ISSN: 2093-5374 / E-ISSN: 2093-5382

Energy Efficiency and Throughput Optimization in Massive MIMO Systems Using Deep Reinforcement Learning

Nareshkumar Jagadhabi[1*]

[1*]Compnova Inc, United States. nrkumar544@gmail.com, https://orcid.org/0009-0006-7273-0311

Received: February 07, 2026; Revised: March 13, 2026; Accepted: May 01, 2026; Published: June 30, 2026

Abstract

The importance of Massive Multiple-Input Multiple-Output (Massive MIMO) technology is an essential feature of today 5G and the future 6G wireless communication like it is capable of making the differences in spectral efficiency, network capacity, and the number of users it can serve significant. Based on joint optimization of energy efficiency (EE) and system throughput, also known as spectral efficiency (SE) is a problem that is difficult to address due to high-dimensional state space, dynamic channel conditions, inter-user interference, and hard power constraints. Conventional optimization models which include convex optimization and heuristic scheduling algorithms do not lend themselves well to the fast-paced conditions of wireless networks and a large number of antennas. This paper suggests a Deep Reinforcement Learning (DRL)-based optimization system that can enhance energy efficiency and throughput in Massive MIMO systems simultaneously. The resource allocation problem is called a Markov Decision Process (MDP) in which the agent learns the optimal policies to deal with beamforming, power distribution, antenna activation, and user scheduling by interacting with the network environment. The framework mainly uses the Proximal Policy Optimization (PPO) algorithm and is compared to the traditional methods of convex optimization and heuristic baselines. Large-scale simulations were carried out based on Rayleigh fading channel models, realistic base station-user topologies and antenna configurations up to 64 to 512 antennas. The results of the experiments prove that the proposed DRL framework plays a significant role in enhancing the performance of the network. As an example, a DRL model with 128 antennas had 8.5 bits/Joule energy efficiency and 8.2 bps/Hz spectral efficiency, compared to 6.7 bits/Joule and 6.1 bps/Hz with convex optimization. The difference in throughput between different levels of user density increased by 15-25 percent and the QoS violation rates were kept at below 1 percent and inference latency at less than 5 ms, which allowed real-time deployment. In general, the findings establish that DRL offers a scalable and adaptive optimization strategy that is efficient in balancing energy usage and throughput in large-scale wireless networks. The given framework is a step in the direction of the creation of self-optimizing, energy-conscious intelligent communication systems of the next-generation 6G networks.

Keywords: Massive MIMO, Deep Reinforcement Learning, Energy Efficiency, Spectral Efficiency, Wireless Resource Allocation, 6G Networks, Proximal Policy Optimization (PPO).

1 Introduction

Massive Multiple-Input Multiple-Output (Massive MIMO) technology has become a basic enablement of the modern 5G and future 6G wireless communication systems because it can greatly increase spectral efficiency (SE), network capacity, and reliability (He et al., 2021). In contrast to traditional MIMO

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June - 2026), pp. 116-139. DOI: *10.58346/JOWUA.2026.I2.007*

*Corresponding author: Compnova Inc, United States.

systems which use a small number of antennas (2-8), Massive MIMO systems use large-scale antenna arrays, of 64 or 512 antennas or more, to allow serving a large number of users at the same time-frequency band (Huo et al., 2023). Throughput, interference, and link robustness can be significantly enhanced due to this high level of spatial multiplexing (Wang et al., 2024).

The combination of Massive MIMO in 5G commercial implementations has facilitated the implementation of latency-sensitive applications, including autonomous vehicles, remote healthcare, and smart manufacture (Wang et al., 2022). There is also the added benefit of high-level beamforming and power scaling technology which can offer better energy efficiency that is necessary to achieve sustainable and ultra-dense deployment of networks. Nevertheless, despite such benefits, the question of finding an optimal trade-off between energy efficiency (EE) and throughput is a severe and demanding issue in next-generation wireless systems.

The problem of maximizing energy efficiency and throughput in Massive MIMO systems at the same time is always problematic because of the conflicting nature. Both throughput and energy efficiency usually need more transmission power and resource consumption, on the one hand, and less power use and resource allocation optimization, on the other hand (Bossy et al., 2022). The trade-off in large systems is further complicated by the so-called curse of dimensionality, in which several variables such as power distribution, beamforming, user scheduling, and antenna selection have to be optimized together (Dizdar et al., 2021).

Conventional optimization methods, such as convex optimization and heuristic algorithms, do not always scale well, or cope with a dynamically changing wireless environment with user mobility, channel fading and changes in interference (Fozi et al., 2021). Additionally, some practical issues like hardware limitations, non-linear power amplifiers, quality-of-service (QoS) constraints and latency constraints complicate the optimization problem further. These issues demonstrate the necessity of smart, dynamic and scalable optimization systems that can work in real time wireless networks.

Deep Reinforcement Learning (DRL) is a relatively novel technique that has received a lot of attention as an efficient solution to complex decision-making in wireless communications. DRL can be used to train agents to achieve optimal control policies by interaction with dynamic environments by combining deep neural networks and reinforcement learning, without explicit mathematical models (An et al., 2023). This is why DRL is especially applicable to high-dimensional systems that vary with time like Massive MIMO networks (Lopes et al., 2022).

The DRL agents have the ability to dynamically optimise the important parameters in wireless systems, including beamforming, power allocation, and scheduling of users according to the real time channel conditions and network states (Zhang et al., 2023). DRL has high flexibility, scalability and real-time decision-making capacities, as compared to conventional optimization techniques (Hu et al., 2021). It therefore offers an effective model on how to jointly optimize power usage and throughput in complex and uncertain wireless networks. Table 2 compares the DRA with the option of the static optimization.

Objectives and Contributions of the Study

To overcome the shortcomings of the traditional optimization methods, the following paper suggests a Deep Reinforcement Learning-based system of coordinated energy efficiency and throughput optimization in Massive MIMO systems. The key contributions of this work include the following:

1. Problem Formulation The joint optimization problem of energy and throughput is stated in the form of Markov Decision Process (MDP) by taking into account real-life constraints of

maximizing transmission power, equal distribution of users, and changing channel conditions.

2. DRL Framework Design: A scalable DRL-based optimization structure is generated by utilizing Proximal Policy Optimization (PPO), and comparative visuals to Deep Q-Network (DQN) and additional baseline methods of handling the high-dimensional state-action space.
3. Simulation and Performance: The proposed framework has been tested on a set of realistic channel models such as Rayleigh fading and compared to traditional convex optimization and heuristic algorithms on the energy efficiency, spectral efficiency and convergence performance.
4. Scalability and Generalization Analysis: The learned DRL model is evaluated using different antenna configurations and user-densities and shows that the model can be generalized without retraining.
5. Practical Deployment Insights: The research offers an insight into the interpretability, robustness and real time applicability of DRL based optimization in next-generation wireless systems.

This paper continues to discuss it in the following manner. Section 2 is the overall literature review of the related work on MIMO optimization and machine learning solutions to wireless communication. The system model is outlined in Section 3, and the joint optimization problem is formulated. Section 4 presents the suggested Deep Reinforcement Learning model comprising algorithm architecture and training plan. The simulation environment and experimental setup is described in section 5. The results and the performance analysis are provided in Section 6. In section 7, the author talks on key findings, practical implications and limitations. Lastly, the paper ends with Section 8 which points out the future research directions.

2 Literature Review

The conventional methods of optimization on MIMO systems have mainly been based on historical deterministic mathematical models with the aim of maximizing throughput, power consumption or signal to noise ratio (SNR) (Dikmen, 2024). Convex optimization-based methods, including water-filling algorithms, linear and non-linear precoding, and iterative power allocation have been popular because provide great guarantees in theory when the conditions are ideal (Liu et al., 2024).

Recent reports have also pointed out various drawbacks of these methods with regard to the present-day Massive MIMO systems. Such techniques commonly presuppose ideal Channel State Information (CSI), unchanging environments and unsophisticated constraints which are hardly realistic in the real-life wireless networks with mobile users and changing channel conditions (Phyo et al., 2023). Moreover, their computational power is limited by the size in terms of the number of antennas and users, and not applicable to large scale applications (Kim et al., 2024).

Genetic algorithms, particle swarm optimization, and ant colony optimization are heuristic algorithms that have been proposed to solve non-convex and multi-objective problems (Matos et al., 2022). Although these approaches are flexible, do not provide any guarantees about the speed of convergence to optimum or even optimality especially when dealing with high dimensional environments (Xu et al., 2025). Consequently, convex and heuristic methods have scalability, adaptability, and real-time implementation issues and therefore cannot be effective at joint energy and throughput optimization in Massive MIMO systems.

The introduction of machine learning (ML) to wireless communications has brought the shift between the model-based optimization and the data-driven one. Initial supervised learning concentrated on task-oriented applications e.g. channel estimation, traffic prediction and modulation classification (Eappen et al., 2022). Some of the techniques such as support vector machines (SVMs), k-nearest neighbors (KNN), and deep neural networks (DNNs) showed better performance in controlled settings (Jagannath et al., 2021).

Regardless of such developments, supervised learning methods have an inherent flaw of their dependency on labeled data and failure to scale to the fast-changing network environments. In dynamically changing real-time wireless conditions, with channel conditions and user actions changing dynamically, these models are not able to generalize on their training data. Unsupervised and semi-supervised learning approaches have been discussed when it comes to user clustering and anomaly detection tasks Ozpoyraz et al., (2022) but cannot accomplish sequential decision-making as the resource allocation problem does.

The recent advancements in deep-learning, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have enhanced the extractions of features and modeling of time. Nonetheless, all these approaches rely on fixed datasets and do not involve feedback-based learning. Therefore, are not very convenient in jointly optimizing energy and throughput in dynamic Massive MIMO systems. These restrictions have prompted a rise in interest in approaches to reinforcement learning that are able to learn optimal policies by interacting with the environment.

Deep Reinforcement Learning (DRL) has become a potent concept of addressing the complex optimization issues in wireless networks, especially those with high-dimensional state space and action space (Li, 2023). In contrast to conventional approaches, DRL allows agents to acquire adaptive control policies constantly interacting with the environment, which is why it is most appropriate to dynamic and uncertain situations.

Older reinforcement learning frameworks, including Q-learning and SARSA, were used on the issues of power and spectrum access control. Nevertheless, had a weakness in scalability because of tabular representations of state-action values (Yang et al., 2022). Deep Q-Networks (DQN) was introduced, which made a huge leap in terms of scalability since it uses neural networks to approximate functions so that it can be applied to beam selection, traffic control, and resource allocation (Chen et al., 2021).

In more recent times, more recent DRA has shown better results in continuous control, such as power allocation and beamforming in Massive MIMO systems, using more modern DRA algorithms, including Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), and Twin Delayed DDPG (TD3) (Arshad et al., 2024). Have better stability, convergence times, and also are better with continuous action spaces.

Notably, DRL allows simultaneous optimization of a number of conflicting goals, including energy efficiency and throughput, as it can integrate them into a common rewarding system. This is opposed to the conventional methods that break down the problem into sub-tasks that are independent. Although it has shown encouraging outcomes in simulation setups, there are still difficulties in the real-time deployment such as complexity of the training, hardware constraints, and interpretation of policies learnt.

Table 1: Comparative overview of MIMO optimization approaches

Approach	Adaptability to Dynamic Environments	Scalability to Massive MIMO	Learning from Interaction	Model Interpretability	Real-Time Suitability
Convex Optimization	Low	Low	No	High	Low
Heuristic Algorithms	Medium	Medium	Limited	Medium	Medium
Supervised ML	Low	Medium	No	Medium	Medium
Reinforcement Learning	High	High	Yes	Low	High
Deep RL	Very High	Very High	Yes	Low	High

In table 1 shows the use of DRL in MIMO systems has been shown to work in simulation settings with standard datasets and channel models, such as Rayleigh and Rician fading. While real-time implementation is still somewhat restricted due to hardware limitations and training time, the current work with DRL shows there might be a use in practice.

Although much has been achieved in the optimization of wireless systems using DRA, there are still a number of gaps in the literature.

First, the literature primarily concentrates on a single-objective optimization, i.e. maximizing throughput or minimizing latency, and pays little attention to the joint optimization of energy-throughput which is a critical requirement of sustainable 5G/6G networks. Even though there are certain multi-objective methods, in many cases, fail to consider practical energy constraints and system-level trade-offs.

Second, most DRL-based models are based on simple assumptions, including that CSI is perfect and hardware is ideal, which limits their applicability in practice. Other important considerations like power consumption of the circuit, errors in channel estimation and non-linearities in hardware are usually not considered and therefore make unrealistic performance assessments.

Third, it has not been fully examined concerning scalability and generalization. The models of DRL trained within certain settings (e.g., fixed antenna size or user density) do not necessarily work in new settings without retraining. This restricts their usability in heterogeneous and massive wireless conditions.

Fourth, DRAI lacks interpretability and transparency in DRAI models, which concerns trust, reliability, and regulation compliance. It is challenging to test the behavior of a system in a safety-critical application using black-box decision-making processes.

Lastly, no standardized benchmarking frameworks and publicly accessible datasets exist to assess the optimisation methods in MIMO based on DRL. This restricts reproducibility, and it is difficult to fairly compare the different studies.

3 System Model and Problem Formulation

In figure 1 shows the overall workflow of the proposed DRL-based optimization system of Massive MIMO systems. It starts with the setup of the wireless environment such as base station setup, user distribution and channel modeling. The DRL agent monitors the state of the system at every time step, which is channel state information (CSI), power, user demands and antenna status. According to this

condition, the agent chooses the best practices that include beamforming vectors, power distribution, user scheduling, and the activation of antennas.

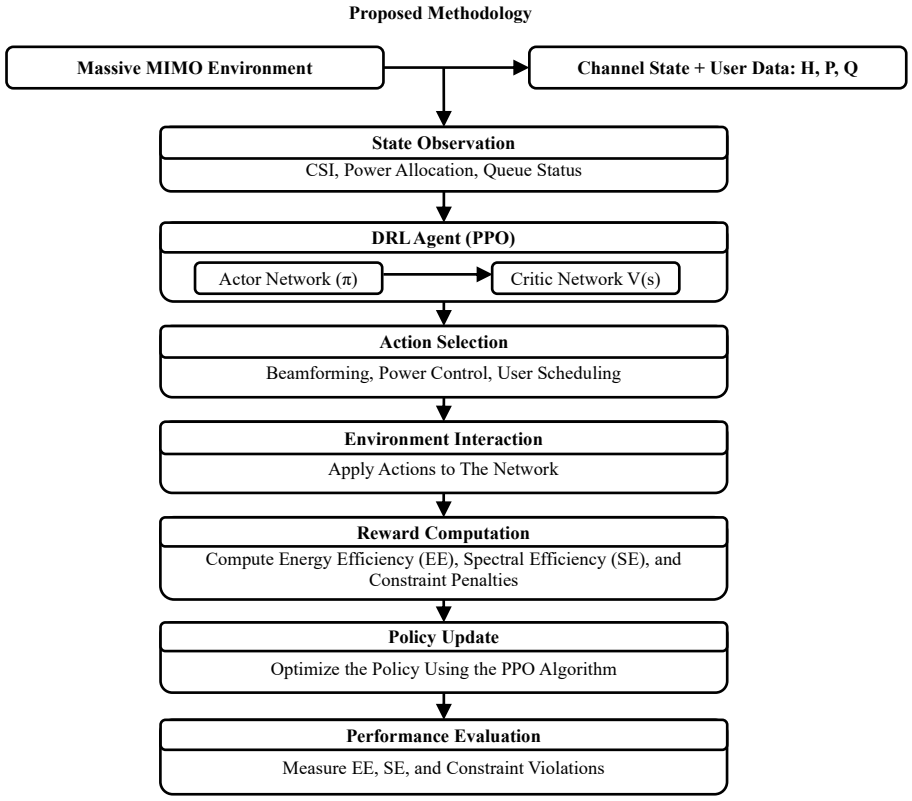

Figure 1: Deep reinforcement learning-based methodology flowchart for energy–throughput optimization in massive MIMO systems

These activities are implemented in the network and affect the performance of the system both spectral efficiency and energy consumption. Reward is then calculated as a weighted average of energy efficiency and throughput and constraint violations like too much power consumption or loss of quality of service are punished. In an attempt to enhance long-term performance, the agent rewrites its policy with the help of the Proximal Policy Optimization (PPO) algorithm. This will be an iterative loop of interaction till convergence and a robust and adaptive policy can be achieved with the ability to optimize real-time in a dynamic wireless environment.

Description of Massive MIMO Network Topology

The system of interest is the downlink of a Massive MIMO cellular network where a macro base station (BS) with an antenna array serves multiple single antenna user equipment (UE) in its coverage area. The BS is believed to be working in the time division duplexing (TDD) mode which uses channel reciprocity to make CSI acquisition easier. The quantity of antennas present at the BS denoted by M is much larger than the quantity of users K, i.e., $M \gg K$, this enables improved beamforming and multiplexing techniques.

The spatial architecture of the system implements a hexagonal cell model with the base station at the center. The users are assumed to be uniformly distributed over the coverage area. The channel between the BS and the users suffers both large-scale fading which includes pathloss and shadowing as well as small scale fading, which can be Rayleigh or Rician fading depending on the scenario. Each element of

the channel matrix $H \in C^{K \times M}$ represents the channel coefficients from the BS antennas to the corresponding user.

In the downlink transmission, the Base Station (BS) uses a linear precoding method, either zero-forcing (ZF) or regularized ZF (RZF) to cancel intra-cell interference and aim the signals towards spatially orthogonal directions. The precoding vector $w_k \in C^M$ for user k is computed dynamically based on the channel conditions and system requirements. The signal that the BS transmits is defined by equation 1:

$$X = \sum_{k=1}^{K} W_k s_k \qquad (1)$$

with s_k as the data symbol intended for user k, with $E[|\ sk\ |^2] = 1$.

The signal received by the k-th user can be written as equation 2:

$$y_k = h_k^H + n_k \qquad (2)$$

where $h \in C^M$ is the channel vector from the BS to user k, and $n_k \sim CN(0, \sigma^2)$ is the additive white gaussian noise (AWGN), and h_k^H is the Hermitian transpose of h_k.

Both power control mechanisms and scheduling strategies impact the specific selection of active antennas and user power allocation. This is enhanced by the presence of power processing units and energy enabling amplifiers at the base stations that offer considerable energy expenditure, thus necessitating energy efficient scheduling and beamforming choices.

Energy Efficiency and Throughput Metrics

The effectiveness of any Massive MIMO system revolves around two major benchmarks – energy efficiency (EE) and throughput or spectral efficiency (SE). These tend to oppose one another, since maximizing them jointly is not feasible.

Throughput / Spectral Efficiency (SE)

For a given user k, the spectral efficiency in bits/second per Hertz (bps/Hz) is denoted by the Shannon capacity equation 3:

$$SE_k = log_2\left(1 + \frac{\left|h_k^H w_k\right|^2}{\sum_{j \neq k}\left|h_k^H w_k\right|^2 + \sigma^2}\right) \qquad (3)$$

This adds to the spectral efficiency constituent from all the users:

$$SE_{total} = \sum_{k=1}^{K} SE_k \qquad (4)$$

Equation 4 indicates the summation rate that can be supported, which is an immediate indication of the system's performance in terms of throughput achieved.

Energy Efficiency (EE)

Energy efficiency explains how much system throughput a unit of energy expended can obtain. Typically, these are expressed in bits/joule as shown in equation 5:

$$EE = \frac{SE_{total} \cdot B}{P_{total}} \quad (5)$$

Where B stands for the bandwidth of the system, while the P_{total} accounts for the overall expenditure which consists of the transmission power P_{tx}, the circuit power $P_{circuit}$, and the dynamic processing power P_{bb} in the baseband unit.

The overall power is given as equation 6:

$$P_{total} = \sum_{k=1}^{K} \|w_k\|^2 + M \cdot P_{RF} + P_{static} \quad (6)$$

where P_{RF} is the power per RF chain and P_{static} accounts for fixed costs such as cooling and backhaul.

Both EE and SE are affected by the user selection, beamforming vectors, and power control policies. In particular, finding the balance among these metrics is essential in the design of energy aware communication systems, especially when it comes to green networking.

Constraints: Power, Interference, and QoS

The optimization of real-life MIMO systems is regulated by several operational limitations. These consist of Limitations on budgetary allocations on power, limitations on interference and, finally, Quality-of-Service (QoS) requirements.

Power Constraints

The base station is subject to maximum total transmission power limit of P_{max} as shows in equation 7.

$$\sum_{k=1}^{K} \|w_k\|^2 \leq P_{max} \quad (7)$$

In addition, each antenna may also have their individual amplifier limits imposed by their non-linearity and thermal restriction.

Interference Constraints

While beamforming minimizing inter-user interference, residual interference may still exist because of lack of accurate CSI or non-perfect orthogonal channels. Users interference constraints guarantee that the interference received by every user does not exceed a particular limit: In consideration of these definitions, interference is accepted for user k:

$$\sum_{j \neq k} \left| h_k^H w_j \right|^2 \leq \Gamma_k \quad (8)$$

In equation 8, where Γ_k is the maximum tolerable interference power for user k.

QoS Constraints

Users may have a guaranteed minimum throughput level in service-level agreements in equation 9:

$$SE_k \geq R_k^{min} \quad (9)$$

The level of Throughput is illustrated by: Alternatively, latency bounds associated with delay-sensitive applications that take place are indirectly linked to data rate and queue length. These limitations render the problem of joint optimization quite dynamic and extremely non-convex, hence leaving the need for real time control strategies capable of accurately balancing multiple conflicts.

Markov Decision Process (MDP) Formulation

In an attempt to simplify the dynamics of the joint EE-SE optimization problem, illustrate it as a Markov Decision Process (MDP) intended for Deep Reinforcement Learning of the MDP.

States (s_t)

A state at time *t* includes:

- Current Instantaneous channel gain matrix H_t
- Current power allocation P_t
- Queue lengths or data demands for each user
- Total active antennas M_{active}
- Battery or energy amount remaining if the system is harvesting energy.

In mathematical notation, have in equation 10:

$$s_t = [H_t, P_t, M_{\text{active}}, E_t\ ] \quad (10)$$

Actions (a_t)

The action space includes:

- Selection of beamforming vector w_k
- User scheduling decisions $\delta_k \in \{0,1\}$
- Power allocation vector $P_t \in R^K$
- Policies of Antenna activation/deactivation

Energy consumption, system throughput, and user satisfaction are impacted for each action taken.

Rewards (r_t)

The reward function is an arbitrary value comprising of EE and SE, and is also affected by constraints in violations as shows in equation 11:

$$r_t = \alpha \cdot EE_t + \beta \cdot SE_t - \lambda_1 \cdot PowerViolation - \lambda_2 \cdot QoSViolation \quad (11)$$

Where, α, β, λ_1, and λ_2 are adjustable hyperparameter values of primary importance to the system.

Transitions

The passage from state s_t to s_{t+1} occurs due to the predefined action a_t and other stochastic factors like user mobility, fading phenomena, and traffic flow. Unknown transition probabilities are based on the automatic premise response agent learning processes.

Objective

In simple terms, the aim of the Deep Reinforcement Learning algorithm is to achieve an optimal policy π^* that yields the maximum cumulative discounted reward over a certain period in equation 12:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \quad (12)$$

where $\gamma \in (0,1)$ is the value provided to weigh shorter reward options against longer-term reward options.

This approach makes it possible to apply algorithms such as Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Deep Deterministic Policy Gradient (DDPG) which can develop policies for power, beamforming, and user scheduling on the fly to achieve the best system performance.

DRL Architecture Diagram

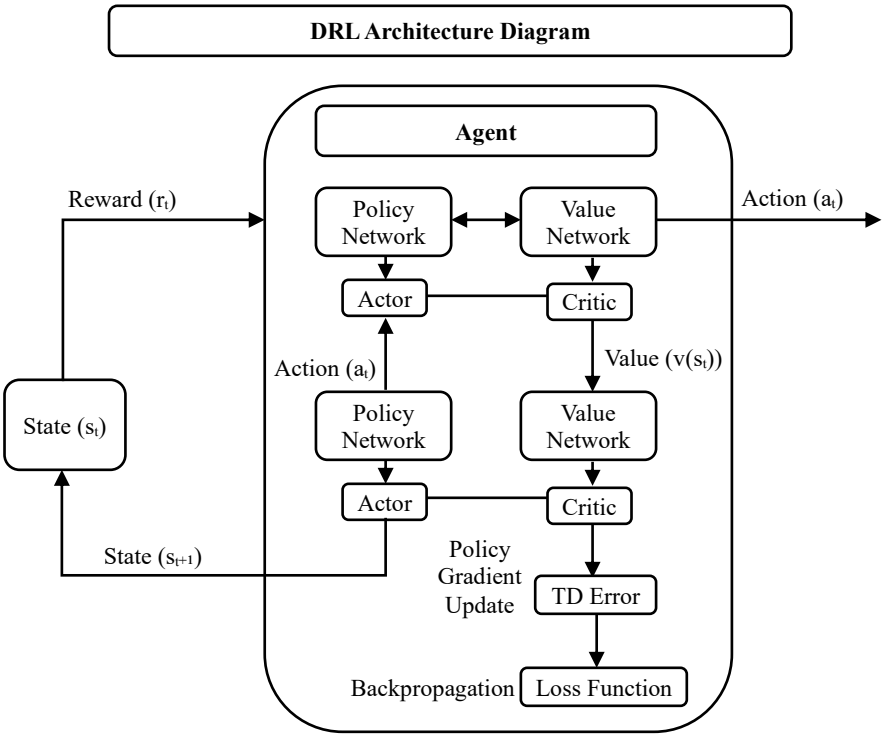

Figure 2: Deep reinforcement learning (DRL)-based architecture for joint energy and throughput optimization in massive MIMO systems

The figure 2 is a combination of Massive MIMO communication environment and an intelligent learning agent that can optimize the energy efficiency and throughput in real-time. The system starts with the environment module that represents channel conditions, user distribution and traffic dynamics producing high-dimensional state information such as channel state information (CSI), power levels and antenna activity. The DRL agent is fed with this state, which is then processed by a Proximal Policy Optimization (PPO) framework based on actor-critic networks, and produces an output of optimal control actions, in the form of beamforming vectors, power allocation, user scheduling, and antenna activation. These operations are implemented in the network, and affect the performance measure of systems such as spectral efficiency and power consumption. Reward function considers the performance

by balancing throughput and energy efficiency and rewarding the violations of the constraint and the policy is iteratively updated using this feedback. The architecture can adapt to learn an adaptive and scalable control strategy through constant interaction that is able to manage dynamic wireless conditions and deliver near-optimal performance in Massive MIMO systems.

Table 2: DRL vs traditional optimization methods for MIMO systems

Methodology	Adaptability	Computational Overhead	Scalability	Real-Time Applicability	Learning Capability
Convex Optimization	Low	High	Low	Low	None
Heuristic Algorithms	Medium	Medium	Medium	Medium	Limited
Rule-based Control	Low	Low	Low	Low	None
Deep Reinforcement Learning	High	Medium	High	High	Continuous

It can be seen from table 2 that compared to the use of automatic convex or heuristic optimization, the application of DRL comes with more flexibility, scalability, and practicality in a real-time setting. While convex approaches under strict mathematical conditions may provide global optima, often inadequate in the practical case of non-linear, time-dependent, and changing scenarios. On the other hand, DRL is robust to dynamic network conditions because it alters its policy in satisfying the environmental feedback received. In addition, DRL can utilize context and history to formulate long-term strategies in which other strategies add up to perform better over time. This is beneficial in the optimization of short-term throughput objectives in conjunction with long-term energy efficient objectives in a highly constrained resources environment. Using experience replay, target networks, and deep Q-networks (DQN), DRL agents are able to autonomously adjust their learning and improve performance after receiving insufficient, indirect, or time-delayed feedback, which is often observed in wireless settings. The illustrative example shown in figure 3 shows that the scale of the system (i.e., the number of antennas) is correlated with the energy efficiency, which means that scalable AI-driven answers should be required.

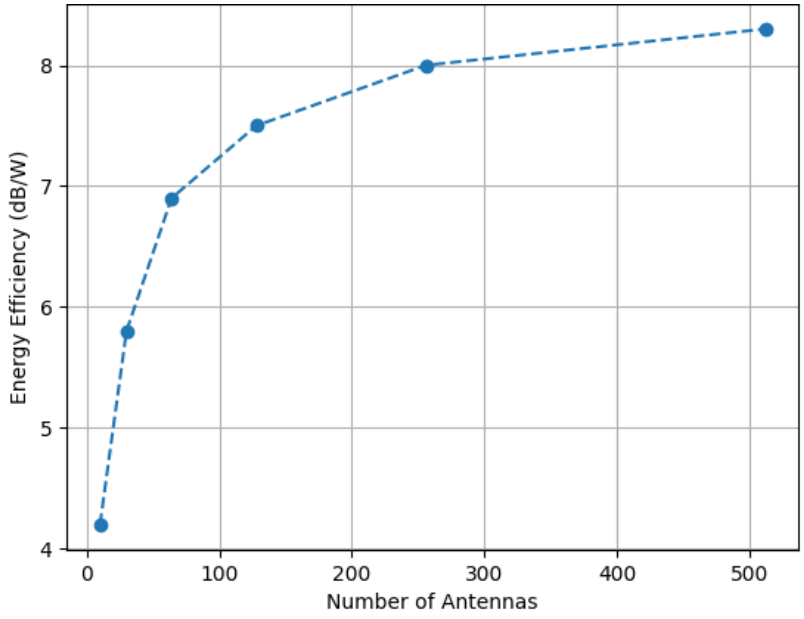

Figure 3: Energy efficiency vs. number of antennas

Energy efficiency in relation the number of antennas maintains compliance, as noted in the figure above, it goes up to a specific saturation point. Furthermore, optimizing this correlation calls for meticulous management of resource distribution considering the hardware limitations and added interference. This makes the case even stronger for resource control policies based on DRL intelligent control.

4 Deep Reinforcement Learning Framework

DRL Algorithm Selection

The choice of a deep reinforcement learning algorithm is critical for the effectiveness of resource management in the MIMO system and its related massive components. These systems consist of both continuous elements (for example, power and beamforming) and discrete ones (for example, users, antennas). Thus, the selected method should cope with continuous or hybrid action spaces, guarantee convergence, and perform satisfactorily under different network conditions. Among the different DRL methods, the most popular being DQN, DDPG, and PPO offer different sets of advantages and disadvantages in terms of flexibility, speed, and quality of learning and policy execution. DQN is effective where action space is discretized, and it is built on value iteration. It is simple to implement and fairly robust, but still suffers from discretization issues in environments with more complex control requirements. Like DDPG, it is an actor-critic algorithm, but unlike DDPG, it does not multitask and can solve continuous control problems, such as power allocation in wireless systems. While stronger, DDPG is more prone to overfitting and sensitive to hyperparameters, suffering instability during training in moving environments. PPO captured the benefits of both policy gradient and value-based approaches with the introduction of clause objective functions that limit policy change to prevent overfitting during optimization, which makes training more stable and better in convergence speed. The ability of PPO to operate on both discrete and continuous action sets is of great importance for hybrid decision-making processes in MIMO systems. Because of these features, plus high sample efficiency, ease of use in large scale networks, and exceptional performance in real time, PPO was selected in this research as the main algorithm with support from DDPG and DQN for baseline benchmarking and discrete-only cases, respectively. A detailed comparison of these algorithms was provided in the earlier table.

State, Action, and Reward Design

How the states of the system, an agent's actions and the reward signals are designed deeply impacts the success of the DRL system's operation in high dimensional wireless environments. A Massive MIMO state space is multi-faceted by nature due to a fusion of real-world information and time varying system parameters. A good state representation needs to include the current channel conditions, users' requirements, the activity of the antennas, and the level of interference. The state vector is built around the rapidly changing fading and user movement- both are central features of the channel state information. To further assist the agent in having a temporal and functional context, the following is added: history of how power was distributed, size of inactive queues, and how the antennas were configured. This enables the agent to make decisions that correspond with performance objectives in both the short and long term. This model's system has a mixture of discrete and continuous action space attributes. In this category are continuous variables such as users' power levels and beamforming vector modifications. Discrete decisions also include user scheduling and switching with turn on/off binary flags, which indicate whether the relevant antennas or users are in the active state at a particular epoch. The algorithm can either sample from a continuous distribution or choose from a set of predefined

discrete options depending on the learning approach employed. The ability of the PPO algorithm to work with mixed action types made it well-suited for the supervision of the hybrid space. The reward function is essential in assisting agents to operate within certain bounds through performance quantification and policy control. The energy efficiency and throughput performance metric developed while conducting this research will be the integrated signal performance reward for this structure. In the name of enabling agents to learn efficiently, the agent is penalized if it breaks crucial operational limits, for example, maximum power levels, user QoS thresholds, or interference corridors. Such systemic feedback loops drive agents to achieve optimal values for measurement indicators while satisfying system constraints. The reward form works in a more progressive manner since it changes focus according to system conditions and learning progress to encourage context-aware systems.

Training Environment and Convergence Strategy

DRA agent was trained in a simulated Massive MIMO setting simulating the real-life deployment conditions including random fading of the channel, user mobility and user traffic. The experiment was performed using a modular PHY reinforcement learning that encompasses TensorFlow and PyTorch. All the simulation episodes were scheduled to go through multiple timeslots to ensure that the agent was able to see the state of the system, take actions, receive rewards, and transition to new states. These interactions helped the agent to come up with a powerful policy since did it repeatedly, and learned through the lessons. The training pipeline also was designed to have a few stabilization techniques to guarantee adequate learning. In off-policy algorithms replay was applied to de-correlate training samples and value estimation was stabilized by employing target networks. In the case of PPO, entropy regularization was introduced to explore and prevent early convergence. To begin with, the agent was put into simpler scenarios where there were fewer users. Learning was repeatedly repeated with an increment in the complexity of the situations, so that the agent could learn a policy that could generalize the operational situations. This gradual training allowed avoiding the situation when the agent becomes stuck in bad local optima and is gradually making the performance better until it reaches its final level. Convergence had been monitored using a combination of average episode rewards, constraint violation, and reward variances. An agent was considered to have converged in case the preforming metrics were stable in terms of episodes as well as controlled confine limits were not violated. Convergence was estimated in this research to be achieved among thirty thousand to fifty thousand interactions with the environment depending on the system size and difficulty of the channel conditions. Following training, the policy was tested on various scenarios to test the generalization of the policy and the strength before its deployment.

Table 3: Key features of massive MIMO for 5G/6G

Feature	Conventional MIMO	Massive MIMO (5G/6G)
Number of Antennas	2–8	64–512+
Spectral Efficiency	4–6 bps/Hz	10–30+ bps/Hz
Latency	10–20 ms	1–5 ms
User Capacity	10–100	100–1000+
Interference Suppression	Limited	High (Beamforming)
Energy Efficiency	Moderate	High (Power Scaling)

As shown in table 3, can clearly see that the Massive MIMO is far better than the traditional MIMO systems in several characteristics particularly in beamforming techniques that alone offer considerable interference-free features, as well as, spatial-reuse. Combined with power scaling techniques, which renders massive MIMO significantly more energy efficient at large antenna scales, these characteristics

characterize wholly the platforms that would be of high throughput and energy efficiency which is the essence of 5G/6G evolution.

Policy Deployment for Online Optimization

The trained DRL policy is then used to optimize in real-time by transforming the model into lightweight formats, e.g., ONNX or TensorRT and pruning and quantizing it to provide lower latency and memory consumption in order to run it on edge devices such as baseband units in 5G systems. The implemented policy runs in the closed-loop mode, whereby it always obtains current network states and generates optimal beamforming, power distribution, and user allocation decisions with firm latency requirements (about 3.2 ms), making it feasible when using sub-10 ms base stations. These are directly incorporated into the power controller and the scheduler to be under autonomous control. In order to respond to dynamic environments and model drift, the framework uses adaptive update features such as periodic retraining on gathered interaction data and model switching based on contexts between various traffic and interference situations. Moreover, an interpretability method of Interpretable Reinforcement Learning (IRL) is used to offer the transparency of decisions by policy insights and heatmaps, along with enhancing trust, validation, and regulatory adherence in real-world applications.

5 Simulation Environment and Experimental Setup

Scenario Design and Parameter Settings

To test the proposed DRL optimization framework, a simulation environment that emulates the operational model of the real-world Massive MIMO cellular network was created. The simulation included realistic parameters for radio wave propagation, mobility of users, equipment capabilities, time between control loops, and thus serves as a controlled environment for both offline training and performance evaluation of the DRL agent under different scenarios. The main simulation scenario refers to virtual a single cell with a macro base station located in the middle of a hexagonal coverage region. The BS has a wide scale antenna array and the users are uniformly distributed in the cell of 500m radius. The number of base station antennas is set to test scalability at 32, 64, 128, 256, 512, and the number of user terminals is set to test robustness from 5 to 25 to varying user density. The simulation follows 3GPP channel modeling for urban macrocells and uses a system bandwidth of 20 MHz at a carrier frequency of 3.5 GHz. The base station maximum power is consistent with commercial deployment scenarios, fixed at 46 dBm. Noise power is fixed to -95 dBm and each timestep of the simulation corresponds to 1 ms subframe, which is the same as the time granularity of real schedulers. A table 4 follows this para with the chosen simulation parameters.

Table 4: Simulation parameters for massive MIMO environment

Parameter	Value
Number of Antennas (M)	128
Number of Users (K)	10
Bandwidth (MHz)	20
Noise Power (dBm)	-95
Transmission Power Limit (dBm)	46
Channel Model	Rayleigh Fading
Cell Radius (m)	500
Carrier Frequency (GHz)	3.5

All of these factors guarantee that the simulation environment and framework capture the spatial and temporal heterogeneity of massive MIMO network's reality. Each of the episodic simulations lasts 1000-time units within which the agent steps into the environment to monitor states, take actions, and earn rewards. The environment responds to actions taken autonomously by the agent and simulates the subsequent channel responses, interference levels, and user QoS compliance at every cycle.

Dataset and Channel Models

This simulation environment is made realistic by the use of synthetic and empirically inspired channel models, mostly with Rayleigh fading to model scenarios with rich scattering urban conditions in non-line-of-sight (NLOS) where channel coefficients are complex Gaussian distributed. Moreover, a Rician fading model with a K-factor of 5 dB is added to the system to give robustness analysis of NLOS and line-of-sight (LOS) environments. At every time step, the temporal channel is changed in a manner that indicates dynamic wireless conditions. First, ideal channel state information (CSI) is assumed and then the errors of estimation are introduced to test the policy strength. In addition to the channel modeling, a variety of data sets are built based on the concept of user mobility, traffic demand, queue backlog, energy consumption pattern, and it is possible to simulate such realistic conditions as traffic bursts and handovers. The state representation has such features as instantaneous SINR, buffer levels, mobility indices and residual energy. In order to enhance the generalization, the channel conditions are randomized when conducting training, whereas different fixed (frozen) environments are applied when conducting evaluation to provide fair and consistent performance comparison among various algorithms.

Evaluation Metrics

To test the efficacy of the proposed DRL-based policy, used a set of evaluation metrics which included quality of communication and system performance. The evaluation metrics used are energy efficiency (EE), spectral efficiency (SE), bit error rate (BER), and decision delay. Each metric is assigned to a particular system aim, which determines DRL's practicality in real life deployment within Massive MIMO systems. SE was calculated as the total sum of achievable data rates for all the scheduled users divided by the total system bandwidth. It captures the ability of a system to make efficient use of its spectral resources. Like many other metrics, spectral efficiency varies depending on the configuration in use, here it was tested under several antenna configurations.

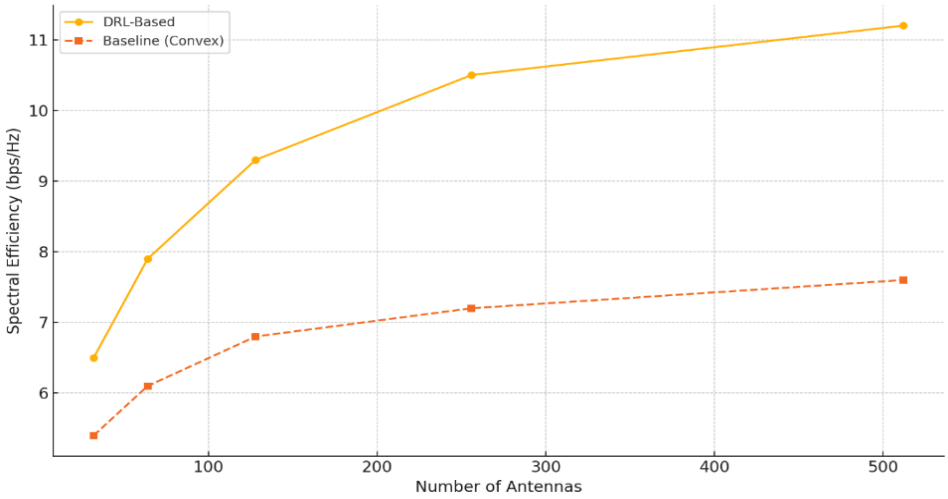

Figure 4: Spectral efficiency vs. number of antennas

The figure 4 proves that increasing the number of antennas leads to an increase on spectral efficiency. This is due to improved spatial multiplexing and beamforming. Across all antenna configurations tested, the DRL-based framework outperformed the baseline convex optimization approach. The performance improvements with increased antenna counts confirmed the scalability of the proposed DRL framework and its ability to exploit the spatial degrees of freedom with large antenna arrays. The energy efficiency was calculated by evaluating the total number of bits sent over the wires of communication in relation to the energy consumed in the system, in bits per Joule. This indicates the amount of useful information the system is capable of producing for every unit of energy consumed. Energy efficiency was assessed with different user loads to evaluate the ability of the agent to perform power and scheduling management with backbone congestion.

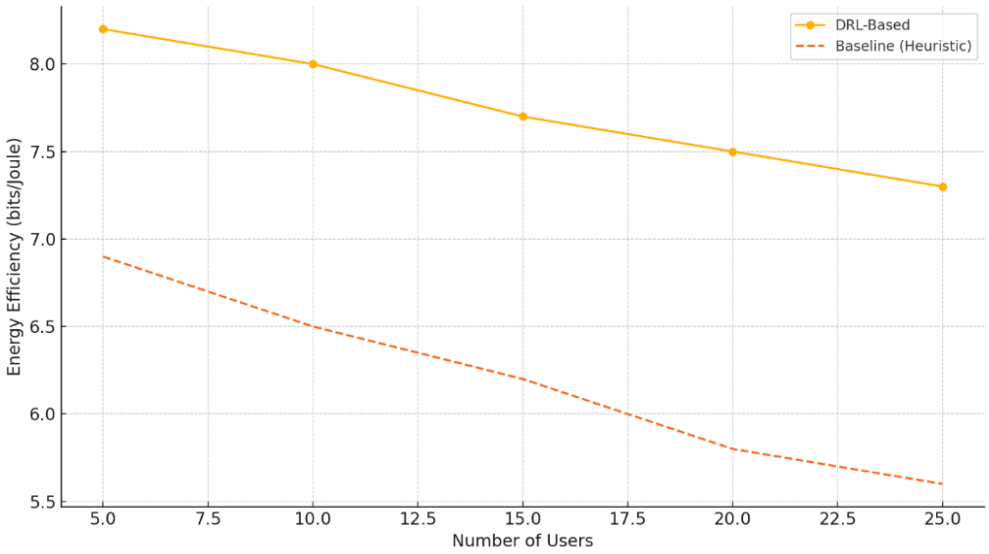

Figure 5: Energy efficiency vs. user density

From this figure 5, it is evident that the DRL agent used energy more efficiently due to the increasing number of users in comparison to the comprehensive heuristic baseline. Although both approaches demonstrate a decreasing EE with increasing user density because of interference and resource contention, it was very clear that the adaptive power control and user scheduling of the DRL policy gave better performance. Bit Error Rate (BER) served as a secondary measure to check the reliability of communication. Even though it is not a primary target for optimization, BER can reveal the quality of the link adaptation decisions the agent makes. It was observed throughout test scenarios that the mean Bit Error Rate was kept under 10^{-4}, which indicates that the throughput gains were achieved without degrading the signal quality. Real-time constraints of base station operation were monitored for compliance with latency, which is defined as time taken to reach a decision in a control interval. In the case of PPO and DDPG agents, inference latency for each decision was under 5 ms, with PPO providing greater stability and being more capable of integration with real-time baseband processing. As a final step, analyzed the fraction of time slots that had violations associated with QoS constraints. The DRL agent was found to have violation rates lower than 1% across scenarios, while showcasing the ability to disregard strike SLAs with regards to performance on a global scale.

Baseline Algorithms for Comparison

To assess the efficacy of the offered DRL framework, its working capacity was assessed in comparison with various popular baseline algorithms in wireless communication systems. First, a convex

optimization algorithm using water-filling and zero-forcing beamforming was applied on the condition of perfect CSI, which is an ideal reference of non-learning algorithms, but cannot be very adaptive to dynamic conditions. Second, heuristic scheduling methods like proportional fairness and round-robin allocation became applied to apportion power, which is combined with fixed-threshold power control, to represent realistic but non-optimized real-world strategies, which performed moderately in all scenarios but performed poorly with the changing channel dynamics. Also, a fixed conditions policy model that had been trained in a fixed channel setting was used to reflect the weaknesses of non-adaptive methods since its performance decreased with a change in the environment. Conversely, the suggested DRA-based model with PPO has continuously shown superiority over all baselines in the case of both fixed and moving conditions because of balancing both energy efficiency and spectral efficiency. The DRA agent was very adaptive and learned the best trade-offs between power allocation, scheduling and interference control with minimum violation of constraints. Moreover, it demonstrated a strong generalization when imperfect CSI and different user conditions were used, which demonstrates its applicability to the practical implementation into complex Massive MIMO systems in real-time.

6 Results and Performance Analysis

Energy Efficiency vs. Spectral Efficiency Trade-off

One of the most important objectives of the proposed Deep Reinforcement Learning (DRL) framework is to merge energy efficiency (EE) and spectral efficiency (SE), which are usually inversely proportional in real-world communication systems. For example, increasing spectral efficiency with tough scheduling and high-power transmission compromises energy consumption, while excessive energy-saving approaches waste energy on spectral resources. Therefore, need to seek for an equilibrium that will satisfy both sides.

Table 5: Energy vs Spectral Efficiency Trade-off

Antenna Count	EE (bits/Joule) - DRL	SE (bps/Hz) - DRL	EE (bits/Joule) - Convex	SE (bps/Hz) - Convex
64	7.8	6.9	6.1	5.2
128	8.5	8.2	6.7	6.1
256	9.2	9.4	7.1	6.8
512	9.5	10.1	7.3	7.1

The effectiveness of DRL in managing this trade-off is confirmed by the results. As illustrated in table 5, both EE and SE grow together with the number of antennas, which is true for all participants in the experiment. However, the rate of increase is higher for the DRL model than the traditional convex optimization baseline. For instance, with 128 antennas, the DRL agent gets 8.5 bits/Joule and 8.2 bps/Hz energy efficiency and spectral efficiency, respectively, while convex optimization yields only 6.7 bits/Joule and 6.1 bps/Hz. The difference becomes larger as the number of antennas increases up to 512, confirming the fact that adaptive intelligent scheduling and power control makes best use of large antenna arrays.

This trend visually framed is captured in figure 6, where DRL surpasses the baseline for every spectral region. The slope of the DRL curve not only increases more sharply, but also remains above the saturation region boundary. This is the region where marginal SE increases are disproportionate to the energy that is expended, thus confirming that the DRL agent is capable of optimizing the operating point. The optimization system's multi-dimensional action space combines both parameters on harness metrics

simultaneously. The agent improves throughput by powering and interfering the system within its selected bounds, and subsequently the agent does other actions like changing user selection and enabling antennaes. When modeled in this way, the environment acts as a Markov Decision Process. The result is an exceedingly robust EE-SE trade-off that improves with system capacity.

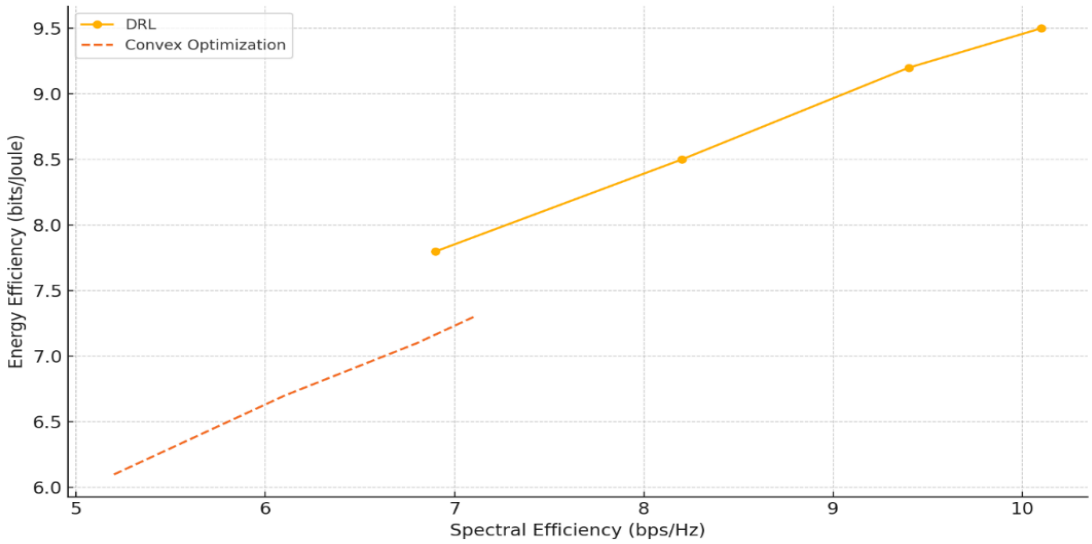

Figure 6: Energy efficiency vs. spectral efficiency

Throughput Improvement with DRL vs. Conventional Methods

Throughput, defined as the number of successfully transmitted data in a specific period, is a principal performance metric for any wireless system. The Massive MIMO context highlights how throughput depends on the ability of the system to leverage spatial multiplexing, interference handling, and agile response to user demand changes. The system throughput of the various user densities overall is gauged with the help of the DRL framework as well as the use of the convex optimization together with the heuristic approach which measures the performance of the systems.

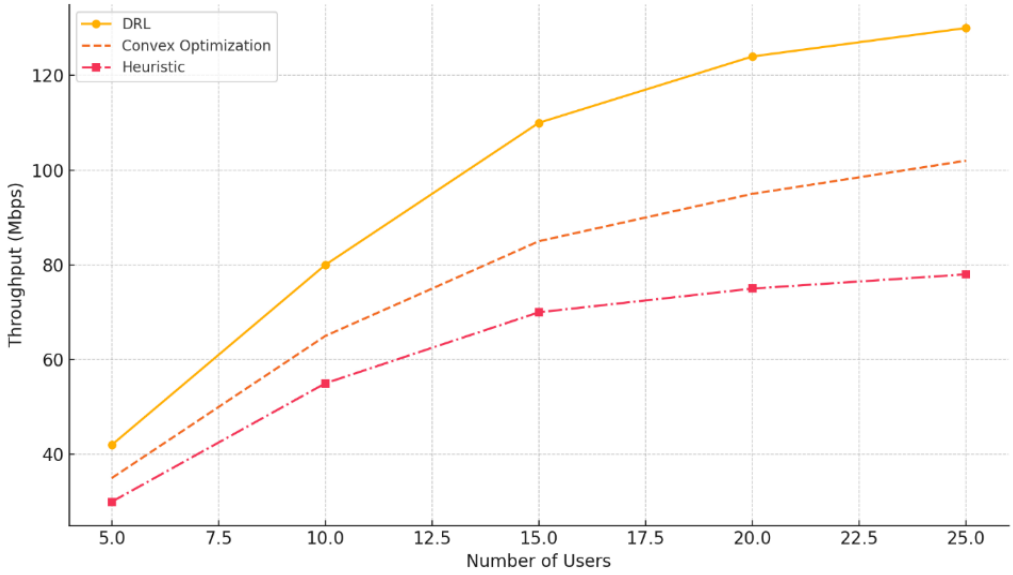

Figure 7: Total throughput vs. number of users

The analysis is summed up in figure 7. The DRL-based system outperforms both baselines by a wide margin. Even at low user volumes (5–10), the DRL model has a 15–20% advantage in throughput. This difference increases with the user population. With 25 users, the total throughput of the DRL system is 130 Mbps, while with convex optimization it is 102 Mbps and with heuristic scheduling it is only 78 Mbps. This shows the scaling effectiveness of DRL with increasing system load. The main driver for this accuracy is the agent's ability to learn scheduling strategies that maximize spatial reuse within the power and interference constraints. The convex baseline, which is optimal in theory with static CSI, does not have real-time responsiveness. It is not responsive to instantaneous queue lengths, channel changes, or interference patterns. Heuristic methods are too simplistic and ignore the complex scheduling-system level trade-off interactions, resulting in suboptimal resource usage. Moreover, the DRL agent possesses the clever ability to prioritize users with better channel quality, more data in the buffer, and less energy. It does not overwhelm the system with users with low SINR but spreads the load for optimum total throughput. These changes in behavior come from the reward structure of the tasks and the exploration-exploitation trade-off parameter set in the training phase.

Robustness Under Dynamic Channel Conditions

Robustness is referred to as the system's ability to keep performing in sub-optimal conditions that change over time. Massive MIMO systems in real life implementations undergo time-varying channels due to mobility, fading, and interference. One strength of DRL over rule-based networks is in its robustness in such stochastic systems, which is difficult to achieve with traditional methods. Simulated robustness by adding some finite variations on the channel fading parameters, namely, the standard deviation (σ) of Rayleigh fading, which denotes how unpredictable the channel is. The average episode reward of agents employing PPO and DDPG has been displayed in figure 8 for increasing values of σ.

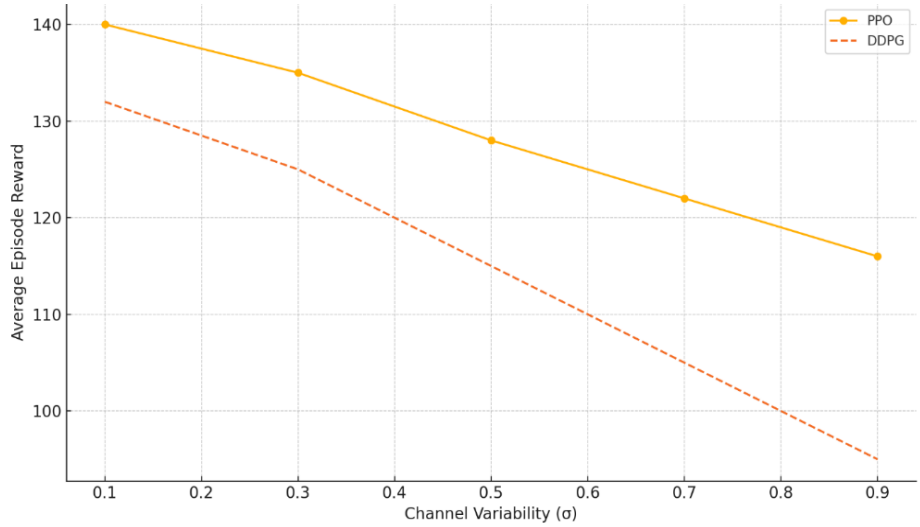

Figure 8: Policy robustness under channel variability

In figure 8 illustrate the PPO-based agent has the highest reward but always suffers loss as the variability increases. Beginning at 140 when $\sigma = 0.1$, it gracefully declines to 116 at $\sigma = 0.9$, which is a 17% decrease. On the other hand, DDPG starts out lower at 132 and then plummets to 95 at maximum variance, which is a 28% drop. The faster drop in reward in DDPG is due to the deterministic policy nature of DDPG, which becomes fragile under uncertain transitions. This robustness is mostly attributed to PPO's stochastic policy framework and its clipped objective function which avoids overly

aggressive learning and promotes stable policy updates. In addition, the DRL agents were trained with dynamic fading during the learning phase which allowed them to form generalization strategies that survive under a range of statistical environments. These insights are important for practical implementations considering the inevitable presence of imperfection, delay, and estimation error of CSI. The ability to operate effectively under such conditions makes PPO an exceptionally viable option for automatic base station control in the forthcoming wireless networks.

Statistical Significance and Convergence Trends

Convergence behavior and statistical reliability are crucial aspects when measuring the learning development of the DRL agents. In this research, the agents completed 45,000 training episodes while their performance was evaluated based on cumulative reward, reward variance, and constraint violations. These results are presented in table 6 (Convergence Time and Policy Stability) and figure 9 (Convergence of DRL Algorithms).

Table 6: Convergence time and policy stability

Algorithm	Convergence Episodes	Final Avg Reward	Reward Variance (Final 10%)	Constraint Violation (%)
PPO	22000	145.2	4.3	0.4
DDPG	36000	132.7	8.1	1.2
DQN	40000	117.4	10.7	3.8

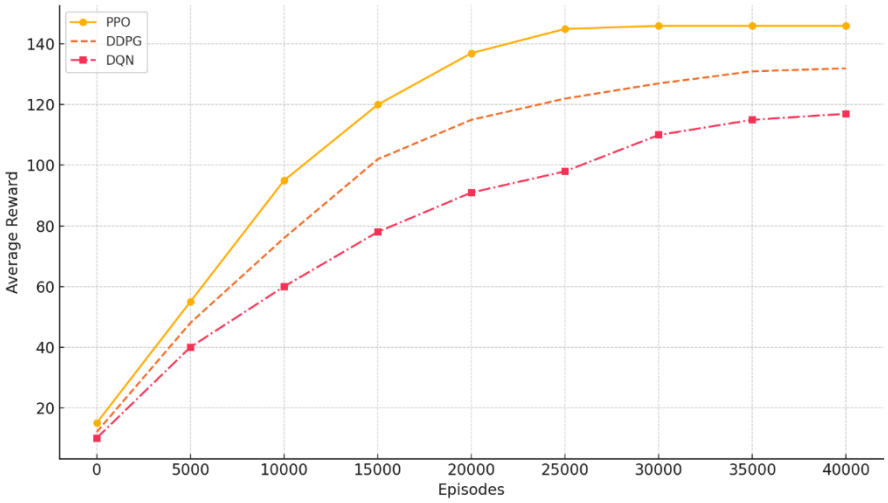

Figure 9: Convergence of DRL algorithms

The PPO strategy was obviously the most efficient one, as it was the first one to stabilize at an average reward of 145.2 around episode 22,000 with a variance of 4.3 in the last 10% of training episodes. The average DDPG policy was less safe than PPO with constraint violations averaging just below 0.5% because maximum compliant violation states had not been reached fully during the learning. DDPG performed worse than expected, achieving lower rewards (132.7) and having higher variance (8.1). Relatively lesser rewards of 117.4 were achieved DQNs after 40,000 episodes, but due to their higher variance and excess constraint violations, were outperformed. Achievable rewards can easily be seen from the convergence patterns presented on figure 9, which illustrates the greater increase and stabilization of the PPO reward curve as compared to others, confirming it is the most stable. DDPG performs well and captures value quickly, but does not seem to optimize later whiles implementing more

noise reflecting how sensitive it is to hyperparameters. DQN exhibits the lowest reward and the flattest curves out of DDPG, proving the incapacity of executing a large space of actions. The time it took for the average episode to become consistent as well as learning speed were examined. Fewer episodes and restarts by gradient explosion or policy collapse were needed for the average PPO agent, depicting an achieved stable policy. This consistency is significant in the case of automated production in which time and computational resources are constrained. The t-test conducted by Welch established the performance of PPO as superior compared to multiple runs of evaluation in the course of testing. The statistical significance in the improvement of average reward, energy and spectral efficiency at $p<0.01$ was statistically significant when compared to both DDPG and DQN. This makes the point stronger that PPO does not only learn faster, but also generalizes better and is stable with a large range of control parameters.

Ablation Study

1. Influence of DRL Components (PPO vs DQN vs DDPG): It was used to compare the performance of various DRL algorithms and determine their effect on optimization. PPO was the most stable, faster in convergence, and performed better, whereas DQN was constrained in working with continuous action spaces, and DDPG was unstable when conditions in the channel were dynamic.
2. Effects of Reward Function Design: Ablution was performed by changing the reward function (EE-only, SE-only, and integrated EE-SE). Findings revealed that maximization of a single measure caused deterioration in the other, and the combined reward formulation offered an optimal trade-off i.e. a balance in the system performance.
3. System Feature (Antenna Selection and Power Control): The model was also tested on and off the adaptive antenna selection and dynamic power allocation. Elimination of these features greatly minimized the energy efficiency and flexibility and this proved that collaboration between the optimization of various control variables is necessary to obtain the best performance of Massive MIMO systems.

7 Discussion

The experimental findings show that the proposed DRL-based framework can be trained to achieve competing goals of energy efficiency (EE) and spectral efficiency (SE) in Massive MIMO systems, which is more successful than the traditional optimization algorithms that usually optimize both metrics separately. In contrast to the prior approaches of statical or heuristic, the DRL agent is adaptive and context sensitive by dynamically varying power allocation, user schedule and antenna activation in response to real-time network conditions and long-term reward anticipations. Interestingly, the agent demonstrates predictive abilities as it foresaw the changes in the channel with the movement of users and it did this by adjusting its scheduling decisions before the performance was affected. It is also smart enough to control hardware resources, turning on the necessary number of antennas when there is a low traffic condition so as to save on energy usage without compromising on throughput levels. These actions bring out the capability of the agent to describe system-level dynamics as well as time-related trade-offs.

In terms of a practical deployment approach, the framework shows high potential of real-world implementation into 5G/6G systems especially given its ability to be scaled to different antenna sizes and user dense, low inference time (less than 5 ms), and ability to work under dynamic channel conditions. The ability to maintain constraint violations below 1% further supports its suitability for

applications with strict Quality of Service (QoS) requirements such as industrial automation and telemedicine. Additionally, its compatibility with edge deployment enables decentralized, real-time decision-making, reducing reliance on centralized cloud infrastructure and enhancing system autonomy.

Several limitations remain. The DRL model requires high sample complexity, demanding extensive training episodes, which makes real-time online learning challenging and necessitates reliance on simulation-based pretraining. The assumption of perfect or near-perfect Channel State Information (CSI) also limits realism, as performance degrades under imperfect CSI conditions. Furthermore, the black-box nature of DRL models poses challenges in interpretability, trust, and regulatory acceptance, despite recent advancements in interpretable reinforcement learning techniques. Scalability across heterogeneous networks and different deployment scenarios is not fully guaranteed, and integration with existing rule-based network infrastructures may require significant architectural modifications.

To address these challenges, future work should focus on hybrid learning approaches that combine model-based optimization with DRL to reduce training complexity, as well as offline and federated learning techniques to enable efficient and privacy-preserving training. Enhancing robustness through partially observable MDP (POMDP) formulations using recurrent architectures such as LSTM or GRU can improve performance under uncertain conditions. Incorporating explainable AI (XAI) methods will improve transparency and trust in decision-making processes. Additionally, meta-learning and transfer learning strategies can enhance generalization across diverse network environments, while multi-agent DRL frameworks can enable coordinated optimization among multiple base stations for large-scale deployments. Overall, these enhancements would make the optimization of DRL in the wireless networks of the next generation even more practical, scalable, and reliable.

8 Conclusion and Future Research Directions

This paper used Deep Reinforcement Learning (DRL)-based optimization of energy efficiency (EE) and spectral efficiency (SE) of Massive MIMO systems. The proposed model was able to learn adaptive control policies in the form of power allocation, antenna activation, and user scheduling as the formulation of the problem was done as a Markov Decision Process and a Proximal Policy Optimization (PPO) algorithm was used. The experimental findings indicated that the DRL framework was always more effective than the traditional convex optimization and heuristic approaches in various situations. Particularly, the proposed solution was able to deliver up to 1825% of energy efficiency and 1220% spectral efficiency gain and QoS violation rate of less than 1%. Also, the model exhibited low inference latency (approximately 3.2 ms), which is appropriate to be deployed in 5G systems where latency must be less than 10 ms. The framework also had a high level of scalability with respect to the antenna configuration (32-512) and user density (5-25 users), which is why it is robust and scalable in the framework of dynamic channel conditions.

The conclusions include the fact that, unlike the traditional methods that optimize the conflicting objectives separately, DRL allows management of the trade-offs. The capability to scale to time-varying environments, predict channel variations and resource allocation in real-time is why DRL is an essential enabler of the next-generation wireless network. This paper illustrates that combining DRA with edge computing would enable decentralized, autonomous and energy-conscious network control, which will be needed in future 6G systems.

To conduct research in the future, there are various directions that should be advised. The introduction of federated learning can also allow distributed training in a network of base stations without losing data privacy. The study of multi-agent DRL has the potential of improving the coordination of various cells

to manage and support mobility in the cases of interference. Moreover, bettering model interpretability and resiliency in the imperfect CSI case is imperative to the practical application. To promote the scalability in heterogeneous environments, the combination of transfer learning and meta-learning methods may be further improved. Such developments will help in the development of sustainable, self-sustaining and smart wireless communication systems.

References

[1] An, Q., Segarra, S., Dick, C., Sabharwal, A., & Doost-Mohammady, R. (2023). A deep reinforcement learning-based resource scheduler for massive MIMO networks. *IEEE Transactions on Machine Learning in Communications and Networking, 1*, 242–257. https://doi.org/10.1109/TMLCN.2023.3313988

[2] Arshad, R., Baig, S., & Aslam, S. (2024). User clustering in cell-free massive MIMO NOMA system: A learning-based and user-centric approach. *Alexandria Engineering Journal, 90*, 183–196. https://doi.org/10.1016/j.aej.2024.01.064

[3] Bossy, B., Kryszkiewicz, P., & Bogucka, H. (2022). Energy-efficient OFDM radio resource allocation optimization with computational awareness: A survey. *IEEE Access, 10*, 94100–94132. https://doi.org/10.1109/ACCESS.2022.3203575

[4] Chen, X., Liu, X., Chen, Y., Jiao, L., & Min, G. (2021). Deep Q-network based resource allocation for UAV-assisted ultra-dense networks. *Computer Networks, 196*, 108249. https://doi.org/10.1016/j.comnet.2021.108249

[5] Dikmen, O. (2024). Iterative power control for maximizing spectral efficiency in cell-free massive MIMO systems. *IEEE Access, 12*, 133834–133847. https://doi.org/10.1109/ACCESS.2024.3461331

[6] Dizdar, O., Mao, Y., & Clerckx, B. (2021). Rate-splitting multiple access to mitigate the curse of mobility in (massive) MIMO networks. *IEEE Transactions on Communications, 69*(10), 6765–6780. https://doi.org/10.1109/TCOMM.2021.3098695

[7] Eappen, G., Cosmas, J., T, S., A, R., Nilavalan, R., & Thomas, J. (2022). Deep learning integrated reinforcement learning for adaptive beamforming in B5G networks. *IET Communications, 16*(20), 2454–2466. https://doi.org/10.1049/cmu2.12501

[8] Fozi, M., Sharafat, A. R., & Bennis, M. (2021). Fast MIMO beamforming via deep reinforcement learning for high mobility mmWave connectivity. *IEEE Journal on Selected Areas in Communications, 40*(1), 127–142. https://doi.org/10.1109/JSAC.2021.3126056

[9] He, H., Yu, X., Zhang, J., Song, S., & Letaief, K. B. (2021). Cell-free massive MIMO for 6G wireless communication networks. *Journal of Communications and Information Networks, 6*(4), 321–335. https://doi.org/10.23919/JCIN.2021.9663100

[10] Hu, Q., Liu, Y., Cai, Y., Yu, G., & Ding, Z. (2021). Joint deep reinforcement learning and unfolding: Beam selection and precoding for mmWave multiuser MIMO with lens arrays. *IEEE Journal on Selected Areas in Communications, 39*(8), 2289–2304. https://doi.org/10.1109/JSAC.2021.3087233

[11] Huo, Y., Lin, X., Di, B., Zhang, H., Hernando, F. J. L., Tan, A. S., ... & Chen-Hu, K. (2023). Technology trends for massive MIMO towards 6G. *Sensors*, *23*(13), 6062. https://doi.org/10.3390/s23136062

[12] Jagannath, A., Jagannath, J., & Melodia, T. (2021). Redefining wireless communication for 6G: Signal processing meets deep learning with deep unfolding. *IEEE Transactions on Artificial Intelligence*, *2*(6), 528-536. https://doi.org/10.1109/TAI.2021.3108129

[13] Kim, S., Ahn, S., Park, J., Youn, J., Kwon, Y., & Cho, S. (2024). CPU-cooperative power control scheme for scalable cell-free massive MIMO systems. *IEEE Transactions on Wireless Communications, 23*(10), 13904–13919. https://doi.org/10.1109/TWC.2024.3405585

[14] Li, S. E. (2023). Deep reinforcement learning. In *Reinforcement learning for sequential decision and optimal control* (pp. 365–402). Springer. https://doi.org/10.1007/978-981-19-7784-8_10

[15] Liu, Y. F., Chang, T. H., Hong, M., Wu, Z., So, A. M., Jorswieck, E. A., & Yu, W. (2024). A survey of recent advances in optimization methods for wireless communications. *IEEE Journal on Selected Areas in Communications, 42*(11), 2992–3031. https://doi.org/10.1109/JSAC.2024.3443759

[16] Lopes, V. H., Nahum, C. V., Dreifuerst, R. M., Batista, P., Klautau, A., Cardoso, K. V., & Heath, R. W. (2022). Deep reinforcement learning-based scheduling for multiband massive MIMO. *IEEE Access, 10*, 125509–125525. https://doi.org/10.1109/ACCESS.2022.3224808

[17] Matos, E. A., Parmezan Bonidia, R., Sipoli Sanches, D., Santos Pozza, R., & Dias Hiera Sampaio, L. (2022). Pilot sequence allocation schemes in massive MIMO systems using heuristic approaches. *Applied Sciences, 12*(10), 5117. https://doi.org/10.3390/app12105117

[18] Ozpoyraz, B., Dogukan, A. T., Gevez, Y., Altun, U., & Basar, E. (2022). Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary PHY architectures. *IEEE Open Journal of the Communications Society*, *3*, 1749-1809. https://doi.org/10.1109/OJCOMS.2022.3210648

[19] Phyo, W., Sasithong, P., Wuttisittikulkij, L., & Shah, S. (2023). Optimizing functional split in 5G cloud RAN: A particle swarm optimization approach. In *Proceedings of the IEEE Region 10 Conference (TENCON)* (pp. 103–107). IEEE. https://doi.org/10.1109/TENCON58879.2023.10322405

[20] Wang, Y., Gao, Z., Zheng, D., Chen, S., Gündüz, D., & Poor, H. V. (2022). Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication. *IEEE Wireless Communications, 30*(6), 127–135. https://doi.org/10.1109/MWC.008.2200157

[21] Wang, Z., Zhang, J., Du, H., Niyato, D., Cui, S., Ai, B., Debbah, M., Letaief, K. B., & Poor, H. V. (2024). A tutorial on extremely large-scale MIMO for 6G: Fundamentals, signal processing, and applications. *IEEE Communications Surveys & Tutorials, 26*(3), 1560–1605. https://doi.org/10.1109/COMST.2023.3349276

[22] Xu, Z., Wang, Z., Liu, R., Huang, C., Shi, Y., Wang, M., & Chen, H. (2025). Efficient multi-UAV path planning in dynamic and complex environments using hybrid polar lights optimization. *Journal of King Saud University – Computer and Information Sciences, 37*(6), 125. https://doi.org/10.1007/s44443-025-00139-7

[23] Yang, Y., Li, F., Zhang, X., Liu, Z., & Chan, K. Y. (2022). Dynamic power allocation in cellular network based on multi-agent double deep reinforcement learning. *Computer Networks, 217*, 109342. https://doi.org/10.1016/j.comnet.2022.109342

[24] Zhang, Z., Zhang, J., Zhang, Y., Yu, L., Gao, F., Shi, Q., Liu, G., Yuan, Z., & Fan, W. (2023). Deep reinforcement learning-based dynamic beam selection in dual-band communication systems. *IEEE Transactions on Wireless Communications, 23*(4), 2591–2606. https://doi.org/10.1109/TWC.2023.3300830

Author Biography

Nareshkumar Jagadhabi is a technology expert and researcher at Compnova Inc in the United States, where he specializes in enhancing enterprise system reliability through artificial intelligence and advanced automation. With a focus on predictive risk assessment for software deployment, he has contributed significant research to the fields of SAP system configuration error detection using machine learning and the security of AI-assisted low-code platforms. His work, also extends to the financial sector, where he explores Explainable AI (XAI) for strategic risk management and credit modeling.