iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: I Month of publication: January 2026

DOI: https://doi.org/10.22214/ijraset.2026.75463

www.ijraset.com

Call: ✆08813907089 | E-mail ID: ijraset@gmail.com

Developing a Stable Control Policy for Vertical Rocket Landing with Deep Reinforcement Learning

Mrs. Snehal Bagal[1], Aditya Godse[2], Prajwal Kumbhar[3], Soham Khule[4]
[1, 2]Dept. of Artificial Intelligence and Data Science AISSMS Institute of Information Technology, Pune, India
[3, 4]Dept. of Artificial Intelligence and Data Science, AISSMS Institute of Information Technology, Pune, India

Abstract: The advent of reusable rockets has fundamentally altered the economics of space exploration, with autonomous powered descent and landing being the critical enabling tech- nology. This maneuver presents a formidable challenge in control theory, characterized by high-dimensional continuous state- action spaces, unstable nonlinear dynamics, and strict terminal constraints. This paper presents a comprehensive theoretical framework for developing a stable control policy for vertical rocket landing using Deep Reinforcement Learning (DRL). We situate the problem within the context of modern policy op- timization algorithms, reviewing the theoretical underpinnings of methods such as Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO), and Soft Actor-Critic (SAC). We analyze the critical role of environment design, reward shaping theory, and numerical simulation in the successful application of DRL to this domain. By systematically comparing existing methodologies, we identify key research gaps, including the sim-to-real transfer problem, the sample efficiency of on- policy methods, and the absence of formal safety guarantees. This paper validates the theoretical feasibility of using model- free DRL to solve high-stakes aerospace control problems and proposes a structured roadmap for future research in robust, verifiable autonomous guidance and control systems.

Index Terms: Deep Reinforcement Learning, Proximal Pol- icy Optimization, Autonomous Control, Aerospace Engineering, Rocket Landing, Theoretical Framework, Control Theory

I. INTRODUCTION

The contemporary era of space exploration is defined by a paradigm shift towards economic sustainability, catalyzed by the advent of reusable launch vehicles. The recovery and reuse of a rocket's first stage, pioneered by organizations such as SpaceX, represents a pivotal technological advancement that has dramatically lowered the cost of space access [18]. This innovation has not only reshaped the commercial launch industry but has also expanded the horizons of scientific and interplanetary missions by making them more economically viable.

Central to this reusability is the autonomous powered descent and vertical landing of the booster stage, a maneuver that constitutes a canonical and exceptionally challenging problem in modern control theory. The task is analogous to balancing an inverted pendulum under the influence of complex, time-varying forces, but is further complicated by high-dimensional, continuous state and action spaces, and unforgiving terminal constraints on velocity and orientation at touchdown. Any failure to meet these constraints results in the catastrophic loss of the vehicle.

The dynamical system of a landing rocket is a formidable subject for control. It is a high-dimensional, underactuated sys- tem governed by inherently unstable, nonlinear dynamics. The vehicle's mass varies significantly as propellant is consumed, altering its inertial properties and response to control inputs. The control agent must contend with complex aerodynamic forces, gravitational effects, and the nonlinearities of engine throttling and attitude control systems. The state of the sys- tem, encompassing its position, velocity, angle, and angular velocity, evolves continuously, as do the required control actions for the main engine throttle and attitude thrusters. The primary control objective is to guide the system from a high-altitude, high-velocity initial state to a terminal state at the landing pad characterized by near-zero velocity and a perfectly vertical orientation. This terminal phase guidance problem is exceptionally difficult due to the tight coupling between translational and rotational dynamics and the strict constraints that define a successful landing.

Traditional control methodologies have long been ap- plied to aerospace guidance problems. Techniques such as Proportional-Integral-Derivative (PID) controllers, Linear- Quadratic Regulators (LQR), and advanced trajectory opti- mization methods based on convex programming [19] have formed the bedrock of classical control. These approaches, however, are fundamentally model-based, presupposing the existence of an accurate analytical model of the system's dynamics. The derivation of such high-fidelity models is a non-trivial and resource-intensive task. More importantly, the performance of these controllers can degrade significantly in the presence of unmodeled dynamics or external perturbations, such as stochastic wind gusts, variations in atmospheric density, or fuel slosh dynamics. While adaptive control techniques can mitigate some of these issues, they often introduce their own complexities. The brittleness of model-based controllers in the face of real-world uncertainty creates a compelling theoretical motivation for exploring alternative control paradigms. Deep Reinforcement Learning (DRL) emerges as a powerful theoretical alternative to classical control. As a model-free paradigm, DRL enables a control agent to learn an optimal policy, a mapping from states to actions, through a process of trial-and-error interaction with an environment [14]. This data-driven approach obviates the need for an explicit, pre- defined system model, allowing the agent to implicitly learn and compensate for complex, nonlinear, and even stochastic dynamics. By leveraging the universal function approximation capabilities of deep neural networks, DRL agents can learn sophisticated control policies for high-dimensional, continuous domains that are often intractable for traditional methods. The agent's learning is guided by a scalar reward signal, an engineering design choice that allows the problem's objectives and constraints to be encoded into a single optimization target. This paper aims to construct a robust theoretical foundation for applying DRL to the vertical rocket landing challenge. We begin by reviewing the evolution of the requisite algorithms, analyzing the theoretical principles of environment and reward design, systematically comparing existing approaches to identify critical research gaps, and ultimately proposing a multi-dimensional agenda for future research. The overarching objective is to reframe the engineering problem of rocket landing as a structured theoretical inquiry within the DRL paradigm, providing a clear and principled roadmap for the development of next-generation autonomous guidance and control systems.

II. LITERATURE REVIEW

The application of DRL to vertical rocket landing draws upon a rich body of work spanning policy optimization algo- rithms, standardized simulation environments, and the theo- retical underpinnings of reward function design. This section provides a focused review of the foundational literature that informs our theoretical framework, with specific attribution to the contributions of each cited work.

According to Schulman et al. (2015), a primary challenge in early policy gradient methods was the instability caused by large, high-variance gradient updates, which could lead to a catastrophic collapse in performance. Their seminal work on Trust Region Policy Optimization (TRPO) addressed this by introducing a formal trust region constraint on the policy update. By limiting the Kullback-Leibler (KL) divergence between the old and new policies, TRPO provides a theo- retical guarantee of monotonic policy improvement. However, this guarantee comes at the cost of significant computational complexity, as it requires solving a constrained optimization problem using second-order methods like the conjugate gra- dient algorithm, making it difficult to implement and scale to large models.

Building directly on this work, Schulman et al. (2017) proposed Proximal Policy Optimization (PPO) as a first- order approximation of TRPO that retains its stability benefits while being substantially simpler to implement. PPO's core theoretical contribution is a novel clipped surrogate objective function. This objective penalizes policy changes that move the probability ratio of an action (under the new versus old policy) outside a small, predefined interval. This mechanism effec- tively creates a soft constraint on the policy update, preventing destructively large changes and ensuring stable learning. The simplicity, robustness, and strong empirical performance of PPO have established it as a benchmark algorithm for a wide range of continuous control tasks.

In parallel with on-policy methods, the field of off-policy actor-critic algorithms has seen significant theoretical advance- ments aimed at improving sample efficiency. Fujimoto, van Hoof, and Meger (2018) conducted a rigorous analysis of actor-critic methods and identified function approximation error as a primary source of overestimated Q-values, which in turn leads to the learning of suboptimal policies. Their algorithm, Twin Delayed Deep Deterministic policy gradient (TD3), introduces a trio of theoretical mechanisms to combat this: a clipped double Q-learning technique that takes the minimum of two independent Q-value estimates to mitigate overestimation bias; delayed policy updates, where the actor is updated less frequently than the critic to allow for lower-variance value estimates; and target policy smoothing, which adds noise to the target action to prevent the policy from exploiting narrow peaks in the value function.

Further advancing the state of off-policy learning, Haarnoja et al. (2018) developed Soft Actor-Critic (SAC), an algorithm grounded in the maximum entropy reinforcement learning framework. The central theoretical principle of SAC is to augment the standard reward maximization objective with an entropy maximization term. This encourages the policy to act as randomly as possible while still successfully completing the task. This entropy bonus has two key benefits: it promotes more extensive exploration, preventing the agent from con- verging to poor local optima, and it results in a more robust final policy that is less sensitive to small perturbations in the environment. SAC's ability to combine high sample efficiency with stable training has made it a state-of-the-art algorithm for continuous control.

The theoretical validity and reproducibility of any DRL application are contingent upon the environment in which the agent is trained. Brockman et al. (2016) were instrumental in standardizing this process with the introduction of OpenAI Gym, a toolkit providing a common API for RL environ- ments. This standardization enabled researchers to benchmark algorithms on a consistent set of tasks, fostering reproducible science. Beyond the interface, Xu and Chen (2021) proposed a theoretical framework for validating the internal design of an RL environment. They argue that for learning to be meaningful and effective, the state features provided to the agent must satisfy two properties: they must be sensitive to the agent's actions (i.e., controllable) and they must be predictive of future rewards. An environment lacking these properties may render the control problem intractable.

For physical simulations, numerical stability is a paramount concern. The Verlet integration algorithm, as originally de- scribed by Verlet (1967) for simulating molecular dynamics, is a numerical integration method with properties that make it highly suitable for simulating physical systems over long time horizons. Its time-reversibility and excellent energy con- servation characteristics ensure that the simulation remains physically plausible and does not diverge due to accumulating numerical errors, a common issue with simpler methods like Euler integration.

A cornerstone of applied reinforcement learning is the design of the reward function. Ng, Harada, and Russell (1999) provided the seminal theoretical foundation for this practice with their work on potential-based reward shaping. They rigorously proved that the optimal policy of a Markov Decision Process remains invariant under an arbitrary reward transformation if and only if the shaping reward, $F(s, a, s')$, is of the form $F = \gamma\Phi(s') - \Phi(s)$. Here, Φ is an arbitrary real-valued potential function over states and γ is the discount factor. This theorem provides a powerful and theoretically sound method for designing dense reward signals that can sig- nificantly accelerate learning without the risk of inadvertently altering the agent's ultimate objective.

Finally, the viability of DRL for complex aerospace control has been established in prior work that serves as a direct prece- dent for this theoretical framework. Vedanta et al. (2019) suc- cessfully applied the PPO algorithm to the problem of three- axis spacecraft attitude control. Their research demonstrated that a DRL agent could learn a control policy that was not only effective but also more robust to significant variations in the spacecraft's mass and inertia properties than a conventionally tuned Quaternion Rate Feedback (QRF) controller. This work confirms that modern policy gradient methods are capable of solving challenging, continuous control problems in the aerospace domain, thereby motivating their application to the more intricate, multi-objective challenge of powered descent and landing.

III. COMPARISON OF METHODOLOGIES

To contextualize the various algorithmic approaches dis- cussed in the literature, a systematic comparison is necessary. Table I analyzes key reinforcement learning algorithms based on their theoretical properties and mechanisms. This com- parison highlights the trade-offs between sample efficiency, stability, and implementation complexity that are central to selecting an appropriate algorithm for a given control problem.

IV. IDENTIFIED RESEARCH GAPS

A thorough review of the existing literature reveals several critical research gaps that must be addressed to advance the application of DRL from simulation to real-world, high-stakes aerospace systems. These gaps represent significant theoretical and practical hurdles in the path toward fully autonomous and verifiable control systems.

First, a significant disparity exists between simulated environments and physical reality, commonly known as the ”sim-to-real” gap. While simulations incorporating advanced numerical methods like Verlet integration [9] can achieve high fidelity, they cannot perfectly capture all real-world complexities. Unmodeled dynamics, such as stochastic wind gusts, fuel slosh, ground effect, and sensor noise, can cause a policy trained exclusively in simulation to fail when deployed on physical hardware. Techniques like domain randomization, where simulation parameters are varied during training, offer a partial solution, but a more robust theoretical gap exists in developing algorithms that are either inherently robust to this reality gap or can adapt efficiently with minimal real-world data.

Second, a fundamental trade-off exists between the stability of on-policy algorithms and the sample efficiency of their off-policy counterparts. As demonstrated by Schulman et al. [7], PPO provides robust and stable learning but is inherently on-policy, meaning it requires new samples for each gradient update, making it data-intensive for complex physical systems where data collection is expensive. In contrast, off-policy algorithms like SAC [8] can reuse past data from a replay buffer, offering superior theoretical sample efficiency. How- ever, they can be more sensitive to hyperparameters and less stable during training. A key research gap lies in developing hybrid algorithms or novel formulations that combine the robust stability of on-policy methods with the high sample efficiency of off-policy learning.

Third, the process of reward function engineering, while the- oretically grounded by the work of Ng et al. [1] on potential- based shaping, remains a highly heuristic and domain-specific art. The performance of the learning agent is critically depen- dent on the design of the potential function $\Phi(s)$, yet there is no systematic methodology for deriving this function. An improperly designed potential function can lead to unintended behaviors, local optima, or slow convergence. The develop- ment of techniques to automatically learn or derive optimal reward structures, for instance through inverse reinforcement learning or by incorporating curriculum learning where the re- ward function evolves with the agent's competence, represents a major open research problem.

Finally, and perhaps most critically for aerospace applica- tions, current DRL frameworks lack formal safety guarantees and methods for verification. The neural network policies learned by DRL agents are opaque, high-dimensional, nonlin- ear functions, making them difficult to analyze with traditional verification tools from control theory. For a safety-critical task like rocket landing, the inability to provide a mathematical proof that a policy will not enter a catastrophic state under a given set of conditions is a major barrier to deployment. There is a profound research gap in the intersection of DRL and formal methods, aimed at developing techniques for certifying the safety, robustness, and stability of learned control policies.

V. FUTURE WORK

Addressing the identified research gaps requires a multi- dimensional research agenda that deepens existing paradigms, broadens the problem scope, and explores novel theoretical directions. We propose a roadmap for future work structured along three conceptual axes, designed to systematically ad- vance the state of the art in autonomous control for aerospace systems.

TABLE I

Comparative Analysis of Reinforcement Learning Algorithms

Algorithm	Author(s)	Type	Core Theoretical Principle	Stability Mechanism	Sample Efficiency
TRPO	Schulman et al. (2015)	On-Policy	Monotonic Policy Improvement	KL Divergence Constraint	Low
PPO	Schulman et al. (2017)	On-Policy	Simplified Trust Region	Clipped Surrogate Objective	Low-to-Medium
TD3	Fujimoto et al. (2018)	Off-Policy	Bias-Variance Trade-off	Clipped Double Q-Learning	High
SAC	Haarnoja et al. (2018)	Off-Policy	Maximum Entropy RL	Entropy Maximization	High

The first axis involves deepening and enhancing current DRL paradigms to improve their efficiency and robustness. A promising direction is the development of hybrid model-based and model-free approaches. While the primary appeal of DRL for this problem is its model-free nature, a learned dynamics model of the rocket could be used to augment the training process. For instance, a learned model could be used for short- horizon planning to refine the actor's actions or to generate synthetic data for the replay buffer, potentially improving the sample efficiency of even on-policy algorithms like PPO. Another area for deeper investigation is the automation of reward engineering. Techniques from inverse reinforcement learning (IRL) could be explored to learn a reward function from expert demonstrations, which could be generated by traditional trajectory optimization solvers. This would replace the heuristic process of designing a potential function with a principled, data-driven approach, leading to more optimal and predictable learning. The second axis focuses on broadening the scope of the control problem to more closely mirror real-world complexity. The current theoretical framework, often validated in 2D, must be extended to a full three-dimensional simulation. This would introduce additional degrees of freedom and control challenges, such as engine gimbaling for thrust vector control and the management of six degrees of freedom (x, y, z, roll, pitch, yaw).

This expanded problem could be framed within a multi-agent reinforcement learning (MARL) context, where different subsystems of the rocket, such as the main engine throttle, the attitude control thrusters, and the landing leg deployment mechanism, are modeled as independent agents that must learn to coordinate their actions to achieve a common goal. Furthermore, the DRL framework could be applied to the entire mission profile, from atmospheric entry through powered descent, requiring the agent to learn a hierarchical policy that is effective across vastly different aerodynamic and dynamic regimes.

The third and most theoretically ambitious axis involves the integration of novel concepts from adjacent fields to address fundamental limitations. To tackle the critical issue of safety, future work should focus on the intersection of DRL and formal methods. Research into ”safe RL” could explore using techniques like reachability analysis or control barrier functions to define a provably safe envelope in the rocket’s state space. The DRL agent would be permitted to optimize its policy freely within this safe envelope but would be constrained by a safety layer that overrides any action predicted to violate the boundary. Additionally, concepts from causal reinforcement learning could be integrated to enhance policy robustness and generalization. By learning a causal model of the environment, an agent could better distinguish spurious correlations from true causal relationships, enabling it to reason more effectively about the consequences of its actions, particularly in response to novel or unforeseen events such as a specific engine failure mode or an unexpected atmospheric disturbance.

VI. CONCLUSION

This paper has presented a comprehensive theoretical frame- work for addressing the autonomous vertical rocket landing problem using Deep Reinforcement Learning. By situating this complex aerospace challenge within the context of modern control theory and machine learning, we have reviewed the evolution of pertinent policy optimization algorithms, analyzed the theoretical requirements for high-fidelity simulation, and underscored the critical role of potential-based reward shaping in guiding the learning process. The systematic comparison of methodologies and the subsequent identification of critical research gaps, spanning sim-to-real transfer, the stability- efficiency trade-off, reward function design, and the profound need for safety verification, collectively illuminate the current limitations and future potential of the field. The proposed multi-axis roadmap for future work, which advocates for hybrid methods, expanded problem scopes, and the integration of formal and causal methods, provides a structured pathway for advancing DRL from a powerful simulation tool to a viable technology for real-world, safety-critical autonomous systems. This work validates the immense theoretical potential of DRL for solving formidable control problems and charts a course for continued innovation in autonomous guidance and control.

REFERENCES

[1] A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in Proc. 16th International Conference on Machine Learning (ICML), 1999, pp. 278–287.

[2] R. Xu and Z. Chen, “A validation tool for designing reinforcement learning environments,” in Proc. 35th AAAI Conference on Artificial Intelligence, 2021, pp. 10475–10483.

[3] G. Brockman et al., “OpenAI Gym,” arXiv preprint arXiv:1606.01540, 2016.

[4] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, “Trust region policy optimization,” in Proc. 32nd International Conference on Machine Learning (ICML), 2015, pp. 1889–1897.

[5] Vedanta, J. T. Allison, M. West, and A. Ghosh, “Reinforcement learning for spacecraft attitude control,” in Proc. AIAA Scitech 2019 Forum, 2019,

[6] p. 1949.

[7] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function ap- proximation error in actor-critic methods,” in Proc. 35th International Conference on Machine Learning (ICML), 2018, pp. 1587–1596.

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Prox- imal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, 2017.

[9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off- policy maximum entropy deep reinforcement learning with a stochastic actor,” in Proc. 35th International Conference on Machine Learning (ICML), 2018, pp. 1861–1870.

[10] L. Verlet, “Computer ’experiments’ on classical fluids. I. Thermodynam- ical properties of Lennard-Jones molecules,” Physical Review, vol. 159, no. 1, p. 98, 1967.

[11] V. Mnih et al., “Human-level control through deep reinforcement learn- ing,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.

[12] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484–489, 2016.

[13] T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” arXiv preprint arXiv:1509.02971, 2015.

[14] V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in Proc. 33rd International Conference on Machine Learning (ICML), 2016, pp. 1928–1937.

[15] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. MIT Press, 2018.

[16] A. Raffin et al., “Stable-Baselines3: A reliable implementation of rein- forcement learning algorithms in Python,” Journal of Machine Learning Research, vol. 22, no. 268, pp. 1–8, 2021.

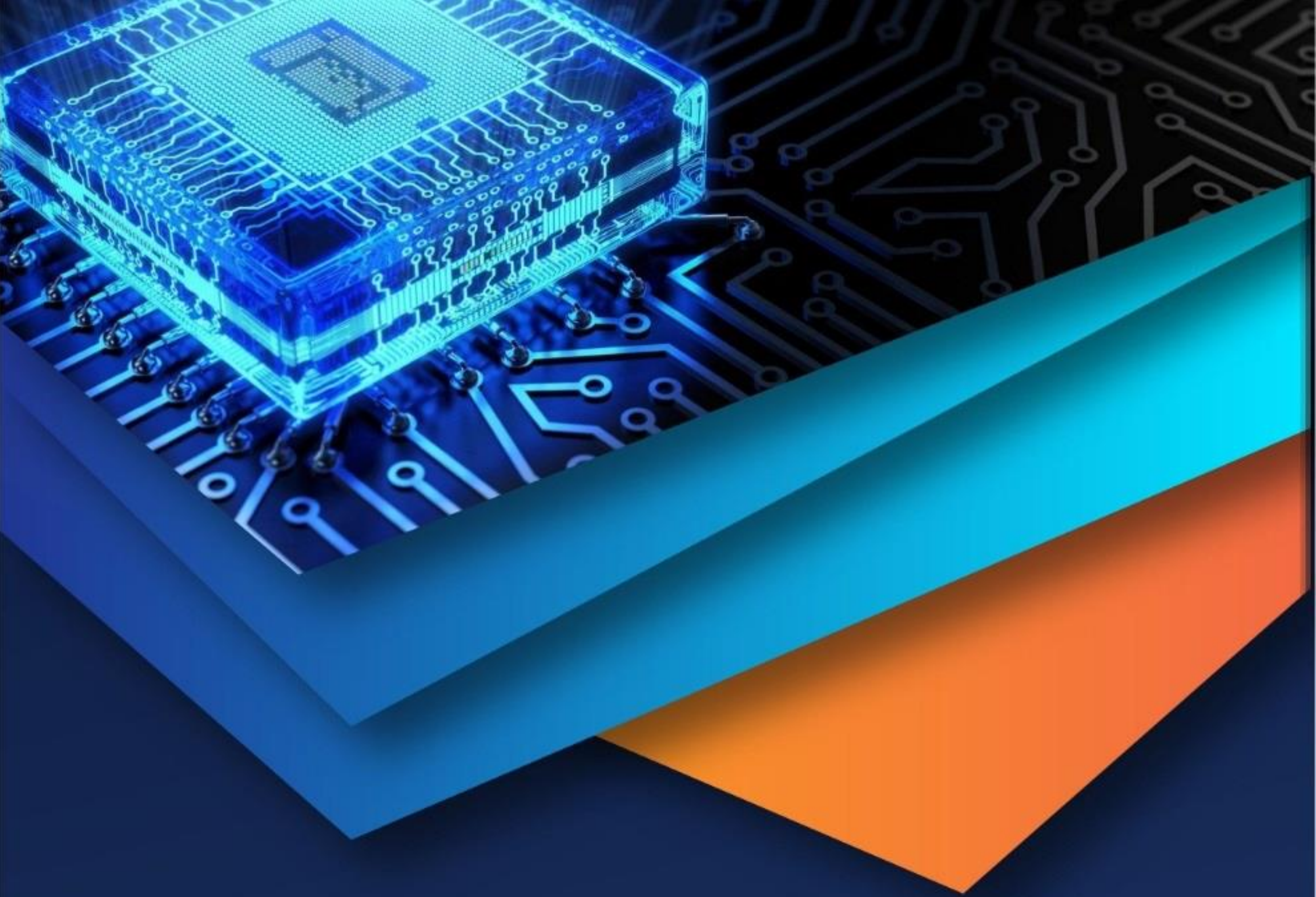
[17] S. P. N. Singh, A. K. Chopra, C. A. Sharma, and S. K. Sharma, “Autonomous drone navigation using deep reinforcement learning,” in Proc. IEEE International Conference on Signal Processing and Communication (ICSPC), 2020, pp. 234–239.

[18] G. Gaudet, R. Furfaro, and R. Linares, “Reinforcement learning for autonomous planetary landing,” in Proc. IEEE Aerospace Conference, 2020, pp. 1–12.

[19] E. Seedhouse, SpaceX’s Dragon: America’s Next Generation Spacecraft. Springer, 2015.

[20] B. Acikmese and S. R. Ploen, “Convex programming approach to powered descent guidance for mars landing,” Journal of Guidance, Control, and Dynamics, vol. 30, no. 5, pp. 1353–1366, 2007.

[21] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” The International Journal of Robotics Research, vol. 32, no. 11, pp. 1238–1274, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)