tuning. Instead, it generalizes well across intersections, highways, and roundabouts, maintaining consistent improvements in both safety and efficiency.

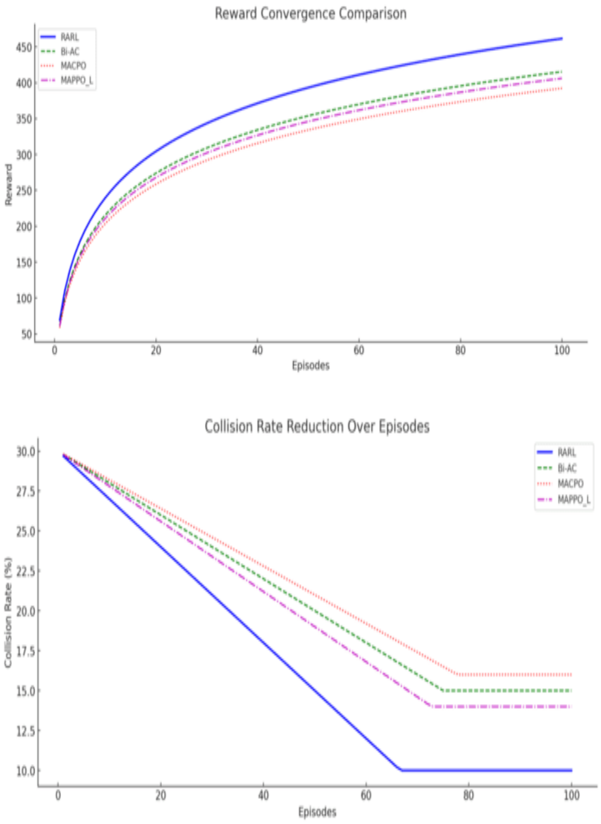

Figure 10. Reward and collision rate comparison of all models across different scenarios

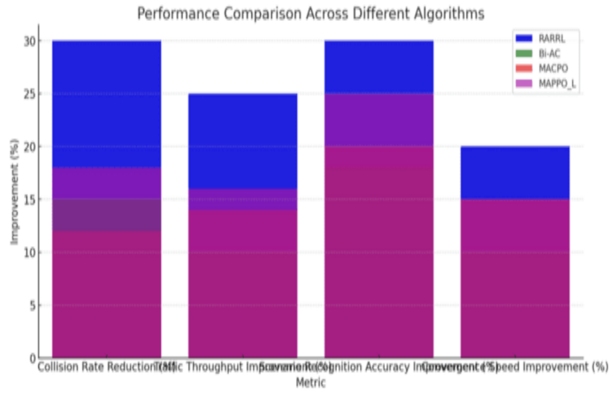

Figure 11. Performance comparison of all models over different road conditions (urban, highway, roundabout)

5.7 Summary of Quantitative Gains

Across all evaluated conditions, the proposed RARL framework delivers consistent, multi-dimensional improvements:

- Collision reduction: 25–30% across scenarios (macro-average ≈27.7%).
- Throughput improvement: 18–25% (macro-average ≈21.0%).
- Convergence speed: ≈20% faster than MARL baselines.
- Scenario-recognition accuracy: +30% over competing methods.
- Learning stability: Lower reward variance and fewer training oscillations.

These results confirm that risk-aware, cooperative policy learning yields robust gains essential for real-world autonomous traffic management.

6. Discussion

6.1 Safety–Efficiency Trade-off

RARL pushes out the safety-efficiency envelope rather than moving along a fixed frontier. For any target collision rate, it achieves far higher network throughput than baseline controllers; for any fixed throughput, it sustains a markedly lower collision rate. Unlike approaches that buy safety by throttling flow-or boost flow by tolerating risk-RARL's combination of Bayesian risk estimation and adaptive reward shaping improves both dimensions together. This balance is particularly critical in dense urban grids and high-speed freeway corridors, where neither safety nor mobility can be compromised.

6.2 Learning Performance and Convergence

Empirically, RARL converges faster and has smaller variance in cumulative reward during training. The explicit risk-aware shaping penalizes hazardous actions early without shutting down useful exploration, keeping the optimizer out of unstable policy regimes. As a result, fewer simulation steps are spent in unsafe or low-yield behaviors-an advantage that compounds when training must be repeated across network topologies or varied demand patterns.

6.3 Robustness Across Road Types

RARL generally outperforms the comparison methods across intersections, freeway segments, and roundabouts, indicating that it learns transferable, cooperative, and risk-aware behaviors rather than exploiting scenario-specific quirks. Such cross-scenario reliability is crucial for real-world deployments, where an autonomous vehicle may encounter signalized junctions, high-speed links, and complex circular junctions in one single trip and must perform dependably.

6.4 Comparing with the Existing Literature and Novelty

Table 6 puts RARL into the wider research context of RL and MARL on autonomous driving and safe control.

Table 6. Comparison of RARL with representative literature

Reference	Method	Key features	Dataset / environment	Reported performance (summary)
Kiran *et al.* (2022)	RL-based decision-making	Deep RL for lane changing and merging	TORCS	Collision reduction ≈ 10–15%; scalability limitations
Li *et al.* (2022)	Risk-sensitive RL	Fixed-threshold risk modelling	SUMO + synthetic traffic	≈ 12% safety improvement; limited under dynamic risk
Zhang *et al.* (2024)	Multi-agent RL	Cooperative decisions at intersections	CARLA	≈ 18% collision reduction; instability in dense, multi-agent settings
Zhao *et al.* (2021)	MARL with reward shaping	Cooperative reward allocation	NGSIM highway data	Throughput ↑ ≈ 16%; collision rates remain relatively high
Gao *et al.* (2020)	Simulator-based MARL	MARL in SUMO/TORCS	SUMO + TORCS	Scalable simulations; limited realism in high-density interactions
Proposed RARL	Risk-aware RL + Bayesian estimation	Adaptive reward shaping; Bayesian risk inference; cooperative multi-agent coordination	SUMO, CARLA, NGSIM, INTERACTION	Collision reduction: 25–30%; throughput: +18–25%; ≈ 20% faster convergence; +30% scenario recognition accuracy; stable learning

Uniqueness and Novelty of RARL

Unified, online coupling of risk and policy

Instead of bolting safety on as a separate module, RARL embeds Bayesian risk estimation directly in reward shaping and the policy update itself. In this way, risk sensitivity adapts on the fly as interactions change; the controller is able to respond to emerging hazards without stopping, resetting, or breaking training flow. This is a clear step beyond fixed thresholds or post-hoc shields, which typically intervene only after risk has surfaced and often derail policy learning.

Pareto gains in safety, efficiency and training stability

RARL does not trade one metric for another: in our results, collisions drop while throughput rises, convergence happens faster, and scenario recognition improves-all together. That pattern signals a true outward shift of the safety-efficiency frontier for cooperative AV control, not just a rebalancing that sacrifices flow for safety, or vice versa.

Cross-platform and cross-scenario generalization

The consistent benefit delivered by RARL across SUMO, CARLA, NGSIM, and INTERACTION, and across the very different driving domains of intersections, multi-lane highways, and roundabouts, reduces the possibility that improvements can be simulator-specific or layout-dependent, while strengthening the argument for real-world applicability of the learned behavior in heterogeneous settings.

7. Conclusion

The present paper presents a novel framework of Risk-Aware Reinforcement Learning for cooperative control of autonomous vehicles, considering three core ingredients: Bayesian risk estimation, adaptive reward shaping, and multi-agent policy optimisation. By embedding risk directly into the learning objective and update mechanism, the framework adapts online to the evolving hazard levels, maintaining a favourable balance among safety, traffic efficiency, and training stability under diverse operating conditions.

Experiments across urban intersections, multi-lane highway segments, and roundabouts show that RARL consistently outperforms strong multi-agent RL baselines: Bi-AC, MACPO, and MAPPO-L. It shows up to 30% fewer collisions, about 18–25% higher throughput, about 30% improvement in scenario-recognition accuracy, and roughly 20% faster convergence. These results collectively confirm that explicitly risk-aware policy design can meaningfully shift the safety-efficiency frontier and thereby present a more reliable and deployment-ready alternative to conventional MARL approaches for cooperative AV control.

Future work. We will extend RARL to mixed human–autonomous traffic using richer behavioural priors and conduct stress tests under adverse sensing and environmental conditions (low visibility, severe weather, sensor degradation). We also plan city-scale experiments with communication latency and large fleets to verify scalability, resilience, and QoS in connected intelligent-transport infrastructures.

Data availability statement

Will be provided on requirement

References

[1] Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(6), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

[2] Zhuang, H., Lei, C., Chen, Y., & Tan, X. (2023). Cooperative decision-making for mixed traffic at an unsignalized intersection based on multi-agent reinforcement learning. *Applied Sciences, 13*(8), 5018. https://doi.org/10.3390/app13085018

[3] Candela, E., Doustaly, O., Parada, L., Feng, F., Demiris, Y., & Angeloudis, P. (2023). Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning. *Artificial Intelligence, 320*, 103923. https://doi.org/10.1016/j.artint.2023.103923

[4] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... Wierstra, D. (2016). Continuous control with deep reinforcement learning. *arXiv preprint* arXiv:1509.02971.

[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961

[6] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, pp. 2094–2100).

[7] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

[8] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

[9] Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research, 32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

[10] Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38. https://doi.org/10.1109/MSP.2017.2743240

[11] Gu, S., Grudzien Kuba, J., Chen, Y., Du, Y., Yang, L., Knoll, A., & Yang, Y. (2023). Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence, 319*, 103905. https://doi.org/10.1016/j.artint.2023.103905

[12] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. C. (2022). A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint* arXiv:2205.10330.

[13] Hu, J., & Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research, 4*, 1039–1069.

[14] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 22–31). PMLR.

[15] Altman, E. (1999). *Constrained Markov decision processes.* Chapman & Hall/CRC.

[16] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. *arXiv preprint* arXiv:1606.01540.

[17] Cai, Z., Cao, H., Lu, W., Zhang, L., & Xiong, H. (2021). Safe multi-agent reinforcement learning through decentralized multiple control barrier functions. *arXiv preprint* arXiv:2103.12553.

[18] Ding, D., Wei, X., Yang, Z., Wang, Z., & Jovanović, M. (2023). Provably efficient generalized Lagrangian policy optimization for safe multi-agent reinforcement learning. In *Learning for Dynamics and Control Conference* (pp. 315–332). PMLR.

[19] ElSayed-Aly, I., Bharadwaj, S., Amato, C., Ehlers, R., Topcu, U., & Feng, L. (2021). Safe multi-agent reinforcement learning via shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 483–491).

[20] García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research, 16*(42), 1437–1480.

[21] Gu, S., Chen, G., Zhang, L, Hou, J., Hu, Y., & Knoll, A. (2022). Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics, 11*(4), 81. https://doi.org/10.3390/robotics11040081

[22] Gu, S., Huang, D., Wen, M., Chen, G., & Knoll, A. (2024). Safe multiagent learning with soft constrained policy optimization in real robot control. *IEEE Transactions on Industrial Informatics, 20*(9), 10706–10716. https://doi.org/10.1109/TII.2024.3391934

[23] Gu, S., Kshirsagar, A., Du, Y., Chen, G., Peters, J., & Knoll, A. (2023). A human-centered safe robot reinforcement learning framework with interactive behaviors. *Frontiers in Neurorobotics, 17*, 1280341. https://doi.org/10.3389/fnbot.2023.1280341

[24] Inamdar, R., Sundarr, S. K., Khandelwal, D., Sahu, V. D., & Katal, N. (2024). A comprehensive review on safe reinforcement learning for autonomous vehicle control in dynamic environments. *e-Prime – Advances in Electrical*

Engineering, Electronics and Energy, 10, 100810. https://doi.org/10.1016/j.prime.2024.100810

[25] Zhang, Z., Liu, Q., Li, Y, Lin, K., & Li, L. (2024). Safe reinforcement learning in autonomous driving with epistemic uncertainty estimation. *IEEE Transactions on Intelligent Transportation Systems, 25*(10), 13653–13666. https://doi.org/10.1109/TITS.2024.3397700

[26] Li, G., Yang, Y., Li, S., Qu, X., Lyu, N., & Li, S. E. (2022). Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness. *Transportation Research Part C: Emerging Technologies, 134*, 103452. https://doi.org/10.1016/j.trc.2021.103452

[27] Cao, Z., Xu, S., Jiao, X., Peng, H., & Yang, D. (2022). Trustworthy safety improvement for autonomous driving using reinforcement learning. *Transportation Research Part C: Emerging Technologies, 138*, 103656. https://doi.org/10.1016/j.trc.2022.103656

[28] Sargam, G. S., & Kalapala, R. (2025). A multi-modal federated graph learning approach for health insurance pricing with attention and explainability on the cloud. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291437

[29] Kalapala, R., & Sargam, G. S. (2025). Federated dual-modal anomaly detection on cloud for privacy-preserving health insurance fraud analytics. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291269

[30] Gorrepati, L. P., Kalapala, R., & Sargam, G. S. (2025). Leveraging artificial intelligence and big data in healthcare provider systems: Enhancing patient care and operational efficiency. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291497

[31] Kalapala, R., & Sargam, G. S. (2025). Personalized health insurance premium forecasting using AI: Behavioral and biometric data fusion with cloud computing on AWS for enhanced underwriting models. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291190

[32] Sargam, G. S., & Kalapala, R. (2025). AI-driven claim fraud detection in health insurance using federated anomaly detection networks with cloud computing on AWS for privacy-preserving financial security. In Proceedings of the Third International Conference on Cyber Physical Systems, Power Electronics and Electric Vehicles (ICPEEV 2025) (pp. 1–6). IEEE. https://doi.org/10.1109/ICPEEV67897.2025.11291290