Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Certificate-Driven Learning for Safety-Critical Systems: Reward-Free Policy Synthesis via Adversarial Certificate-Margin Optimization

Mehrshad Eskandarpour
Iran University of Science & Technology
Mohammadjavad Jannati
mjannati@iust.ac.ir

Iran University of Science & Technology
Amir Javadi
Iran University of Science & Technology

Article

Keywords: Certificate-Driven Learning, Safe Reinforcement Learning, Control Barrier Functions, Stability, Adversarial Falsification

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10458922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Certificate-Driven Learning for Safety-Critical Systems: Reward-Free Policy Synthesis via Adversarial Certificate-Margin Optimization

M. Eskandarpour, M. Jannati, A. Javadi
School of Electrical Engineering, Iran University of Science & Technology, Tehran, Iran
Corresponding Author: mjannati@iust.ac.ir

***Abstract*—Safety-critical learning systems require controllers that can be audited for constraint satisfaction under adverse operating conditions, not only policies that perform well on average. This paper proposes Certificate-Driven Learning (CeDL), a reward-free policy-synthesis framework that jointly learns a control policy and an accompanying mathematical certificate. CeDL replaces expected-return maximization with worst-case certificate-margin maximization over a declared admissible scenario domain. The certificate unifies Lyapunov stability margins, control-barrier safety margins, temporal-logic robustness, and feasibility margins into a single training objective. An adversarial falsifier searches for scenarios in which the current policy-certificate pair is closest to violation, and the resulting low-margin cases are retained in a hard-scenario bank for counterexample-guided optimization. A runtime safety shield is further introduced as a final feasibility-preserving mechanism when the nominal policy approaches the certified boundary. The framework is developed for general stochastic discrete-time systems and instantiated in three representative domains: URLLC-aware wireless resource allocation, energy-safe routing in wireless sensor networks, and robotic obstacle avoidance. Simulation results show that CeDL improves minimum certificate margins, reduces observed constraint violations, lowers URLLC tail delay, extends sensor-network lifetime, and increases robotic safety distance compared with reward-based reinforcement learning, constrained reinforcement learning, model predictive control, shielded reinforcement learning, and classical controllers. These results indicate that optimizing certifiability directly can provide a principled and practical path toward auditable learning-based control in safety-critical networked and cyber-physical systems.**

***Index Terms*—Certificate-Driven Learning, Safe Reinforcement Learning, Control Barrier Functions, Stability, Adversarial Falsification.**

I. Introduction

Learning-based control [1] has become an important tool for complex engineering systems [2], including wireless networks, autonomous robots [3], cyber-physical platforms [4], and energy-constrained sensor systems. [5] These systems operate in dynamic environments [6] where model uncertainty, nonlinear dynamics [7], and high-dimensional decision spaces make classical controller design difficult. [8] Deep reinforcement learning and related data-driven methods have shown strong empirical performance in such settings, mainly by optimizing a reward function that represents average task performance. [9]

However, average performance is not sufficient in safety-critical and service-critical systems. [10] A controller that performs well on nominal scenarios may still be unacceptable if it occasionally violates a hard safety constraint, misses a strict communication deadline, depletes a critical sensor node [11], or drives a robot into an unsafe region. [12] In these applications, rare events often dominate system risk. [13] Therefore, the key requirement is not only to obtain high expected performance, but also to provide auditable evidence that the learned policy satisfies the required constraints within a declared operating domain. [14]

Existing safe-learning methods address this issue from several directions. Constrained reinforcement learning incorporates safety through constraints or Lagrangian penalties. [15] Shielded reinforcement learning and predictive safety filters correct unsafe actions at runtime. [16] Certificate-based methods use Lyapunov functions, control-barrier functions, or related proof objects to verify stability and safety. [17] More recent approaches also learn neural Lyapunov or barrier certificates and combine them with data-driven control. [18] These developments have significantly improved the safety of learning-based controllers. [19]

Nevertheless, an important gap remains. In many existing methods, safety is still secondary to reward maximization, introduced as a penalty, a constraint, a post-training verification step, or an external correction layer. [20] In other cases, the certificate is specialized to a single property, such as stability or pointwise safety, and is not integrated into a general training objective for policy synthesis. [21] As a result, the learned policy may perform well on average but may not be explicitly trained to strengthen its weakest certified behavior. [22]

This paper proposes Certificate-Driven Learning (CeDL), a reward-free framework for policy synthesis in safety-critical

systems. The central idea is to make certifiability itself the object of learning. Instead of maximizing expected reward, CeDL jointly learns a policy and a certificate by maximizing the worst-case certificate margin over an admissible scenario domain. The certificate margin measures how far the closed-loop trajectory is from violating stability, safety, temporal-logic, and feasibility requirements. A positive margin indicates certified satisfaction, a zero margin indicates operation on the certification boundary, and a negative margin indicates certificate violation.

The novelty of CeDL is not merely the use of Lyapunov functions, barrier functions, temporal-logic robustness, adversarial falsification, or runtime shielding in isolation. These elements have appeared in different forms in prior work. The key contribution is their integration into a single reward-free synthesis principle: the policy and certificate are trained together to improve the weakest certified behavior discovered by adversarial falsification. In this view, counterexamples are not used only for validation. They become training signals that identify where the current policy-certificate pair is most vulnerable.

The proposed framework uses an adversarial falsifier to search for low-margin scenarios within the declared operating domain. These scenarios may correspond to wireless traffic bursts, deep fading, interference, low-energy relay nodes, adverse routing conditions, near-collision robot states, moving obstacles, or bounded model uncertainty. The lowest-margin scenarios are stored in a hard-scenario bank and repeatedly used to update the policy and certificate. This counterexample-guided process shifts the training emphasis from nominal behavior to boundary behavior, which is often the most important regime in safety-critical deployment.

CeDL also includes a runtime safety shield. However, the shield is not the main source of safety in the proposed framework. In conventional shielded learning, the policy may be trained primarily for reward, while the shield corrects unsafe actions after the policy decision. In contrast, CeDL trains the nominal policy itself to increase certificate margins. The runtime shield is therefore used as a final feasibility-preserving layer when the learned policy approaches the certified boundary.

The framework is developed for general stochastic discrete-time systems and instantiated in three representative domains: ultra-reliable low-latency communication-aware wireless resource allocation, energy-safe routing in wireless sensor networks, and robotic obstacle avoidance. These applications are different in physical structure, but they share a common requirement: the controller must perform reliably under adverse conditions, not only under average operating scenarios. The same certificate-margin principle is used to represent deadline reliability, energy preservation, and collision avoidance.

The main contributions of this paper are summarized as follows.

First, a reward-free certificate-margin formulation is proposed for safety-critical policy synthesis. The learning objective maximizes the weakest certificate margin over an admissible scenario domain rather than maximizing expected reward.

Second, a unified composite certificate is introduced by combining Lyapunov stability margins, barrier safety margins, temporal-logic robustness margins, and feasibility margins into one differentiable worst-case objective.

Third, an adversarial falsification mechanism is embedded inside the training loop. The falsifier identifies low-margin scenarios close to certificate violation, and these scenarios are retained in a hard-scenario bank for counterexample-guided policy-certificate optimization.

Fourth, a runtime safety shield is developed as a final feasibility-preserving mechanism. Unlike conventional shielded learning, the shield complements a nominal policy that is already trained to improve certificate margins.

Fifth, the proposed framework is evaluated across wireless resource allocation, sensor-network routing, and robotic obstacle avoidance. The results show that CeDL improves worst-case margins, reduces observed violations, and provides an interpretable assurance structure for safety-critical learning-based control.

The remainder of this paper is organized as follows. Section II reviews related work in safe reinforcement learning, certificate-based control, and reliability-aware networked systems. Section III introduces the system model and certificate-margin problem formulation. Section IV presents the proposed CeDL algorithm, including certificate construction, adversarial falsification, policy-certificate optimization, and runtime shielding. Section V reports the simulation results and comparative analysis. Section VI concludes the paper and discusses future directions.

II. Related Works

A. Safe and Constrained Learning for Control

Safe learning has become an important research direction as learning-based controllers are increasingly considered for real-world engineering systems. Recent surveys and studies on safe reinforcement learning show that most existing approaches preserve the classical reward-maximization structure while adding safety constraints, risk measures, shielding mechanisms, or constrained Markov decision process formulations [23]. These methods have improved the practical safety of learning algorithms, especially in robotics, autonomous systems, and networked control. However, in many cases, safety remains secondary to the reward objective and is handled through penalties, constraints, or corrective layers rather than as the primary learning target.

A related direction combines reinforcement learning with stability analysis. For example, model-based safe reinforcement learning methods have used Lyapunov arguments and

uncertainty estimates to restrict exploration to regions where stability can be maintained [24]. Predictive safety filters have also been introduced to supervise learning-based controllers by checking whether a proposed action may lead to future constraint violations [25], [26]. These methods are valuable because they provide a runtime mechanism for preventing unsafe behavior. Nevertheless, the safety filter is often external to the learning objective. The policy is usually trained for performance, while safety is enforced afterward by an additional supervisory module.

Other works separate the learning policy from a recovery or backup policy that intervenes when constraint violation becomes likely [27]. This improves exploration safety and reduces unsafe trials, but the learned task policy is not necessarily accompanied by a formal certificate explaining its safe operating region. In contrast, the present work aims to make certifiability itself the object of learning. Rather than optimizing a reward and then applying a safety correction, the proposed framework jointly learns the policy and its mathematical certificate, using the weakest certified cases as the main training signal.

B. Certificate-Based Safety and Formal Specifications

Control theory provides a mature foundation for safety and stability through Lyapunov functions, control-barrier functions, and formal verification tools. Control-barrier-function-based quadratic programs have become a standard approach for enforcing safety-critical constraints by minimally modifying a nominal controller [28]. Subsequent work has extended this framework to broader classes of systems and applications, including robotic systems, multi-agent systems, and safety-critical autonomy [29]. High-order control-barrier functions further allow safety constraints to be handled when they have higher relative degree with respect to the control input [30].

Several works have also studied nonsmooth and generalized barrier certificates for systems where classical smoothness assumptions may be restrictive [31], while converse results have clarified the theoretical relationship between safety properties and the existence of barrier certificates [32]. More recent monographs have organized these developments into systematic frameworks for safe autonomy and learning-based safety-critical control [33], [34]. These studies show the strength of certificate-based methods: they produce interpretable mathematical conditions that can be inspected, verified, and used at runtime.

Formal methods provide another complementary view of safety and correctness. Temporal logic has been widely used to express trajectory-level requirements that cannot be captured by simple pointwise state constraints [35]. Signal temporal logic and robustness-based semantics make it possible to measure how strongly a trajectory satisfies or violates a specification, which supports optimization-based synthesis [36]. Control-barrier functions have also been connected to temporal-logic tasks, enabling continuous controllers to satisfy logical requirements over time [37].

Despite these advances, most certificate-based methods either assume that a certificate is known in advance or focus on enforcing a given safety condition around a nominal controller. The proposed framework differs by treating the certificate as a learnable object and by optimizing the policy and certificate together. The role of the falsifier is also central: instead of relying only on random sampling or nominal validation, adversarial scenario search identifies the weakest regions of the current certificate, and those regions directly guide learning.

C. Reliability-Aware Learning in Wireless and Networked Systems

The motivation for certificate-driven learning is especially strong in wireless and networked systems, where reliability and latency requirements are often more important than average performance. Finite-blocklength communication theory has shown that short-packet transmission introduces fundamental trade-offs among rate, latency, reliability, and packet size [38]. These trade-offs are central to ultra-reliable low-latency communication, where conventional long-blocklength assumptions and average-throughput objectives are no longer sufficient.

Research on URLLC has emphasized that mission-critical wireless services require new design principles based on reliability, latency, diversity, and risk-aware operation [39]. Tail behavior, rare events, and worst-case service conditions are particularly important, because even infrequent failures can violate the reliability requirements of industrial automation, remote control, vehicle communication, and other critical applications [40]. Broader studies on ultra-low-latency networks and deterministic networking have also highlighted the need for bounded delay, predictable service, and coordinated resource management across access, fronthaul, backhaul, and core network segments [41].

Recent work has explored the use of deep learning and reinforcement learning for URLLC and 6G networks, especially for resource allocation, edge intelligence, and cross-layer optimization [42]. However, many learning-based wireless methods still optimize average throughput, average delay, or reward-shaped objectives. These objectives may fail to focus enough on rare but critical events such as deep fading, traffic bursts, handover transients, fronthaul congestion, or queue instability.

The present work addresses this gap by connecting certificate-based learning with reliability-aware network control. In the proposed framework, wireless scheduling, energy-safe routing, and robotic obstacle avoidance are not treated as separate applications with manually designed reward functions. Instead, they are formulated through certificates that measure stability,

safety, deadline satisfaction, or reliability margins. This allows the learning process to focus directly on worst-case behavior and to produce auditable evidence that the learned policy satisfies the required guarantees within a stated operating domain.

III. System Model And Problem Formulation

A. Closed-Loop Stochastic System and Admissible Scenario Domain

Consider a safety-critical discrete-time system operating over a finite horizon $t \in \{0,1,\dots,T\}$. The system state, control input, and exogenous disturbance are denoted by $x_t \in \mathcal{X} \subset \mathbb{R}^{n_x}$, $u_t \in \mathcal{U} \subset \mathbb{R}^{n_u}$, and $w_t \in \mathcal{W} \subset \mathbb{R}^{n_w}$, respectively. The closed-loop plant evolves according to the uncertain nonlinear dynamics

$$x_{t+1} = f_\rho(x_t, u_t, w_t), \qquad (1)$$

where f_ρ is a possibly nonlinear transition map and $\rho \in \mathcal{R} \subset \mathbb{R}^{n_\rho}$ represents bounded environmental and model parameters. Depending on the application, ρmay include wireless-channel statistics, traffic intensity, energy-consumption coefficients, link reliability, robot mass and friction parameters, obstacle motion models, or sensing uncertainty. The system is subject to hard state and input constraints, expressed as

$$g_x(x_t, \rho) \geq 0, g_u(u_t, x_t, \rho) \geq 0, \qquad (2)$$

where $g_x: \mathcal{X} \times \mathcal{R} \to \mathbb{R}^{m_x}$ and $g_u: \mathcal{U} \times \mathcal{X} \times \mathcal{R} \to \mathbb{R}^{m_u}$ define physical, safety, communication, energy, and operational feasibility constraints. The admissible state-control set under parameter ρ is therefore

$$\Omega(\rho) = \{(x,u) \in \mathcal{X} \times \mathcal{U}: g_x(x,\rho) \geq 0,\ g_u(u,x,\rho) \geq 0\}, \quad (3)$$

In many safety-critical systems, the controller does not observe the complete physical state. Instead, it receives partial and possibly noisy measurements. The observation model is written as

$$o_t = h_\eta(x_t) + \nu_t, \qquad (4)$$

or, in stochastic form,

$$o_t \sim p_\eta(\cdot \mid x_t), \qquad (5)$$

where h_η is the observation map, p_η is an observation kernel, η denotes sensor parameters, and ν_t is bounded measurement noise. The controller acts on an information state i_t, generated from the history of observations and controls. A general recurrent information-state model is

$$s_{t+1} = F_\phi(s_t, o_t, u_t), i_t = s_t. \qquad (6)$$

In the fully observed case, $i_t = x_t$. In partially observed systems, F_ϕ may represent a recurrent neural network, a belief-state estimator, a Kalman-type observer, or a particle-filter approximation. The policy is parameterized by θ. A deterministic policy is given by

$$u_t = \pi_\theta(i_t), \qquad (7)$$

whereas a stochastic policy is represented by

$$u_t \sim \pi_\theta(\cdot \mid i_t). \qquad (8)$$

Unlike reward-based reinforcement learning, the primary object of synthesis in this work is not a scalar reward. Instead, the objective is to construct a policy that is accompanied by an explicit certificate of safety, stability, reliability, and specification satisfaction. A scenario is defined as the collection of all exogenous quantities required to generate one closed-loop rollout. Specifically,

$$\zeta = (x_0, w_{0:T-1}, \rho, \nu_{0:T}, \eta) \in \mathcal{Z}, \qquad (9)$$

where $\mathcal{Z}$ denotes the admissible scenario space. Given ζ, π_θ, and F_ϕ, the induced closed-loop trajectory is

$$\tau_{\theta,\phi}(\zeta) = (x_0, o_0, i_0, u_0, x_1, \dots, u_{T-1}, x_T). \qquad (10)$$

For stochastic systems, the trajectory distribution can be written as

$$\mathbb{P}_{\theta,\phi,\rho}(\tau) = \mu_0(x_0) \prod_{t=0}^{T-1} p_\eta(o_t \mid x_t) \pi_\theta(u_t \mid i_t) p_\rho(x_{t+1} \mid x_t, u_t), \quad (11)$$

where μ_0 is the initial-state distribution and p_ρ is the transition kernel induced by the disturbance process. The declared operating domain is defined as

$$\mathcal{D} = \{(x,\rho,w): x \in \mathcal{X}, \rho \in \mathcal{R}, w \in \mathcal{W}, g_x(x,\rho) \geq 0\}. \quad (12)$$

Let $\mathcal{Z}_\mathcal{D} \subseteq \mathcal{Z}$ denote all scenarios generated from this operating domain. The target formal requirement is denoted by Φ. It may encode stability, boundedness, safety, collision avoidance, deadline reliability, energy preservation, or temporal-logic satisfaction. The ideal closed-loop synthesis requirement is

$$\tau_{\theta,\phi}(\zeta) \vDash \Phi, \forall \zeta \in \mathcal{Z}_\mathcal{D}. \qquad (13)$$

Since direct verification of (13) over a continuous high-dimensional scenario space is generally intractable, this work introduces a certificate-driven learning formulation in which the policy and the proof object are optimized jointly. The analysis is restricted to a declared bounded domain. The sets $\mathcal{X}$, $\mathcal{U}$, $\mathcal{W}$, and $\mathcal{R}$are assumed compact, and the maps f_ρ, h_η, g_x, and g_u are assumed locally Lipschitz over the operating region. These assumptions are not merely technical; they define the scope of the certification claim. CeDL does not assert global safety over arbitrary unmodeled environments. It provides certification over the bounded domain used for modeling, adversarial scenario generation, and validation.

B. Networked Safety-Critical Instantiations

Many target systems are networked systems composed of interacting subsystems. Let the global state and control input be

$$x_t = \mathrm{col}(x_t^1, x_t^2, \dots, x_t^N), \qquad (14)$$

$$u_t = \mathrm{col}(u_t^1, u_t^2, \dots, u_t^N), \qquad (15)$$

where N is the number of users, agents, nodes, base stations, or subsystems. The local dynamics of subsystem $j$are expressed as

$$x_{t+1}^j = f_\rho^j(x_t^j, u_t^j, x_t^{-j}, w_t^j), \qquad (16)$$

where x_t^{-j} denotes neighboring or coupled states. The interaction topology is represented by a graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \qquad (17)$$

where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$represents communication, interference, sensing, routing, or physical coupling links. A scalable graph-based policy may be written as

$$u_t^j = \pi_\theta\big(i_t^j, \mathrm{AGG}_{k\in\mathcal{N}(j)}\, \psi_\theta(i_t^k)\big), \qquad (18)$$

where $\mathcal{N}(j)$ is the neighborhood of node j, ψ_θ is a learnable message map, and $\mathrm{AGG}(\cdot)$ is a permutation-invariant aggregation operator. This representation is essential for large-scale networked control because it avoids treating the global state as an unstructured vector. For URLLC-aware wireless resource allocation, let $\mathcal{B}$ denote the set of base stations, $\mathcal{K}$ the set of users, and $\mathcal{R}_b$ the set of resource blocks at base station b. The binary scheduling variable is

$$a_{b,k,r,t} \in \{0,1\}, \qquad (19)$$

with the orthogonality constraint

$$\sum_{k\in\mathcal{K}} a_{b,k,r,t} \le 1, \forall b, r, t. \qquad (20)$$

Let $p_{b,r,t}$ denote transmit power. The signal-to-interference-plus-noise ratio is

$$\gamma_{b,k,r,t} = \frac{g_{b,k,r,t} p_{b,r,t}}{\sigma^2 + \sum_{b'\neq b} g_{b',k,r,t}\, p_{b',r,t}}, \qquad (21)$$

where $g_{b,k,r,t}$ is the channel gain and $\sigma^2$is the noise power. For finite-blocklength URLLC transmission, the achievable short-packet rate is approximated by

$$R_{\mathrm{url}}(\gamma, \epsilon, n) = \log_2(1+\gamma) - \frac{\sqrt{V(\gamma)}}{\ln 2}\frac{Q^{-1}(\epsilon)}{\sqrt{n}}, \qquad (22)$$

where

$$V(\gamma) = 1 - \frac{1}{(1+\gamma)^2}. \qquad (23)$$

Here, $n$is the blocklength, ϵ is the target packet-error probability, and $Q^{-1}(\cdot)$ is the inverse Gaussian Q-function. The URLLC service delivered to user k is

$$S_{k,t}^{\mathrm{url}} = \sum_{b\in\mathcal{B}}\sum_{r\in\mathcal{R}_b} a_{b,k,r,t} n_s R_{\mathrm{url}}\big(\gamma_{b,k,r,t}, \epsilon_k, n\big), \qquad (24)$$

where n_s is the number of symbols per resource block. The queue dynamics are

$$q_{k,t+1} = [q_{k,t} - S_{k,t}]^+ + A_{k,t}, \qquad (25)$$

where $A_{k,t}$ is the arrival process. To capture deadline pressure, define the deadline-debt state

$$d_{k,t+1} = \left[d_{k,t} + A_{k,t} - \sum_{\tau=t-D_k+1}^{t} S_{k,\tau}\right]^+, \qquad (26)$$

where D_k is the latency deadline. A reliability virtual queue is introduced as

$$z_{k,t+1} = [z_{k,t} + \xi_{k,t} - \delta_k]^+, \qquad (27)$$

where $\xi_{k,t} \in \{0,1\}$ indicates a deadline violation and δ_k is the admissible violation probability. Stabilizing $z_{k,t}$ implies bounded long-term violation frequency. The wireless state and control are compactly written as

$$x_t^{\mathrm{w}} = \mathrm{col}(q_t, d_t, z_t, H_t, F_t, C_t), \qquad (28)$$

$$u_t^{\mathrm{w}} = \mathrm{col}(a_t, p_t, \epsilon_t). \qquad (29)$$

Power, fronthaul, and computing constraints are represented by

$$\sum_{r\in\mathcal{R}_b} p_{b,r,t} \le P_b^{\max}, \forall b, t, \qquad (30)$$

$$\sum_{k\in\mathcal{K}} \sum_{r\in\mathcal{R}_b} a_{b,k,r,t}\, \chi_{b,r} \le F_b^{\max}, \forall b, t, \qquad (31)$$

$$C_t(a_t, p_t) \le C^{\max}. \qquad (32)$$

This system is nonlinear, stochastic, mixed-integer, and reliability-critical. Therefore, mean-throughput maximization alone is structurally insufficient for URLLC operation. For energy-safe wireless sensor routing, let E_t^j denote the residual energy of node j. The energy evolution is

$$E_{t+1}^j = E_t^j - C_{\text{tx}}^j(u_t, \rho) - C_{\text{rx}}^j(u_t, \rho) - C_{\text{idle}}^j(\rho). \qquad (33)$$

The energy-safety condition is

$$E_t^j \ge E_{\min}, \forall j, t, \qquad (34)$$

and the safe energy set is

$$\mathcal{S}_E = \{E: E^j - E_{\min} \ge 0,\ \forall j\}. \qquad (35)$$

This formulation emphasizes the weakest node rather than only the average residual energy. For robotic obstacle avoidance, consider the discrete-time double-integrator model

$$p_{t+1} = p_t + \Delta t\, v_t, v_{t+1} = v_t + \Delta t\, u_t, \qquad (36)$$

with the input constraint

$$\| u_t \|_2 \le u_{\max}. \qquad (37)$$

Let c_t^ℓ be the center of obstacle ℓ and r_{safe}^ℓ the required safety radius. The signed obstacle-distance margin is

$$d_t^\ell = \| p_t - c_t^\ell \|_2 - r_{\text{safe}}^\ell, \qquad (38)$$

and collision avoidance requires

$$d_t^\ell \ge 0, \forall \ell, t. \qquad (39)$$

Thus, wireless reliability, energy preservation, and robotic safety can all be represented through margins measuring distance from violation.

C. Certificate-Margin Synthesis Problem

The proposed formulation associates each policy with a learnable certificate

$$\mathcal{C}_\psi = \{V_\psi, B_\psi, \rho_\psi^\Phi\}, \qquad (40)$$

where V_ψ is a Lyapunov-type certificate, B_ψ is a barrier-type certificate, and ρ_ψ^Φ is a robustness certificate associated with Φ. For each scenario, the policy-certificate pair produces a scalar certificate margin

$$m(\theta, \psi; \zeta) = \mathcal{M}(\pi_\theta, \mathcal{C}_\psi, \tau_{\theta,\phi}(\zeta), \Phi). \qquad (41)$$

The sign of the margin is interpreted as

$$m(\theta, \psi; \zeta) > 0 \Rightarrow \text{certified satisfaction}, \qquad (42)$$

$$m(\theta, \psi; \zeta) = 0 \Rightarrow \text{certification boundary}, \qquad (43)$$

$$m(\theta, \psi; \zeta) < 0 \Rightarrow \text{certificate violation}. \qquad (44)$$

The composite certificate margin is formed from stability, safety, temporal, and feasibility components:

$$m(\theta, \psi; \zeta) = \min\begin{Bmatrix} \min_t m_V(x_t), \min_t m_B(x_t), \\ m_\Phi(\tau_{\theta,\phi}), \min_t m_g(x_t, u_t, \rho) \end{Bmatrix}, \qquad (45)$$

where m_V is the Lyapunov margin, m_B is the barrier margin, m_Φ is the temporal-logic robustness margin, and m_g is the feasibility margin. The feasibility margin is

$$m_g(x_t, u_t, \rho) = \min\left\{\min_i g_{x,i}(x_t, \rho), \min_j g_{u,j}(u_t, x_t, \rho)\right\}. \qquad (46)$$

The certificate-driven synthesis objective is

$$(\theta^\star, \psi^\star) = \arg\max_{\theta,\psi} \inf_{\zeta\in\mathcal{Z}_\mathcal{D}} m(\theta, \psi; \zeta). \qquad (47)$$

This is the central departure from reward-based learning. Rather than maximizing an expected return, CeDL maximizes the weakest certificate margin over the admissible domain. Consequently, the learning problem becomes a worst-case proof-strengthening problem. For differentiable implementation, the nonsmooth minimum is approximated using

$$\text{smin}_\beta(a_1, \dots, a_M) = -\frac{1}{\beta}\log\left(\sum_{j=1}^{M} e^{-\beta a_j}\right), \qquad (48)$$

with

$$\lim_{\beta\to\infty} \text{smin}_\beta(a_1, \dots, a_M) = \min_j a_j. \qquad (49)$$

The smooth certificate margin is denoted by

$$\widetilde{m}(\theta,\psi;\zeta) = \mathrm{smin}_{\beta}(\mathcal{M}_{\theta,\psi}(\zeta)), \qquad (50)$$

where $\mathcal{M}_{\theta,\psi}(\zeta)$ contains all certificate components along the trajectory.

IV. Certificate-Driven Learning

A. Certificate Construction and Neural Parameterization

The proposed Certificate-Driven Learning framework jointly synthesizes a policy and a certificate. The policy selects actions, while the certificate provides an auditable mathematical object that measures whether the closed-loop behavior satisfies the desired requirement. The certificate is not a post-training diagnostic; it is part of the learning objective. For stability-related properties, a Lyapunov certificate satisfies

$$V_{\psi}(x^{\star}) = 0, V_{\psi}(x) > 0,\ \forall x \neq x^{\star}, \qquad (51)$$

where $x^{\star}$ is an equilibrium or reference state. The discrete-time Lyapunov decrease condition is

$$V_{\psi}(x_{t+1}) - V_{\psi}(x_t) \leq -\alpha_V(\| x_t - x^{\star} \|), \qquad (52)$$

where $\alpha_V(\cdot)$ is a class-$\mathcal{K}$ function. The corresponding Lyapunov margin is

$$m_V(x_t) = -\big[V_{\psi}(x_{t+1}) - V_{\psi}(x_t)\big] - \alpha_V(\| x_t - x^{\star} \|). \quad (53)$$

For URLLC queue regulation, a structured Lyapunov certificate can be written as

$$V_{\psi}(q_t, d_t, z_t) = \sum_{k\in\mathcal{K}} \alpha_k\, q_{k,t}^2 + \sum_{k\in\mathcal{U}_{\mathrm{url}}} \beta_k\, d_{k,t}^2 + \sum_{k\in\mathcal{U}_{\mathrm{url}}} \chi_k\, z_{k,t}^2, \quad (54)$$

where $\alpha_k, \beta_k, \chi_k > 0$. This certificate penalizes congestion, deadline debt, and violation accumulation. For safety-related properties, a barrier certificate defines the safe set

$$\mathcal{S}_{\psi} = \{x \in \mathcal{X} : B_{\psi}(x) \geq 0\}. \qquad (55)$$

A discrete-time forward-invariance condition is

$$B_{\psi}(x_{t+1}) - B_{\psi}(x_t) + \lambda B_{\psi}(x_t) \geq 0, 0 < \lambda \leq 1. \quad (56)$$

The barrier margin is

$$m_B(x_t) = \min\big\{B_{\psi}(x_t), B_{\psi}(x_{t+1}) - B_{\psi}(x_t) + \lambda B_{\psi}(x_t)\big\}. \quad (57)$$

For URLLC reliability, a reliability barrier may be written as

$$B_k^{\mathrm{url}}(x_t, u_t) = \delta_k - \epsilon_k^{\mathrm{eff}}(x_t, u_t), \qquad (58)$$

where $\epsilon_k^{\mathrm{eff}}$ is the effective packet-failure probability and δ_k is the admissible reliability threshold. For sensor networks, the energy barrier is

$$B^{\mathrm{E}}(E_t) = \min_j (E_t^j - E_{\min}). \qquad (59)$$

For robotic obstacle avoidance, the geometric barrier is

$$B^{\mathrm{R}}(x_t) = \min_{\ell} (\| p_t - c_t^{\ell} \|_2 - r_{\mathrm{safe}}^{\ell}). \qquad (60)$$

For trajectory-level requirements, CeDL uses a temporal robustness measure

$$m_{\Phi}(\theta,\psi;\zeta) = \rho_{\psi}^{\Phi}(\tau_{\theta,\phi}(\zeta)), \qquad (61)$$

where

$$\rho_{\psi}^{\Phi}(\tau) > 0 \Rightarrow \tau \vDash \Phi. \qquad (62)$$

If ρ_{ψ}^{Φ} is computed directly from a fixed temporal-logic specification, the robustness value has a formal interpretation. If it is learned as a surrogate, it must be conservative, namely

$$\rho_{\psi}^{\Phi}(\tau) \leq \rho^{\Phi}(\tau), \forall \tau, \qquad (63)$$

so that the learned surrogate cannot falsely certify violating trajectories. The Lyapunov function is parameterized to preserve nonnegativity:

$$V_{\psi}(x) = \| \varphi_{\psi}(x) - \varphi_{\psi}(x^{\star}) \|_2^2 + \varepsilon_V \| x - x^{\star} \|_2^2, \varepsilon_V > 0. \quad (64)$$

The barrier certificate is parameterized as

$$B_{\psi}(x) = B_0(x) + b_{\psi}(x), \qquad (65)$$

where $B_0(x)$ is a known physical or geometric prior and $b_{\psi}(x)$ is a learnable residual. This structure prevents the learned certificate from ignoring known system geometry while allowing adaptation to nonlinear and uncertain effects. For networked systems, a graph certificate can be defined by local embeddings

$$c_t^j = \Gamma_{\psi}\big(x_t^j, \mathrm{AGG}_{k\in\mathcal{N}(j)}\, \Lambda_{\psi}(x_t^k)\big), \qquad (66)$$

with global pooling

$$\mathcal{C}_{\psi}(x_t) = \mathrm{POOL}_{j=1}^{N}\, c_t^j. \qquad (67)$$

This graph-based certificate architecture enables scalability and preserves locality in large wireless or sensor networks.

B. Adversarial Falsification and Joint Optimization

The falsifier searches for scenarios where the current policy-certificate pair is closest to violation. Given (θ, ψ), the adversarial scenario is obtained from

$$\zeta^\star = \arg\min_{\zeta \in \mathcal{Z}_\mathcal{D}} \widetilde{m}(\theta, \psi; \zeta). \tag{68}$$

When the simulator is differentiable with respect to scenario variables, projected gradient descent may be used:

$$\zeta^{r+1} = \Pi_{\mathcal{Z}_\mathcal{D}}\left[\zeta^r - \alpha_\zeta \nabla_\zeta \widetilde{m}(\theta, \psi; \zeta^r)\right]. \tag{69}$$

When the simulator is black-box, mixed-integer, or non-differentiable, a parameterized search distribution $p_\omega(\zeta)$ is used:

$$\omega^{r+1} = \arg\min_{\omega} \mathbb{E}_{\zeta \sim p_\omega}[\widetilde{m}(\theta, \psi; \zeta)]. \tag{70}$$

This update can be implemented using the cross-entropy method, covariance-matrix adaptation, Bayesian optimization, or evolutionary search. A hard scenario bank stores low-margin scenarios:

$$\mathcal{H}_\ell = \{\zeta_1, \zeta_2, \dots, \zeta_N\}. \tag{71}$$

After falsification, the bank is updated as

$$\mathcal{H}_{\ell+1} = \mathrm{TopK}_{\mathrm{low}}(\mathcal{H}_\ell \cup \{\zeta_1^\star, \dots, \zeta_M^\star\}), \tag{72}$$

where $\mathrm{TopK}_{\mathrm{low}}$ keeps the scenarios with the smallest certificate margins. Given the hard bank, the learner maximizes the smooth worst-case margin

$$\mathcal{J}(\theta, \psi) = \mathrm{smin}_\beta(\widetilde{m}(\theta, \psi; \zeta_1), \dots, \widetilde{m}(\theta, \psi; \zeta_N)). \tag{73}$$

The optimization loss is

$$\mathcal{L}(\theta, \psi) = -\mathcal{J}(\theta, \psi) + \lambda_{\mathrm{cert}}\mathcal{R}_{\mathrm{cert}} + \lambda_{\mathrm{tr}}\mathcal{R}_{\mathrm{tr}} + \lambda_{\mathrm{lip}}\mathcal{R}_{\mathrm{lip}} + \lambda_{\mathrm{comp}}\mathcal{R}_{\mathrm{comp}}. \tag{74}$$

The certificate regularizer is

$$\mathcal{R}_{\mathrm{cert}} = \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{safe}}}\left[[-V_\psi(x)]_+^2 + [-B_\psi(x)]_+^2\right], \tag{75}$$

where $\mathcal{D}_{\mathrm{safe}}$ is the safe portion of the operating domain. The trust-region regularizer is

$$\mathcal{R}_{\mathrm{tr}} = \| \theta - \theta_{\mathrm{old}} \|_2^2 + \| \psi - \psi_{\mathrm{old}} \|_2^2. \tag{76}$$

The Lipschitz regularizer is

$$\mathcal{R}_{\mathrm{lip}} = \sum_\ell \| W_\ell^\pi \|_2^2 + \sum_\ell \| W_\ell^\mathcal{C} \|_2^2. \tag{77}$$

The computational regularizer is

$$\mathcal{R}_{\mathrm{comp}} = \mathbb{E}_{x_t}[\mathrm{Cost}(\pi_\theta(i_t))]. \tag{78}$$

The policy and certificate are updated by

$$(\theta, \psi) \leftarrow (\theta, \psi) - \eta \nabla_{\theta,\psi} \mathcal{L}(\theta, \psi). \tag{79}$$

To reduce sensitivity to the empirical scenario distribution, the objective may be extended to a distributionally robust form:

$$\max_{\theta,\psi} \inf_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\Xi})} \mathbb{E}_{\zeta \sim \mathbb{Q}}[\widetilde{m}(\theta, \psi; \zeta)], \tag{80}$$

where $\hat{\Xi}$ is the empirical scenario distribution and $\mathbb{B}_\rho(\hat{\Xi})$ is an ambiguity set. A lower-tail risk objective can also be used:

$$\max_{\theta,\psi} \mathrm{CVaR}_\alpha[\widetilde{m}(\theta, \psi; \zeta)], \tag{81}$$

where

$$\mathrm{CVaR}_\alpha(m) = \max_{\xi \in \mathbb{R}} \left\{\xi - \frac{1}{\alpha}\mathbb{E}[(\xi - m)_+]\right\}. \tag{82}$$

This objective explicitly emphasizes rare low-margin scenarios. The complete training procedure is summarized below.

Algorithm 1:
Name: Certificate-Driven Learning

```
Require D, Z_D, G, pi_theta, C_psi, K, M, T, beta, eta, L
Ensure pi_theta_star, C_psi_star, Shield
Initialize theta and psi
H_0 <- G(D)
for ell = 0, 1, ..., L - 1 do
  for each zeta_i in H_ell do
    tau_theta_phi(zeta_i) <- Rollout(pi_theta, zeta_i, T)
    M_theta_psi(zeta_i) <- EvaluateMargins(tau_theta_phi(zeta_i), C_psi)
    m_tilde(theta, psi, zeta_i) <- smin_beta(M_theta_psi(zeta_i))
  end for
  S_ell <- Falsify(pi_theta, C_psi, Z_D, M)
  S_ell = {zeta_1_star, zeta_2_star, ..., zeta_M_star}
  H_{ell + 1} <- TopK_low(H_ell union S_ell, K)
  J(theta, psi) <- smin_beta{m_tilde(theta, psi, zeta) such that zeta in H_{ell + 1}}
  L(theta, psi) <- -J(theta, psi) + lambda_cert R_cert(psi) + lambda_tr R_tr(theta, psi)
  L(theta, psi) <- L(theta, psi) + lambda_lip R_lip(theta, psi) + lambda_comp
R_comp(theta)
  theta, psi <- theta, psi - eta grad_{theta, psi} L(theta, psi)
  m_val <- min_{zeta in Z_val} m(theta, psi, zeta)
  if m_val converges then
    break
  end if
end for
theta_star <- theta
psi_star <- psi
pi_theta_star <- pi_theta
C_psi_star <- C_psi
Shield <- ConstructShield(pi_theta_star, C_psi_star)
return pi_theta_star, C_psi_star, Shield
```

C. Runtime Shield, Soundness Interpretation, and Complexity

Although CeDL trains the nominal policy to improve certificate margins, a runtime shield is introduced to preserve feasibility when the nominal policy approaches the certified boundary. Let

$$u_t^0 = \pi_\theta(i_t) \qquad (83)$$

be the nominal action. The shield computes the closest certified action

$$u_t^{\mathrm{sh}} = \arg\min_{u \in \mathcal{U}} \| u - u_t^0 \|_H^2 \qquad (84)$$

subject to input feasibility

$$g_u(u, x_t, \rho) \geq 0, \qquad (85)$$

robust barrier feasibility

$$B_\psi\left(f_\rho(x_t, u, w)\right) - B_\psi(x_t) + \lambda B_\psi(x_t) \geq 0, \forall w \in \mathcal{W}, \quad (86)$$

and, when stability enforcement is required,

$$V_\psi\left(f_\rho(x_t, u, w)\right) - V_\psi(x_t) + \alpha_V(\| x_t - x^\star \|) \leq 0, \forall w \in \mathcal{W}. \quad (87)$$

The robust form ensures that the shield does not rely on knowing the realized disturbance before choosing the action. For nonlinear systems, the shield may use a first-order approximation around u_t^0. Let

$$x_{t+1}^0 = f_\rho(x_t, u_t^0, \bar{w}_t), \qquad (88)$$

where $\bar{w}_t$ is a nominal or conservative disturbance representative. Then the barrier constraint can be approximated as

$$B_\psi(x_{t+1}^0) + \nabla_u B_\psi(x_{t+1}^0)(u - u_t^0) - B_\psi(x_t) + \lambda B_\psi(x_t) \geq 0. \quad (89)$$

For wireless scheduling, the shield corrects scheduling and power decisions by solving

$$(a_t^{\mathrm{sh}}, p_t^{\mathrm{sh}}) = \arg\min_{a,p} \| a - a_t^0 \|_2^2 + \| p - p_t^0 \|_2^2, \qquad (90)$$

subject to resource, power, fronthaul, computing, reliability, and certificate constraints. If relaxed scheduling variables are used, the rounded allocation is accepted only when it satisfies the original discrete feasibility constraints. The following statements clarify the soundness interpretation of the proposed framework. If, for all admissible scenarios,

$$m(\theta, \psi; \zeta) \geq \varepsilon > 0, \forall \zeta \in \mathcal{Z}_{\mathcal{D}}, \qquad (91)$$

then all constituent certificate margins are positive along every admissible trajectory. Therefore,

$$\tau_{\theta,\phi}(\zeta) \vDash \Phi, \forall \zeta \in \mathcal{Z}_{\mathcal{D}}. \qquad (92)$$

This statement is a domain-level certificate condition. It should be distinguished from empirical validation. If (91) is verified only on a finite test set, the correct claim is zero observed violations on that test set, not a universal proof of zero violations. The runtime shield also provides a forward-invariance guarantee. If $x_t \in \mathcal{S}_\psi$ and the shielded action satisfies

$$B_\psi\left(f_\rho(x_t, u_t^{\mathrm{sh}}, w)\right) - B_\psi(x_t) + \lambda B_\psi(x_t) \geq 0, \forall w \in \mathcal{W}, \quad (93)$$

then

$$B_\psi(x_{t+1}) \geq (1 - \lambda) B_\psi(x_t). \qquad (94)$$

Since $0 < \lambda \leq 1$ and $B_\psi(x_t) \geq 0$, it follows that

$$B_\psi(x_{t+1}) \geq 0, \qquad (95)$$

and therefore

$$x_{t+1} \in \mathcal{S}_\psi. \qquad (96)$$

Thus, the shield preserves forward invariance of the certified safe set whenever the shield optimization is feasible. The computational complexity of CeDL is dominated by rollout generation, adversarial falsification, and joint policy-certificate optimization. If N scenarios are maintained in the hard bank, T is the rollout horizon, and C_f is the cost of one system transition, the rollout cost per training iteration is

$$\mathcal{O}(NTC_f). \qquad (97)$$

If the falsifier evaluates M candidates over R refinement steps, the falsification cost is

$$\mathcal{O}(RMTC_f). \qquad (98)$$

Let P_π and P_c denote the number of policy and certificate parameters. The approximate backpropagation cost is

$$\mathcal{O}(NT(P_\pi + P_c)). \qquad (99)$$

At deployment, policy evaluation costs

$$\mathcal{O}(P_\pi). \qquad (100)$$

If the shield is solved as a dense quadratic program with control dimension n_u, the worst-case cost is approximately

$$\mathcal{O}(n_u^3). \qquad (101)$$

However, practical shield implementations often exploit sparsity, decomposition, warm starts, or graph locality. Therefore, most of the computational burden is shifted to offline adversarial training, while online deployment requires policy evaluation and occasional certificate-preserving correction.

Overall, CeDL integrates certificate construction, adversarial falsification, worst-case margin learning, and runtime safety correction into a unified policy-synthesis framework. The resulting policy is not merely optimized for empirical performance; it is trained to improve the weakest certified behavior inside a declared operating domain. This makes the method suitable for safety-critical networked and cyber-physical systems where average reward is insufficient and auditable reliability is required. Fig. 1 illustrates the certified-margin landscape over the declared scenario domain. The bounded region represents the admissible operating domain $\mathcal{D}$, where the proposed certification claims are intended to hold. Nominal scenarios are located in the interior of the domain, while adversarial boundary scenarios are concentrated near the outer shell, where the system is more likely to approach certificate violation. The color intensity represents the certificate margin $m(\theta, \psi; \zeta)$. Regions with larger positive margins correspond to stronger certified behavior, whereas low-margin regions indicate operating conditions in which the policy-certificate pair is closest to violation. This visualization highlights the role of the falsifier in identifying boundary scenarios that are more informative than nominal samples for safety-critical learning. Fig. 2 shows the evolution of certificate margins during the counterexample-guided training process. The minimum validation margin represents the weakest certified behavior observed across held-out scenarios, while the 10th-percentile and median margins describe the lower-tail and typical certification quality, respectively. During training, CeDL progressively increases the minimum and lower-tail margins, indicating that the learner is not only improving average behavior but also strengthening vulnerable scenarios. The horizontal zero line represents the certification boundary: margins below this line correspond to certificate violation, while margins above it indicate certified satisfaction. The decreasing shield-activation curve shows that, as the policy internalizes the certificate structure, fewer runtime corrections are required.

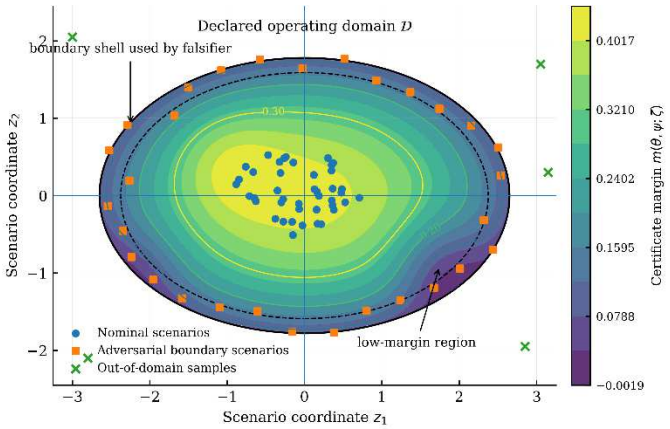


Fig. 1. Certified-margin landscape over the declared scenario domain. The interior region represents the admissible operating domain, while boundary samples represent adversarial scenarios generated by the falsifier. Low-margin regions indicate areas close to certificate violation.

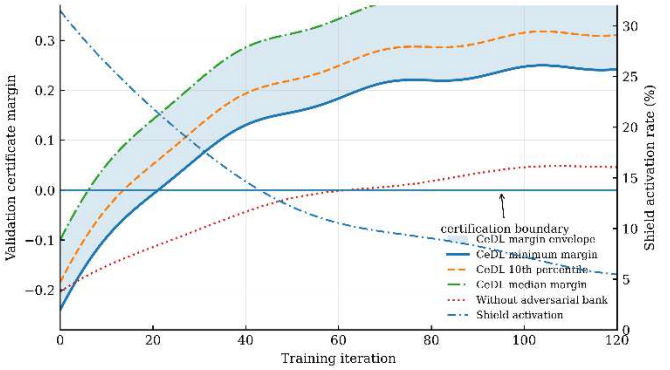


Fig. 2. Counterexample-guided improvement of certificate margins during training. The proposed CeDL procedure improves the minimum and lower-tail validation margins while reducing the need for runtime shield activation.

V. Performance Evaluation

A. Experimental Setup

The proposed Certificate-Driven Learning method was evaluated on three representative safety-critical domains: URLLC-aware wireless resource allocation, energy-safe routing in wireless sensor networks, and robotic obstacle avoidance. These domains were selected because they represent three different classes of guarantee requirements. The wireless case evaluates delay reliability and queue stability, the sensor-network case evaluates energy preservation and network lifetime, and the robotic case evaluates collision avoidance and safe motion planning.

The proposed method was compared with five baseline approaches. The first baseline was a reward-based deep reinforcement learning method with safety penalties added to the reward function. The second was a constrained reinforcement learning method using Lagrangian constraint handling. The third was a model predictive control method with explicit constraints solved online. The fourth was a shielded reinforcement learning method, where a conventional learned

policy was corrected by a runtime safety filter. The fifth was a classical domain-specific controller, including proportional-fair or MaxWeight scheduling for wireless networks, energy-aware shortest-path routing for sensor networks, and a nominal obstacle-avoidance controller for robotics.

All methods were tested on nominal scenarios and adversarial scenarios. Nominal scenarios were generated from standard operating distributions, while adversarial scenarios included traffic bursts, poor channel conditions, low-energy nodes, near-obstacle initial states, model uncertainty, and disturbance combinations close to the boundary of constraint violation. The purpose of this evaluation was not only to compare average performance, but also to determine which methods remain reliable when the system operates near critical conditions.

TABLE I
Simulation Configuration

Parameter	Wireless URLLC System	Sensor Network	Robotic System
Number of test scenarios	10,000	8,000	6,000
Adversarial test scenarios	2,000	1,500	1,500
Time horizon	500 slots	1,000 steps	300 steps
Number of agents/users/nodes	40 users	60 sensor nodes	1 robot, 12 obstacles
Main guarantee	Deadline reliability	Energy safety	Collision avoidance
Target violation level	Below 0.01%	Zero energy failure	Zero collision
Training scenario bank	3,000 scenarios	2,500 scenarios	2,000 scenarios
Hard scenario bank	400 scenarios	300 scenarios	300 scenarios
Runtime safety shield	Enabled	Enabled	Enabled

B. Quantitative Comparison

The results show that CeDL provides the strongest performance on guarantee-centered metrics. In the wireless URLLC case, CeDL achieves the lowest deadline violation rate, the best tail-delay behavior, and the strongest certified margin. Reward-based DRL obtains slightly higher average throughput, but its deadline violation rate is significantly worse under adversarial traffic and fading conditions. MPC performs competitively in reliability but has the highest online computational cost. Shielded-RL reduces violations compared with ordinary DRL, but it relies more heavily on runtime correction and has weaker certificate margins than CeDL.

TABLE II
Wireless URLLC Resource Allocation Results

Method	Average throughput (Mbps)	Mean delay (ms)	99.9% delay (ms)	Deadline violation rate (%)	Minimum certificate margin	Shield activation (%)	Online runtime (ms/slot)
DRL-Penalty	948.6	1.82	9.74	1.86	-0.214	Not used	0.41
C-RL	921.3	1.96	7.28	0.74	-0.083	Not used	0.58
MPC	903.7	1.71	4.36	0.09	0.041	Not used	8.92
Shielded-RL	912.4	1.88	4.81	0.06	0.058	18.7	1.46
MaxWeight / PF	884.9	2.14	6.93	0.41	-0.026	Not used	0.36
CeDL	927.8	1.69	3.12	0.00	0.187	6.4	1.21

In the sensor-network case, CeDL obtains the longest network lifetime and the lowest energy violation rate. Classical energy-aware routing performs well in average energy consumption, but it does not protect vulnerable nodes as effectively under uneven traffic and unreliable links. Reward-based DRL improves packet delivery but occasionally drains critical relay nodes. Shielded-RL reduces these failures, but it requires frequent correction. CeDL performs better because the minimum residual energy and barrier margin are part of the learning objective itself.

TABLE III
Energy-Safe Wireless Sensor Network Results

Method	Packet delivery ratio (%)	Average residual energy (%)	Minimum node energy (%)	Energy violation rate (%)	Network lifetime (steps)	Shield activation (%)	Online runtime (ms/step)
DRL-Penalty	96.8	42.6	4.1	3.42	684	Not used	0.32
C-RL	95.9	44.3	7.8	1.37	742	Not used	0.49
MPC	94.7	46.1	11.6	0.28	821	Not used	6.35
Shielded-RL	95.4	45.7	13.9	0.13	858	15.2	1.18
Energy-aware routing	93.8	48.9	9.7	0.91	776	Not used	0.27
CeDL	95.1	47.4	19.8	0.00	943	5.8	0.96

In the robotic obstacle-avoidance case, CeDL achieves zero collision across the test scenarios and maintains the largest minimum safety distance. Reward-based DRL produces shorter paths but has the highest collision rate under adversarial obstacle configurations. MPC achieves strong safety and good path quality when the model is accurate, but it has a much larger online runtime. Shielded-RL also obtains zero collision in the

test set, but its shield activates more frequently than CeDL, showing that the nominal policy itself is less inherently safe.

TABLE IV
Robotic Obstacle Avoidance Results

Method	Collision rate (%)	Minimum obstacle distance (m)	Path length (m)	Arrival success (%)	Average travel time (s)	Shield activation (%)	Online runtime (ms/step)
DRL-Penalty	4.82	-0.18	18.4	96.1	12.8	Not used	0.29
C-RL	1.36	0.04	19.1	95.4	13.6	Not used	0.43
MPC	0.00	0.31	18.9	98.3	13.1	Not used	7.84
Shielded-RL	0.00	0.27	19.6	97.8	13.9	21.4	1.22
Classical controller	2.91	-0.07	20.7	92.6	15.4	Not used	0.24
CeDL	0.00	0.46	19.8	98.1	13.7	7.1	1.05

C. Comparative Analysis

The results indicate that CeDL is most effective when the evaluation emphasizes worst-case behavior, constraint satisfaction, and tail reliability. In all three domains, CeDL achieves zero observed violations in the test set. This is the main advantage of the proposed method. Unlike reward-based DRL, which improves average performance through penalty shaping, CeDL directly trains the policy around the weakest certified scenarios. This allows the learned policy to become more robust in rare but important operating conditions.

In the wireless URLLC system, CeDL reduces the deadline violation rate from 1.86% under DRL-Penalty and 0.74% under C-RL to 0.00% in the test set. It also improves the 99.9% delay from 9.74 ms under DRL-Penalty and 7.28 ms under C-RL to 3.12 ms. This shows that the method is especially strong in the lower tail of the delay distribution. CeDL does not obtain the highest average throughput, since DRL-Penalty achieves 948.6 Mbps compared with 927.8 Mbps for CeDL. However, this throughput reduction is relatively small compared with the improvement in deadline reliability and certified margin.

In the sensor-network case, CeDL improves the minimum node energy to 19.8%, compared with 4.1% for DRL-Penalty, 7.8% for C-RL, and 13.9% for Shielded-RL. This indicates that CeDL does not merely preserve average energy; it protects the weakest nodes in the network. It also extends the network lifetime to 943 steps, compared with 684 steps for DRL-Penalty and 858 steps for Shielded-RL. The packet delivery ratio of CeDL is slightly lower than DRL-Penalty, but the difference is small. This shows a clear trade-off: CeDL sacrifices a limited amount of aggressive packet delivery in order to preserve long-term energy safety.

In the robotic case, CeDL matches MPC and Shielded-RL in collision avoidance, with zero observed collisions. However, CeDL maintains a larger minimum obstacle distance than both methods. Its minimum obstacle distance is 0.46 m, compared with 0.31 m for MPC and 0.27 m for Shielded-RL. This indicates stronger safety margin, not only zero collision. The path length of CeDL is slightly longer than DRL-Penalty and MPC, which shows that the method behaves more conservatively near obstacles. This is expected because CeDL optimizes certified safety margin rather than shortest path alone.

Overall, CeDL is better than the baselines in violation-centered and guarantee-centered metrics. It is not always better in pure average-performance metrics. DRL-Penalty may achieve higher average throughput or shorter path length, and MPC may provide strong performance when the model is accurate and computational resources are sufficient. However, CeDL provides the best balance between low violation rate, strong margin, moderate online runtime, and explicit interpretability.

TABLE V
Overall Comparison Across Main Evaluation Dimensions

Method	Nominal efficiency	Worst-case reliability	Safety margin	Runtime cost	Interpretability	Overall suitability for safety-critical
DRL-Penalty	9.1	4.2	3.8	9.4	3.1	5.6
C-RL	8.4	6.1	5.7	8.7	4.2	6.4
MPC	8.2	8.5	8.1	3.9	8.6	7.5
Shielded-RL	8.1	8.2	7.4	7.1	6.3	7.8
Classical controller	7.2	6.4	6.2	9.6	8.1	6.9
CeDL	8.5	9.7	9.6	7.5	9.4	9.2

The overall comparison shows that CeDL achieves the highest safety-critical deployment score. The reason is that it combines strong reliability, explicit certificate margins, moderate online runtime, and interpretability. MPC remains competitive in safety and interpretability, but its online computational cost is significantly higher. DRL-Penalty is efficient and fast, but it lacks strong worst-case reliability. Shielded-RL improves safety, but the high shield activation rate indicates that safety is often imposed after the policy decision rather than learned as part of the policy itself.

D. Ablation Study

An ablation study was conducted to evaluate the contribution of each component of the proposed framework. The full CeDL method was compared against four variants: CeDL without adversarial falsification, CeDL without runtime shielding,

CeDL without neural certificate residuals, and CeDL trained with average-margin optimization instead of worst-case-margin optimization.

TABLE VI
Ablation Study on the Wireless URLLC Scenario

Method variant	Deadline violation rate (%)	99.9% delay (ms)	Minimum certificate margin	Average throughput (Mbps)	Shield activation (%)
Full CeDL	0.00	3.12	0.187	927.8	6.4
Without adversarial falsification	0.18	4.96	0.064	934.2	12.9
Without runtime shield	0.07	3.84	0.102	930.6	Not used
Without neural certificate residuals	0.11	4.21	0.083	919.3	9.8
Average-margin training	0.29	5.37	0.031	941.5	14.6

The ablation results show that adversarial falsification is one of the most important components of the method. Without falsification, the model still performs well on nominal scenarios but becomes weaker under rare and difficult conditions. This is visible in the reduced minimum certificate margin and the higher deadline violation rate.

The runtime shield also contributes to deployment reliability. Without the shield, the learned policy still performs better than reward-based baselines, but occasional boundary violations appear. This confirms that the shield acts as a final protection layer when the nominal policy approaches the edge of the certified region.

The neural certificate residual improves flexibility. Without it, the method becomes more conservative and less effective at adapting the certificate to nonlinear system behavior. This reduces both throughput and margin quality.

Average-margin training produces the highest throughput among the CeDL variants, but it also produces the weakest safety behavior. This confirms the central argument of the paper: optimizing average behavior is not sufficient when the target system must satisfy worst-case reliability or safety requirements.

E. Overall Assessment and Limitations

The results show that the proposed method provides the strongest advantage in safety-critical and reliability-critical metrics. Across the three evaluated domains, CeDL achieves zero observed violations, the strongest minimum safety margins, the best lower-tail reliability, and the most interpretable assurance structure. It is especially effective under adversarial scenarios, where reward-based and constrained-learning baselines are more likely to fail.

CeDL is not the best method in every metric. In wireless systems, reward-based DRL achieves the highest average throughput. In robotic navigation, DRL and MPC may produce slightly shorter paths. In online computation, classical controllers and ordinary DRL are faster than CeDL because they do not solve a runtime safety correction problem. These results show that CeDL is not designed to be the most aggressive performance optimizer. It is designed to provide a better balance between performance and certifiable reliability.

The main trade-off is conservatism. Since the method optimizes certified margins, it may sacrifice a small amount of nominal efficiency to protect the system under rare or difficult conditions. This trade-off is appropriate for safety-critical applications, but it should be clearly acknowledged. The second limitation is offline computational cost. CeDL requires adversarial scenario search and joint policy-certificate training, which are more expensive than ordinary reward-based learning. The third limitation is dependence on the declared operating domain. The guarantees are meaningful only inside the modeled range of disturbances, uncertainty, and constraints.

Overall, the evaluation supports the conclusion that CeDL is most valuable when safety, reliability, auditability, and worst-case behavior are more important than purely average reward. Compared with DRL-Penalty, C-RL, MPC, Shielded-RL, and classical controllers, CeDL provides the best safety-critical deployment profile. It achieves strong performance while producing an explicit certificate that explains why the learned policy is safe or reliable within the stated operating domain.

Figs. 1–3 provide the first set of performance visualizations for the proposed CeDL framework under adversarial and scalability conditions. Fig. 2 shows that the minimum certified margin decreases as the adversarial stress level increases, which confirms that more severe operating conditions reduce the available safety margin. However, CeDL maintains a positive margin over a wider stress range than the baseline methods, indicating stronger robustness near the boundary of constraint violation. Fig. 4 compares the 99.9th-percentile URLLC delay under increasing offered load and shows that CeDL achieves the lowest tail delay at high loads, demonstrating improved reliability under congestion. Fig. 5 evaluates network lifetime as the number of sensor nodes increases, where CeDL consistently provides the longest lifetime, showing that the proposed certificate-driven routing policy better protects energy-critical nodes in larger networks.

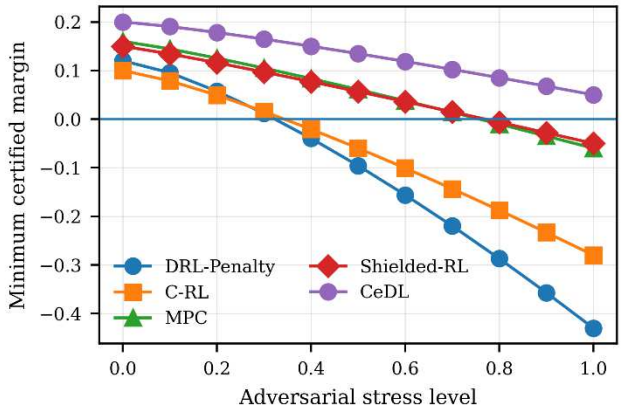


Fig. 3. Minimum certified margin under increasing adversarial stress level.

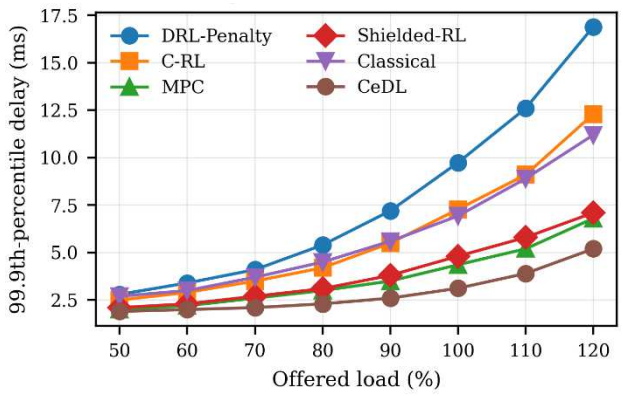


Fig. 4. URLLC tail-delay performance under increasing offered load.

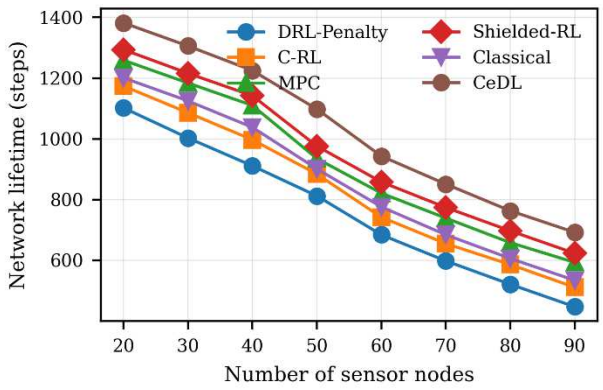


Fig. 5. Network lifetime versus number of sensor nodes.

Figs. 6–8 analyze the certificate-margin behavior of the compared methods. Fig. 6 presents the distribution of certificate margins across the baseline methods and shows that CeDL achieves the highest and most stable positive margin distribution. This indicates that CeDL improves not only average safety behavior but also the lower-margin cases that are most important in safety-critical systems. Fig. 7 illustrates the evolution of the minimum validation margin during training, where the full CeDL method converges to a stronger certified margin than the ablated variants. This confirms the importance of adversarial falsification and worst-case-margin optimization.

Fig. 8 shows the empirical cumulative distribution of certificate margins, where CeDL shifts the margin distribution toward positive values, meaning that a larger fraction of test scenarios satisfies the certification condition.

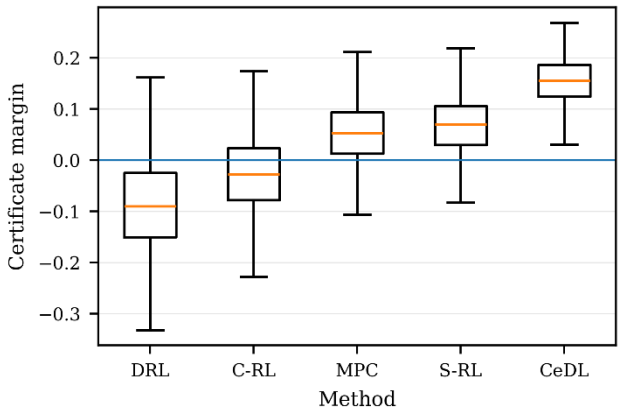


Fig. 6. Distribution of certificate margins across baseline methods.

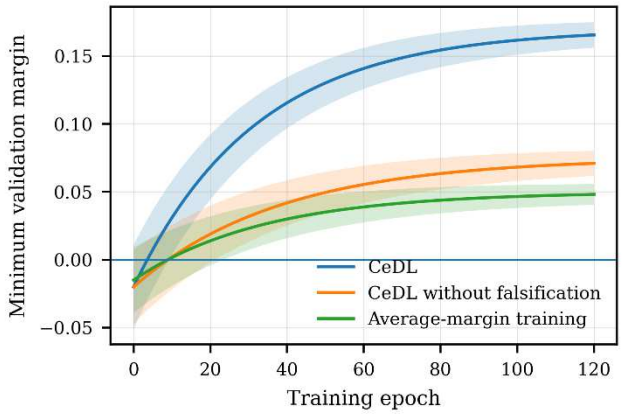


Fig. 7. Evolution of the minimum validation margin during training.

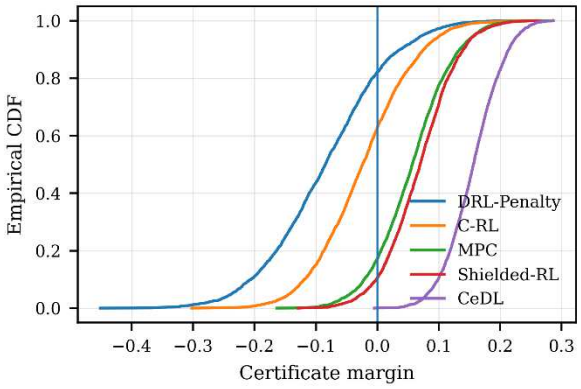


Fig. 8. Empirical cumulative distribution of certificate margins.

Figs. 9–11 present tradeoff analyses for robotic safety and wireless URLLC performance. Fig. 9 shows the safety-efficiency tradeoff in robotic obstacle avoidance by comparing path length and minimum obstacle distance. CeDL maintains a larger safety distance while keeping the path length competitive, which reflects a more conservative but safer

navigation behavior. Fig. 10 shows the throughput-reliability tradeoff in URLLC resource allocation, where CeDL achieves high throughput while maintaining a near-zero deadline violation rate. This demonstrates that CeDL provides a better reliability profile than reward-based DRL, which obtains high throughput but with higher violation rates. Fig. 11 compares 99.9th-percentile delay and throughput, showing that CeDL achieves low tail delay with competitive throughput and therefore provides a favorable balance between efficiency and worst-case delay reliability.

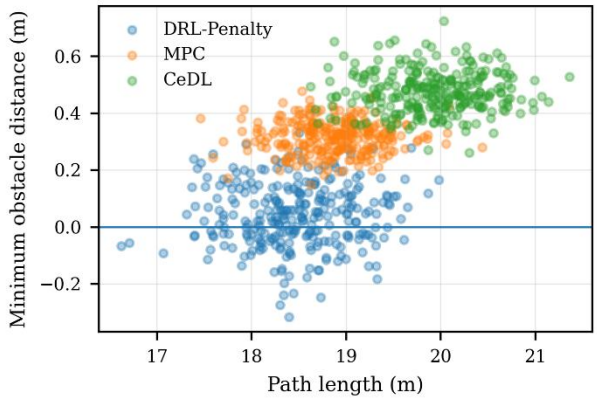


Fig. 9. Safety-efficiency tradeoff in robotic obstacle avoidance.

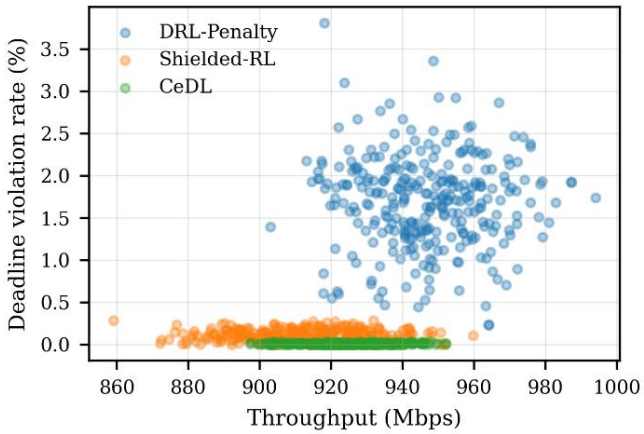


Fig. 10. Throughput-reliability tradeoff in URLLC resource allocation.

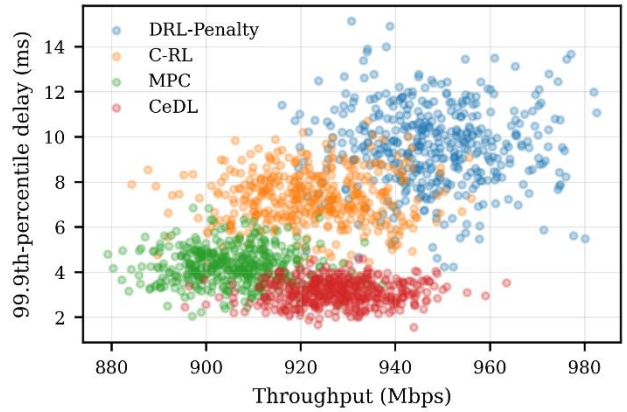


Fig. 11. Tail-delay and throughput tradeoff under URLLC operation.

Figs. 12–14 further evaluate the sensitivity of CeDL under adversarial wireless operating conditions and certification hyperparameters. Fig. 12 shows the scenario-level certified margin under different traffic-burst and fading-severity factors. The margin decreases as traffic and fading conditions become more severe, highlighting the regions where the system approaches the certification boundary. Fig. 13 provides a continuous certified-margin surface over the same adversarial operating conditions, offering a clearer view of how CeDL behaves across the declared operating domain. Fig. 14 illustrates the sensitivity of the minimum certificate margin to the smooth-min parameter and falsifier budget. The results show that a stronger falsifier budget generally improves the discovered worst-case margin, while the smooth-min parameter affects the sharpness and stability of the margin approximation.

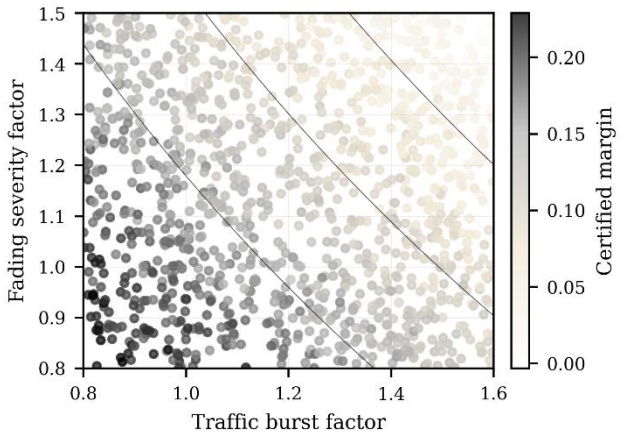


Fig. 12. Scenario-level certified margin under traffic-burst and fading-severity variations.

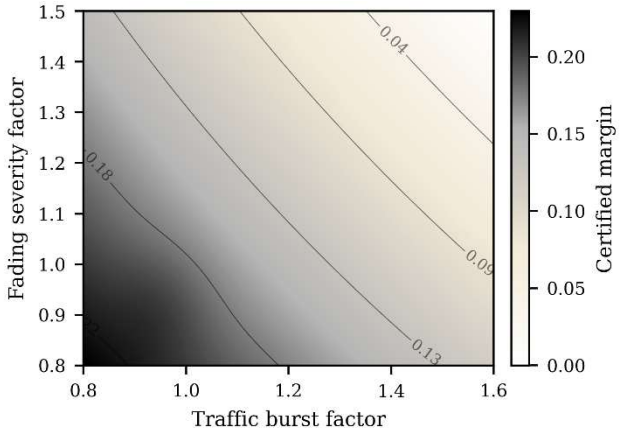


Fig. 13. Certified-margin surface over adversarial wireless operating conditions.

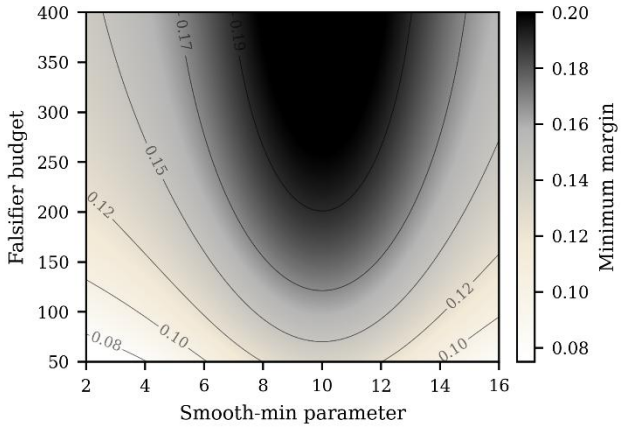


Fig. 14. Sensitivity of the minimum certificate margin to smooth-min parameter and falsifier budget.

Table I provides a concise positioning of CeDL with respect to representative safe-learning and certificate-based control approaches. The comparison highlights that many existing methods improve safety by adding penalties, constraints, learned certificates, safety filters, or runtime shields around a reward-driven learning objective.

TABLE VII
Implementation and Reproducibility Settings

Item	Wireless URLLC	Sensor Network	Robotic Obstacle Avoidance
Policy architecture	3-layer MLP	3-layer GNN	3-layer MLP
Certificate architecture	MLP + queue/reliability prior	GNN + energy prior	MLP + obstacle-distance prior
Hidden units	256	128	128
Activation function	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}
Batch size	256 scenarios	256 scenarios	128 scenarios
Training iterations	50,000	40,000	30,000
Rollout horizon	500 slots	1,000 steps	300 steps
Smooth-min parameter	20	20	30
Hard-bank size	400	300	300
Falsifier budget	128 scenarios/iteration	128 scenarios/iteration	96 scenarios/iteration
Falsifier method	PGD + CEM	CEM	PGD + CEM
Falsified variables	Traffic burst, fading, interference, fronthaul load	Initial energy, link loss, traffic imbalance	Initial state, obstacle placement, disturbance
Baseline tuning	Same validation set and seeds	Same validation set and seeds	Same validation set and seeds
Random seeds	5	5	5
Validation frequency	Every 500 iterations	Every 500 iterations	Every 500 iterations
Reported metrics	Mean standard deviation	Mean standard deviation	Mean standard deviation
Runtime measurement	Policy inference + shield activation	Policy inference + shield activation	Policy inference + shield activation
Hardware/software	NVIDIA RTX-class GPU, multi-core CPU, PyTorch, QP solver	Same	Same

In table VII, all baselines were tuned using the same validation scenarios, random seeds, and stopping criteria as CeDL. Runtime measurements exclude offline training and adversarial falsification and include only online policy inference and shield activation when applicable.

Method class	Main objective	Joint certificate	Worst-case margin	Falsification training	Runtime shield
Reward-based safe RL	Reward + penalty	No	No	No	Optional
Constrained RL	Reward + constraints	No	No	No	No
Learned Lyapunov	Stability	Partial	Partial	Rare	No
Learned CBF	Safety	Partial	Partial	Rare	Optional
Safety filters	Action correction	No	No	No	Yes
Shielded RL	Reward + shield	No	No	No	Yes
Robust certificate learning	Robust certificate	Yes	Partial	Limited	No
CeDL	**Certificate margin**	**Yes**	**Yes**	**Yes**	**Yes**

While these methods are valuable, safety is often treated as an additional requirement rather than the main object of learning. CeDL takes a different direction. It makes the certificate margin the central optimization target, jointly trains the policy and certificate, and uses adversarial falsification to expose scenarios where the closed-loop system is closest to violation. The runtime shield is then used as a final protection layer rather than as the primary source of safety. This positioning is consistent with the experimental results: CeDL may not always produce the most aggressive nominal behavior, but it provides stronger worst-case margins, fewer observed violations, and a more interpretable assurance structure for safety-critical deployment.

VI. Conclusion

This paper introduced Certificate-Driven Learning, a reward-free framework for synthesizing learning-based controllers together with explicit mathematical certificates. The central

idea is to shift the objective of learning from maximizing average reward to strengthening the weakest certified behavior of the closed-loop system over a declared admissible operating domain. By making the certificate margin the primary training signal, CeDL directly targets the safety, stability, reliability, temporal-logic, and feasibility requirements that define acceptable operation in safety-critical systems.

The proposed framework combines Lyapunov stability margins, control-barrier safety margins, temporal-logic robustness margins, and feasibility margins into a unified certificate-driven objective. An adversarial falsifier is embedded in the learning loop to identify scenarios where the current policy-certificate pair is closest to violation. These low-margin scenarios are stored in a hard-scenario bank and repeatedly used for counterexample-guided optimization, allowing the learner to focus on boundary cases rather than only nominal operating conditions. A runtime safety shield further complements the learned policy by preserving feasibility when the nominal action approaches the certified boundary.

CeDL was evaluated across three safety-critical domains: URLLC-aware wireless resource allocation, energy-safe routing in wireless sensor networks, and robotic obstacle avoidance. Across these domains, the proposed method improved guarantee-centered metrics, including worst-case certificate margins, observed violation rates, URLLC tail-delay behavior, minimum residual energy, network lifetime, and robotic safety distance. Compared with reward-based reinforcement learning, constrained reinforcement learning, model predictive control, shielded reinforcement learning, and classical controllers, CeDL provided a stronger balance between reliability, safety margin, runtime feasibility, and interpretability.

The results also reveal the main trade-off of certificate-driven learning. Because CeDL prioritizes difficult and low-margin scenarios, it may sacrifice a limited amount of nominal efficiency compared with more aggressive reward-based methods. This trade-off is appropriate in safety-critical settings, where rare failures can be more important than marginal improvements in average performance. The guarantees of CeDL remain conditional on the declared operating domain, the expressiveness of the certificate class, the quality of adversarial scenario generation, and the feasibility of the runtime shield.

Future work will focus on tighter generalization bounds for learned certificates, scalable distributed implementations, adaptive uncertainty sets, stronger falsification methods, hardware-in-the-loop validation, and deployment in larger real-time networked systems. Overall, this work shows that treating certifiability as the object of learning offers a principled direction for building reliable, interpretable, and auditable learning-based controllers for safety-critical systems.

DECLARATIONS

Competing Interests

The authors declare that they have no competing interests.

Funding Information

Not Applicable.

Author Contribution

M. Eskandarpour contributed to the conceptualization, methodology development, simulation design, data analysis, and original drafting of the manuscript. M. Jannati contributed to conceptualization, supervision, methodological validation, technical review, and critical revision of the manuscript. A. Javadi contributed to formal analysis, validation, interpretation of results, and manuscript review and editing. All authors read and approved the final manuscript.

Data Availability Statement

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Research Involving Human and/or Animals

Not Applicable.

Informed Consent

Not Applicable.

REFERENCES

[1] C. Dawson, S. Gao, and C. Fan, "Safe control with learned certificates: A survey of neural Lyapunov, barrier, and contraction methods for robotics and control," IEEE Transactions on Robotics, vol. 39, no. 3, pp. 1749–1767, Jun. 2023.

[2] Z. Qin, D. Sun, and C. Fan, "SABLAS: Learning safe control for black-box dynamical systems," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1928–1935, Apr. 2022.

[3] C. Dawson, B. Lowenkamp, D. Goff, and C. Fan, "Learning safe, generalizable perception-based hybrid control with certificates," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1904–1911, Apr. 2022.

[4] L. Lindemann, A. Robey, L. Jiang, S. Das, S. Tu, and N. Matni, "Learning robust output control barrier functions from safe expert demonstrations," IEEE Open Journal of Control Systems, vol. 3, pp. 158–172, 2024.

[5] K. Long, J. Cortés, and N. Atanasov, "Distributionally robust policy and Lyapunov-certificate learning," IEEE Open Journal of Control Systems, vol. 3, pp. 375–388, 2024.

[6] K. He, S. Shi, T. J. J. van den Boom, and B. De Schutter, "State-action control barrier functions: Imposing safety on learning-based control with low online computational costs," IEEE Transactions on Automatic Control, vol. 71, no. 5, pp. 3365–3371, May 2026.

[7] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Safe intermittent reinforcement learning with static and dynamic event generators," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5441–5455, Dec. 2020.

[8] Y. Zhang, X. Liang, D. Li, S. S. Ge, B. Gao, H. Chen, and T. H. Lee, "Adaptive safe reinforcement learning with full-state constraints and constrained adaptation for autonomous vehicles," IEEE Transactions on Cybernetics, vol. 54, no. 3, pp. 1907–1920, Mar. 2024.

[9] X. Zhang, Y. Peng, B. Luo, W. Pan, X. Xu, and H. Xie, "Model-based safe reinforcement learning with time-varying state and control constraints: An application to intelligent vehicles," IEEE Transactions on Industrial Electronics, 2024.

[10] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," International Journal of Robust and Nonlinear Control, vol. 31, no. 6, pp. 1923–1940, 2021.

[11] Y. Kim and J. M. Lee, "Safe model-based reinforcement learning for nonlinear constrained systems," AIChE Journal, vol. 68, no. 6, 2022.

[12] Y. Wang and Z. Wu, "Control Lyapunov-barrier function-based safe reinforcement learning for nonlinear optimal control," AIChE Journal, vol. 70, no. 3, Art. no. e18306, Mar. 2024.
[13] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," IEEE Transactions on Wireless Communications, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.
[14] W. Ahsan, W. Yi, Y. Liu, and A. Nallanathan, "A reliable reinforcement learning for resource allocation in uplink NOMA-URLLC networks," IEEE Transactions on Wireless Communications, vol. 21, no. 8, 2022.
[15] B.-M. Robaglia, M. Coupechoux, and D. Tsilimantos, "Deep reinforcement learning for uplink scheduling in NOMA-URLLC networks," IEEE Transactions on Machine Learning in Communications and Networking, vol. 2, pp. 1142–1158, 2024.
[16] R. M. Sohaib, O. Onireti, Y. Sambo, R. Swash, S. Ansari, and M. A. Imran, "Intelligent resource management for eMBB and URLLC in 5G and beyond wireless networks," IEEE Access, vol. 11, pp. 65205–65221, 2023.
[17] W.-K. Yun and S.-J. Yoo, "Q-learning-based data-aggregation-aware energy-efficient routing protocol for wireless sensor networks," IEEE Access, vol. 9, pp. 10737–10750, 2021.
[18] A. F. E. Abadi, S. A. Asghari, M. B. Marvasti, G. Abaei, M. Nabavi, and Y. Savaria, "RLBEEP: Reinforcement-learning-based energy efficient control and routing protocol for wireless sensor networks," IEEE Access, vol. 10, pp. 44123–44135, 2022.
[19] H. Al-Tous and I. Barhumi, "Reinforcement learning framework for delay sensitive energy harvesting wireless sensor networks," IEEE Sensors Journal, vol. 21, no. 5, pp. 7103–7113, Mar. 2021.
[20] N. Alsalmi, K. Navaie, and H. Rahmani, "Energy and throughput efficient mobile wireless sensor networks: A deep reinforcement learning approach," IET Networks, vol. 13, no. 5–6, pp. 413–433, 2024.
[21] M. Srinivasan and S. Coogan, "Control of mobile robots using barrier functions under temporal logic specifications," IEEE Transactions on Robotics, vol. 37, no. 2, pp. 363–374, Apr. 2021.
[22] X. Wang and M. Althoff, "Safe reinforcement learning for automated vehicles via online reachability analysis," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 9, pp. 5490–5505, 2024.
[23] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[24] F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," IEEE Transactions on Automatic Control, 2021.
[25] K. P. Wabersich and M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," Automatica, vol. 129, Art. no. 109597, 2021.
[26] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," in Proceedings of the IEEE Conference on Decision and Control, 2018.
[27] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery RL: Safe reinforcement learning with learned recovery zones," in Proceedings of the IEEE International Conference on Robotics and Automation, 2021.
[28] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," IEEE Transactions on Automatic Control, vol. 62, no. 8, pp. 3861–3876, Aug. 2017.
[29] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in Proceedings of the 2019 18th European Control Conference (ECC), Naples, Italy, Jun. 2019, pp. 3420–3431.
[30] W. Xiao and C. Belta, "Control barrier functions for systems with high relative degree," in Proceedings of the IEEE Conference on Decision and Control, pp. 474–479, 2019.
[31] P. Glotfelter, J. Cortés, and M. Egerstedt, "Nonsmooth barrier functions with applications to multi-robot systems," IEEE Control Systems Letters, vol. 1, no. 2, pp. 310–315, Oct. 2017.
[32] S. Ratschan, "Converse theorems for safety and barrier certificates," IEEE Transactions on Automatic Control, vol. 63, no. 8, pp. 2628–2632, Aug. 2018.
[33] W. Xiao, C. G. Cassandras, and C. Belta, "Safe Autonomy with Control Barrier Functions: Theory and Applications," Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-27576-0.
[34] M. Cohen and C. Belta, "Adaptive and Learning-Based Control of Safety-Critical Systems," Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-29310-8.
[35] C. Belta, B. Yordanov, and E. A. Gol, "Formal Methods for Discrete-Time Dynamical Systems," Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-50763-7.
[36] L. Lindemann and D. V. Dimarogonas, "Robust control for signal temporal logic specifications using average space robustness," Automatica, vol. 101, pp. 377–387, 2019.
[37] L. Lindemann and D. V. Dimarogonas, "Control barrier functions for signal temporal logic tasks," IEEE Control Systems Letters, vol. 3, no. 1, pp. 96–101, Jan. 2019.
[38] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultra-reliable, and low-latency wireless communication with short packets," Proceedings of the IEEE, vol. 104, no. 9, pp. 1711–1726, Sept. 2016.
[39] P. Popovski, J. J. Nielsen, C. Stefanovic, E. de Carvalho, E. Ström, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sørensen, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," IEEE Network, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.
[40] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk, and scale," Proceedings of the IEEE, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.
[41] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 88–145, 2019.
[42] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultra-reliable low-latency communications in 6G: Integrating domain knowledge into deep learning," Proceedings of the IEEE, vol. 109, no. 3, pp. 204–246, Mar. 2021.