Network Environment-Aware Drone Path Planning in Last-Mile Medical Supply Delivery

Aneesh Calyam, Subrahmanya Chandra Bhamidipati, Sharan Srinivas
University of Missouri-Columbia, USA; Email: {ackfw, sb5q6, srinivassh}@missouri.edu

***Abstract*—Drones have immense potential in medical supply delivery during disaster incident response. However, drones often face dynamic environments with uncertainties (e.g., obstacles and wind), and intermittent network connectivity, requiring navigation adaptation and rerouting to ensure safety and mission success under a given constrained energy budget. In this paper, we present a novel cross-layered network environment-aware Drone Path Planning (DPP) model based on Q-Learning, designed to adapt drone delivery missions in dynamic environments given a battery budget to optimize deliveries. Our model utilizes network awareness in state representations including the drone's position, proximity to wind, telemetry status, and its general direction to adapt to uncertainties. We conduct simulations within a realistic grid environment to evaluate the performance of our DPP model. The results demonstrate DPP model's robust performance in optimizing path decisions and ensuring timely, efficient deliveries, achieving better rewards compared to the state-of-the-art A* path-finding algorithm for medical supply delivery.**

***Index Terms*—Drone Trajectory Planning, Dynamic Decision-making, Reinforcement Learning, Network Uncertainties**

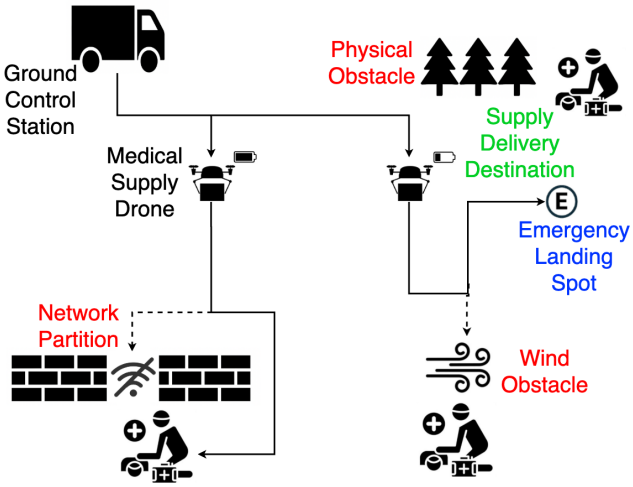

Fig. 1. Illustration of drone path planning in a dynamic environment with obstacles, wind and low connectivity.

I. Introduction

During medical emergencies in disaster response, prompt and timely intervention is crucial in saving patient lives. Drones, in conjunction with first responders, show great potential in delivering essential medical supplies such as defibrillators (AEDs), vaccines, and drugs. Drones can specifically speed up delivery time and operate in inaccessible physical spaces with proper aerial guidance [1]. However, when faced with uncertain terrain and intermitted network connectivity that are common in disaster scenes, the safety and reliability of drones remains a significant challenge for ground operators [2].

To address these challenges, there is a need for dynamic path-planning to adapt drone navigation to ensure mission success and safety. For instance, a drone's path needs adaptation in real-time due to physical environment obstacles (e.g., trees, buildings) as well as sudden adverse weather conditions such as wind gusts that disrupt preplanned drone trajectory as shown in Figure 1. Further, in disaster scenes, drones can lose telemetry data (e.g., location awareness data) due to poor network connectivity causing the drone to significantly deviate from its preplanned flight path. Common approaches such as A* use static information when determining flight path [3]. In disaster scenarios, these approaches are not suitable because they do not have network environment awareness.

Based on above issues, there is a need to develop network environment-aware path planning approaches that can maximize medical supply delivery, especially under constrained battery conditions. The approach must address geographical distances between the ground control station (GCS), patients at disaster scenes and varying sizes of physical obstacles under a set of given energy constraints. The approach must be able aware of these dynamic factors and adapt in a timely manner to sudden changes in e.g., wind, loss of telemetry data.

In this paper, we address the problem of drone path planning when faced with dynamic environments and uncertainties. Specifically, we present a novel Drone Path Planning (DPP) model designed to adapt drone delivery missions to adapt to dynamic environments to optimize deliveries under a constrained battery budget. The DPP model determines whether to deliver supplies, return to the GCS in case of an issue, or divert to an emergency landing location, as shown in Figure 1. The DPP model uses a Q-learning based Reinforcement Learning (RL) approach [4]. This approach leverages a novel RL state encoding that jointly models wind turbulence, network connectivity, battery depletion, and mission status, enabling cross-layer adaptive decision making.

We evaluate the performance of our DPP model in extensive simulations involving realistic disaster scenarios and medical supply delivery mission requirements. Our experiments show the drone's ability to navigate efficiently and safely in a dynamic environment with a give energy budget, and maximize mission success in terms of successful deliveries to multiple destinations. Using reward metrics, we compare

our DPP model against the state-of-the-art A* solution [5].

The remainder of this paper is organized as follows: Section II provides an overview of the related works for drone path planning/adaptation. Section III details the problem formulation relevant to medical supply delivery involving drone path planning in uncertain environments. Section IV describes the decision-making framework featuring core mechanics of the proposed Q-learning approach to adapt the drone path plan. Section V presents the performance evaluation of our DPP model in extensive simulation scenarios. Finally, Section VI concludes the paper.

II. Related Work

The potential of drone based medical supply delivery is tremendous in its ability to promptly and efficiently deliver medical supplies especially in rural and inaccessible areas [6]. Drones can deliver life saving supplies such as AEDs, vaccines, and drugs when minutes make the difference in patient survival. However, the areas in which drones must navigate present many challenges to successful delivery. Authors in [2] addressed the medical supply delivery problem in a rural area. They found that drones often lose telemetry data such as drone localization mid-flight due to the low network connectivity in remote areas. The team also found that sudden adverse weather conditions inhibited the preplanned drone path. The authors in [7] identified that drone energy consumption is a major issue to be solved in the context of drone medical supply delivery. These challenges motivate our solution by presenting real world barriers to implementation of drone medical supply delivery that must be addressed when creating a drone path planning model.

Current common solutions involve path planning using the A* path finding algorithm [5]. In static environments, A* has been proven to always find the fastest and most efficient routes for drone delivery. However, A* is not suitable in situations where there are partially unknown or uncertain environments. A* is unable adapt to these changes in its environment. The D* path finding algorithm was presented as a solution to the issue with the A* algorithm [8]. However, this solution faces challenges when dealing with high-dimensional problems due to the long computing time [9]. This time lost is impractical in the medical supply delivery case where providing timely care is essential to saving lives. To address these challenges, recent solutions primarily use RL-based algorithms to ensure adaptable path planning in dynamic environments. The authors in [7] used RL to optimize drone deliveries because of its ability to outperform heuristic approaches such as A* and some variants of D* and its adaptability to dynamic environments. However current state-of-the-art RL solutions fail to consider real world environmental changes such as wind and low intermitted network connectivity. The lack of a cross-layer state representation presents a barrier to the use of RL as a means of dynamic path planning. These issues motivate our solution that uses a novel cross-layered RL algorithm to maximize delivery while ensuring safety using network environment awareness under an energy budget while considering real world environmental changes.

III. Problem Formulation

In this section, we formulate the medical supply delivery problem and define the parameters for mission success in the presence of physical obstacles and network communication uncertainties.

A. Assumptions

We assume that the disaster scene environment can be represented in the form of a grid. This allows us to place physical obstacles and determine distances between the GCS and supply delivery destinations where first responders and patients are located. The grid environment allows us to track the drones position, movement, and velocity in a measurable way quantitatively. Dynamic factors such as the wind and low network connectivity can be represented as zones in the grid environment. Physical obstacles (e.g., trees, buildings) are also represented as zones. The sizes of the zones indicate areas where the flight trajectory of the drone may be affected. We represent supply delivery destinations and drone landing spots, both emergency and ground control stations in the grid environment. We also assume that the drone moves at a constant velocity, and once a drone reaches a supply delivery destination, the medical supply handoff occurs.

In our model, each zone is an area the drone may navigate through, however the drone will face consequences or obstructions when traversing a zone in a preplanned path. Hence, the goal is to use network environment awareness to evade these zones and avoid mission failure and ensure drone safety. We modeled three types of harsh zones the drone must navigate around. Firstly, we modeled physical obstacles that the drone must evade to avoid obstacle collision and damage to the drone. Secondly, we model wind as a zone factor which randomly appears, disappears, reappears throughout the drones mission. Upon entering such wind zones, the drone must use more energy, which drains limited battery resources. Lastly, we model low network connectivity zones which also randomly appear, disappear, and reappear throughout the drone's mission. When the drone enters such zones, it runs the probability of losing control of its actions, causing the drone to be unable to continue its mission.

B. Research Questions

The driving research question from the problem context is - *How can the drone adapt its path when faced with dynamic factors in the form of zones to ensure drone safety and mission success while also working under an energy budget*?

From the context of the solution approach - *How can an RL-based algorithm be configured for it to be effective in completing mission requirements when compared with state-of-the-at algorithms such as A**?

From the context of the contribution significance - *How can we optimize medical supply deliveries to save lives given a set of patient locations, constrained drone resources, and network environment awareness between the GCS and the destinations*?

IV. Learning-Based Mission Adaptation Strategy

This section outlines the novel RL-based approach used to enable *cross-layered network environment-awareness* for mission adaptation in drone-based medical supply delivery. The objective is to train the drone to make timely and context-aware routing decisions in disaster response environments characterized by uncertainty and limited energy availability. Unlike prior work that primarily focused on maintaining network connectivity [10], our proposed model emphasizes *endurance-aware* and *energy-budgeted* planning, where the agent learns to optimize the number and success of deliveries under a finite energy budget. Further, our model is cross layered in that it jointly integrates physical disturbances (e.g., wind), network connectivity uncertainties (e.g., telemetry loss), and application-layer mission constraints (e.g., energy budget and delivery progress) into a single RL state representation. Through repeated interaction with a dynamic environment containing wind disturbances and low network signal strength regions, the model acquires an adaptive decision policy that balances timeliness, efficiency, and safety. This cross-layered integration is absent in existing drone path planning frameworks for medical supply delivery. The following subsections describe the complete learning architecture, including the state formulation, control space, reward structure, and energy-constrained decision process.

A. State Representation and Action Selection

The proposed model formulates the medical supply delivery problem as a sequential decision-making process, where the drone must make time-critical navigation choices under dynamic and uncertain conditions. The operational environment is modeled as a discretized two-dimensional grid in which each position encodes not only spatial information but also contextual factors that affect the timeliness of deliveries. The state simplistically for two patient destinations at time t can be defined as -

$$s_t = \{x_t, y_t, f_1, f_2, b_t, \omega_t, \lambda_t\}$$

where (x_t, y_t) represents the drone's current coordinates, f_1 and f_2 are binary indicators signifying whether deliveries to two patients have been respectively completed, b_t denotes the remaining battery level, ω_t indicates whether the drone is currently affected by wind interference, and λ_t reflects low-signal exposure. The cross-layered state representation thus encodes both the physical position, environmental hazards, and the temporal urgency associated with each mission phase—capturing whether the drone is on schedule to safely complete time-sensitive medical deliveries within its given limited energy budget. In this approach, state transitions evolve not only as a function of distance traveled or obstacles encountered but also with respect to elapsed time, energy consumption, and progress toward mission completion. This enables the RL agent to develop policies that prioritize *timely*, *energy-efficient*, and *situation-aware* decision-making for medical supply delivery.

The action space consists of a set of discrete motion primitives, encompassing both directional movements and goal-oriented navigation. The available actions are drawn from -

$$\begin{aligned}\mathcal{A} = \{&\text{up},\ \text{down},\ \text{left},\ \text{right},\ \text{up-left},\ \text{up-right},\\ &\text{down-left},\ \text{down-right},\ \text{to-patient1},\ \text{to-patient2},\\ &\text{to-GCS},\ \text{to-emergency}\}.\end{aligned}$$

Each action updates the drone's position deterministically within the grid while incurring a variable energy cost dependent on the local environment. This hybrid control design allows both precise obstacle avoidance and high-level navigation toward mission-critical locations. The RL agent incrementally learns which of these actions yield optimal long-term rewards given the combined influence of distance, energy, and hazard exposure.

B. Energy-Dependent Adaptation Logic

Battery depletion serves as the central constraint that regulates the RL agent's decision-making process, effectively defining an operational *energy budget* for each mission. Energy consumption is modeled as -

$$b_{t+1} = b_t - (0.01\, v_t + 0.5\, \mathbb{I}_{\omega_t}),$$

where v_t denotes the instantaneous movement magnitude, and $\mathbb{I}_{\omega_t}$ equals 1, when the drone traverses a wind-affected region. This formulation integrates environmental influence directly into the depletion dynamics, ensuring that every navigational choice incurs a measurable cost against the finite energy budget. The coupling of the reward and energy models reinforces budget-awareness: aggressive maneuvers in turbulent or resistive zones accelerate depletion, decreasing the likelihood of timely mission completion. Conversely, conservative routing and steady pacing preserve the remaining budget, extending the feasible delivery horizon. Through repeated interactions, the agent learns to manage its limited energy reserve as a strategic resource when balancing speed, safety, and mission coverage under the imposed energy budget to maximize overall medical supply delivery success.

C. Reward Feedback and Policy Optimization

To ensure that the learned policy reflects the operational goals of *timely*, *reliable*, and *energy-aware* medical delivery, the reward function is redesigned using a structured and physically grounded formulation rather than large, ad hoc constants. The reward $R(s_t, a_t)$ combines mission outcomes, energy expenditure, hazard exposure, and a potential-based shaping term that continuously encourages progress toward delivery targets. This structure ensures that the agent receives feedback proportional to the operational significance of its actions, allowing it to prioritize timely deliveries while avoiding unsafe or inefficient behavior.

$$\begin{aligned}R(s_t, a_t) = R_{\text{event}} - c_{\text{energy}}\,\Delta E_t - c_{\text{wind}}\, I_{\text{wind},t} - c_{\text{signal}}\, I_{\text{signal},t}\\ - c_{\text{step}} + \eta\big(\Phi(s_{t+1}) - \Phi(s_t)\big).\end{aligned}$$

Each component serves a distinct operational purpose, and the chosen magnitudes are justified by the relative impact of these factors on real medical missions.

Mission Event Rewards: Mission-critical events are given moderate but dominant rewards or penalties to ensure that the policy is driven primarily by successful delivery and safety-related outcomes:

$$R_{\text{event}} = \begin{cases} +250, & \text{successful and timely delivery,} \\ +100, & \text{safe return to GCS,} \\ +50, & \text{controlled emergency landing (low battery),} \\ -250, & \text{collision or energy depletion,} \\ 0, & \text{otherwise.} \end{cases}$$

These values are deliberately chosen to be an order of magnitude larger than stepwise movement costs but smaller than the scale of cumulative shaping feedback. This reflects real-world priorities: completing deliveries and maintaining safety dominate over intermediate concerns such as local turbulence or temporary signal degradation.

Energy and Hazard Costs: Energy and hazard penalties are proportional to their operational burden rather than being fixed extreme constants. The term $c_{\text{energy}}\Delta E_t$ penalizes energy consumption based on the drone's actual power usage at time t, encouraging energy-efficient flight paths. Wind and low-connectivity penalties,

$$c_{\text{wind}}\, I_{\text{wind},t}, \qquad c_{\text{signal}}\, I_{\text{signal},t},$$

reflect the real consequences of environmental exposure: wind increases energy draw and delays progress, while low connectivity elevates the risk of telemetry loss. Their magnitudes are set higher than the per-step cost yet lower than mission-critical penalties, capturing their intermediate operational importance. These coefficients were selected through bounded parameter sweeps to identify stable learning behavior while preserving realistic mission trade-offs.

A small constant c_{step} discourages unnecessary movement and directly supports timely goal completion by preventing idle wandering.

Potential-Based Shaping: To provide dense and consistent guidance toward timely deliveries, a potential-based shaping term is used:

$$\Phi(s) = -d_{\text{goal}}(s),$$

where $d_{\text{goal}}(s)$ is the Euclidean distance to the active delivery target or return location. The difference

$$\Phi(s_{t+1}) - \Phi(s_t)$$

rewards incremental reductions in distance and penalizes regress. This shaping term ensures the learning process remains forward-directed and time-aware without overshadowing mission event rewards. The coefficient η is tuned such that shaping accelerates convergence while preserving the correct priority ordering of mission-critical outcomes.

Policy Optimization: The resulting reward integrates mission urgency, safety, and energy awareness into a unified learning signal. The control policy $\pi(s)$ is optimized using Q-learning, with the action-value function updated via the Bellman equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big[R(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big]. \tag{1}$$

The learning rate α governs the influence of new experience, while the discount factor γ balances immediate progress toward *timely delivery* against long-horizon mission success. Exploration is regulated through an ϵ-greedy strategy, with ϵ decaying over training to transition from exploration to exploitation. Each episode terminates upon successful delivery, safe landing, or energy depletion. Over repeated episodes, this structured and justified reward design yields a stable policy capable of fast, energy-efficient, and hazard-aware routing in dynamic environments.

D. Computational Efficiency Analysis

The computational feasibility of the proposed learning framework is evaluated in terms of its time and space complexity. For a discretized environment characterized by a state space $\mathcal{S}$ and an action space $\mathcal{A}$, the asymptotic complexities of the tabular Q-learning implementation are expressed as

$$T = \mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|\,N), \qquad M = \mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|),$$

where N denotes the number of training episodes required for convergence. The time complexity stems from iterative Q-value updates across all state–action pairs, while the space complexity arises from storing the corresponding entries within the Q-table.

In the medical delivery domain, the state space expands proportionally with spatial grid resolution, the number of binary delivery indicators, and the environmental awareness parameters representing wind and signal variations. The action space, however, remains fixed at twelve discrete movement primitives, thereby bounding the action-dependent growth of the Q-table. For the simulated environment, a discretization of 20×20 pixels within an 800×800 grid yields approximately 1.6×10^3 distinct spatial states prior to incorporating delivery and hazard attributes—maintaining tractability for real-time computation on standard hardware.

The inclusion of awareness-based features such as wind zones, signal interference, and dynamic energy tracking does introduce additional computational overhead. However, this overhead directly enhances the agent's contextual understanding and mission reliability. By modeling real-world uncertainties within the learning loop, the framework achieves more accurate energy forecasting, route stability, and hazard avoidance. These capabilities ultimately improve the timeliness and success rate of medical deliveries—an outcome that far outweighs the marginal cost in processing and storage.

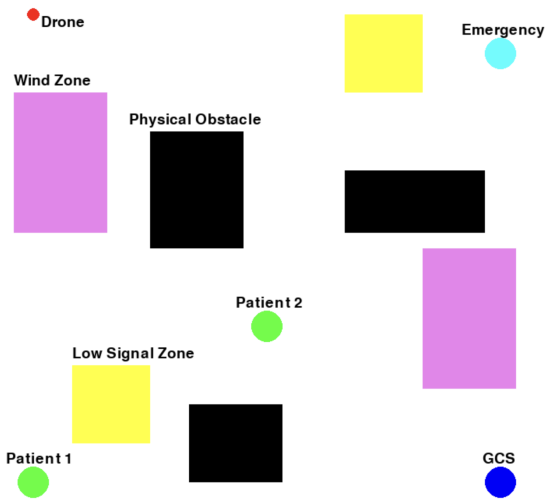

Fig. 2. Grid Environment Setup

While the time complexity reflects iterative learning over multiple episodes, the trade-off is justified by the resulting convergence toward policies that minimize mission duration and ensure timely delivery under energy constraints. In critical medical logistics, where every second of delay can endanger patient outcomes, such overhead is a necessary investment in operational precision. Thus, the modest increase in computational demand is offset by significant gains in situational awareness, efficiency, and reliability, affirming the practicality of the proposed reinforcement learning framework for real-world time-sensitive missions.

V. Performance Evaluation

In this section, we first describe the experimental testbed and simulation parameters used to evaluate our proposed DPP approach. We then outline the objectives of the experiments, the selected performance metrics, and the baselines used for comparison. Finally, we present the key experimental results and salient findings derived from the simulations.

A. Experimental Setup

The experimental evaluation is conducted in a controlled grid-based simulation environment designed to emulate realistic operational challenges for drone-based medical supply delivery. The simulation testbed is represented as an 800×800 pixel domain, corresponding to the drone's operational area. Within this environment, the drone encounters static and dynamic entities, including physical obstacles, wind zones, low-connectivity regions, multiple patient delivery locations, a GCS, and an emergency landing zone. The simulation layout is designed to maintain a consistent and repeatable configuration while preserving sufficient variability for adaptive learning.

Grid Environment: The simulation grid is discretized into 20×20 pixel cells to simplify the action and state spaces. Each cell encodes environmental features relevant to decision-making, such as obstacle presence, wind interference, or connectivity loss. Obstacles are placed in fixed positions to simulate urban structures, while navigable corridors ensure the existence of feasible routes to both patients and landing sites. This configuration maintains a balance between controlled experimentation and realistic path complexity.

Wind Zones: Wind zones are dynamically generated regions that simulate turbulent or energy-intensive airspaces. A new configuration of wind zones is initialized at the start of each episode, and zones may shift or disappear as the episode progresses. The stochastic nature of these zones introduces environmental unpredictability, requiring the drone to continuously assess whether traversing a wind zone is worthwhile relative to the mission timeline and remaining energy budget. This dynamic modeling captures the operational reality where atmospheric conditions can change mid-mission.

Low-Connectivity Zones: Low network connectivity zones represent areas of intermittent communication with the GCS. These zones appear and vanish randomly during each episode, forcing the drone to develop awareness of signal reliability when planning routes. Although the drone can choose to traverse such areas, prolonged operation within them may increase risk due to reduced telemetry and delayed control feedback. This element introduces realistic communication constraints commonly encountered in remote or disaster-affected environments.

Patients: Multiple fixed patient locations are defined within the environment, representing delivery destinations for medical supplies. Each patient is modeled as a circular region with a predefined radius. When the drone enters this radius, the delivery is registered as successful, and the state variables e.g., assuming two patient locations f_1 or f_2 are updated accordingly. The spatial separation of patients introduces a multi-objective routing challenge, requiring the RL agent to plan drone path sequences that optimize both timeliness and energy expenditure.

Landing Zones: Two designated landing sites are integrated into the environment for our study purposes i.e., the GCS and an emergency landing area. Both are represented as circular safe zones. A mission is considered complete when the drone successfully returns to either site after completing deliveries or initiates an emergency landing due to low battery. The inclusion of multiple recovery options allows the agent to learn context-dependent termination strategies based on energy level, distance, and remaining delivery tasks.

This simulation testbed thus provides a controlled yet dynamic platform to evaluate our DPP approach under conditions that reflect the trade-offs faced in real-world medical supply delivery missions, while balancing safety, timeliness, and energy efficiency in the presence of unpredictable environmental disruptions.

B. Evaluation Methodology

The proposed framework is evaluated through extensive simulations comprising 100,000 training episodes to ensure robust policy convergence and time-efficient learning behavior. Each episode represents a complete delivery mission in which the drone interacts with a dynamic environment containing

obstacles, wind disturbances, and low-signal regions. The learning process is governed by three key parameters—the learning rate α, the discount factor γ, and the exploration rate ϵ—which together control the balance between exploration, exploitation, and temporal optimization of deliveries. The selected parameter values are tuned to achieve convergence within a feasible training horizon.

During each episode, the drone follows an ϵ-greedy policy, selecting an action a_t based on the current state s_t, observing the transition to s_{t+1}, receiving a reward $R(s_t, a_t)$, and updating the corresponding Q-value. The ϵ parameter decays gradually across episodes, reducing random exploration and increasing reliance on the learned Q-table. This decay mechanism enables extensive exploration early in training while shifting toward exploitation in later stages, refining trajectory selection for timely and energy-efficient delivery.

The exploration–exploitation balance is achieved using a linear decay schedule defined as -

$$\epsilon = \max(0.01,\ 0.9 - \frac{\text{episode}}{50000}),$$

where the initial exploration rate is 0.9, gradually decreasing to a minimum threshold of 0.01 as training progresses. This formulation allows the agent to explore extensively during early episodes to discover feasible routes and environmental responses, and later exploit the learned Q-values to perform consistent, time-optimal navigation within the given energy budget. The gradual reduction in ϵ thus ensures that the policy transitions smoothly from stochastic exploration to deterministic decision-making, yielding an adaptive and convergent control strategy for *timely*, *energy-efficient*, and *mission-aware* medical supply delivery operations.

C. Experimental Results

This subsection presents a comparative evaluation between the proposed DPP approach that uses network environment-awareness vs. the classical A* baseline. The analysis focuses on three major aspects of performance: reward convergence, environmental hazard avoidance, and adaptive mission behavior under uncertainty. The results collectively demonstrate that the RL agent learns to balance timeliness, energy efficiency, and safety—achieving superior adaptability compared to the deterministic A* planner.

1) Reward Progression and Convergence: Figure 3 shows the cumulative reward trajectory for the RL-based scheme used in the DPP approach over $100{,}000$ episodes. The reward trajectory shows an initial period dominated by negative values due to collisions, unnecessary movement, and repeated exposure to wind and low-signal regions. As training progresses and the agent acquires policies that minimize energy usage and avoid hazardous zones, the cumulative rewards steadily increase and stabilize. This stabilization reflects a transition from exploratory, failure-prone behavior to consistent execution of safe and efficient routing decisions. The smoothed moving average of the reward curve confirms stable convergence of the learned policy.

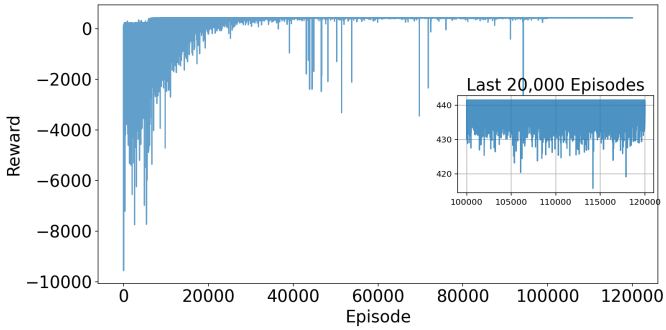

Fig. 3. Cumulative reward progression for RL-based delivery over training episodes, showing convergence toward stable policy behavior.

In contrast, Figure 4 illustrates the performance of the A* baseline, which remains nearly constant after initialization. As a deterministic planner, A* produces static path costs and cannot adapt its route based on environmental dynamics such as wind or low network signal strength related interference.

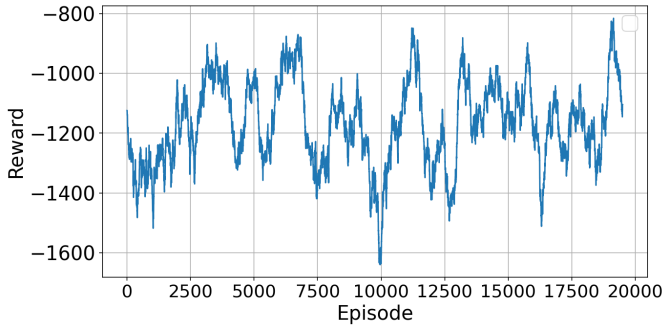

Fig. 4. A* baseline reward progression remaining static across episodes, reflecting its non-adaptive behavior.

2) Wind Zone Avoidance: Figures 5 and 6 compare the number of wind zone entries across episodes for the RL-based DPP and A* methods, respectively. The RL agent initially enters turbulent regions frequently but learns to avoid them as training progresses, resulting in a clear decline in entries over time. This adaptive avoidance directly reflects the learned risk–reward balance: the drone only traverses high-resistance zones when time constraints justify the energy expenditure. In contrast, the A* planner demonstrates a near-uniform distribution of wind encounters, indicating its lack of network environmental awareness, and inability to adjust drone route selection in a dynamic manner.

3) Low-Signal Zone Awareness: Figures 5 and 6 show similar trends for low-network-connectivity encounters. The RL agent rapidly reduces its exposure to low-signal regions as training advances, maintaining reliable connectivity and improving communication safety. The A* baseline again exhibits little variation, entering low-signal areas with nearly constant frequency throughout. This contrast confirms that the RL-based DPP internalizes network environment constraints as part of its decision process, learning to prioritize reliable data link continuity while adhering to a given energy budget.

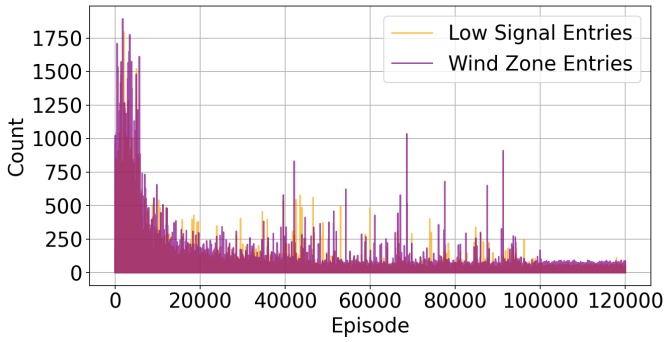

Fig. 5. Wind zone and low-signal entries per episode for the RL framework, showing a sharp decline as the agent learns to avoid them.

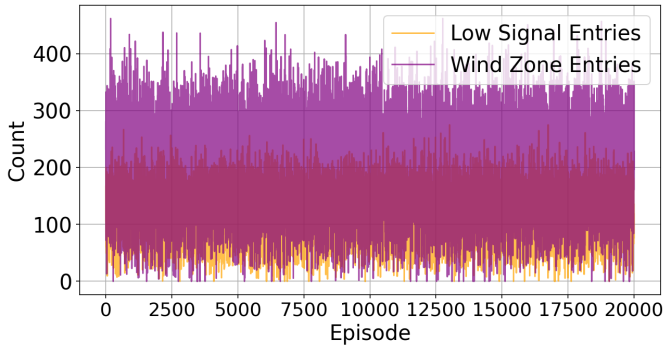

Fig. 6. Wind zone and low-signal zone entries for A* baseline, showing no adaptive avoidance across episodes.

4) Overall Performance Interpretation: Across all indicators, the RL-based DPP demonstrates progressive refinement of operational efficiency and network environment awareness. The RL agent autonomously learns to minimize risk exposure, balance energy expenditure, and achieve timely medical supply deliveries without explicit human supervision. The deterministic A* baseline, while spatially optimal does not have network environment awareness. Hence, it remains insensitive to real-world dynamics, often traversing hazardous zones to minimize the path length in a detrimental manner for mission success. These findings substantiate the claim that RL-based DPP enables *mission-aware*, *time-sensitive*, and *energy-efficient* decision-making—key attributes for real-world medical drone supply delivery.

VI. Conclusion

In this paper, we addressed the problem of adapting the DPP for medical supply delivery when a drone is faced with an uncertain terrain and a dynamic environment during a disaster response scenario. Specifically, we presented a novel cross-layered Q-learning based algorithm for adapting the DPP to ensures drone safety, mission success, and timeliness given a constrained battery budget during medical supply delivery.

By integrating RL into our DPP algorithm, the supply delivery drone becomes aware of its dynamic environment, allowing it to navigate through uncertainties such as physical obstacles (e.g., buildings, trees), wind zones, and low network connectivity zones. The drone is also aware of its battery budget and battery depletion, creating more realism and complexity in the DPP flight adaptation.

We conducted a comparative study where our RL-based DPP performs in a practical manner in comparison with the state-of-the-art A* algorithm in terms of reward accumulation and drone safety in disaster incident scenes for medical supply delivery. While A* fails in dynamic environments, our Q-learning in the DPP algorithm enabled continuous adaptation due to its network environment awareness, making it more robust when faced with dynamic environment factors. Additionally, our DPP approach integrates drone energy-awareness, enabling the drone to adapt the flight path given a battery budget to ensure drone safety during medical supply delivery missions. Our findings thus demonstrate successful and timely delivery of supplies, even in a dynamic environment with obstacles, wind and low network connectivity, which in turn helps save more lives in disaster scenarios.

Future work can involve testing the DPP benefits in real-world testbeds with realistic zone, energy, and signal-strength patterns, as well as extending the evaluation to include dynamic replanning baselines such as D* to assess performance against online, environment-aware path planners.

Acknowledgments

This research was supported by the National Science Foundation (NSF) under Award No. CNS-2313887. The views presented are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] A. A. Nyaaba and M. Ayamga, "Intricacies of medical drones in healthcare delivery: Implications for africa," *Technology in Society*, vol. 66, p. 101624, 2021.

[2] S. Aggarwal, P. Gupta, N. Mahajan, S. Balaji, K. J. Singh, B. Bhargava, and S. Panda, "Implementation of drone based delivery of medical supplies in north-east india: experiences, challenges and adopted strategies," *Frontiers in Public Health*, vol. 11, p. 1128886, 2023.

[3] A. Gasparetto, P. Boscariol, A. Lanzutti, and R. Vidoni, "Path planning and trajectory planning algorithms: A general overview," *Motion and operation planning of robotic systems: Background and practical approaches*, pp. 3–27, 2015.

[4] C. Qu, R. Singh, A. E. Morel, F. B. Sorbelli, P. Calyam, and S. K. Das, "Obstacle-aware and energy-efficient multi-drone coordination and networking for disaster response," pp. 446–454, 2021.

[5] A. K. Guruji, H. Agarwal, and D. Parsediya, "Time-efficient a* algorithm for robot path planning," *Procedia Technology*, vol. 23, pp. 144–149, 2016.

[6] S. Sharma and H. Sharma, "Drone a technological leap in health care delivery in distant and remote inaccessible areas: A narrative review," *Saudi journal of anaesthesia*, vol. 18, no. 1, pp. 95–99, 2024.

[7] G. Wu, M. Fan, J. Shi, and Y. Feng, "Reinforcement learning based truck-and-drone coordinated delivery," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 754–763, 2021.

[8] C. Saranya, M. Unnikrishnan, S. A. Ali, D. Sheela, and V. Lalithambika, "Terrain based d algorithm for path planning," *IFAC-PapersOnLine*, vol. 49, no. 1, pp. 178–182, 2016.

[9] G. Gugan and A. Haque, "Path planning for autonomous drones: Challenges and future directions," *Drones*, vol. 7, no. 3, p. 169, 2023.

[10] S. C. Bhamidipati, A. Maxwell, E. Pham, J. Zhang, Z. Murry, A. E. Morel, C. Qu, S. Srinivas, and P. Calyam, "Q-learning-based dynamic drone trajectory planning in uncertain environments," in *2025 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2025, pp. 709–715.