Reinforcement Learning-Driven Adaptive Spectrum Utilization in Cognitive B5G Networks

by

Sadia KHAF

MANUSCRIPT-BASED THESIS PRESENTEDTO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JULY 4, 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS:

M. Georges Kaddoum, Thesis supervisor
Department of Electrical Engineering, École de technologie supérieure

M. Aris Leivadeas, Chair, Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

M. Michel Kadoch, Member of the Jury
Department of Electrical Engineering, École de technologie supérieure

M. Ming Zeng, External Examiner
Department of Electrical and Computer Engineering, Université Laval

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JUNE 30, 2026

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

FOREWORD

This thesis presents research conducted during my doctoral studies at the Department of Electrical Engineering, École de technologie supérieure (ETS), Université du Québec, under the supervision of Prof. Georges Kaddoum.

The research addresses spectrum management in cognitive and B5G wireless networks. In this work, I developed learning-based methods for cooperative spectrum sensing, SU-side access optimization, and end-to-end coordination involving spectrum-consuming and spectrum-granting nodes.

The thesis is organized as a manuscript-based thesis comprising three articles published in peer-reviewed scientific journals.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my research supervisor, Prof. Georges Kaddoum, for his invaluable guidance, continuous support, and encouragement throughout my doctoral studies. His expertise and insights have been instrumental in shaping the direction of this research.

I would also like to thank my co-authors, Dr. Mohammad T. Alkhodary, Dr. Joao V. C. Evangelista, and Dr. Majid Altamimi, for their valuable contributions to the research presented in this thesis.

This work was supported by the Fonds de recherche du Québec – Nature et technologies (FRQNT), scholarship number 305094, and the Canada Research Chair Program tier-II entitled "Towards a Novel and Intelligent Framework for the Next Generations of IoT Networks."

Finally, I would like to thank my partners Vincent and Nick, my friends Franziska, Faus, Nara, and Dawoon, and all my therapists for their emotional and logistical support and encouragement that enabled completion of my work.

Utilisation adaptative du spectre guidée par l'apprentissage par renforcement dans les réseaux B5G cognitifs

Sadia KHAF

RÉSUMÉ

La croissance des services sans fil et des appareils connectés a accru la demande en spectre radio. L'allocation statique du spectre ne pouvait pas répondre pleinement à cette demande, car les bandes licenciées peuvent demeurer sous-utilisées selon le lieu et le moment. Les réseaux radio cognitifs (CRN) ont permis aux utilisateurs secondaires (SU) d'accéder au spectre disponible tout en préservant la priorité des utilisateurs primaires (PU). Une utilisation efficace du spectre dans ces réseaux exigeait une détection fiable, des décisions d'accès efficaces, une coordination du côté infrastructure, de la résilience et une gestion de la confiance.

Dans ces travaux, nous avons amélioré l'utilisation du spectre en abordant la détection coopérative du spectre, l'accès au spectre du côté SU et la coordination entre les nœuds qui consomment le spectre et ceux qui l'allouent. Ces tâches impliquaient des décisions séquentielles sous incertitude et des interactions couplées entre utilisateurs et niveaux du réseau. Nous avons donc utilisé l'apprentissage par renforcement pour soutenir des décisions adaptatives de détection, d'accès et de coordination.

Premièrement, nous avons développé l'apprentissage par renforcement multi-agent partiellement coopératif (PCMARL) pour la détection coopérative du spectre dans les réseaux radio cognitifs de l'Internet des objets soumis à des attaques de falsification des données de détection du spectre. PCMARL a permis aux SU de coopérer de manière sélective et a amélioré la fiabilité de la détection et l'efficacité énergétique sans coopération complète de tout le réseau.

Deuxièmement, nous avons étendu l'optimisation du côté SU de la détection vers la détection et l'accès conjoints. Le cadre IMPACT a optimisé la participation à la détection, la coopération, la sélection des canaux et la puissance de transmission sous information imparfaite sur l'état du canal. Comme le contrôle d'accès exigeait à la fois des décisions discrètes et un contrôle continu de la puissance, nous avons introduit une formulation d'apprentissage à actions hybrides pour l'accès coopératif au spectre.

Troisièmement, nous avons ajouté le côté allocation et coordination du réseau. Le cadre d'apprentissage par renforcement hiérarchique fédéré (F-HRL) a coordonné les SU, les stations de base de nouvelle génération, les véhicules aériens sans pilote et les stations à plateformes de haute altitude. Cette architecture a soutenu l'allocation des accès, la priorité des PU, l'ordonnancement tenant compte de la confiance, la fédération préservant la confidentialité et la continuité du service lors de perturbations de l'infrastructure.

Dans l'ensemble, les résultats ont montré que l'amélioration de l'utilisation du spectre exigeait l'optimisation de la détection et de l'accès du côté SU, ainsi que la coordination du côté allocation, la résilience de l'infrastructure, l'agrégation fédérée et le contrôle tenant compte de la confiance.

Ces travaux ont développé un cadre intégré fondé sur l'apprentissage pour une gestion résiliente du spectre dans les futurs réseaux sans fil.

Mots-clés: radio cognitive, apprentissage par renforcement, gestion du spectre, apprentissage fédéré, résilience

Reinforcement Learning-Driven Adaptive Spectrum Utilization in Cognitive B5G Networks

Sadia KHAF

ABSTRACT

In recent years, the growth of wireless services and connected devices increased the demand for radio spectrum. Static spectrum allocation could not fully address this demand, because licensed bands may remain underused across time and location. Cognitive radio networks (CRNs) allowed secondary users (SUs) to access available spectrum while preserving primary user (PU) priority. In these networks, efficient spectrum use required reliable sensing, effective access decisions, infrastructure-side coordination, resilience, and trust management.

Accordingly, in the research reported in the present thesis, we improved spectrum utilization by addressing cooperative spectrum sensing, SU-side spectrum access, and coordination between spectrum-consuming and spectrum-granting nodes. These tasks involved sequential decisions under uncertainty and coupled interactions across users and network tiers. Therefore, we used reinforcement learning to support adaptive sensing, access, and coordination decisions.

First, we developed Partially Cooperative Multi-Agent Reinforcement Learning (PCMARL) for cooperative spectrum sensing in cognitive radio Internet-of-Things networks under spectrum sensing data falsification attacks. PCMARL allowed SUs to selectively cooperate and improved sensing reliability and energy efficiency without full-network cooperation.

Second, we extended SU-side optimization from sensing to joint sensing and access. The proposed IMPACT framework optimized sensing participation, cooperation, channel selection, and transmission power under imperfect channel state information. Since access control required both discrete decisions and continuous power control, we introduced a hybrid-action learning formulation for cooperative spectrum access.

Third, we added the spectrum-granting and coordination side of the network. The proposed Federated Hierarchical Reinforcement Learning (F-HRL) framework coordinated SUs, next-generation Node Bs, uncrewed aerial vehicles, and high-altitude platform stations. This architecture supported grant allocation, PU priority, trust-aware scheduling, privacy-preserving federation, and service continuity during infrastructure disruption.

The results revealed that improved spectrum utilization required SU-side sensing and access optimization together with granting-side coordination, infrastructure resilience, federated aggregation, and trust-aware control. Overall, in this research, we developed an integrated learning-based framework for resilient spectrum management in future wireless networks.

Keywords: Cognitive radio, reinforcement learning, spectrum management, federated learning, resilience

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
6G	Sixth Generation
A2C	Advantage Actor-Critic
ADMM	Alternating Direction Method of Multipliers
AC	Actor-Critic
B5G	Beyond 5G
CR	Cognitive Radio
CRN	Cognitive Radio Network
CR-IoT	Cognitive Radio Internet of Things
CSS	Cooperative Spectrum Sensing
CTDE	Centralized Training with Decentralized Execution
CSI	Channel State Information
CDF	Cumulative Distribution Function
D3QN	Dueling Double Deep Q-Network
DDPG	Deep Deterministic Policy Gradient
DL	Deep Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning

DSS	Dynamic Spectrum Sharing
EGT	Evolutionary Game Theory
eMBB	Enhanced Mobile Broadband
ESA	Exhaustive Search Agent
ETS	École de technologie supérieure
FC	Fusion Center
F-DDPG	Federated Deep Deterministic Policy Gradient
F-HRL	Federated Hierarchical Reinforcement Learning
FeDRL	Federated Deep Reinforcement Learning
FL	Federated Learning
FRL	Federated Reinforcement Learning
FSTL	Federated Split Transfer Learning
GAIL	Generative Adversarial Imitation Learning
GBA	Greedy Bandit Agent
GEO	Geostationary Earth Orbit
GFSTL	Generalized Federated Split Transfer Learning
gNB	Next-Generation Node B
HAPS	High-Altitude Platform Station
HDRL	Hierarchical Deep Reinforcement Learning
HMM	Hidden Markov Model

HRL	Hierarchical Reinforcement Learning
ICSI	Imperfect Channel State Information
IMPACT	Intelligent Multi-User Participatory Algorithm for Cooperative Transmission
IMPACT-D	Discretized IMPACT
IQR	Interquartile Range
IoT	Internet of Things
IRS	Intelligent Reflective Surface
ISAC	Integrated Sensing and Communication
ITS	Intelligent Transportation Systems
KPI	Key Performance Indicator
LEO	Low Earth Orbit
LTE-LAA	Long-Term Evolution License-Assisted Access
LOS	Line-of-Sight
LSTM	Long Short-Term Memory
MAC	Medium Access Control
MAFDRL	Multi-Agent Federated Deep Reinforcement Learning
MA-PPO	Multi-Agent Proximal Policy Optimization
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MIMO	Multiple-Input Multiple-Output

ML	Machine Learning
MMPP	Markov-Modulated Poisson Process
mMTC	Massive Machine Type Communication
NB-IoT	Narrowband Internet of Things
NLOS	Non-Line-of-Sight
NR	New Radio
NTN	Non-Terrestrial Network
O2I	Outdoor-to-Indoor
ODE	Ordinary Differential Equation
OSA	Opportunistic Spectrum Access
PCMARL	Partially Cooperative Multi-Agent Reinforcement Learning
POMDP	Partially Observable Markov Decision Process
PPO	Proximal Policy Optimization
PSD	Power Spectral Density
PU	Primary User
QoS	Quality of Service
Q-UCB	Q-Learning with Upper Confidence Bound
Q-UCB-H	Q-Learning with Upper Confidence Bound-Hoeffding
RA	Random Agent
RB	Resource Block

RF	Radio Frequency
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
SAC	Soft Actor-Critic
SARSA	State-Action-Reward-State-Action
SDF	Spectrum Sensing Data Falsification
SE	Spectral Efficiency
SFA	Sensing Falsification Attack
SINR	Signal-to-Interference-plus-Noise Ratio
SMDP	Semi-Markov Decision Process
SNR	Signal-to-Noise Ratio
SU	Secondary User
TL	Transfer Learning
TN-NTN	Terrestrial Network–Non-Terrestrial Network
TN	Terrestrial Network
TRPO	Trust Region Policy Optimization
UAV	uncrewed Aerial Vehicle
UCB	Upper Confidence Bound
UCB-H	Upper Confidence Bound-Hoeffding
UE	User Equipment

UMa	Urban Macro
UPA	Uniform Planar Array
uRLLC	Ultra-Reliable Low-Latency Communication
V2X	Vehicle-to-Everything
WOA	Whale Optimization Algorithm

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

α	Learning rate
γ	Discount factor
γ_{snr}	Average received signal-to-noise ratio
$\gamma_{u,k}$	Signal-to-interference-plus-noise ratio of UE u on channel k
$\bar{\gamma}_b$	Average signal-to-interference-plus-noise ratio at gNB b
$\Gamma_{\min}$	Minimum signal-to-interference-plus-noise ratio threshold
ϵ	Exploration rate (epsilon-greedy)
δ	Number of discretization intervals
δ_p	Transmit-power discretization factor
δ_ψ	Cooperation-time discretization factor
ζ^{uc}	Channel allocation indicator for UE u on subchannel c
ζ	Channel allocation indicator
κ^{uc}	Sensing indicator for UE u on subchannel c
κ	Channel sensing indicator
λ_b	Spectrum demand at gNB b
$\bar{\lambda}_b$	Aggregate spectrum demand at gNB b
Λ_d	Traffic load from UAV d
θ	Neural network parameters (actor)
$\theta(\cdot)$	Actor network output

ϕ	Neural network parameters (critic)
$\phi(\cdot)$	Critic network output
μ_θ	Actor network output (policy)
μ_{noise}	Noise mean
∇	Gradient
ψ^u	Cooperation time for UE u
ψ	Cooperation ratio
ρ^{uc}	Throughput of UE u on subchannel c
ρ	Throughput
σ^2	Variance
σ^2_{noise}	Noise variance
σ^2_e	Estimation error variance
τ	Federated aggregation period
$\mathbb{B}$	Binary space
$\mathbb{N}$	Discrete space
$\mathbb{R}$	Continuous space
$\mathcal{A}$	Action space
$\mathcal{A}_d$	Discrete action space
a	Action
a_d	UAV activation binary

$a_j^{s_i}$	Action of SU s_i for channel j
B_c	Bandwidth of subchannel c
B_k	Bandwidth of channel k
B	Channel bandwidth
C	Coalition
$\mathcal{C}$	Set of channels
c_j	Channel j
C_d^{BH}	UAV backhaul capacity
C_h^{BH}	HAPS backhaul capacity
$C_{\min}$	Minimum backhaul capacity
$\text{cov}_d(u)$	UAV coverage indicator for UE u
d	Distance or terminal state done flag
$\mathcal{D}$	Agent memory or set of UAVs
$\mathcal{D}_h$	UAVs coordinated by HAPS h
E	Energy consumption
E_d	Residual energy of UAV d
$E_{\min}$	Minimum energy threshold for UAVs
E_{total}	Total energy consumption
EE	Energy efficiency
f_c	Carrier frequency

f_s	Sampling frequency
$g_{u,k}$	Channel gain of UE u on channel k
h	Channel gain
h^{uc}	Channel gain of UE u on subchannel c
I	Interference power
I_b	Interference indicator at gNB b
I_k^{PU}	Interference observed at PU channel k
$I_{u,k}$	Aggregate co-channel interference at UE u on channel k
I_l	Interference threshold at primary user l
$I_k^{\max}$	Maximum interference threshold on channel k
k_u	Channel selected by SU u
L	Loss or PU transmission length
M	Number of subchannels or licensed channels
N_0	Noise power spectral density
N_C	Number of coalitions
N_{ep}	Number of episodes
N_h	Number of agents in HAPS region h
N_L	Number of primary users per coalition
N_o	Gaussian noise power
N_p	Number of PUs

N_s	Number of SUs
N_U	Number of UEs per coalition
o	Observation
$\mathbf{o}_i^t$	Noisy observation received by tier-i agent at step t
o_t	Observation at time t
o'	Next observation
P	Probability of falsification due to SDF attack
$P^{\max}$	Maximum SU transmit power
P_d	Probability of detection
P_f	Probability of false alarm
P_{od}	Overall probability of detection
P_{of}	Overall probability of false alarm
P_j	Coalition probability of incorrect decision
p	Transmit power
p_u	Transmit power of SU u
p^u	Transmit power of UE u
p_θ	Power threshold
$\mathbf{p}_d$	UAV position
PL	Path loss
Q	Estimated Q-value

Q_ϕ	Q-value function (critic network output)
$R_{u,k}$	Achievable rate of UE u on channel k
r	Reward
$\mathcal{R}_h$	Agent set participating in HAPS h federation round
$\mathcal{S}$	Set of SUs
s^c	Channel status of subchannel c
s_i	SU i
$SINR^{uc}$	Signal-to-interference-plus-noise ratio of UE u on subchannel c
T	Horizon or number of steps per episode
t	Time slot
$\mathcal{U}_P$	Set of PUs
$\mathcal{U}_S$	Set of SUs
$\mathcal{U}_{S,b}$	SUs associated with gNB b
$\mathcal{U}_d$	UEs served by UAV d
w_u	Priority weight for UE u
$x_{b,h}$	Routing indicator from gNB b via HAPS h
$x_{d,h}$	Routing indicator from UAV d via HAPS h
$x_{u,k}$	Scheduling indicator for SU u on channel k

INTRODUCTION

The growth of wireless services, connected devices, and Beyond 5G (B5G) networks continues to increase the demand for radio spectrum. Conventional static allocation assigns frequency bands to licensed users even when those bands are not continuously occupied in every location or at each time. This creates a persistent mismatch between spectrum scarcity and underutilization. In Canada, for example, several licences in the 700 MHz, 2500 MHz, and 3500 MHz bands were reported as apparently unused in different service areas (Joseph, 2018). These conditions motivate adaptive access methods that would use available spectrum more efficiently while respecting licensed-user priority.

Cognitive radio (CR) enables wireless devices to observe their radio environment and adapt their transmission decisions accordingly. In cognitive radio networks (CRNs), secondary users (SUs) may opportunistically access spectrum assigned to primary users (PUs) while maintaining PU priority. Here, spectrum sensing is a foundational function: before accessing a channel, an SU must determine whether or not the channel is occupied by a PU. Since individual sensing decisions can be unreliable under fading, noise uncertainty, hardware limitations, or malicious reporting, cooperative spectrum sensing is frequently used to improve detection reliability.

Cooperation also introduces overhead and security risks. With the growth of the network, full cooperation can increase energy consumption and signalling load. Spectrum sensing data falsification (SDF) attacks can further corrupt collective decisions when malicious SUs report false sensing information. Reliable sensing requires selective cooperation, where SUs identify useful cooperation opportunities while limiting overhead and reducing the influence of unreliable reports.

Sensing alone does not complete the CRN decision process. Once spectrum opportunities are identified, SUs must decide which channels to access and how to transmit. These decisions involve quality of service, energy consumption, interference constraints, and overall spectrum use. In addition, practical access requires power control, which introduces a continuous decision variable alongside discrete decisions such as sensing, cooperation, and channel selection. The move from sensing to access leads to a hybrid action-space problem, especially when channel state information (CSI) is imperfect.

The first two articles in the present thesis consider the spectrum-consuming side, represented by the SUs. In the first article, we rewarded SUs for effective cooperative sensing under SDF attacks. In the second article, we extended SU-side decision making to cooperative access, including channel selection and transmission-power control under imperfect CSI. These two studies moved from identifying spectrum opportunities to efficiently using them.

End-to-end spectrum management also requires the granting and coordinating side of the network. SU-side optimization can improve sensing and access behavior. However, it does not by itself address access authorization, scheduling, policy enforcement, or network-level coordination. In B5G networks, these functions may be supported by ground next-generation Node Bs (gNBs). Adding gNB-side coordination expands the scope from local SU behavior to interaction between spectrum-consuming and spectrum-granting nodes.

This expansion introduces reliability, scalability, trust, and privacy challenges. A gNB may become unavailable during a disaster, localized outage, or overload event. Centralized coordination may be constrained by backhaul failure, latency, or excessive signalling overhead. SUs may be malicious or unreliable, which brings about the need for trust-aware scheduling and aggregation. Privacy constraints may also prevent raw observations from being centralized. These challenges motivate the use of uncrewed aerial vehicles (UAVs) as backup support for impaired ground infrastructure and high-altitude platform stations (HAPS) as higher-level coordinators for federated aggregation and wider-area resilience.

In the present thesis, we used reinforcement learning (RL) because the decisions considered here are sequential, uncertain, and coupled across users. RL agents can learn policies through interaction with the environment, rather than entirely rely on fixed models or complete prior knowledge. We used multi-agent RL (MARL), hybrid-action RL, hierarchical RL (HRL), and federated learning (FL) according to the scope of each problem: cooperative sensing, SU-side sensing and access, and end-to-end hierarchical coordination.

0.1 Problem statement

The objective of the present thesis is to improve spectrum utilization in B5G wireless networks. Reliable sensing, PU protection, efficient access, and resilient coordination support this objective. This thesis addresses these functions across the following three levels:

First, SUs must reliably sense spectrum while limiting cooperation overhead and resisting malicious sensing behavior. While cooperative sensing can improve PU detection, full cooperation is not always scalable or energy-efficient, and SDF attacks can corrupt collective decisions. Therefore, a practical sensing method must support selective cooperation while maintaining reliable spectrum-occupancy decisions.

Second, SUs must move beyond sensing and access available spectrum efficiently. This requires joint decisions over sensing, cooperation, channel selection, and transmission power. Since these decisions contain both discrete and continuous components, and CSI may be imperfect, SU-side access requires a method that can handle hybrid action spaces under uncertainty.

Third, spectrum management must include the coordinating and spectrum-granting side of the network. SU-only optimization does not fully address access authorization, scheduling, policy enforcement, trust management, privacy preservation, or federation. While adding gNB-side coordination addresses these network-level functions, it also introduces resilience concerns when gNBs fail, become overloaded, or lose reliable backhaul. Therefore, end-to-end spectrum management must coordinate SUs with gNBs while using UAVs and HAPS to support service continuity, federation, and resilience.

0.2 Research objectives

The overarching objective of the present research is to improve spectrum utilization by linking cooperative sensing among SUs, SU-side access decisions, and spectrum-granting nodes within an end-to-end coordination framework.

The specific objectives are as follows:

1. **Improving cooperative spectrum sensing among SUs:** The main challenges are sensing reliability, cooperation overhead, energy efficiency, and SDF attacks in CR-IoT networks. This objective is addressed through selective SU cooperation using partially cooperative MARL.
2. **Optimizing SU-side sensing and access decisions:** Here, the main challenges are joint channel sensing, cooperation, channel selection, and transmission-power control under imperfect CSI. This objective is addressed through hybrid-action learning combining discrete access decisions with continuous power-control decisions.
3. **Extending spectrum management to granting and coordinating nodes:** The main challenges are access authorization, scheduling, trust management, privacy preservation, federation, and resilience to gNB outage or overload. This objective is addressed by incorporating gNBs for spectrum-granting and coordination functions, UAVs for backup support, and HAPS for wider-area aggregation through federated hierarchical learning.

0.3 Research approach

The present thesis expands the decision scope from spectrum sensing to SU-side access and then to infrastructure-side coordination.

In the first article, we address cooperative spectrum sensing. Specifically, we propose Partially Cooperative Multi-Agent Reinforcement Learning (PCMARL), where SUs learn to participate in cooperative sensing without requiring full cooperation across the entire network. The reward structure is found to promote accurate sensing, efficient channel use, and resistance to unreliable or malicious sensing information.

In the second article, we keep the focus on SUs, but expand their responsibility from sensing to sensing and access. We propose the Intelligent Multi-User Participatory Algorithm for Cooperative Transmission (IMPACT), together with its distributed variant, to optimize sensing, cooperation, channel selection, and transmission-power control under imperfect CSI. Hybrid-action optimization follows from the access problem: realistic access control requires both discrete choices and continuous power decisions.

In the third article, we expand the scope from SU-side optimization to end-to-end spectrum management. Specifically, we propose Federated Hierarchical Reinforcement Learning (F-HRL), where SUs learn local access behavior, gNBs and UAVs coordinate trust-aware scheduling, and HAPS nodes perform federated aggregation across network regions. This architecture addresses the additional challenges introduced by the granting and coordinating side of the network, including gNB outage, centralized-control failure, privacy preservation, and malicious participation.

Across the three articles, simulation studies evaluate the proposed methods under controlled assumptions. The evaluations compare learning performance, spectrum utilization, energy efficiency, resilience, and fairness across relevant CRN and NTN-inspired scenarios.

0.4 Thesis contributions

The main contributions of this thesis can be summarized as follows:

1. **Scalable and attack-resilient cooperative sensing:** The present research improves spectrum sensing in CR-IoT networks under sensing uncertainty, energy constraints, and SDF attacks using partially cooperative MARL, sensing coalitions, and local majority decisions.
2. **Joint SU-side sensing and access optimization:** This thesis extends SU-side spectrum use by jointly optimizing sensing, cooperation, channel selection, and transmission-power control under imperfect CSI.
3. **End-to-end coordination across consuming and granting nodes:** This research improves spectrum management beyond SU-only optimization by incorporating gNB-side coordination, UAV support, and HAPS-based federation for resilient and trust-aware access.
4. **Progressive integration of sensing, access, and coordination:** We establish cooperative sensing, cooperative access, and infrastructure-side coordination as stages of end-to-end optimization for improving spectrum utilization in B5G wireless networks.

0.5 Thesis organization

This manuscript-based thesis is organized as follows. The literature review in Chapter 1 summarizes the state of the art related to CRNs, cooperative spectrum sensing, learning-based spectrum access, hybrid-action optimization, federated learning, hierarchical coordination, and resilient NTN-assisted spectrum management. It also identifies the research gaps addressed by the three articles.

Chapter 2 presents the article titled “Partially Cooperative Scalable Spectrum Sensing in Cognitive Radio Networks under SDF Attacks,” published in The *IEEE Internet of Things Journal,* vol. 9, no. 11, pp. 8901–8912, 2022. This work introduces PCMARL for cooperative spectrum sensing among SUs under SDF attacks.

Chapter 3 presents the article titled “Partially Cooperative RL for Hybrid Action CRNs with Imperfect CSI,” published in The *IEEE Open Journal of the Communications Society,* vol. 5, pp. 3762–3774, 2024. This article extended SU-side learning from cooperative sensing to joint sensing, access, and power control under imperfect CSI.

Chapter 4 presents the article titled "Federated Hierarchical Reinforcement Learning for Resilient Spectrum Sharing in 6G Non-Terrestrial Networks," published in The *IEEE Open Journal of the Communications Society,* in 2026; DOI: 10.1109/OJCOMS.2026.3693148. This work introduces an end-to-end federated hierarchical framework involving SUs, gNBs, UAVs, and HAPS for resilient and trust-aware coordination.

The **Conclusion** chapter synthesizes the articles and summarizes contributions of the results to cooperative sensing, SU-side access optimization, and end-to-end coordination.

CHAPTER 1

LITERATURE REVIEW

Prior work on spectrum sensing, spectrum access, and end-to-end coordination in cognitive and beyond 5G (B5G) wireless networks reveals limitations in secondary-user (SU)-side decision making, infrastructure-side coordination, hierarchical control, federated learning, resilience, and trust-aware spectrum management.

1.1 Secondary user-side spectrum sensing and access in cognitive radio networks

Previously, spectrum sensing in cognitive radio internet-of-things (CR-IoT) networks had to balance detection reliability, sensing time, energy consumption, scalability, and robustness to sensing data falsification (SDF) attacks (Ali & Hamouda, 2017; Zou *et al.*, 2015). Traditional sensing methods, including energy detection, matched filtering, and cyclostationary feature detection, identified spectrum holes; however, their energy cost remained a concern for low-power internet-of-things (IoT) devices (Sharma *et al.*, 2015; Ali & Hamouda, 2017; Sun *et al.*, 2018). Periodic sensing, passive listening, and sensing-time reduction strategies were studied to improve this trade-off (Sun *et al.*, 2018; Zhang *et al.*, 2017; Sarikhani & Keynia, 2020). However, passive listening and frequent sensing still consumed energy that could be prohibitive for constrained SUs (Mastronarde & Van Der Schaar, 2011).

Cooperative spectrum sensing (CSS) improved detection reliability and reduced collision rates relative to non-cooperative sensing. Its benefit was particularly relevant under fading, local sensing errors, and sparse malicious behavior. However, with the growth of the number of SUs, full cooperation increased signalling overhead and energy consumption. While previous studies explored optimization and heuristic methods for CSS and spectrum handoff (Quan *et al.*, 2008; Oueis *et al.*, 2015; Kim & Ko, 2015; Xu *et al.*, 2020; Pham *et al.*, 2014), these methods often required tractable environment models, limited network sizes, or centralized coordination (Scutari & Pang, 2013; Alfa *et al.*, 2016; Verma *et al.*, 2017).

Game-theoretic approaches were also applied to dynamic spectrum allocation and cooperative sensing (Ni *et al.*, 2013). Evolutionary game theory was used to model SU participation in sensing and reduce energy consumption (Li *et al.*, 2017; Dai *et al.*, 2016), and adaptive random access was proposed for sensing-data collection (Lee, 2015). However, although these methods provided useful cooperation models, their reward structures frequently depended on detection and false-alarm probabilities, rather than on direct interaction with the radio environment. They could also require homogeneous assumptions or game-rule distribution that limited practical scalability (Tan *et al.*, 2011).

Machine learning and RL-based methods addressed some of these limitations by adapting to changing spectrum conditions without complete statistical knowledge of PU activity (Bkassiny *et al.*, 2013; Thilina *et al.*, 2013; Sarikhani & Keynia, 2020; Liu *et al.*, 2020; Alghorani *et al.*, 2015). RL was applied to sensing strategies, coexistence with LTE-LAA and Wi-Fi systems, SU selection, and multi-agent CSS (Oksanen *et al.*, 2012; Zhang *et al.*, 2019; Mosleh *et al.*, 2020; Sarikhani & Keynia, 2020; Lunden *et al.*, 2013, 2011). Nevertheless, many RL-based sensing studies assumed full cooperation, perfect channel information, or centralized information exchange. These assumptions left room for partially cooperative approaches that reduced sensing overhead while preserving resilience to SDF attacks.

Sensing alone was also insufficient for efficient spectrum utilization. On identifying a spectrum opportunity, SUs had to decide which channel to access and how to transmit. This introduced channel selection, cooperation, and transmission-power control as coupled decisions. Hybrid action spaces arose because sensing, cooperation, and channel selection were discrete, whereas power control was continuous. Furthermore, recent studies considered RL architectures for continuous or hybrid action spaces (Koike *et al.*, 2023; Guo & Zhao, 2022; Li *et al.*, 2020). In parallel, imperfect channel state information (CSI) remained a practical limitation for low-power devices and dynamic CRNs (Khoshkbari & Kaddoum, 2023; Khoshkbari *et al.*, 2023; Tan *et al.*, 2022). Finally, existing SU-side methods did not fully close the gap between cooperative sensing and practical access under imperfect CSI.

1.2 Limits of secondary user-only optimization

Previous research on CRNs largely emphasized SU behavior: sensing, cooperation, channel selection, and transmit-power adaptation. However, albeit necessary, this focus was incomplete. SU-only optimization did not directly address access authorization, scheduling, policy enforcement, trust management, or coordination across network infrastructure. These functions became important when spectrum access had to be managed beyond local SU decisions.

Fusion-center (FC) based sensing and mitigation methods partially addressed coordination by collecting sensing information at a central point. Previously, FC-based approaches were used for probabilistic SDF mitigation, sequential fusion, and secondary control functions (Ahmadfard *et al.*, 2017; Wu *et al.*, 2020; Bendouda *et al.*, 2018). However, centralized fusion created a single point of failure and increased information-sharing overhead. In addition, it did not represent the broader granting and scheduling role that infrastructure nodes could perform in B5G systems.

To reduce the effect of malicious or selfish SUs, trust-based and distributed mechanisms were proposed (Haddadou *et al.*, 2015; Sajid *et al.*, 2020). However, while these studies improved parts of the sensing or reporting process, they typically remained focused on SU reports, rather than on the joint operation of spectrum-consuming nodes and infrastructure-side granting nodes. This gap motivated a move from SU-only decision making toward end-to-end coordination.

1.3 Hierarchical coordination for multi-tier wireless networks

Multi-tier B5G and non-terrestrial network (NTN)-assisted systems introduced larger state spaces, heterogeneous node roles, and cross-layer coordination requirements. Hierarchical reinforcement learning (HRL) and hierarchical deep reinforcement learning (HDRL) were studied to reduce complexity by decomposing large control problems into smaller decision layers or options (Khoshkbari *et al.*, 2025; Li *et al.*, 2022; Lee, 2021). This decomposition was relevant to high-altitude platform station (HAPS)-integrated and NTN-assisted systems, where large coverage areas and heterogeneous network elements created high-dimensional resource-management problems (Khoshkbari *et al.*, 2025; Abbasi *et al.*, 2025).

Previous studies that applied HDRL to resource scheduling, user association, and channel assignment in HAPS or NTN settings showed that hierarchical control could reduce complexity while maintaining spectral efficiency (Umer *et al.*, 2025). Similarly, several surveys and design studies also emphasized the need to balance cross-tier coordination, communication overhead, and energy objectives in NTN deployments (Naous *et al.*, 2023; Farajzadeh *et al.*, 2025). However, many hierarchical approaches considered resource allocation without explicitly modeling distinct CR roles such as PUs, SUs, and granting nodes. They also frequently separated local access decisions from infrastructure-side trust and scheduling functions.

The key limitation was not only the size of the action space–the more important issue was role separation across tiers: SUs consumed spectrum, gNBs coordinated and granted access, UAVs supported impaired infrastructure, and HAPS nodes aggregated across regions. Existing hierarchical approaches did not fully integrate these roles into a single spectrum-management process.

1.4 Federated learning for distributed spectrum management

Federated learning (FL) was proposed for wireless systems where centralized training was constrained by backhaul, latency, privacy, or communication overhead. To date, FL and federated reinforcement learning (FRL) were studied in satellite, low Earth orbit (LEO), terrestrial IoT, and reconfigurable intelligent surface (RIS)-assisted settings (Qureshi *et al.*, 2023; Hlophe *et al.*, 2021; Malik *et al.*, 2023; Pala *et al.*, 2025). Federated deep reinforcement learning (DRL) methods were also reported to approach centralized-learning performance while reducing execution or coordination costs (Ahmadzadeh *et al.*, 2024; Xu *et al.*, 2020).

However, while these studies supported the use of distributed model aggregation, they frequently treated federation primarily as a training architecture. In spectrum access, federation also had to support operational requirements: privacy preservation, cross-region coordination, non-stationary learning, and trust-aware aggregation. Vertical FL and related privacy-preserving methods reduced data leakage by partitioning features across participants (Zhang *et al.*, 2020b), but privacy alone did not address malicious sensing reports, poisoned updates, or infrastructure failures.

Federated approaches for NTNs and aerial networks needed to be coupled with spectrum-control objectives and tier-specific roles. HAPS-assisted federation was relevant because HAPS nodes could provide a higher-level aggregation tier across heterogeneous regions. However, prior FL studies did not fully integrate HAPS aggregation with SU access learning, gNB-side scheduling, UAV failover, and trust-aware control.

1.5 Resilience and infrastructure failure in beyond 5G non-terrestrial networks

B5G networks had to maintain service under congestion, mobility, backhaul impairment, disaster conditions, or infrastructure loss. NTN components such as HAPS, satellites, and uncrewed aerial vehicles (UAVs) supported coverage extension, emergency recovery, and multi-tier connectivity (Rahman *et al.*, 2023; Saad *et al.*, 2024; 5G Americas, 2023). With the growing integration of terrestrial and non-terrestrial networks, spectrum access had to account for interference, quality of service (QoS), security, spectral efficiency, and energy efficiency across heterogeneous network layers (Rahman *et al.*, 2023).

Dynamic spectrum sharing based on FL and DRL was studied for emergency CRNs after disasters (Yang *et al.*, 2023). Federated UAV coordination was also explored for dense IoT deployments, and transfer-learning methods were applied to UAV swarms (Yang *et al.*, 2024; Zhou *et al.*, 2024). These studies highlighted the relevance of aerial support and distributed learning under changing network conditions.

However, many resilience-oriented studies assumed static infrastructure, trusted relays, or stable ad-hoc coordination. In addition, they did not fully examine how spectrum access should adapt when a granting node such as a gNB failed, became overloaded, or lost reliable backhaul. Finally, past studies also tended to treat recovery separately from trust-aware spectrum scheduling. This left a gap for approaches that connected infrastructure failure, UAV support, HAPS federation, and spectrum access in one control loop.

1.6 Trust, security, and adversarial robustness

Security affected both sensing and distributed learning. In CRNs, individual sensing reports could be falsified due to device inaccuracy or SDF attacks (Zou *et al.*, 2015). Malicious reports could increase interference to PUs or reduce spectrum utilization (Chen *et al.*, 2016). While majority voting could mitigate sparse SDF attacks, centralized majority decisions increased overhead and created failure points.

Trust models, blockchain-based mechanisms, tensor-based detectors, and weighted reporting schemes were proposed to reduce the adverse impact of malicious or unreliable nodes (Sajid *et al.*, 2020; Haddadou *et al.*, 2015; Getu *et al.*, 2019b,a). Yet, while these methods addressed sensing trust, they were frequently not integrated with access scheduling or federated model aggregation.

Federated and distributed learning introduced additional adversarial risks, including poisoned updates, noisy gradients, and malicious clients. Anti-jamming FRL, Byzantine-resilient aggregation, communication-efficient FRL, and NTN security studies addressed parts of this problem (Sharma *et al.*, 2023; Das *et al.*, 2024; Zuo *et al.*, 2025; Krouka *et al.*, 2022; Alharbi *et al.*, 2025). Existing countermeasures tended to treat sensing attacks, poisoning attacks, and infrastructure recovery separately. End-to-end spectrum optimization required these trust mechanisms to be connected across sensing, scheduling, aggregation, and recovery.

1.7 Research gaps

The studies reviewed above highlight the following three main gaps previously unaddressed in the literature:

1. **Scalable SU-side sensing under adversarial conditions:** Existing CRN sensing methods did not jointly provide scalable cooperation, reduced sensing overhead, energy efficiency, and SDF resilience for low-power CR-IoT devices.

2. **Joint SU-side sensing and access under imperfect CSI:** Existing methods tended to treat sensing, channel access, and power control separately, while practical CRN access required joint discrete-continuous decisions under imperfect CSI.
3. **End-to-end coordination across consuming and granting nodes:** Existing hierarchical, federated, and resilience-oriented approaches did not fully integrate SUs, gNBs, UAVs, and HAPS for trust-aware, privacy-preserving, and resilient spectrum optimization.

Chapter 2 addresses scalable cooperative sensing under SDF attacks. Chapter 3 focuses on joint SU-side sensing and access under imperfect CSI. Chapter 4 explores end-to-end coordination across spectrum-consuming and spectrum-granting nodes using federated hierarchical learning.

CHAPTER 2

PARTIALLY COOPERATIVE SCALABLE SPECTRUM SENSING IN COGNITIVE RADIO NETWORKS UNDER SENSING DATA FALSIFICATION ATTACKS

Sadia Khaf[1] , Mohammad T. Alkhodary[1] , Georges Kaddoum[1]

[1]Department of Electrical Engineering, École de technologie supérieure, Montreal, QC, H3C 1K3, Canada

Published in IEEE Internet of Things Journal, June 2022.

Abstract

Massive Internet of Things (IoT) connectivity requires addressing spectrum congestion caused by spectrum scarcity in wireless communications. Over the past decade, cognitive radio has been proposed as a promising solution to utilize the licensed spectrum efficiently. Conventional spectrum sensing approaches are complex and require statistical information about the behavior of licensed users, which is impractical. To overcome this limitation, several reinforcement learning (RL)-based spectrum sensing approaches have been proposed that are also highly adaptable to the dynamics of IoT environments. Additionally, cooperative RL-based spectrum sensing approaches have been widely used because they are more accurate than non-cooperative approaches. However, the advantage comes at the cost of scalability due to increased information sharing overhead. Furthermore, cooperative spectrum sensing (CSS) approaches suffer from attacks on the network, such as sensing data falsification (SDF) attacks, which deteriorate sensing accuracy dramatically. In this paper, we present a scalable, partially CSS algorithm that is highly resilient to SDF attacks. The novelty of the proposed algorithm lies in partial cooperation through coalition formation, which reduces sensing and information sharing overhead while improving sensing accuracy. Moreover, the algorithm learns to adapt the sensing participation percentage and selects the most rewarding channel for sensing to maximize rewards while minimizing energy consumption. The proposed algorithm outperforms state-of-the-art CSS algorithms in terms of sensing accuracy and overhead. Contrary to centralized CSS algorithms,

the proposed algorithm's performance is directly proportional to the number of devices; hence, it is suitable for massive connectivity.

2.1 Introduction

The number of Internet of Things (IoT) devices is expected to grow to 41.6 billion devices, generating 79.4 zettabytes (ZB) of data, in 2025 (MacGillivray & Reinsel, 2019). Therefore, the need for intelligent resource management and efficient bandwidth utilization is greater than ever before (Khan *et al.*, 2016). The radio spectrum is divided into licensed bands, which are used mainly by cellular, radio and television networks, and unlicensed bands occupied primarily by IoT devices. The rapid growth in the number of IoT devices will quickly consume all available unlicensed bands, leading to congestion and packet drops. The concept of cognitive radio (CR) has become much more popular in recent years to solve the problem of congestion in the unlicensed spectrum (Mitola & Maguire, 1999). In CR, secondary (unlicensed) users (SUs) are allowed to utilize the licensed spectrum when it is not occupied by primary (licensed) users (PUs). Ordinary IoT devices would inevitably need to be transformed into cognitive IoT devices to solve spectrum congestion (Khan *et al.*, 2016). Introducing CR to IoT opened the door to massive connectivity in IoT networks and promised support for an unprecedented number of sensors and devices. The concept of CR-IoT emerged as a promising solution that leverages the vacant spaces in the licensed spectrum to solve the congestion problem in the unlicensed one. The authors of (Khan *et al.*, 2017) survey the CR-IoT architectures and frameworks and propose potential applications. However, CR algorithms face significant challenges in the context of IoT, such as sensing accuracy, energy consumption, and network (Cheng *et al.*, 2020).

Spectrum sharing in the context of CR-IoT relies heavily on spectrum sensing accuracy to minimize harmful interference to PUs caused by SUs. Sensing accuracy is commonly compromised by device accuracy and sensing data falsification (SDF) attacks. The former is caused by hardware limitations of the sensing device, and the latter is due to malicious attackers falsifying the sensing results of SUs, leading to the SUs learning incorrect channel statuses (Zou *et al.*, 2015). An attack of the latter type affects a few SUs, and majority vote is an effective strategy to mitigate this type of attack, and hence improve sensing accuracy. Cooperative spectrum sensing (CSS) was introduced to improve the sensing accuracy of individual spectrum sensing devices (Vu-Van & Koo, 2011). In this vein, several multi-agent deep learning approaches have been proposed to provide a high level of accuracy using CSS (Sarikhani & Keynia, 2020; Zhang *et al.*, 2019; Alghorani *et al.*, 2015). The authors of (Sarikhani & Keynia, 2020) propose an SU selection strategy that chooses an SU to sense the channel status based on energy detection and shares the results with other SUs. The authors of (Zhang *et al.*, 2019) also use deep reinforcement learning to efficiently explore the radio environment through CSS. However, these approaches focus on full cooperation among SUs to improve accuracy and are not scalable to massive IoT networks, and hence have limited performance under high load factors. Therefore, an energy efficient approach that maintains a high level of sensing accuracy under various load factors is needed to support the future demands of IoT networks.

Current state-of-the-art spectrum sensing techniques consume a lot of time and energy, which makes it impractical to implement them with low-powered IoT devices (Hussain *et al.*, 2020). To overcome this limitation, spectrum sensing using evolutionary game theory (EGT) approaches gives SUs the freedom to be free riders to conserve their energy (Li *et al.*, 2017), but suffers from inaccurate sensing due to probabilistic environment modeling. On the other hand, spectrum sensing using multi-agent reinforcement learning (RL) trains the agents (SUs) through rewards from interaction with the environment to provide a realistic environment and accurate sensing (Sarikhani & Keynia, 2020). However, the aforementioned approaches consume a lot of time and energy under a high load factor (Jagannathan *et al.*, 2012). Thus, reducing the energy consumption (for sensing and collaboration) of multi-agent CSS is an open research problem.

To tackle CR-IoT's sensing accuracy, energy consumption and scalability challenges, this paper presents a novel partially cooperative multi-agent RL (PCMARL) spectrum sensing algorithm. Partial cooperation is achieved through coalition formation, which restricts sensing and information sharing to small subsets of SUs. The SUs in these subsets collaborate and use majority voting to fight SDF attacks and improve sensing accuracy as well as reduce sensing overhead. Our approach teaches the SUs to optimize their sensing participation by learning to be free riders to conserve energy while maximizing rewards. We also give SUs the freedom to select a channel to sense based on reward history, which gives them the flexibility to switch channels when a channel is frequently busy or due to SU mobility.

Novelty and contribution

This paper presents the first scalable, energy efficient, RL-based algorithm that takes advantage of coalition formation and majority voting for reliable spectrum sensing in the presence of SDF attacks. The contribution of this paper can be summarized as follows:

- The PCMARL algorithm provides agents with a realistic model of the radio environment, rather than a probabilistic one as in the case of EGT-based algorithms. In other words, the environment is more realistic in the sense that it does not rely on SUs' probability of detection and probability of false alarm, but rather calculates the rewards through the agent's interaction with the radio environment.
- We provide a coalition-based partial cooperation strategy that reduces the amount of energy spent physically sensing the channel, and sharing the channel status with other SUs.
- Contrary to traditional CSS approaches, the PCMARL algorithm does not rely on a fusion center to calculate majority vote. Instead, the SUs in each coalition compute majority vote locally, which makes the PCMARL algorithm scalable to massive connectivity.
- This is the first approach that gives SUs the freedom to choose the channel for sensing based on channel-specific reward history and teaches them to be free riders to conserve energy.
- In contrast to state-of-the-art spectrum sensing algorithms, the PCMARL algorithm is the first approach whose performance is directly proportional to the number of devices in the network, making it suitable for massive connectivity.

The remainder of this paper is organized as follows. Section 2.2 provides a literature review of the most relevant works. The system model and the proposed PCMARL spectrum sensing algorithm are presented in Section 2.3. Section 2.4 discusses the simulation results. Finally, Section 2.5 concludes this work.

2.2 Related works

The challenges of spectrum sensing in the context of IoT networks can be summarized as spectrum sensing scheduling, sensing time minimization, energy consumption minimization, enabling massive connectivity, and maintaining a high level of accuracy in the presence of SDF attacks (Ali & Hamouda, 2017; Zou *et al.*, 2015). The solutions proposed to address some of these challenges can be classified into three categories based on the approach used: A) Optimization and Heuristic Approaches, B) Game-Theoretic Approaches, and C) Machine Learning Approaches. A brief comparison of these approaches in terms of methodology, advantages, and disadvantages, is presented in Table 2.1 with corresponding references.The key aspects of each approach are discussed below.

Table 2.1 Literature on spectrum sensing: benefits and drawbacks

	Optimization and Heuristic Approaches	Game Theoretic Approaches	Machine Learning Approaches
Methodology	Simplified model of the environment to formulate spectrum sensing in classic optimization problem. Optimization is done using genetic algorithms, e.g., particle swarm optimization, etc.	Distributing resources among competing players to reach a Nash equilibrium	Learning from data or the environment to make decisions autonomously. Improves by experience.
Advantages	Fast implementation with a closed-form solution.	Fair distribution of resources. Dynamically varying in time	No prior environment model is needed. Learns complex patterns in the data. Reacts to dynamic environment
Disadvantages	Incomplete environment model. Difficult to scale in dynamic IoT environments.	Does not support heterogeneous players. Slow convergence. May not reach equilibrium. Difficult to scale in dynamic IoT environments.	Rely on labeled data. Need training using large datasets in case of deep learning (DL). Curse of dimensionality. Supervised machine learning is difficult to scale in dynamic IoT environments.
References	(Quan *et al.*, 2008; Oueis *et al.*, 2015; Kim & Ko, 2015; Xu *et al.*, 2020; Pham *et al.*, 2014)	(Ni *et al.*, 2013; Li *et al.*, 2017; Lee, 2015; Dai *et al.*, 2016)	(Hlophe & Maharaj, 2019; He & Jiang, 2019; Shi *et al.*, 2020; Bkassiny *et al.*, 2013; Thilina *et al.*, 2013)

2.2.1 Optimization and heuristic approaches

Optimization and heuristic approaches were among the earliest approaches used for CSS (Quan *et al.*, 2008). More recently, authors of (Oueis *et al.*, 2015; Kim & Ko, 2015; Xu *et al.*, 2020) have tried sub-carrier allocation subject to delay, maximum power, and interference constraints. The proposed methods are computationally expensive. They work well with a limited number of IoT devices; however, scaling them to a large number of devices makes them too complex. The authors of (Pham *et al.*, 2014) used a Hidden Markov Model (HMM) for CSS and spectrum hand-off in a CR network (CRN); however, the fully cooperative nature of the solution makes it unscalable. In short, classical optimization methods are not suitable for CR-IoT applications since a complete model of the environment is not available (Scutari & Pang, 2013), they are not scalable to support massive IoT connectivity (Alfa *et al.*, 2016; Verma *et al.*, 2017), and they cannot adapt to the dynamic and evolving environment (Verma *et al.*, 2017).

2.2.2 Game-theoretic approaches

The use of game theoretic approaches for dynamic spectrum allocation is studied in (Ni *et al.*, 2013), and both cooperative and non-cooperative game models are explored. The application of EGT for CSS is presented in (Li *et al.*, 2017). The authors present a novel concept of “free will" for the SUs to participate in spectrum sensing to reduce energy consumption. An extension of the work that allows SUs to join multiple coalitions at the same time is studied in (Dai *et al.*, 2016). Adaptive random access for collecting sensing data is proposed in (Lee, 2015) to optimize sensing time.

In all the game theoretic approaches mentioned above, the rewards are calculated from SUs' probability of detection and probability of false alarm, which provide an unrealistic representation of the environment. Additionally, game theoretic approaches assume homogeneous rewards, are highly complex, and have poor convergence and low flexibility. Furthermore, game theoretic approaches can be challenging to implement in practical scenarios due to the difficulty of distributing the game rule for the players (Tan *et al.*, 2011).

2.2.3 Machine learning approaches

Several deep learning frameworks for spectrum sensing are proposed in (Hlophe & Maharaj, 2019; He & Jiang, 2019; Shi *et al.*, 2020) to improve the energy efficiency of the system and provide a low-complexity alternative to classical optimization methods. A survey of supervised and unsupervised machine learning methods for spectrum sensing can be found in (Bkassiny *et al.*, 2013). Our earlier works (Ali *et al.*, 2019, 2020a,b) provide a foundation for resource sharing among competing IoT devices using deep learning. Similarly, the authors of (Sami *et al.*, 2020) propose a reinforcement learning method, SARSA, for resource provisioning and horizontal container scaling in fog computing. The authors further extend their work to deep reinforcement learning in (Sami *et al.*, 2021a,b) proposing deep Q-learning as an alternative to heuristic methods for service provisioning to IoT devices due to the NP hardness of the problem and demonstrating the algorithm's efficiency in making proactive placement and scaling decisions. Several supervised and unsupervised machine learning approaches such as K-means clustering, Gaussian mixer model, support vector machine, and K-nearest neighbors are proposed for opportunistic spectrum access in (Thilina *et al.*, 2013). Nevertheless, it should be highlighted that data-driven machine learning approaches suffer from a number of inherent drawbacks such as lack of sufficient labeled data, partial observability of the environment, and lack of support for incorporating delayed feedback from the environment. Such shortcomings prevent machine learning algorithms from being an ideal solution for spectrum allocation in a dynamic radio environment and motivate the use of RL.

The authors of (Oksanen *et al.*, 2012) propose to use RL for spectrum sensing in a dynamic IoT environment and analyze and compare the performance of ϵ-greedy and upper confidence bound (UCB) exploratory strategies. The idea of using RL with UCB is further explored in (Zhang *et al.*, 2019), where the authors created several PU traffic models such as bursty, legacy, and frequency hopping patterns to show the adaptability of RL-based spectrum sensing methods. The use of RL to facilitate the coexistence of long-term evolution license-assisted access (LTE-LAA) and IEEE 802.11 Wi-Fi systems is studied in (Mosleh *et al.*, 2020), where agents compete for channel access, considering the effects of MAC and physical layers. The authors of (Sarikhani & Keynia,

2020) proposed a spatial-correlation-based SU selection mechanism that selects the most suitable SU for local sensing.

An RL-based multi-agent CSS problem is studied in (Zhang *et al.*, 2019). A distributed model is used that assumes perfect channel information. Since the approach is distributed, it is easily scalable to massive IoT networks. It is assumed that there will always be only one PU in six channels with a 20% probability that the PU remains on the same channel and an 80% probability that it switches channel. This is an unrealistic assumption since, in practice, there can be more than one PU transmitting at the same time. The authors of (Lunden *et al.*, 2013, 2011) also propose a multi-agent RL framework for spectrum sensing in CR-IoT. The use of RL for spectrum sensing is also advocated in more recent works due to the RL agents' ability to quickly adapt to a highly dynamic IoT sensor environment (Shin *et al.*, 2020; Sarikhani & Keynia, 2020; Chang *et al.*, 2019; Panayiotou *et al.*, 2019; Zia *et al.*, 2019; Xu *et al.*, 2019).

Cooperative RL approaches exhibit superior performance in terms of accuracy. To implement cooperative RL algorithms for low-power CR-IoT devices, a significant reduction in energy consumption, sensing overhead, and sensing time is needed. This can be achieved by using coalition formation and majority vote, and sharing sensing belief with a limited number of neighbors as shown herein.

Sensing falsification

The sensing data of individual SUs is vulnerable to sensing falsification due to device sensing inaccuracies and SDF attacks. Therefore, the belief-sharing aspect of CSS introduces new security threats. In an SDF attack, malicious attackers send false local sensing results to mislead the cooperating SUs (Zou *et al.*, 2015). Attacks may result in either excessive interference in the PU network or a decrease in spectrum utilization (Chen *et al.*, 2016). Attacks of this type are carried out by sparse agents, and majority vote is an effective strategy to mitigate them. In addition to SDF attacks, SU sensing results are subject to falsification due to device-hardware sensing inaccuracies. The authors of (Sajid *et al.*, 2020) suggest using block chains to mitigate the effect of falsified sensing. The authors of (Haddadou *et al.*, 2015) propose a distributed trust model to discredit the malicious and selfish nodes, thereby reducing the weight of their signals. In other works (Ahmadfard *et al.*, 2017), the authors discuss probabilistic SDF attacks and a centralized fusion center (FC)-based mitigation strategy. Similarly the authors of (Wu *et al.*, 2020) propose a sequential fusion-based mitigation strategy that relies on an FC. The authors of (Bendouda *et al.*, 2018) also propose an FC-based network topology with three levels of control, where the secondary controller is responsible for resource sharing and management strategies as well as attack detection and mitigation. Although FC is optimized overall, it creates a single point of failure and makes the approach non-scalable. Moreover, the performance of FC-based approaches degrades in terms of energy efficiency due to large information sharing overhead. In contrast, we propose an energy efficient partially cooperative SDF mitigation strategy that computes the majority vote locally and improves, both energy efficiency and accuracy with a greater number of devices.

2.3 The proposed algorithm

This section presents the proposed system model, introduces the notation used, explains the structure and flow of the PCMARL algorithm, lists performance indicators, shows the algorithm's resilience against sensing falsification, and analyzes its complexity.

2.3.1 System model

The proposed CR environment consists of N_P PU, N_s SU, and M licensed channels, and we assume $M = N_p$. The internal clocks of all PUs and SUs are synchronized and sufficiently accurate (Zhang & Su, 2011). The stochastic process governing PUs' access to channels can be modeled as a Markov decision process (MDP). Each PU transmits randomly with transmission length $L = kt$, where k is a random number and t is the transmission unit "time slot". The SUs have no prior knowledge of the transmission pattern or transmission length of the PUs, and should be sure to not interfere with the PUs. Each SU can sense only K channels, where ($K \leq M$), in each time slot due to energy and hardware limitations. Each SU's sensing results are subject to falsification with with probability P due to SDF attacks and hardware sensing inaccuracy. The CR model is distributed, and there is no centralized unit to share sensing information among SUs.

Let $\mathcal{S} = \{s_i, ..., s_{N_s}\}$ be the SU set with N_s user "agents", and $C = \{c_i, ..., c_M\}$ be a set of M channels occupied by N_p PUs. Each agent interacts with the environment and with other agents, a subset of $\mathcal{S}$ that sense the same channel, to collect information about channel occupancy and decide on its sensing policy. The sets of *observations* and *actions* are donated by $\mathcal{O} = \{0, 1, ..., T\}$ and $\mathcal{A} = \{\text{not-sensed, idle, busy}\}$, respectively, with size $|\mathcal{O}|$ and $|\mathcal{A}|$. The "not-sensed" *action* represents the agent's action if it decides to be a free rider to conserve energy. The notation $a_j^{s_i} \in \mathcal{A}$ represents the action taken by the agent s_i for the channel j. The set $\mathcal{D}_j = \{a_j^{s_1}, ..., a_j^{s_{N_j}}\}$ represents the multi-set of actions of all agents in collation j, where N_j is the number of agents sensing channel j. Similarly, $r_j^{s_i}(o_t) \in \{0, 1\}$ donates the reward of agent s_i as a result of selecting channel j for observation $o_t \in \mathcal{O}$.

Interaction with the environment happens through coalition formation, the sharing of sensing belief $\mathcal{D}_j$, majority vote, and the exchange of rewards. Each agent keeps a record of its own actions and its rewards from each channel. In each time slot, the agent picks a channel to sense based on its reward history. All agents sensing channel j are considered coalition C_j, and they can join or leave a coalition based on their quality of service (QoS), as depicted in Figure 2.1. After joining a coalition, each agent calculates the best action, i.e., channel status, based on the Q-table and shares it with all members in the coalition. All agents in the coalition then calculate the majority vote of the coalition locally using the channel status shared by other members and update their actions based on the majority vote. Lastly, the agents receive a reward from the environment if their predicted channel status matches the actual status and update their action-reward history.

The work of this paper (PCMARL algorithm) is inspired by EGT and RL in the following manners. From an EGT perspective, the work uses the concept of coalition formation, belief sharing, and majority vote. From RL perspective, the proposed algorithm uses belief sharing,

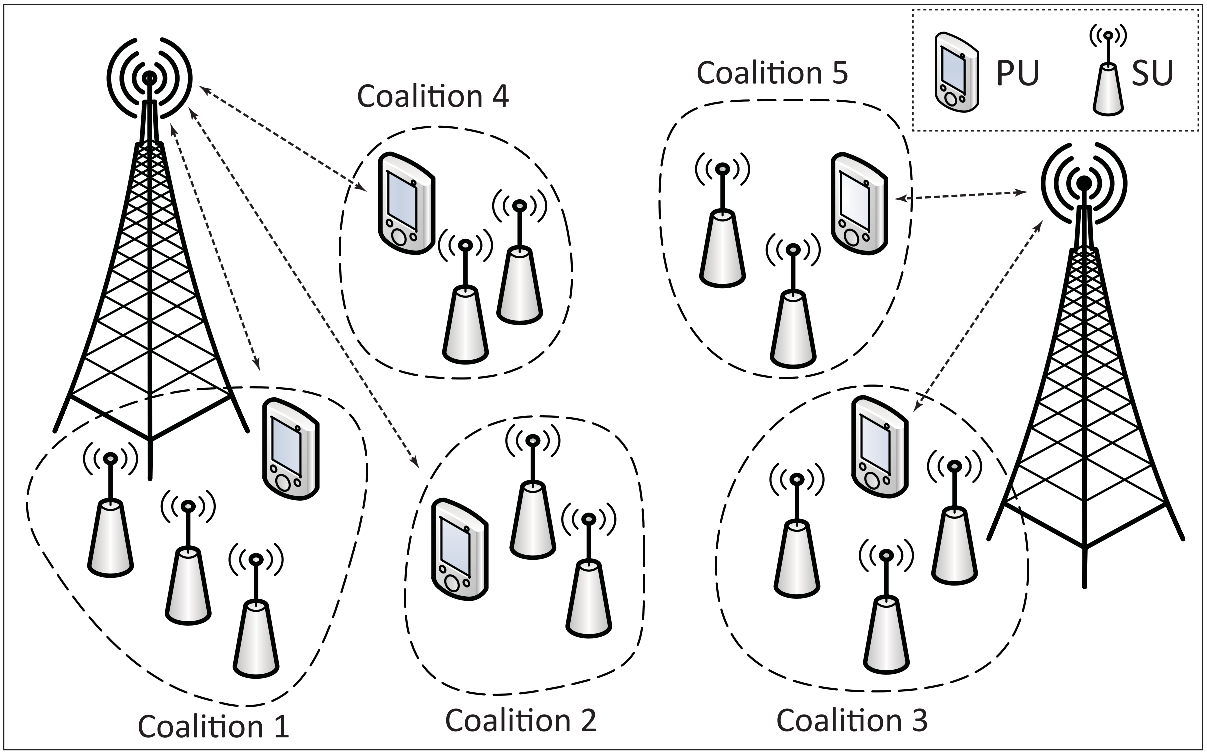

Figure 2.1 The proposed cognitive radio environment

Q-learning, and rewards through interaction with the environment. Thus, the algorithm uses belief sharing, which is common to cooperative games and cooperative RL approaches, to provide a unified and more realistic model of the radio environment.

2.3.2 The proposed algorithm

Algorithm 2.1 highlights the important steps in the PCMARL spectrum sensing algorithm. Additionally, Figure 2.2 depicts the flow of the algorithm. The input parameters α and γ define the learning rate and the weight associated with the temporal difference in the Q-learning algorithm, respectively. The Q-table and agent histories are initialized to zero so that the initial actions of all agents are always exploratory.

The PCMARL algorithm uses ϵ-greedy to balance exploration and exploitation such that the agent takes exploratory action with probability ϵ and chooses the greedy action, i.e., the action

Algorithm 2.1 Partially cooperative multi-agent reinforcement learning-based spectrum sensing

Algorithm parameters: Learning rate $\alpha \in (0,1]$, γ, majority-vote $v \in [0,1]$;
Initialize $Q(o,a)$, for all $o \in \mathcal{O}, a \in \mathcal{A}$, ;
Initialize $R(o,s)$, for all $o \in \wr, s \in \mathcal{S}$, ;
foreach *episode* **do**
    Reset Environment;
    foreach *step of episode* **do**
        if *random float* $\leq \epsilon(episodecount)$ **then**
            Sample a from $\mathcal{A}$;
        else
            Choose channel with $\max(\overline{R_c})$;
            Choose action for channel from Q-table;
        end if
        if $v == 1$ **then**
            Form coalition C_j; $s_i \in C_j$ if s_i sensed channel j ;
            Calculate majority decision for each coalition ;
            Update actions of each s_i based on majority-vote ;
        foreach $s_i \in \mathcal{S}$ **do**
            Take action a, observe R, o;
            $Q(o,a) \leftarrow Q(o,a) + \alpha[R + \gamma \max_a Q(o',a) - Q(o,a)]$;
        end foreach
        $o \leftarrow o'$;
    end foreach
end foreach

with the highest Q-value, with probability $1-\epsilon$. However, the value of ϵ decreases exponentially from 1 to 0 with the number of episodes rather than being constant.

The Q-table holds Q-values for all actions and all observations for each channel. In the case of a random action, the agent randomly selects the channel and the channel status, whereas in the case of a greedy action, the agent picks the best channel to sense based on channel reward history and the channel status based on the Q-table. The total reward $R_j^{s_i}$ that SU s_i obtains from channel j is

$$R_j^{s_i} = \sum_{o_t|t=1}^{o_t|t=t_o} r_j^{s_i}(o_t), \tag{2.1}$$

where o_t and t_o are the observation at time t and the current time, respectively. The average channel reward $\bar{R}_j^{s_i}$ of s_i for channel j is calculated as

$$\bar{R}_j^{s_i} = \frac{R_j^{s_i}}{K_j^{s_i}}, \tag{2.2}$$

where $K_j^{s_i}$ is the number of times that channel j was sensed in all past observations up to the current observation. In each time slot, each SU picks the channel with highest average sensing reward to sense as follows:

$$c_j = \arg\max_j \bar{R}_j^{s_i}. \tag{2.3}$$

After selecting the best channel, the agent needs to choose the best action for that channel. The action, i.e., the channel status for channel j, is selected by SU s_i based on

$$a_j^{s_i} = \arg\max_{a \in \mathcal{A}} Q(o_t, a). \tag{2.4}$$

Since the action space contains “free riding", “idle", and “busy” as actions, the one with the highest Q-value is selected according to Eq. 2.4.

In a traditional Q-learning approach, an individual action chosen based on the Q-table is first compared to the PU transmission, and then the agent receives a reward accordingly. SUs’ individual sensing results can be falsified with probability P due to SDF attacks and hardware sensing inaccuracy. Coalition formation and majority vote provide a strong defense against falsification, as explained in detail in Section 2.3.4.

Once all agents in all coalitions have chosen their individual actions, they share the channel status with the other members of their coalition (Zhang *et al.*, 2019). Each coalition member in coalition C_j thus receives a multi-set $\mathcal{D}_j = \{a_j^{s_1}, ..., a_j^{s_{N_j}}\}$ containing the actions of all SUs in the coalition, some of which may have been falsified. Then, each agent calculates the majority action of the coalition as

$$d_j = \text{mode}(\mathcal{D}_j), \tag{2.5}$$

and all members of the coalition update their actions as $a_j^{s_i} = d_j$. Once all members have updated their actions, their actions are compared to the actual transmission pattern of the PUs and the agents receive a reward is $r_j^{s_i} \in \{0, 1\}$ from the radio environment for predicting the channel status. The reward $r_j^{s_i} = 1$ for correctly predicting channel status or $r_j^{s_i} = 0$ for not sensing or incorrectly predicting it. The goal of the SUs is to maximize their rewards. Each agent s_i then updates its Q-table as follows:

$$Q(o_t, a_j^{s_i}) = Q(o_t, a_j^{s_i}) + \alpha * (r_j^{s_i} + \gamma * (\max_{a \in \mathcal{A}} Q(o_{t+1}, a) - Q(o_t, a_j^{s_i}))). \tag{2.6}$$

The agents continue to repeat the above steps until they reach the terminal observation, which marks the end of the episode. At that point, the environment is reset to the initial observation, a new episode begins, and the process repeats itself. The convergence of the algorithm is well established and can be found in (Tsitsiklis, 1994; Strehl *et al.*, 2006; Sutton & Barto, 2018). The notations used to describe the above process are listed in Table 2.2.

2.3.3 Performance indicators

Sensing accuracy is the first performance indicator used in this work. It is defined as the number of correct predictions normalized by the number of times the SU participated in spectrum sensing, as shown below:

$$\text{Sensing Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total sensing contribution}} \times 100. \tag{2.7}$$

Sensing accuracy is calculated per SU, and the total sensing contribution is an indicator of the energy consumed by the SU for spectrum sensing and collaboration. The SU's goal is therefore to minimize this while maintaining a high level of accuracy. In terms of RL, this is achieved by

Table 2.2 Table of notations

$\mathcal{S} = \{s_i, ..., s_{N_s}\}$	set of SUs
$C = \{c_i, ..., c_M\}$	set of channels
s_i	SU i
$N_s = \|\mathcal{S}\|$	number of SUs
$M = \|C\|$	number of channels
$N_p = M$	number of PUs
T	number of steps per episode, horizon
t	time
$o_t : t = 1, ..., T$	observation (function of time)
P	probability of falsification due to SDF attack
$a_j^{s_i} \in \mathcal{A}$	action of SU s_i for channel j
$\mathcal{A} = \{0, 1, 2\}$	action space
$O = \{o_1, o_2, \dots o_T\}$	observation space
$r_j^{s_i}(o_t) \in \{0, 1\}$	reward of SU s_i from channel j in o_t
$R_j^{s_i}$	total reward of SU s_i from channel j
R^{s_i}	total reward of SU s_i regardless of the channel
$\bar{R}_j^{s_i}$	average channel reward of SU s_i
$\hat{R}^{s_i}$	total reward-per-contribution of SU s_i
$K_j^{s_i}$	number of times channel j was sensed by s_i
K^{s_i}	number of times SU s_i participated in sensing
$Q^{s_i}(o_t, a_j^{s_i})$	Q-table of SU s_i
$\mathcal{D}_j = \{a_j^{s_1}, ..., a_j^{s_{N_j}}\}$	multi-set of actions of all SUs in coalition C_j
$\{s_{N_1}, ..., s_{N_j}\} \subset \mathcal{S}$	set of SUs in coalition C_j
d_j =mode$(\mathcal{D}_j)$	majority decision of coalition C_j
N_j	number of SUs sensing channel j
$\eta = N_s/N_p$	load factor
$L_j \in [1, 3]$	transmission length of PU j
H^{s_i}	sensing overhead of SU s_i

assigning a small negative "step cost" to all actions other than those “not-sensed". However, two agents with different sensing contributions can have the same sensing accuracy but different sensing rewards. Therefore, the second performance indicator used is the total rewards per episode for each SU (regardless of the channel) R^{s_i}, which is defined as

$$R^{s_i} = \sum_{j=1}^{|C|} R_j^{s_i}. \tag{2.8}$$

Since total reward alone does not reflect SU's contribution, reward-per-contribution $\hat{R}^{s_i}$ is another performance indicator, defined as

$$\hat{R}^{s_i} = R^{s_i}/K^{s_i}, \tag{2.9}$$

where K^{s_i} is the number of times s_i participated in spectrum sensing. The last performance indicator used is sensing overhead H^{s_i}, which shows the amount of energy SUs consume in spectrum sensing without receiving any reward. Sensing overhead is defined as

$$H^{s_i} = 1 - R^{s_i}/K^{s_i}. \tag{2.10}$$

PCMARL algorithm's performance is evaluated in more detail using the aforementioned indicators in Section 2.4.

2.3.4 Resilience of the proposed algorithm

CR approaches are vulnerable to sensing inaccuracy due to hardware limitations and SDF attacks. Majority vote provides a strong defense against such inaccuracies since the coalition's probability of incorrect decision P_j is always lower than P, where P is the probability that the sensing result reported by an SU has been falsified. We calculate P_j as

$$\begin{aligned} P_j &= \Pr(\text{more than half of the agents in coalition } C_j \text{ report} \\ &\quad \text{falsified results}) \\ &= \sum_{k=\left\lceil \frac{1+|C_j|}{2} \right\rceil}^{|C_j|} \Pr(k \text{ agents in } C_j \text{ falsified}). \end{aligned} \tag{2.11}$$

The k agents from C_j who reported a falsified channel status form subset C_j^k. Hence, there are $\binom{|C_j|}{k}$ possibilities of such subsets. Therefore, the coalition's probability of an incorrect decision can be calculated as

$$P_j = \sum_{k=\lceil \frac{1+|C_j|}{2} \rceil}^{|C_j|} \binom{|C_j|}{k} P^k (1-P)^{|C_j|-k}. \tag{2.12}$$

Since $\forall |C_j| > 1$ and $P \in (0, 0.5)$, $P_j < P$, majority vote is an effective remedy against SDF attacks. Figure 2.3 shows this phenomenon by plotting a coalition's probability of incorrect decision P_j for various $|C_j|$ against P. It also demonstrates that larger coalitions provide stronger defense against SDF attacks, which is advantageous to support scalability. Furthermore, Eq. 2.12 provides a relationship between $|C_j|$ and P_j where $|C_j|$ is the number of SUs that actually participated in sensing. The relationship helps determine the number of contributors required in a coalition to achieve a certain probability of an incorrect decision. Hence, coalition formation and majority vote not only provide a defense against SDF attacks but also enable greater energy savings by providing a threshold of contributors required to achieve a certain level of accuracy. If there are more SUs in a coalition than required, the remaining SUs can be free riders.

2.3.5 Complexity analysis

The authors of (Kearns & Singh, 2002) determine the complexity of general MDPs to be bounded by polynomials in experiment time, and the size of the state-space and action-space. The time complexity of model-free algorithms, including Q-learning with ϵ-greedy, UCB, and UCB-H exploration strategies is shown in (Jin *et al.*, 2018) to be $O(E)$, where E denotes the total experiment time. The space complexity of Q-learning is proven in (Strehl *et al.*, 2006) to be $O(|\mathcal{S}||\mathcal{A}|)$, where $|\mathcal{S}|$ is the size of the state-space, and $|\mathcal{A}|$ is the size of the action-space. The complexity of the PCMARL algorithm differs in only two aspects:

1. Reward history-based channel selection, which is a max operation over the number of channels and, hence, has a complexity of $O(M)$.
2. Majority vote, which is a *mode* operation over the action set $\mathcal{D}_j$ and has a complexity of $O(|\mathcal{D}_j|)$. Note that free-riders reduce this complexity, since the *mode* is performed only on contributors.

Therefore, the additional complexity is linear in channel quantity and coalition size but constant in load factor since each agent performs all computations locally.

The simulation results are discussed in the next section.

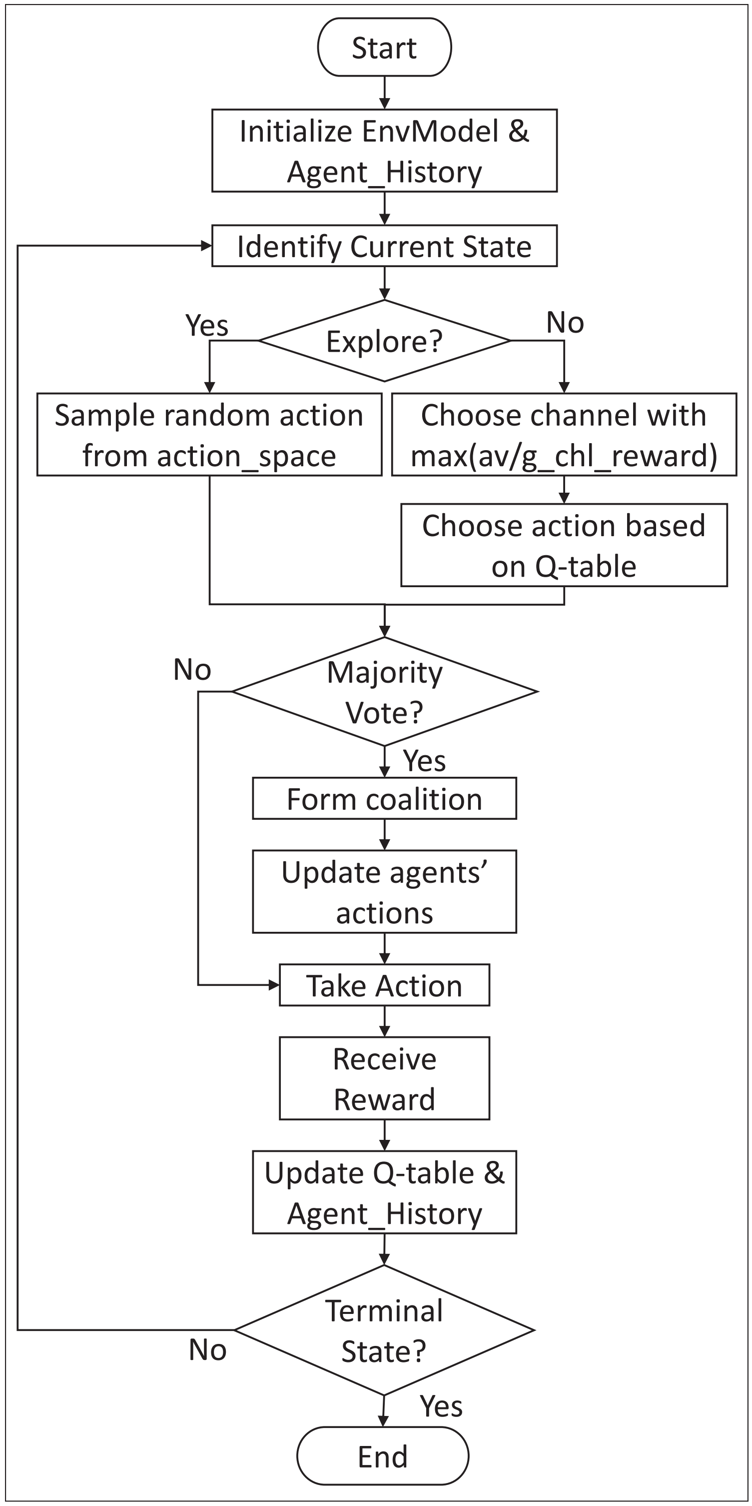

Figure 2.2 Flowchart of the proposed cooperative sensing algorithm

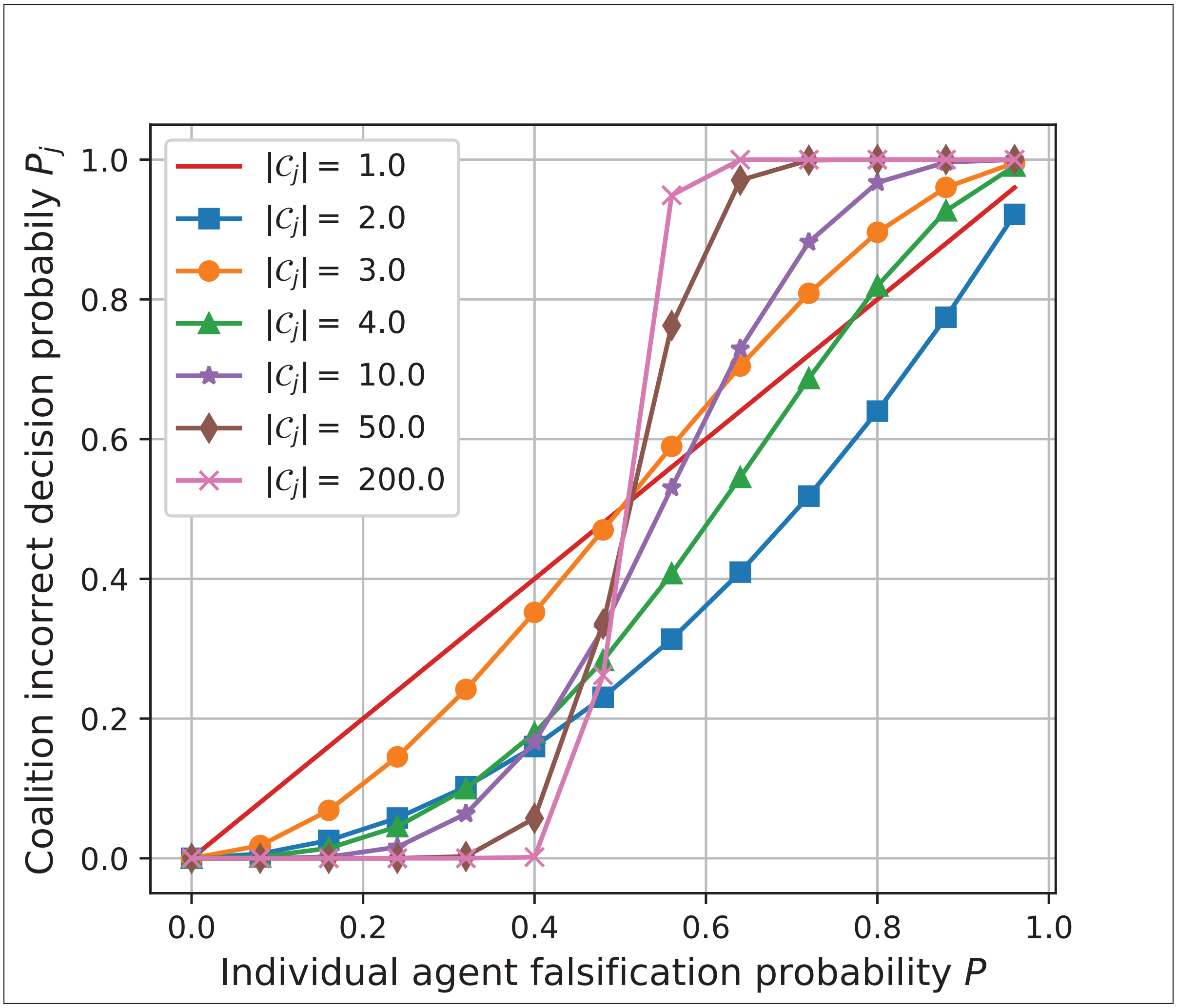

Figure 2.3 Coalition's probability of incorrect decision against individual falsification probability

2.4 Performance evaluation

In this section, we present the simulation results, discuss the PCMARL sensing algorithm and evaluate various performance indicators under several load factors. The PCMARL algorithm's performance is compared to that of traditional Q-learning with constant ϵ proposed in (Zhang *et al.*, 2019), Q-learning with UCB as used in (Sutton & Barto, 2018), and UCB-H as used in (Zhang *et al.*, 2019) for CSS.

The sensing capacity K of each SU is assumed to be 1, i.e., each SU can sense only one channel in one time slot, which is a realistic assumption for most low-powered IoT devices. The SUs are assumed to be capable of switching between available channels. We use $N_p = 3$ PUs for all simulations and $N_s = \eta N_p$ where $\eta \in \{10, 25, 50, 100, 200\}$. The simulations are carried out for $1,000$ episodes, and each episode has $T = 20$ steps. If an SU participates in spectrum sensing, their reported channel status is subject to change due to SDF attacks and sensing inaccuracy with probability P. The simulation parameters are given in Table 2.3.

Table 2.3 Simulation parameters

Parameter	Value(s)	Parameter	Value(s)
η	$[2, 10, 25, 50, 100, 200]$	N_S	ηN_P
P	$[0.01, 0.05, 0.1, 0.2, 0.3]$	$\mathcal{A}$	$\{0, 1, 2\}$
Number of episodes	$[100, 500, 1000]$	T	20
N_P	3	σ	5
γ	0.1	α	0.1

Figure 2.4 compares the sensing accuracy of the proposed PCMARL algorithm with three related approaches, as mentioned above, under load factor $\eta = 5$ and falsification probability due to SDF attack $P = 0.3$. Since the PCMARL algorithm uses majority voting, it is significantly more accurate than traditional Q-learning with constant ϵ, the Q-UCB approach of (Sutton & Barto, 2018), and the Q-UCB-H approach of (Zhang *et al.*, 2019). Similarly, Figure 2.5 shows the sensing accuracy of all four approaches for $\eta = 5$ and $P = 0.1$. Lower values for η and P

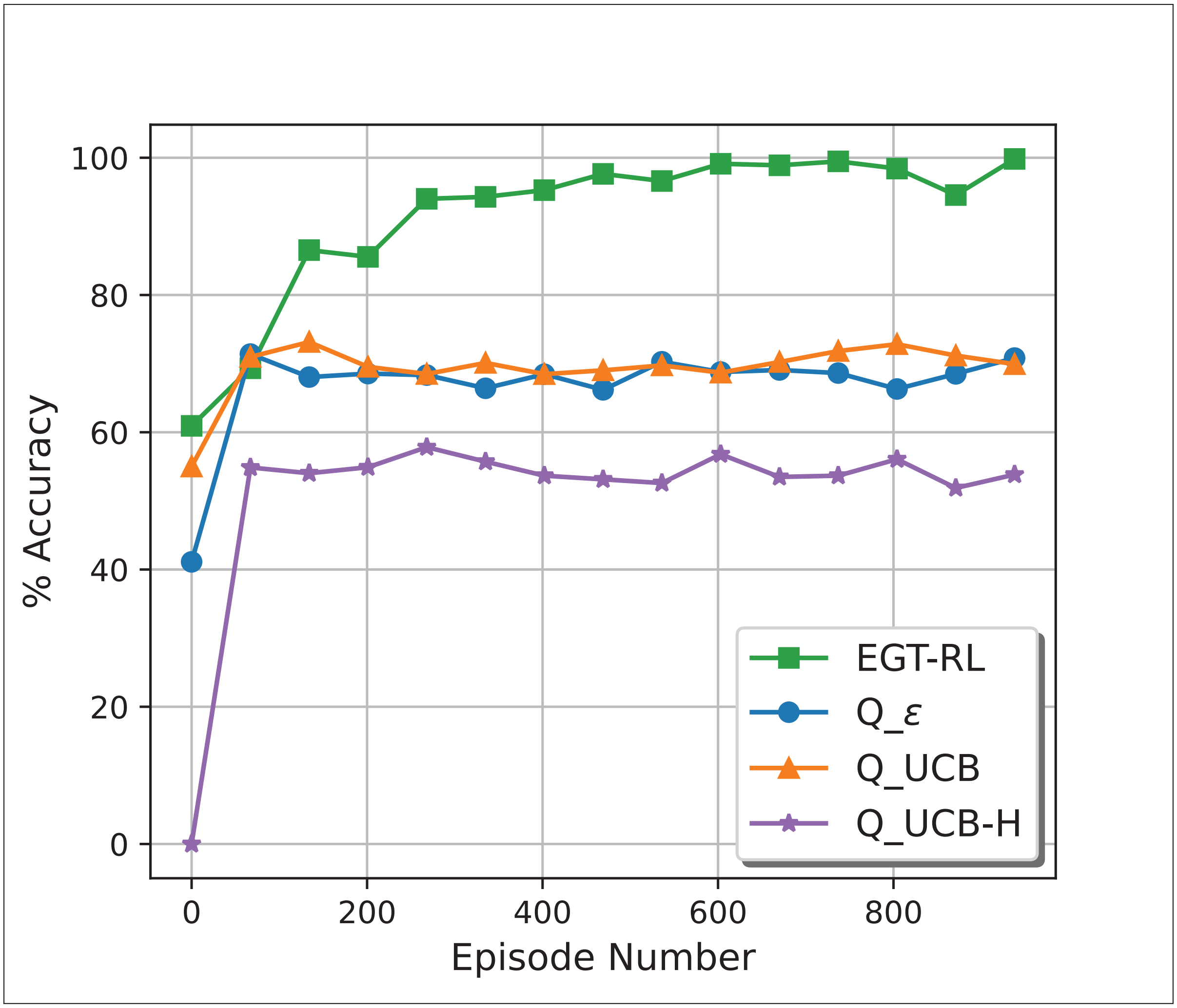

Figure 2.4 Average sensing accuracy under $\eta = 10$, $P = 0.3$ for all approaches

compared to Figure 2.4 brings the performance of other approaches closer to that of the PCMARL algorithm, nevertheless the PCMARL algorithm is still significantly more accurate.

In order to show the effect of the load factor on sensing accuracy, Figure 2.6 plots the sensing accuracy of all approaches for two different load factors: $\eta = 2$ and $\eta = 100$. With a high load factor, the PCMARL algorithm converges to 100% accuracy in under 200 episodes due to cooperation among a large number of SUs, which provides a strong defense against SDF attacks.

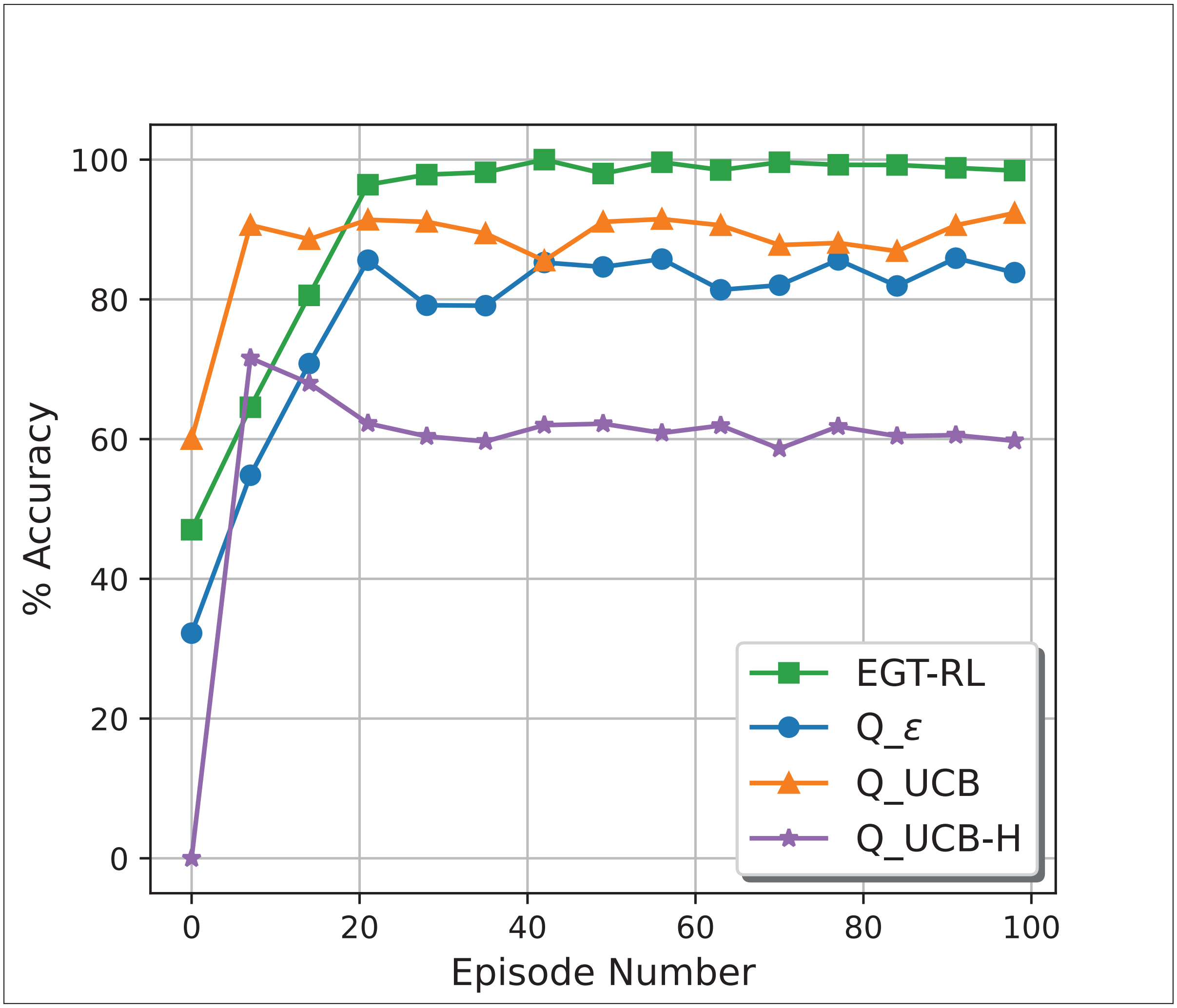

Figure 2.5 Average sensing accuracy under $\eta = 5$, $P = 0.1$ for all approaches

Figure 2.7 shows the sensing accuracy of the PCMARL algorithm under various load factors to show scalability. As the load factor increases, majority vote heavily influences individual decisions. This figure also shows our approach's resilience against SDF attacks when the probability of falsification is high, i.e., $P = 0.3$. The system was tested with various load factor values up to $\eta = 200$, and it achieved 100% sensing accuracy in about half as many episodes as the other approaches for all load factors.

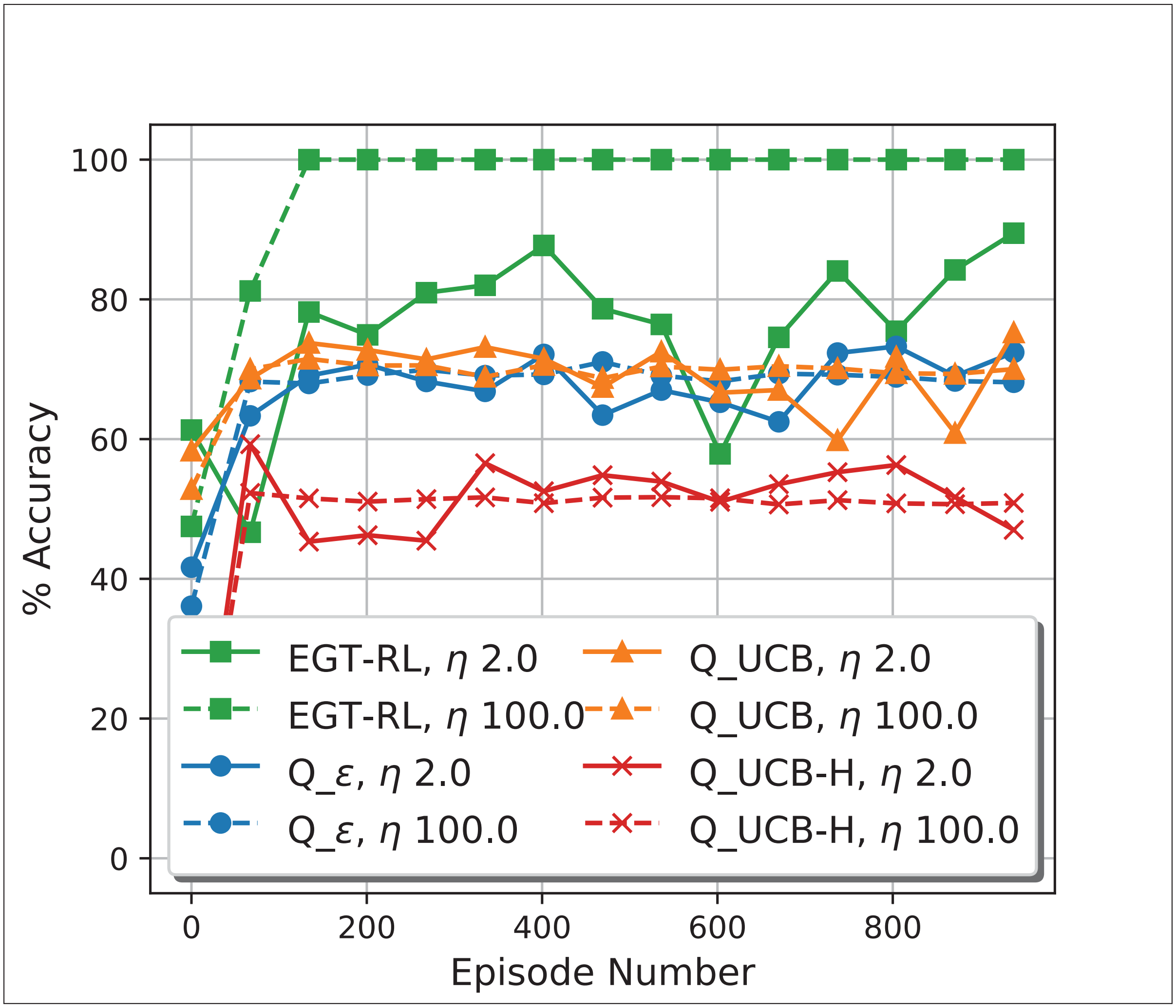

Figure 2.6 Average sensing accuracy under $\eta = 2$ and $\eta = 100$ for $P = 0.3$ for all approaches

Figure 2.8 compares the accuracy of all approaches with different probabilities of falsification in order to show the proposed algorithm's resilience. The PCMARL algorithm is the most resilient to SDF attacks even when 30% of the reported channel statuses have been falsified. The performance of all approaches degrades with increasing P. Figure 2.9 shows the PCMARL algorithm's resilience to SDF attacks by plotting its accuracy for various probabilities of sensing falsification.

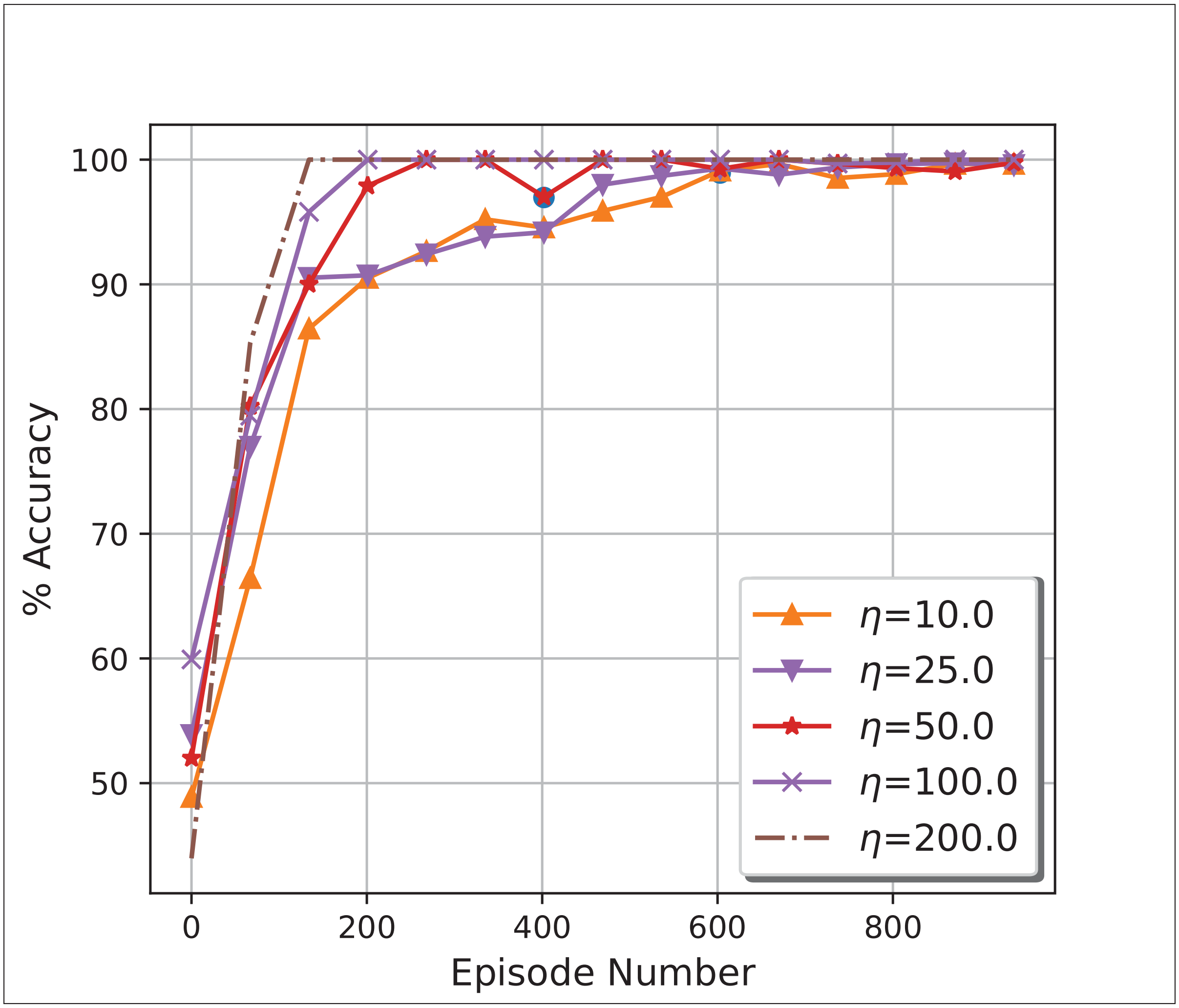

Figure 2.7 Average sensing accuracy of the proposed algorithm for different load-factors

Figure 2.10 shows the average rewards per episode over all SUs for all approaches. Each episode has $T = 20$ slots, and the SU s_i receives reward $r_j^{s_i}(o_t) = 1$ for sensing and correctly predicting the status of channel j in state o_t. The minimum and maximum quantity of rewards per episode is therefore $[0, 20]$, which represents the y-axis in all rewards-related figures.

Figure 2.11 shows the average reward per episode for all approaches under various load factors. The PCMARL algorithm converges faster to higher rewards than the other approaches despite the probability of falsification being $P = 0.3$. Similarly, Figure 2.12 shows the average reward

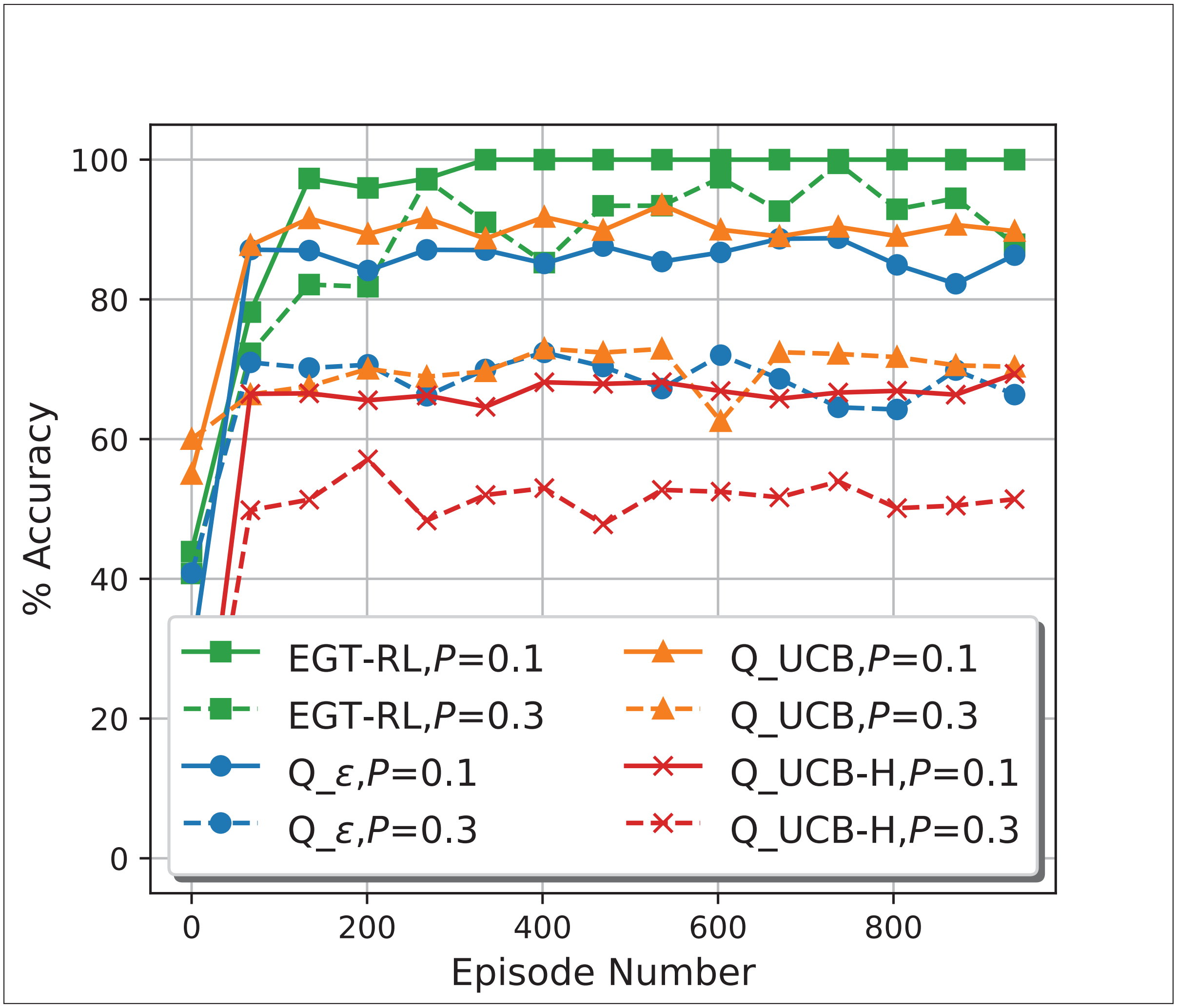

Figure 2.8 Average sensing accuracy of all approaches under various falsification probability

per episode with different probabilities of falsification for $\eta = 10$. It emphasizes the PCMARL algorithm's resilience against SDF attacks, as its average reward even when $P = 0.3$ is higher than that of other approaches when $P = 0.1$.

One of the contributions of this work is giving SUs the freedom to choose a channel for sensing based on their reward history. If a channel frequently becomes busy, an SU can decide to stop sensing it, leave the coalition and join another coalition (the one corresponding to the channel from which it has historically received the highest reward) for sensing in the hope of finding a

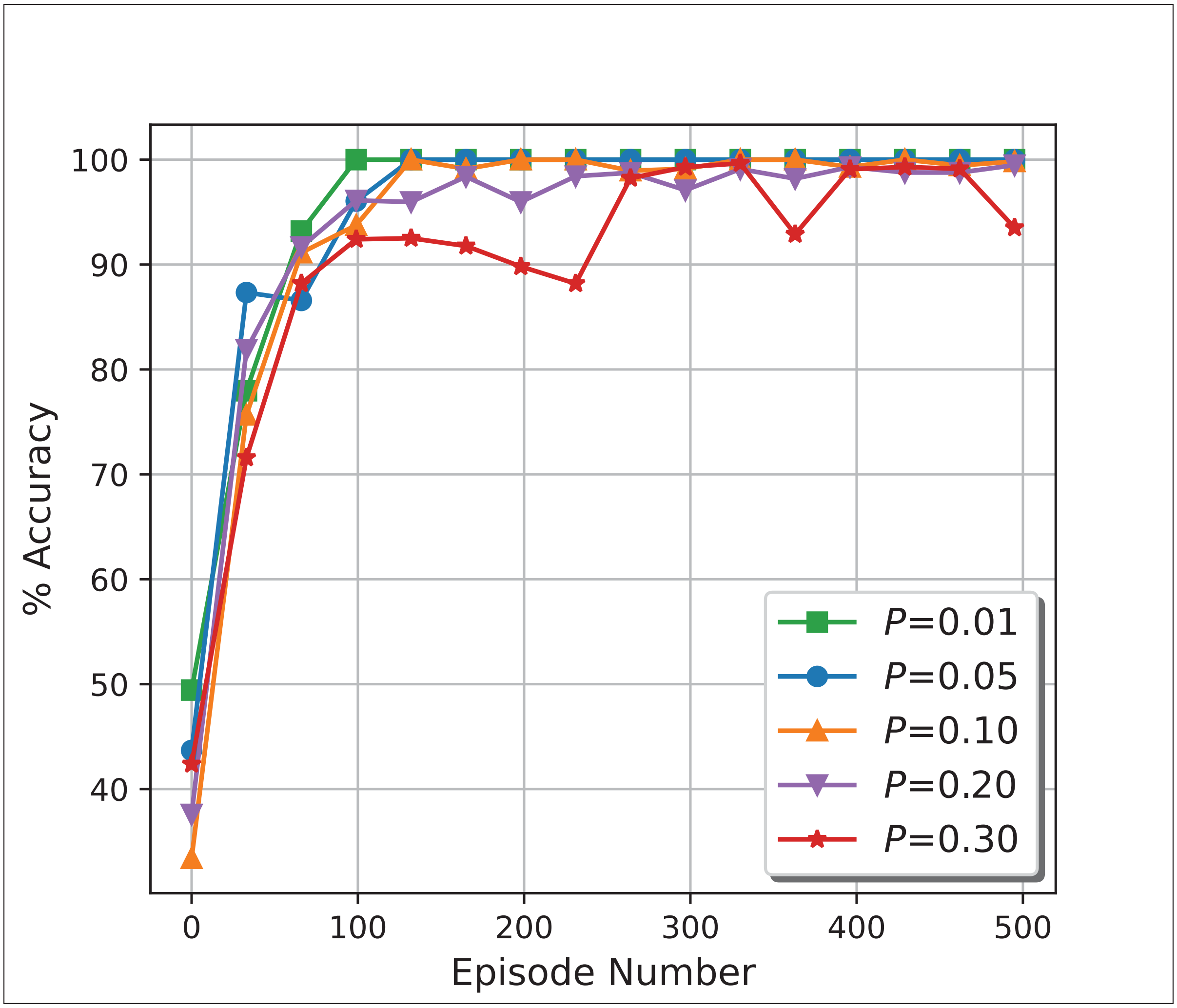

Figure 2.9 Average sensing accuracy of the proposed algorithm under various falsification probability

more frequently available channel. Another reason for SUs to switch between coalitions is their mobility, which affects their sensing accuracy. The SUs can thus pick a channel for which they are able to predict channel status more accurately , thereby leading to higher rewards. In other words, the SUs also learn to join the most suitable coalition for higher rewards. This concept is reflected in the PCMARL algorithm's behavior as shown in Figures 2.10, 2.11 and 2.12.

The SUs have the free will to be contributors or free riders, and there is a small per-step negative cost associated with participating in spectrum sensing to encourage the SUs to learn to participate

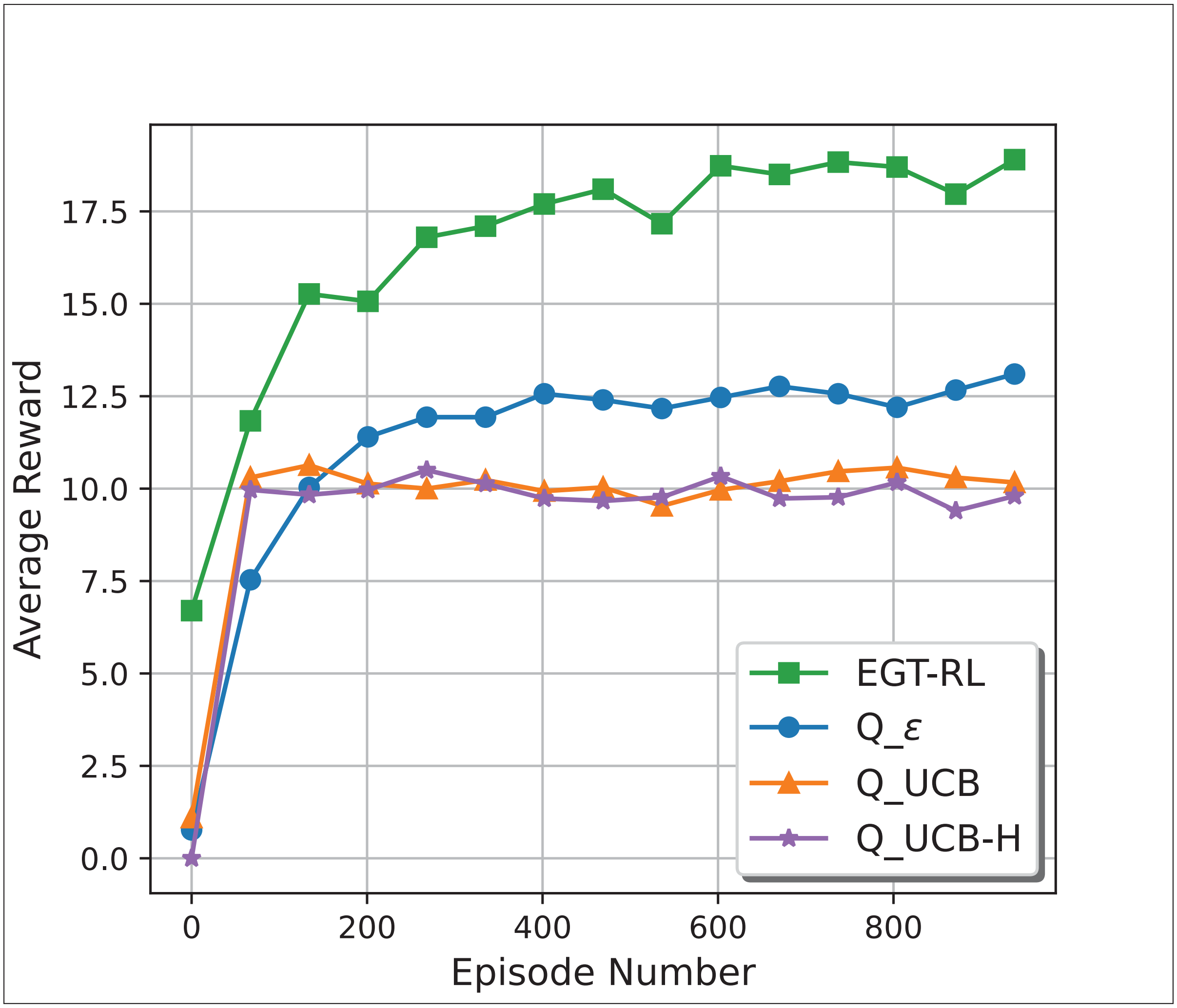

Figure 2.10 Average reward over all secondary users per episode for all approaches

in sensing only when a high reward is expected in order to conserve their energy. In other words, SUs can learn to settle for a lower reward to conserve their energy. The number of times that SUs participate in spectrum sensing per episode is indicated by their contribution percentage, shown in Figure 2.13, which also reflects their total energy consumption. We assume that an SU spends one unit of energy whenever it participates in spectrum sensing in a particular time slot and zero units of energy whenever it is a free rider. Note that the epsilon-greedy algorithm exhibits better exploration of the environment than the UCB and UCB-H algorithms that prioritize saving energy.

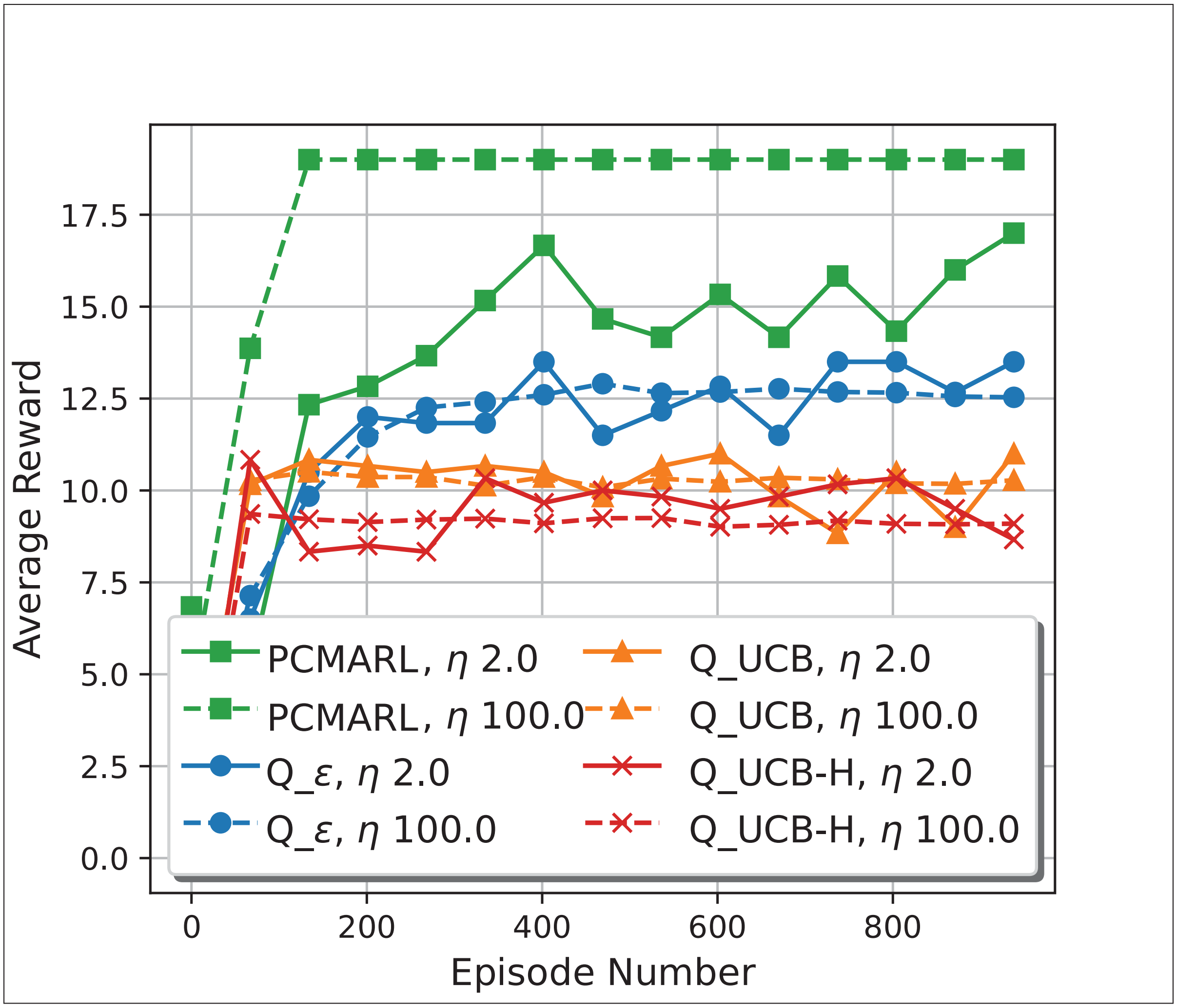

Figure 2.11 Average reward per episode under varying load factors

However, total rewards and contribution percentage alone do not give a complete measure of energy efficiency for SUs since they do not reflect whether the reward is worth the energy spent for spectrum sensing. For example, an SU may have participated in spectrum sensing only three times per episode to conserve energy and predicted correctly each time. SU's total reward per episode might still be lower than that of the other SUs that participated more in sensing. Thus, rewards and contribution percentage, when looked at separately, do not fully represent energy efficiency.

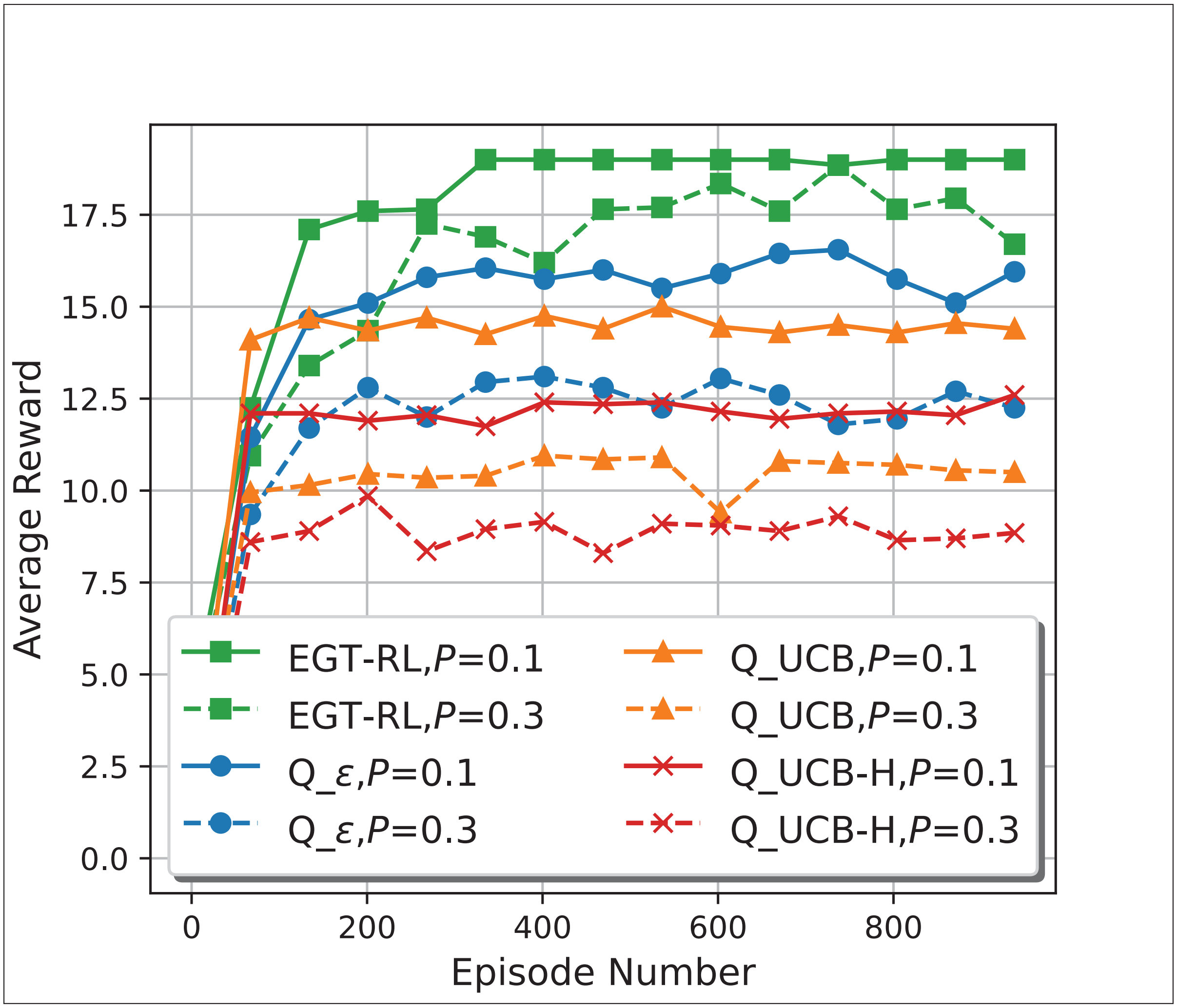

Figure 2.12 Average reward per episode for all approaches under various falsification probability

A better indicator of energy-reward balance is energy-per-reward, which reflects the total energy the SU consumed per episode normalized by the total rewards it received per episode. Figure 2.14 shows total energy-per-reward for the SUs for all approaches. A lower energy-per-reward ratio reflects higher energy efficiency since it shows that the SUs learned to participate in sensing only when they expected a reward. Ideally, the energy-per-reward ratio should be equal 1. The PCMARL algorithm achieves a near ideal energy-per-reward ratio, whereas UCB-H is the most energy inefficient approach and does not learn to lower the participation percentage to conserve energy. Figure 2.15 shows the PCMARL algorithm's energy-per-reward ratio under various

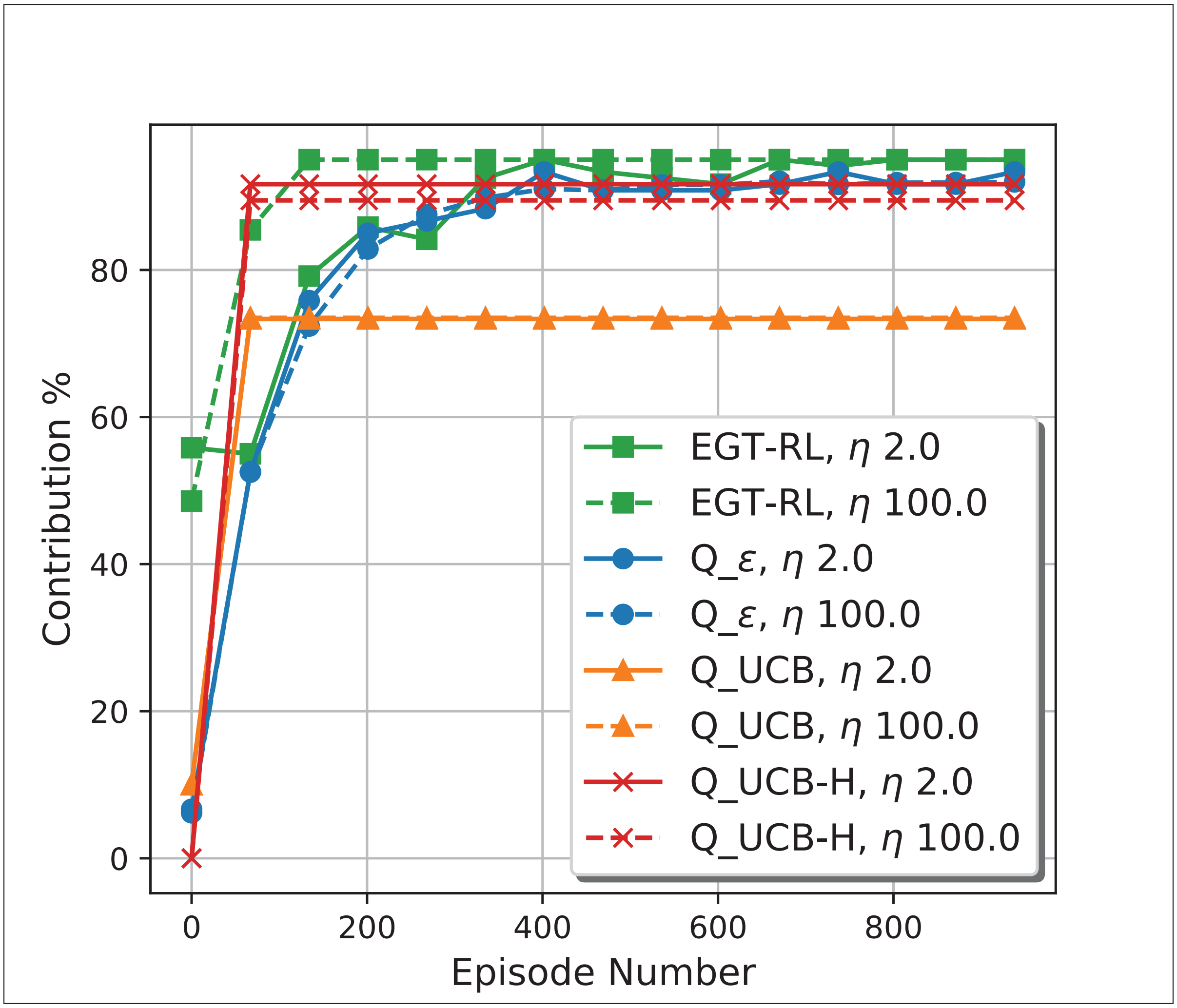

Figure 2.13 Average contribution percentage for all approaches

load factors, and it is clear that as the load factor increases, the SUs have more opportunities to conserve their energy and still achieve a near ideal energy-per-reward ratio.

A related measure of energy wastage is sensing overhead, which indicates the amount of energy the SUs spent sensing a channel when they did not receive any reward. Figure 2.16 shows the average sensing overhead of all approaches. The PCMARL algorithm has extremely low sensing overhead after half as many episodes as the other approaches. Sensing overhead is high at the beginning when the SUs are exploring the environment, but drops as the number of episodes increases. In an ideal system with zero wasted energy, the sensing overhead will be zero. Figure

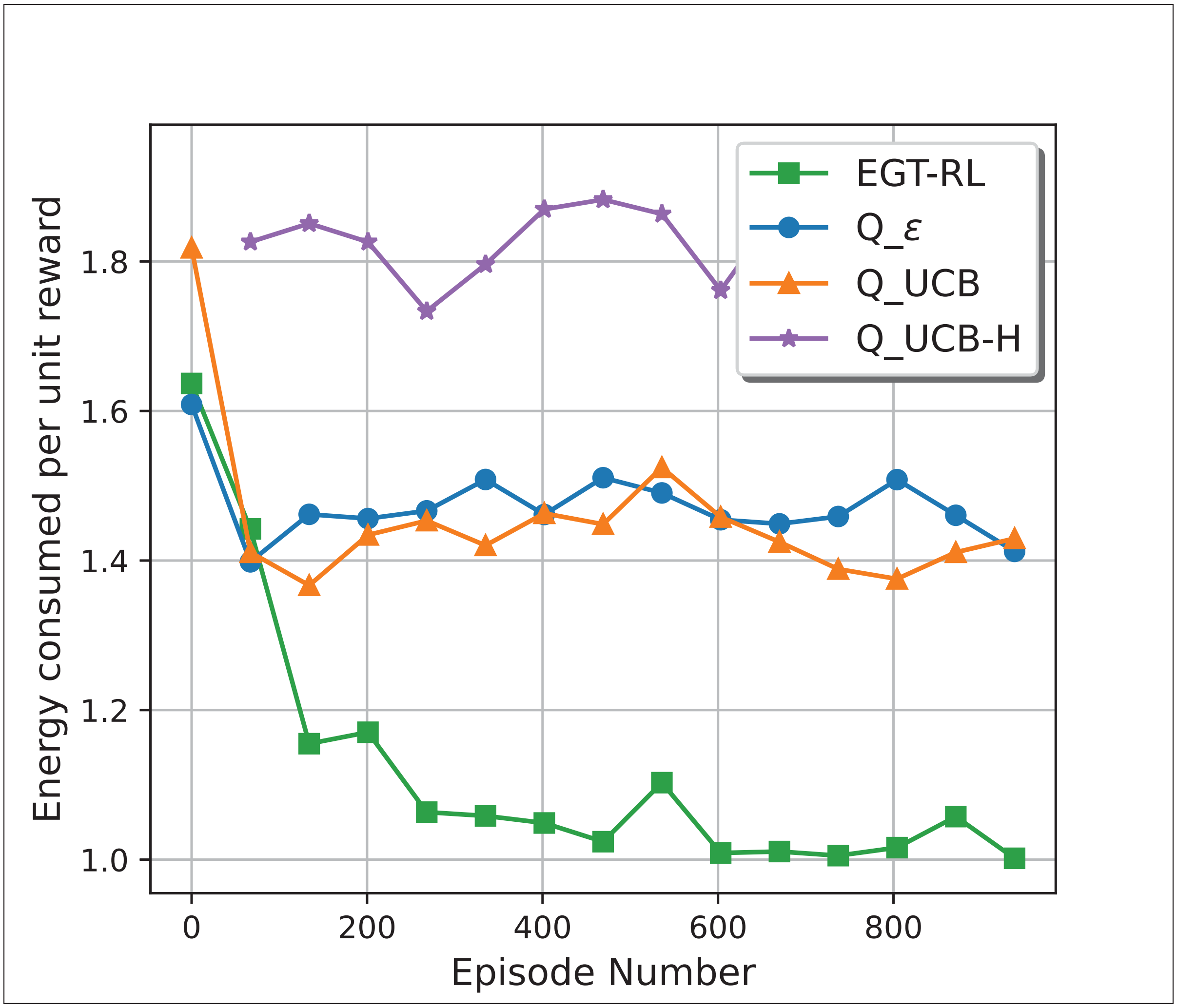

Figure 2.14 Average energy consumption per unit reward for all approaches

2.17 shows the PCMARL algorithm's sensing overhead with various P values and that it has very low overhead even with high P values.

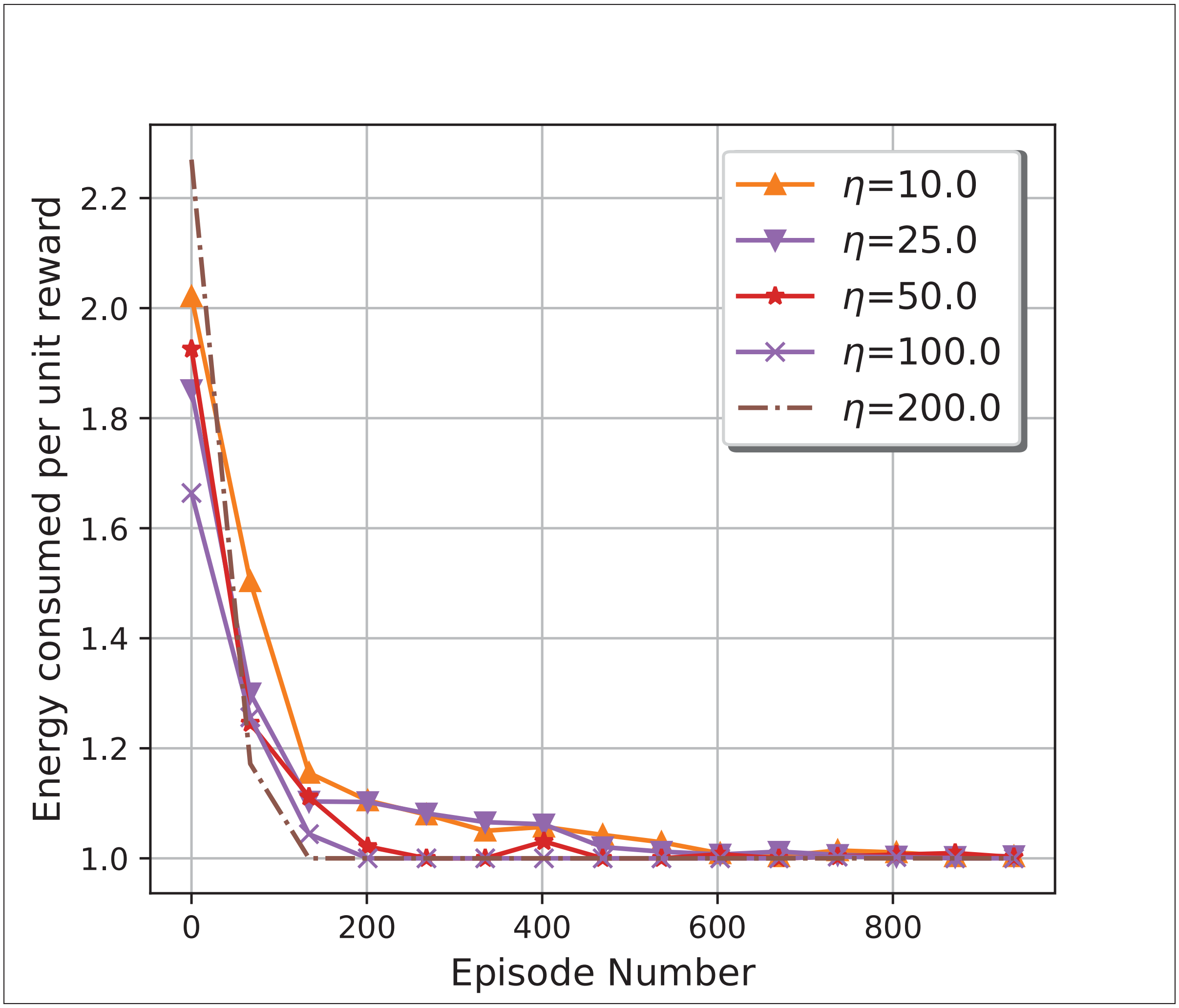

Figure 2.15 Average energy consumption per unit reward for the proposed algorithm under various load factors

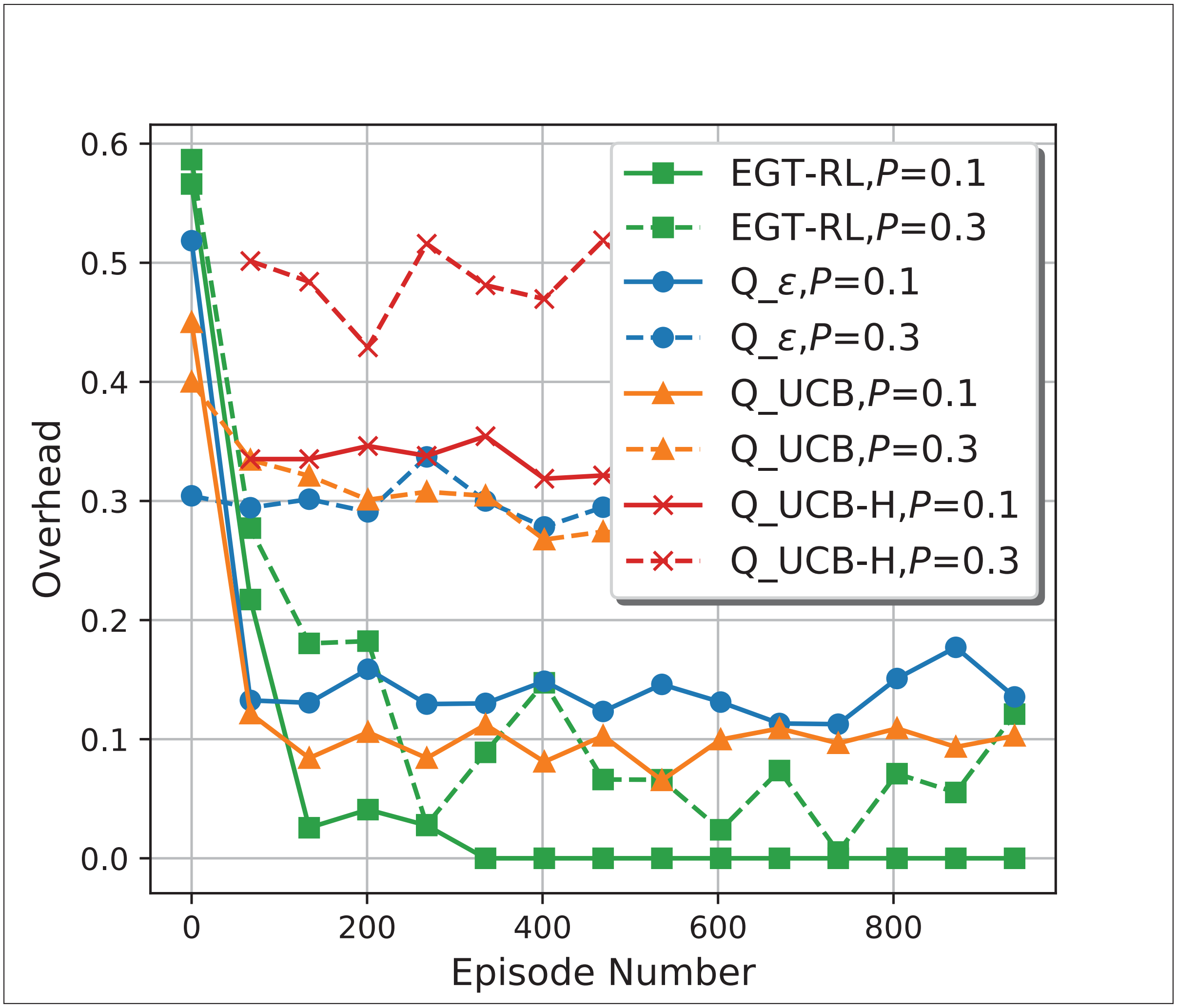

Figure 2.16 Average sensing overhead of all approaches under various falsification probabilities

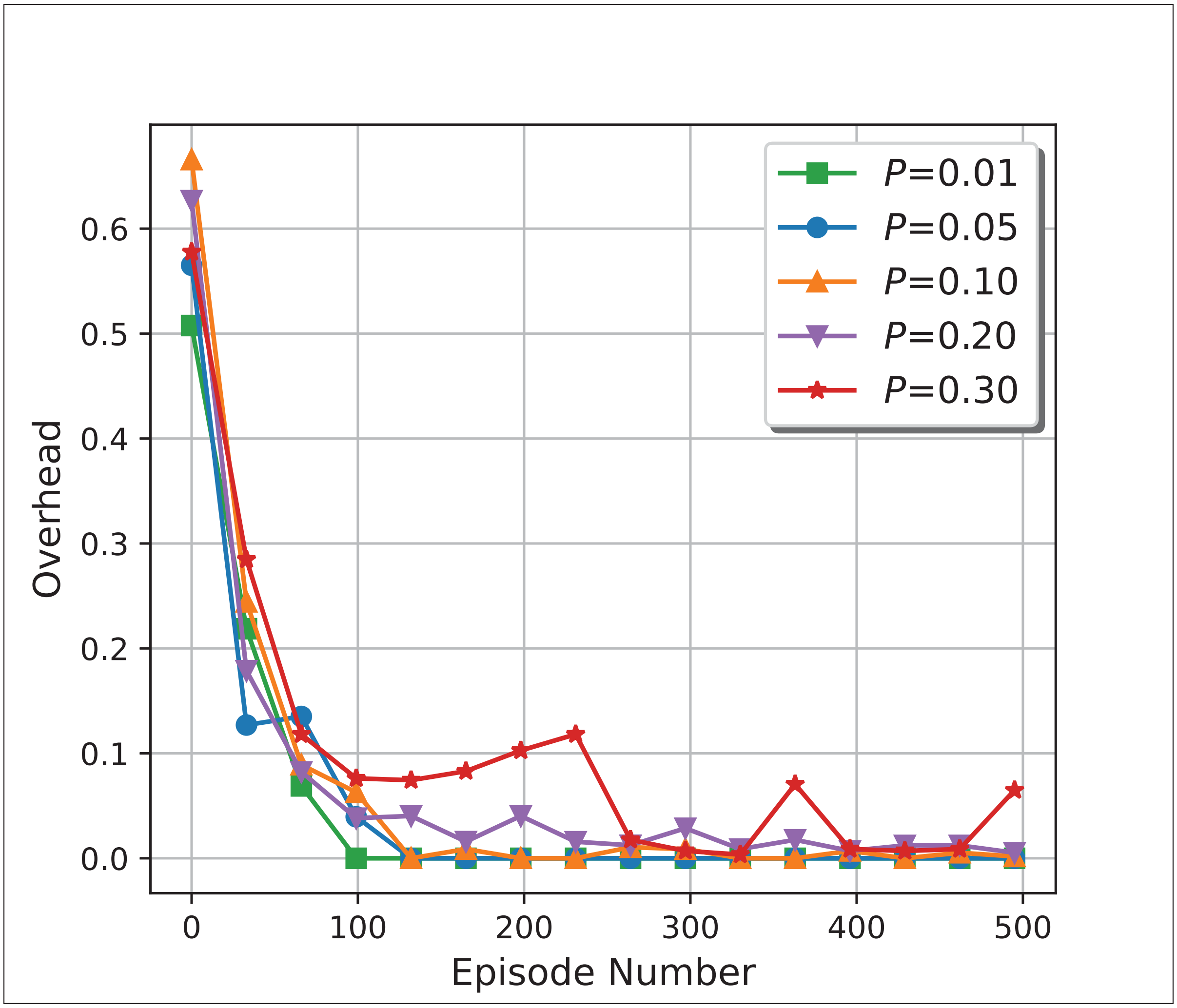

Figure 2.17 Average sensing overhead of the proposed algorithm under various falsification probabilities

2.5 Conclusion

In this paper, we propose a PCMARL spectrum sensing algorithm that takes advantage of coalition formation and majority voting to combat SDF attacks in CR-IoT. The proposed PCMARL algorithm outperforms state-of-the-art RL-based sensing algorithms, namely Q-ϵ, Q-UCB, and Q-UCB-H, in terms of sensing accuracy and energy efficiency. The adopted majority vote mechanism makes the PCMARL algorithm resilient to SDF attacks even when 30% of the reported sensing results have been falsified. Unlike centralized fully CSS approaches, the proposed PCMARL algorithm has a coalition-specific partially cooperative sensing and belief-sharing mechanism that makes its performance directly proportional to the number of devices in the network. Moreover, the algorithm forces agents to learn to optimize their participation percentage to conserve energy. Agents also learn to switch channels based on their reward history to maximize their rewards. Future works may consider approximating the Q-table to extend the work to very large state-spaces and implement it in low-powered devices.

CHAPTER 3

PARTIALLY COOPERATIVE REINFORCEMENT LEARNING FOR HYBRID-ACTION COGNITIVE RADIO NETWORKS WITH IMPERFECT CHANNEL STATE INFORMATION

Sadia Khaf[1] , Georges Kaddoum[1] , Joao V. C. Evangelista[2]

[1]Department of Electrical Engineering, École de technologie supérieure, Montreal, QC H3C 1K3, Canada
[2]Ericsson Canada

Published in IEEE Open Journal of the Communications Society, June 2024.

Abstract

Cognitive radio networks (CRNs) mitigate spectrum scarcity by leveraging the holes in the licensed spectrum to enable internet of things (IoT) devices to opportunistically access the spectrum. However, IoT devices need to sense the spectrum before they can access it, which is an energy-intensive process and hinders the practical implementation of opportunistic spectrum access for energy-constrained IoT devices. In this context, reinforcement learning (RL)-based algorithms that encourage cooperation among IoT devices to eliminate the need for constant sensing are promising candidates for practical CRN implementation. As exciting as the application of RL to CRNs is, benchmarking the performance of different algorithms is a huge challenge due to a lack of standardized comparison metrics, especially for hybrid action spaces that comprise both discrete and continuous actions. We propose a hybrid discrete-continuous space deep RL algorithm that maximizes the energy efficiency of CRNs by optimizing sensing, cooperation, and transmission by IoT devices. We also analyze the algorithm's performance by setting the theoretical upper bound for throughput and find that it reaches 99.4% of the theoretical upper bound, while its discrete action-space version reaches 96% and other baseline algorithms range between 70% and 86%.

3.1 Introduction

The exponential growth of the internet of things (IoT), which is expected to reach 41.6 billion devices and generate 79.4 zettabytes (ZB) of data in 2025, has resulted in increased demand for bandwidth. Cognitive radio networks (CRNs) have emerged as a promising solution to address the limited spectrum issue (Khan *et al.*, 2016). CRNs enable unlicensed IoT devices, which are termed secondary users (SUs), to opportunistically access the underutilized frequency bands that are licensed to primary users (PUs). However, the effective coexistence of heterogeneous devices in CRNs poses significant challenges for optimal spectrum utilization and network performance (Zhang *et al.*, 2015).

One primary condition for SUs to access the licensed spectrum is to avoid causing interference to PUs. Spectrum sensing, which is a mechanism that enables SUs to access the licensed spectrum without interfering with PU traffic, detects spectrum holes—free slots where PUs are not transmitting—but is energy-intensive (Zhang *et al.*, 2020a). The lack of energy-efficient sensing and transmission mechanisms is a dominant barrier to practical CRN implementation.

Reinforcement learning (RL)-based sensing and scheduling algorithms offer promise for addressing the energy constraints of CRNs (Khaf *et al.*, 2022; Sarikhani & Keynia, 2020; Liu *et al.*, 2020). Unlike heuristic or game-theoretic methods, RL-based algorithms adapt to PU traffic patterns and eliminate the need for constant sensing by learning in real time and facilitating SU cooperation to conserve energy (Zhang *et al.*, 2020a). Nevertheless, applying RL in CRNs and maintaining scalability becomes complicated when there are a variety of devices.

Hybrid action spaces, which integrate discrete and continuous actions, pose another challenge to RL-based CRNs. Sophisticated algorithms are required to manage the different actions and handle the heterogeneity effectively (Li *et al.*, 2023).

Furthermore, the limitations of low-powered IoT devices hinder the gathering of perfect channel state information (CSI), which complicates decision-making with imperfect CSI (ICSI) for single-agent RL (Kaur *et al.*, 2023; Kaur & Kumar, 2022). Single-agent sensing and decision making also suffer from spatial false alarms (SFA) due to active PUs outside the sensing region and require either adjusting the decision threshold based on aggregate interference (Furtado *et al.*, 2016) or reliance on spatio-temporal sensing (Zhang *et al.*, 2018; Khalid & Yu, 2019).

Cooperative RL presents promise for complex decision-making in multi-agent systems such as CRNs. However, current approaches often rely on centralized structures, which poses scalability challenges as networks grow in complexity and have a greater number of agents.

Developing decentralized cooperative RL algorithms is an open challenge to enable scalable and efficient CRN operations. These algorithms must facilitate cooperative learning without having overwhelming computational or memory requirements and address challenges like hybrid action spaces and ICSI. Successful development could optimize spectrum allocation, alleviate the curse of dimensionality, and enhance spectral and energy efficiency in CRNs.

This paper is organized as follows: The next section provides a comprehensive review of the related literature and highlights existing methodologies and approaches in the field. Section 3.3 presents the system model and problem formulation. Section 3.4 presents the proposed algorithms in detail. Section 3.5 describes the experimental setup and discusses the simulation results obtained. Finally, the paper concludes with a summary of our findings and avenues for future research in Section 3.6.

3.2 Related works

Research in the domain of CRNs has evolved significantly in recent years and aims to tackle the multifaceted challenges associated with spectrum access, interference mitigation, and network optimization. Spectrum sensing techniques play a pivotal role in identifying spectrum opportunities for SUs while avoiding causing interference to PUs (Hussain *et al.*, 2020). Traditional methods like energy detection, matched filtering, and cyclostationary feature detection have been extensively investigated for spectrum hole identification. However, SFA and the energy-intensive nature of these approaches remains a pressing concern, particularly for resource-constrained IoT devices operating in CRNs (Sharma *et al.*, 2015; Ali & Hamouda, 2017; Sun *et al.*, 2018; Furtado *et al.*, 2016). In response to this challenge, recent research endeavors have increasingly focused on harnessing machine learning approaches to enhance spectrum sensing efficiency, adaptability, and energy conservation in CRNs (Khaf *et al.*, 2022; Sarikhani & Keynia, 2020; Liu *et al.*, 2020).

Efforts to mitigate the time-intensiveness of spectrum sensing have led to various approaches. Some strategies involve trade-offs between sensing accuracy and transmission opportunities, while others advocate periodic sensing to reduce overall sensing time (Sun *et al.*, 2018; Zhang *et al.*, 2017). Passive listening schemes have also been explored to maximize transmission opportunities while minimizing the interference caused to PUs and SU energy consumption (Sarikhani & Keynia, 2020). Despite these attempts, passive listening still consumes relatively a lot of energy (Mastronarde & Van Der Schaar, 2011), which poses challenges for implementation in low-powered IoT devices. Moreover, cooperative spectrum sensing (CSS) has been shown to exhibit more accurate sensing and lower collision rates compared to non-cooperative mechanisms. CSS is more resilient to malicious attacks like sensing data falsification (SDF) attacks, which further emphasizes the importance of minimizing SU energy consumption in CSS approaches.

Real spectrum data analysis indicates there are certain patterns in PU traffic that SUs can use to their advantage (Sun *et al.*, 2022). However, exploiting these patterns requires prior knowledge of the statistical characteristics of the spectrum data (Li *et al.*, 2017), which are seldom available to SUs (Mastronarde & Van Der Schaar, 2011). RL algorithms are gaining traction (Sarikhani & Keynia, 2020; Liu *et al.*, 2020; Alghorani *et al.*, 2015) because they are able to adapt in real-time without the need for prior statistical knowledge of spectrum data (Li *et al.*, 2017). However, their adaptability in CRNs when the CSI is delayed, noisy, falsified, or unavailable remains an open research area (Khoshkbari & Kaddoum, 2023; Khoshkbari *et al.*, 2023).

The coexistence of discrete and continuous action spaces in CRNs poses challenges for conventional RL algorithms. Recent research has addressed this challenge by exploring novel algorithmic architectures that are capable of effectively handling continuous action spaces

(Koike *et al.*, 2023; Guo & Zhao, 2022; Li *et al.*, 2020). Furthermore, due to the limitations of low-powered IoT devices, gathering accurate CSI becomes challenging. Consequently, adapting RL-based decision-making with ICSI remains a crucial focus area for improving RL performance in CRNs (Khoshkbari & Kaddoum, 2023; Khoshkbari *et al.*, 2023; Tan *et al.*, 2022).

To address these challenges, we propose the intelligent multi-user participatory algorithm for cooperative transmission (IMPACT). IMPACT leverages a deep RL-based approach in a decentralized cooperative framework to optimize the SUs' actions in order to maximize the CRN's energy efficiency. The algorithm, which employs a deep deterministic policy gradient (DDPG) structure with four deep neural networks, can handle hybrid action spaces and is benchmarked against exhaustive search and other algorithms. Participatory spectrum sensing and results sharing enable partial cooperation among agents to conserve energy and improve sensing accuracy while avoiding causing interference to PUs (Khaf *et al.*, 2022). The proposed coalition scheme enhances the SUs' learning with ICSI and mitigates SFA and SDF attacks for granular optimization of the sensing, cooperation, and transmission processes.

3.2.1 Novelty and contribution

IMPACT is novel in a number of ways and makes several contributions:

- **Handling hybrid action space:** It efficiently manages both discrete and continuous action sets, which enables effective decision-making in heterogeneous CRN environments.
- **Addressing ICSI:** It incorporates mechanisms to adapt RL-based decision-making strategies with ICSI, which mitigates the impact of delayed, inaccurate, falsified, or unavailable channel information in CRN environments.
- **Facilitating decentralized cooperation:** It introduces a decentralized cooperative learning framework that enables agents in CRNs to cooperate in select small groups and make informed decisions without relying on centralized structures, which helps to improve scalability and efficiency.
- **Enhanced scalability:** It leverages decentralized cooperation to reduce the dependency on extensive shared resources and enable efficient operation in large-scale CRN scenarios.

- **Benchmarking against various methods:** It is benchmarked against exhaustive search, bandit-based approaches, random selection, and a variant termed IMPACT-D, which discretizes the action space using two actor-critic deep Q-networks (DQNs). This comprehensive evaluation makes it possible to assess IMPACT's performance against that of various baseline methods that are commonly used in CRNs.

Therefore, the proposed algorithm has the potential to significantly improve the energy efficiency and performance of CRNs, and we hope that it will inspire further research on applying RL to CRNs.

3.3 System model

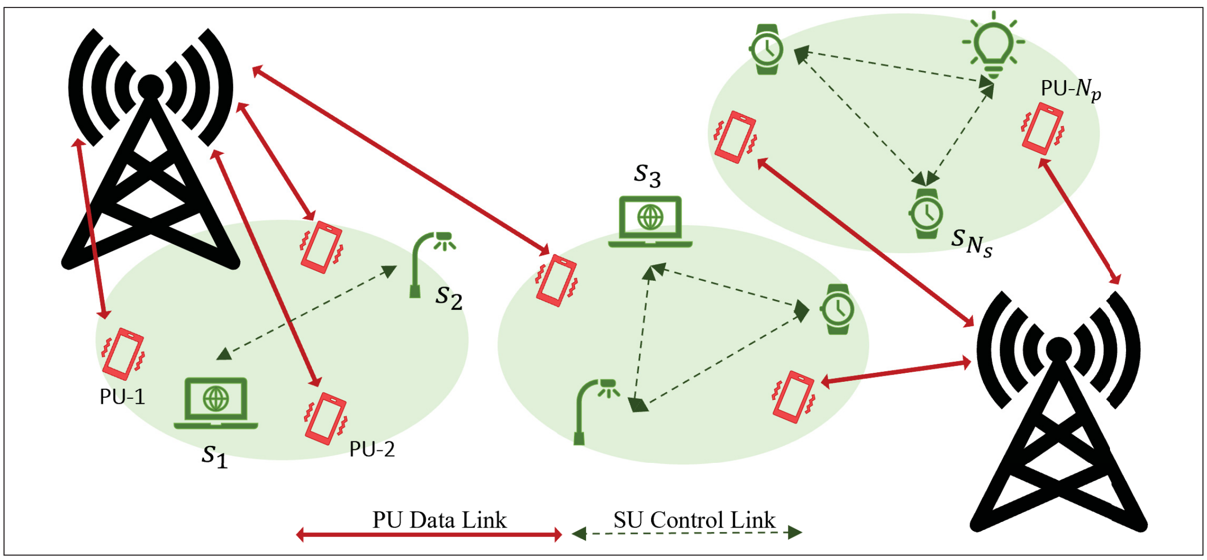

Figure 3.1 System model

We consider a cellular system consisting of N_p PUs surrounded by N_s SUs, as shown in Figure 3.1. Each SU s_i can independently decide the portion of time slot t it spends in sensing and cooperating with other SUs to improve its sensing accuracy. The SUs have limited power and can sense only a limited number of channels in a given time slot. SUs communicate over common control channel for handshake, neighbour discovery, channel access negotiation, topology change, and collaboration (Lo, 2011), and opportunistically access the spectrum holes in licensed PU channels after sensing and/or collaboration. Moreover, both active sensing and passive listening to cooperate with others are energy-consuming processes, and the SUs' objective is to minimize the amount of energy spent on learning over time.

3.3.1 Time slots

Each time slot t can consist of a sensing t^{sn}, a cooperation t^{cp}, and a transmission t^{tr} period, as shown in Figure 3.2. SUs that do not participate in sensing or cooperation can conserve time and energy and use the entire time slot for transmission if they so choose; however, they then run a greater risk of causing harmful interference to PUs by transmitting when the PU channel is busy. Therefore, SUs' objective is to maximize their throughput and energy efficiency while staying within the PU interference thresholds.

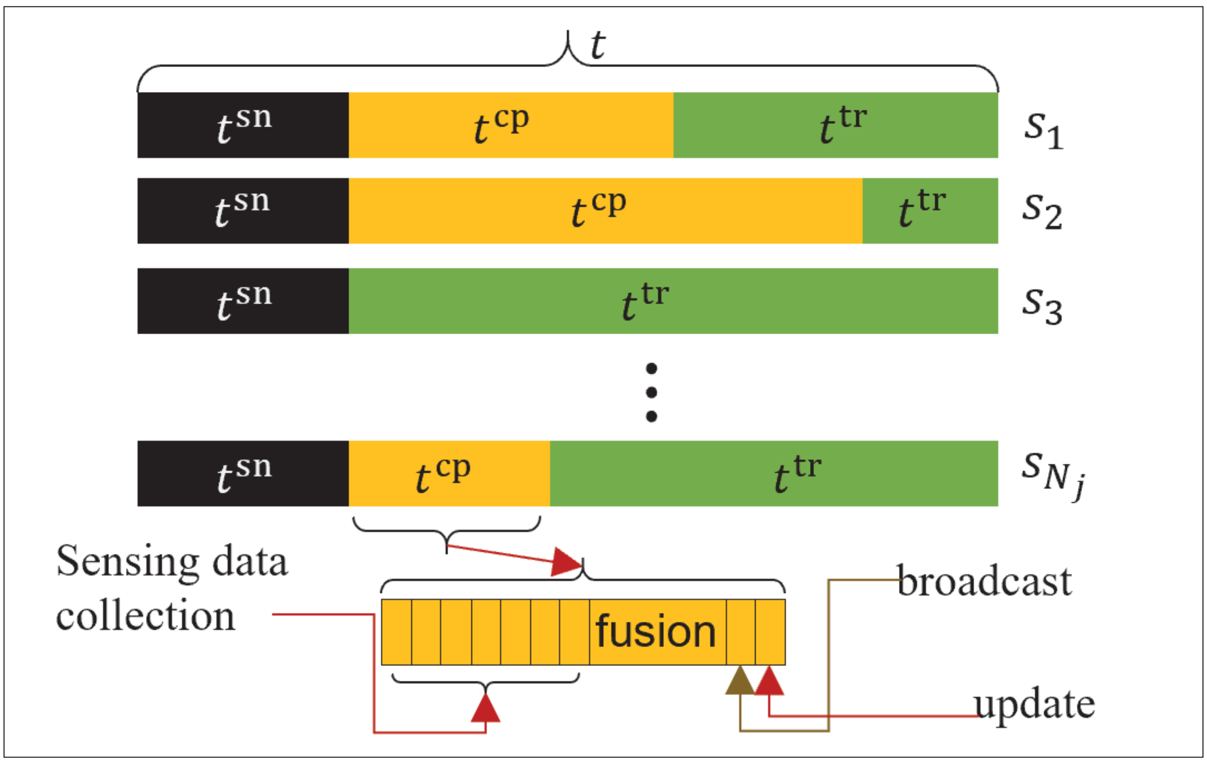

Figure 3.2 Time slot model

Participation in sensing is a binary decision, with a minimum amount of time allocated for sensing. If an SU decides to sense a channel, it will receive the PU's signal power for that channel and infer the channel status by thresholding (Koçkaya & Develi, 2020). We consider both the perfect CSI scenario in which PU signal detection is assumed to be accurate and the ICSI scenario in which the PU's detected signal power is noisy, which can lead to the inaccurate inference of channel status $\{idle, busy\}$. Other forms of ICSI explored include delayed channel state and wrong channel state.

3.3.2 Transmission patterns

Transmission patterns represent different channel usage behaviors in a CRN. They offer channel utilization models to facilitate the evaluation of CRN protocols and algorithms. Exploring different transmission patterns is particularly important in CR, since SUs with simpler learning algorithms might be able to learn well from the environment with periodic channel utilization, but may suffer in more challenging scenarios. A variety of transmission patterns representing channel utilization by humans (bursty), sensor networks (periodic), medical equipment (frequently busy, random), streetlights (frequently idle), public transport systems (periodic, prolonged bursty) etc. is important in order to analyze the learning capabilities of SUs. Figure 3.3 shows a brief snapshot of a detected PU signal and its corresponding inferred channel status with different types of transmission patterns. The inferred channel status shown is assuming perfect CSI. With ICSI, the SUs need to adapt their decisions to base them not only on the current time slot's CSI but also on their previous experience with the environment. The amalgamation of these patterns forms a comprehensive and realistic representation of channel usages, which is crucial for running simulations with the proposed algorithms and testing them on realistic scenarios.

3.3.3 Training and cooperation

An SU can decide not to sense if it is confident enough about the channel status from its previous interactions with the environment after a certain amount of training, or if it decides to be a passive listener and rely on other SUs to sense the channels and transmit their results in the

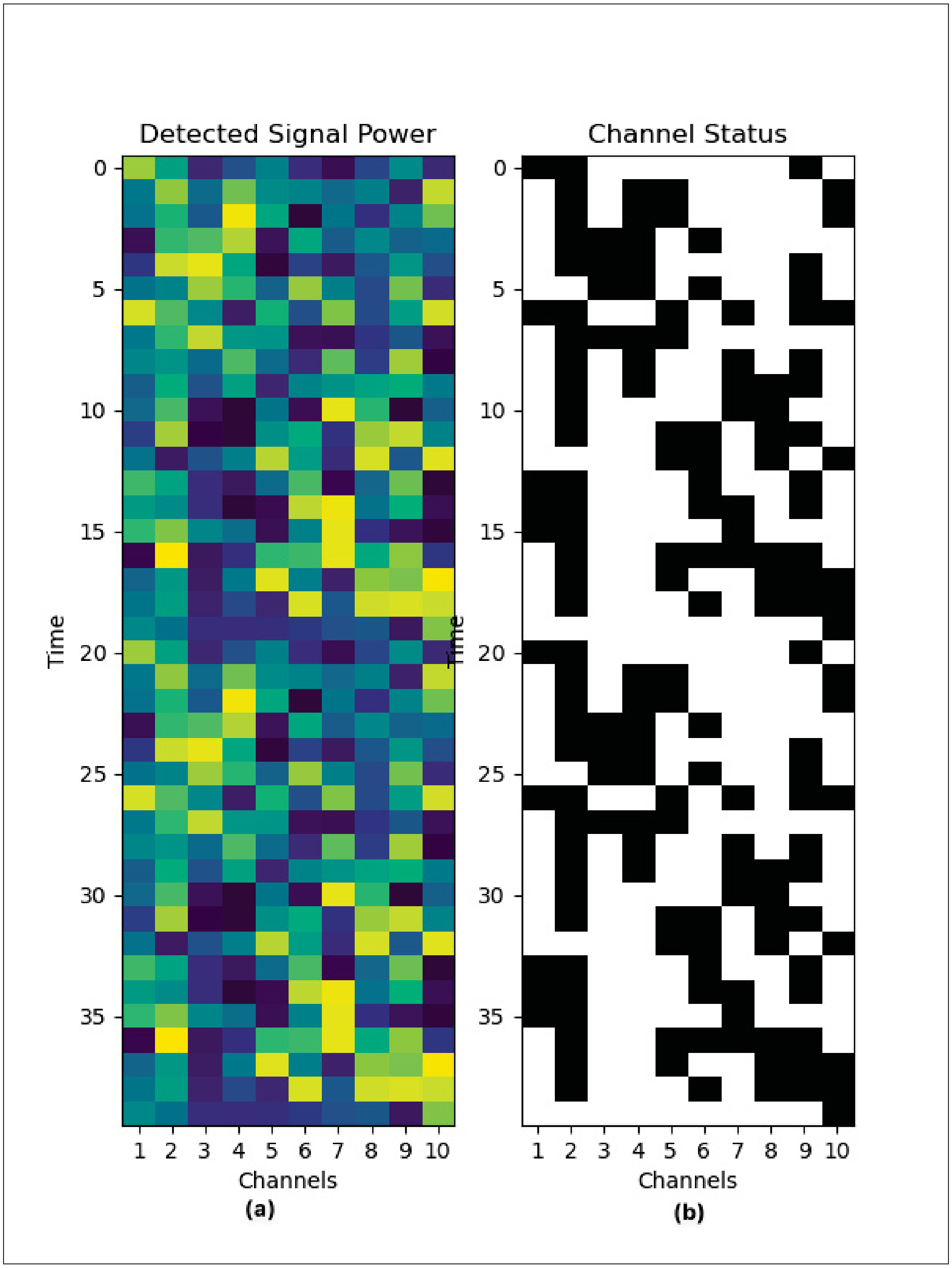

Figure 3.3 Primary user traffic patterns: (a) received signal, and (b) inferred channel status

cooperation time slot. The SUs that participate in cooperation transmit only their inferred channel statuses. Additionally, the SUs can choose to optimize their collaboration time by trading off accuracy for more transmission time. The cooperation time is divided into mini-slots for gathering sensing results from other SUs, applying a local data fusion rule to determine the channel status, and deciding whether or not to transmit. The majority rule is used as the local data fusion rule as it has been shown to be effective at improving sensing accuracy and mitigating SDF attacks (Khaf *et al.*, 2022).

3.3.4 SU transmission

Each SU can access one or more of the M PU channels in the set of channels $C = \{c_i, ..., c_M\}$ and transmit with transmit power p^{s_i} to achieve a desired throughput. Channel selection is a discrete decision variable, whereas transmit power is a continuous variable that the SU aims to optimize in order to maximize its throughput without causing harmful interference to PUs. Together, the decisions about sensing participation, cooperation time, channel selection, and transmit power optimization create a hybrid discrete-continuous action space.

3.3.5 The environment model

The CR-IoT system is modeled as a partially observable Markov decision process (POMDP), as established in previous works (Sarikhani & Keynia, 2020; Liu *et al.*, 2020; Mastronarde & Van Der Schaar, 2011), which enables the SUs to determine the state of a channel through sensing and cooperation. The observation o_t, action a_t, reward r_t tuple is represented as

$$\mathcal{X} = \{o_t, a_t, r_t\}.$$

At each time slot, each SU chooses an action $\mathbf{a}_t \in \mathcal{A}$, performs the action, and receives a reward $r_t \in R$.

3.3.5.0.1 Observation space

The complete state space consists of the PU signal powers in each of the M channels at each time slot $\{t, p_{t,c_1}, p_{t,c_2}, \cdots, p_{t,c_M}\}$. The SUs do not have access to the complete state space. Instead, they observe a part of it $o_t = \{t, p_{t,c_j}, c_j \subseteq C\}$ depending on which channels they decide to sense in a given time slot. If they decide not to sense or participate in cooperation, they may either use o_{t-n}, where $t-n$ is the last time slot in which they sensed the channel in question or approximate o_t from their previous interactions with the environment.

Table 3.1 Table of notations

$\mathbb{1}$	interference indicator	$\mathcal{F}$	update frequency	P_{od}	overall probability of detection
α	soft update parameter	f_s	sampling frequency	P_{of}	overall probability of false alarm
$\mathcal{A}$	action space	γ	discount factor	p	transmit power
$\mathcal{A}_d$	discrete action space	γ_{snr}	average received SNR	p_θ	power threshold
a	action	h	channel gain	ψ	cooperation ratio
$\mathbb{B}$	binary space	I	interference power	Q	estimated Q-value
$\mathcal{B}$	mini-batch sampled from memory	κ	channel sensing indicator	$\mathbb{R}$	continuous space
B	channel bandwidth	L	loss	r	reward
C	set of channels	μ_{noise}	noise mean	ρ	throughput
c_j	channel j	$\mathbb{N}$	discrete space	$\mathcal{S}$	set of SUs
δ	number of discretization intervals	N_{ep}	number of episodes	σ^2_{noise}	noise variance
∇	gradient	N_o	Gaussian noise	s_i	SU i
$\mathcal{D}$	agent memory	N_p	number of PUs	T	horizon
d	terminal state done flag	N_s	number of SUs	t	time slot
E	energy consumption	o	observation	$\theta(.)$	$ActorNet$ output
$\phi(.)$	$CriticNet$ output	o'	next observation	ζ	channel allocation indicator

3.3.5.0.2 Perfect and imperfect CSI

In practical systems, SUs don't always have access to perfect CSI due to SU receiver accuracy limitations, environmental conditions, channel conditions, and the presence of malicious agents in the environment (Khaf *et al.*, 2022), and this can impact their decisions. It is, therefore, crucial to model a system that allows for ICSI and ensures that SUs adapt their decisions to avoid causing interference to PUs. The system model considered includes the following forms of ICSI:

- Delayed CSI: SUs receive CSI in a delayed manner, i.e., instead of receiving o_t, they receive o_{t-n}.
- Noisy CSI: SUs receive a noisy PU signal power, i.e., instead of receiving $o_t = \{t, p_{t,c_j}, c_j \subseteq C\}$, they receive $o_t = \{t, p_{t,c_j} + p_{n_i} \sim \mathcal{N}(\mu_{\text{noise}}, \sigma^2_{\text{noise}})\ ., c_j \subseteq C\}$, where p_{n_i} is the noise signal power with $\mu_{\text{noise}} = 0,\ \sigma^2_{\text{noise}} = 1$.
- Falsified CSI: Malicious agents in the system create sensing data falsification attacks during the cooperation period as outlined in (Khaf *et al.*, 2022).
- Unavailable CSI: SUs do not receive any discernible channel information despite participating in sensing, e.g., they cannot infer a channel status $\{idle, busy\}$ from the received signal and detection threshold (Koçkaya & Develi, 2020).

3.3.5.0.3 Action space

The action space $\mathcal{A}$ is defined as the Cartesian product of all possible combinations of actions an SU can choose from:

$$\mathcal{A} = \prod_{s_i=1}^{N_s} \left(\prod_{c_j=1}^{N_p} \left(\mathbb{B}_\kappa \times \mathbb{R}_\psi \times \mathbb{B}_\zeta \times \mathbb{R}_p \right) \right), \tag{3.1}$$

where $\mathbb{B}$ represents a binary space and $\mathbb{R}$ represents a continuous real space. Note that the notation $(.)_t^{s_i,c_j}$ will be used to represent variable/parameter $(.)$ for SU s_i and channel c_j at time t. Omitting the superscript s_i indicates that $(.)$ is a vector containing values for all s_i, and omitting c_j indicates a constant value of $(.)$ for all c_j.

An action $a_t^{s_i,c_j} \in \mathcal{A}$ represents the action of SU s_i for channel c_j at time t and is defined as:

$$a_t^{s_i,c_j} = \left(\kappa_t^{s_i,c_j}, \psi_t^{s_i}, \zeta_t^{s_i,c_j}, p_t^{s_i}\right), \tag{3.2}$$

where

- $\kappa_t^{s_i,c_j} \in \mathbb{B}_\kappa$ represents s_i's sensing decision for channel c_j at time t.
- $\psi_t^{s_i} \in \mathbb{R}_\psi$ represents the fraction of cooperation time t^{cp} that s_i decides to spend cooperating in time slot t. $\psi_t^{s_i}$ varies between $\psi_{\min}$ and $\psi_{\max}$.
- $\zeta_t^{s_i,c_j} \in \mathbb{B}_\zeta$ represents s_i's transmission decision for channel c_j at time t.
- $p_t^{s_i} \in \mathbb{R}_p$ represents s_i's transmit power at time t, which is restricted to the range $[0, p_{\max}^{s_i}]$, where $p_{\max}^{s_i}$ represents s_i's maximum transmit power.

Similarly, $a_t^{s_i}$ represents s_i's action for all channels at time t and is defined as:

$$a_t^{s_i} = \left(\kappa_t^{s_i,c_j}\ \forall c_j \in \mathcal{C}, \psi_t^{s_i}, \zeta_t^{s_i,c_j}\ \forall c_j \in \mathcal{C}, p_t^{s_i}\right). \tag{3.3}$$

Note that $\mathbb{B}_\kappa$ and $\mathbb{B}_\zeta$ are binary spaces representing sensing and transmission decisions. And $\mathbb{R}_\psi$ and $\mathbb{R}_p$ are continuous spaces representing cooperation time and transmit power. Hence, action space $\mathcal{A}$ is a hybrid discrete-continuous space. A discretized version of $\mathcal{A}$, $\mathcal{A}_d$ will be referenced in parts of this paper where a discretization function f maps $\mathcal{A}$ to $\mathcal{A}_d$ using the discretization factors δ_t and δ_p such that:

- $f_\psi : \mathbb{R}_\psi \to \mathbb{N}^{\delta_\psi}$
- $f_p : \mathbb{R}_p \to \mathbb{N}^{\delta_p}$

The cooperative CRN system's goal is to maximize the overall system's energy efficiency while minimizing interference to PUs by choosing the optimal combination of aforementioned actions.

3.3.5.0.4 Reward

Each SU receives a reward from the environment for its actions based on the success or failure of transmission, the amount of energy it consumes, and its achieved throughput. Since an SU must not cause harmful interference to the PU channel it is accessing, such interference is penalized in the reward function. We take all these factors into account and define the reward $r_t^{s_i}$ for SU s_i at time t as follows:

$$r_t^{s_i} = \frac{1}{E_t^{s_i}} \sum_{c_j=1}^{M} \mathbb{1}_{sp,t}^{s_i,c_j} \rho_t^{s_i,c_j}, \tag{3.4}$$

where $\mathbb{1}_{sp,t}^{s_i,c_j}$ is is the SU-to-PU collision indicator, $\rho_t^{s_i,c_j}$ is s_i's throughput from channel c_j at time t, and $E_t^{s_i}$ is s_i's energy consumption at time t. We define $E_t^{s_i}$ as:

$$E_t^{s_i} = \psi_t^{s_i} E_{cp}^{s_i} + \sum_{c_j=1}^{M} \kappa_t^{s_i,c_j} E_{sn}^{s_i} + \zeta_t^{s_i,c_j} p_t^{s_i} t^{s_i,tr}, \tag{3.5}$$

where $t^{s_i,tr}$ is s_i's transmission time, and $E_{sn}^{s_i}$ and $E_{cp}^{s_i}$ are s_i's sensing, and cooperation energy consumption per slot, respectively. We define the collision indicator $\mathbb{1}_{sp,t}^{s_i,c_j}$ as:

$$\mathbb{1}_{sp,t}^{s_i,c_j} = \begin{cases} -1 & \text{if } \zeta_t^{s_i,c_j} p_t^{s_i} \geq P_{th} \ \& \ p_t^{c_j} \geq P_{th}^{ch} \\ 0 & \text{otherwise} \end{cases}, \tag{3.6}$$

where P_{th} is the power threshold to prevent interference with PU traffic, and P_{th}^{ch} is the power threshold for channel being busy.

SU s_i's throughput, which is given by Shannon's capacity theorem, is determined by:

$$\rho_t^{s_i}\left(a_t^{s_i}, o_t\right) = \sum_{c_j=1}^{M} B_{c_j} \log_2 \left(1 + \frac{|h|^2 p_t^{s_i}}{N_0 B_{c_j} + \mathbb{1}_{ss,t}^{c_j} I}\right), \tag{3.7}$$

where $\mathbb{1}_{ss,t}^{c_j}$ is the SU-to-SU interference indicator, I is the interference power from other SUs, B_{c_j} is the channel bandwidth, N_o is the Gaussian noise power, h is the channel gain, and $p_t^{s_i}$ is s_i's transmission power. I follows Gaussian distribution with mean -80 dbm and variance

$1e-9$, and h follows a Rayleigh distribution. We define the interference indicator $\mathbb{1}_{ss,t}^{c_j}$ as:

$$\mathbb{1}_{ss,t}^{c_j} = \sum_{s_i \in \mathcal{S}} \zeta_t^{s_i,c_j} \ \forall c_j \in \mathcal{C}. \tag{3.8}$$

As the sensing time increases, the probability of detection increases and the probability of a false alarm decreases (Ahmed *et al.*, 2021; Lorincz *et al.*, 2022) as follows:

$$\begin{aligned} P_{od} =& 1-(1-P_d)^M \\ =& 1-\left(1-Q\left[\left(\frac{\epsilon}{\sigma_s^2}-\gamma_{snr}-1\right)\sqrt{\frac{t^{sn} f_s}{2\gamma_{snr}+1}}\right]\right)^M, \end{aligned} \tag{3.9}$$

where $Q[\ .\ ]$ is the Gaussian Q-function, f_s is the sampling frequency, and ϵ is the energy detection threshold. Moreover, t^{sn} denotes time spent sensing, and γ_{snr} is the average signal-to-noise ratio received from the PU at the SU antenna. This relationship is used to model SUs' probability of detecting PU's presence when they participate in sensing under perfect CSI conditions. With delayed, noisy, falsified, or unavailable CSI, they receive o_t as described in Section 3.3.5.0.2.

The overall objective of cooperative CRN is to maximize the SU energy efficiency while limiting the SU-to-PU and SU-to-SU interference and meeting SU throughput constraints, which can be formulated as follows:

$$\begin{aligned} \text{P1}: \max_{\kappa_t^{s_i,c_j},\psi_t,\zeta_t^{s_i,c_j},p_t} & \sum_{s_i \in \mathcal{S}} \left(r_t^{s_i}\left(\kappa_t^{s_i,c_j},\psi_t^{s_i},\zeta_t^{s_i,c_j},p_t^{s_i}\right)\right) \\ & C1 \sum_{c_j=1}^{M} |h|^2 p_t^{s_i} \leq I_{pu},\ \forall s_i \in \mathcal{S},\ \forall c_j \in \mathcal{C} \\ & C2 \sum_{s_i \in \mathcal{S}} \zeta_t^{s_i,c_j} \leq I_{su}\ \forall c_j \in \mathcal{C}, \\ & C3 \sum_{c_j=1}^{M} \rho_t^{s_i}(a_t^{s_i},o_t) \geq \rho_{min}^{s_i},\ \forall s_i \in \mathcal{S} \end{aligned} \tag{3.10}$$

where $C1$ constrains the SU-to-PU interference to remain under the PU interference threshold I_{pu}, $C2$ constrains the SU-to-SU interference to remain under the SU interference threshold I_{su}, and $C3$ constrains each SU's achieved throughput to be above the minimum required SU throughput $\rho_{min}^{s_i}$. Additionally, we assume that the SUs have a full buffer, i.e., they always have some data to transmit.

The problem in (3.10) is extremely hard to solve. It is noted that if we rewrite the argument in the summation of the objective function, which is given in (3.7), as:

$$\begin{aligned}\rho_t^{s_i} = \sum_{c_j=1}^{M} B_{c_j} \log_2 \left(|h|^2 p_t^{s_i} + N_0 B_{c_j} + \mathbb{1}_{ss,t}^{c_j} I\right) \\ - \log_2 \left(N_0 B_{c_j} + \mathbb{1}_{ss,t}^{c_j} I\right),\end{aligned} \tag{3.11}$$

we see that (3.11) consists of the maximization of a double summation of differences of concave functions with integer arguments and integer constraints, and therefore a non-convex integer programming problem, which is non-trivial to solve with conventional optimization methods.

3.3.6 The dimensionality of the action space

The dimensionality of the action space in the system model considered is expansive, and accounts for multiple decision-making facets. With N_p PUs (channels) and N_s SUs, the SUs' individual decisions regarding channel sensing, cooperation time (discretized into δ_t bins), channel selection for transmission, and transmit power (discretized into δ_p bins) result in a substantial action space. The total number of possible action combinations is given by:

$$\begin{aligned}N_{act} &= (N_s \times 2^{N_p}) \times (\delta_t^{N_s}) \times (N_s \times 2^{N_p}) \times (\delta_p^{N_s}) \\ &= N_s^2 \times 2^{2N_p} \times \delta_t^{N_s} \times \delta_p^{N_s}\end{aligned}$$

For example, when $N_s = 2$, $N_p = 5$, $\delta_t = 10$, and $\delta_p = 10$, the resulting action space encompasses $40,960,000$ potential combinations, making methods like exhaustive search and Q-learning impractical and thus motivate using the DRL algorithms with a lower complexity.

3.4 The proposed algorithms

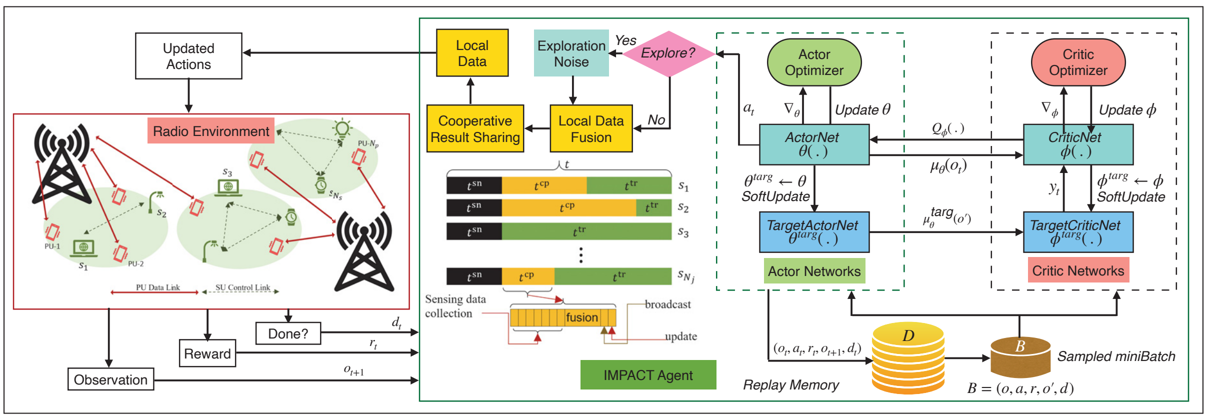

Figure 3.4 Flowchart of the proposed algorithm

The POMDP can maximize the SU's throughput and energy efficiency in many ways, one of which is exhaustive search, where each SU searches through all possible actions for each channel to maximize its throughput. However, the complexity of exhaustive search $O(N_s^2 2^{2N_p} \delta_t^{N_s} \delta_p^{N_s})$ makes it far from practical. We propose the intelligent multi-user participatory algorithm for cooperative transmission (IMPACT) that uses deep RL to maximize the SU's throughput while minimizing energy consumption. For the sake of comparison, we also propose a less complex algorithm called discretized IMPACT (IMPACT-D) that uses half as many neural networks as IMPACT and a discretized action space $\mathcal{A}_d$ that can still achieve performance higher than simpler algorithms such as greedy bandits.

3.4.1 Algorithm 1

IMPACT is shown as a flowchart in Figure 3.4, and its main steps are highlighted in Algorithm 3.1. It uses a DDPG architecture with four identical neural networks, called *ActorNet* $\theta(.)$, *CriticNet* $\phi(.)$, *TargetActorNet* $\theta^{\text{targ}}(.)$ and *TargetCriticNet*$\phi^{\text{targ}}(.)$. The algorithm's advantage lies in

Algorithm 3.1 Intelligent multi-user participatory algorithm for cooperative transmission

Input: $ActorNet\ \theta$ and $CriticNet\ \phi$ parameters, α; Initialize replay memory $\mathcal{D}$;
Initialize $TargetActorNet\ \theta^{targ}$ and $TargetCriticNet\ \phi^{targ}$ as:
$\theta^{targ} \leftarrow \theta, \quad \phi^{targ} \leftarrow \phi$;
for $episode = 1$ *to* N_{ep} **do**
  Reset Environment, observe o_t;
  for $t = 1$ *to* T **do**
    Choose $a_t = \mu_\theta(o_t) + \psi \sim \mathcal{N} + p \sim \mathcal{N}$;
    Cooperative agents transmit inferred channel status ;
    Every participatory agent performs local data fusion to to update a_t using majority rule on inferred channel status;
    Execute action a_t, obtain reward r_t, the next observation o', and terminal flag *done* d ;
    Store tuple (o_t, a_t, r_t, o', d) in replay memory $\mathcal{D}$;
    Sample mini-batch of transitions $\mathcal{B} = (o, a, r, o', d)$ from $\mathcal{D}$;
    Compute $y = r + \gamma(1 - d)Q_\phi^{\text{targ}}(o', \mu_\theta^{\text{targ}}(o'))$;
    Update $CriticNet$ by one step of gradient descent as

$$\nabla_\phi \frac{1}{\mathcal{B}} \sum_{\mathcal{B}} (Q_\phi(o, a) - y(r, o', d))^2.$$

    Update the $ActorNet$ by one step of gradient ascent as

$$\nabla_\theta \frac{1}{\mathcal{B}} \sum_{\mathcal{B}} Q_\phi(o, \mu_\theta(o)),$$

    Update target networks using a soft update as

$$\phi^{\text{targ}} \leftarrow \alpha\phi^{\text{targ}} + (1 - \alpha)\phi$$
$$\theta^{\text{targ}} \leftarrow \alpha\theta^{\text{targ}} + (1 - \alpha)\theta$$

  end for
end for

its ability to use $ActorNet$ to learn a deterministic policy that directly maps observations to continuous actions. During training, the network learns to approximate the optimal deterministic policy by adjusting its weights through gradient descent, so the network's aim is to keep adjusting the mapping of input o_t to output a_t through training until it learns the optimal mapping.

At the beginning of each time slot t all SUs choose which channels, if any, they are going to sense. Afterwards, each cooperative SU shares its channel inferences with the other SUs, and each passive listener gets channel inferences from the other cooperative SUs. During the cooperation slots, the SUs perform local data fusion on the channel inferences to improve their sensing accuracy and mitigate SDF attacks. The SUs choose their transmission power and which channels they want to use for transmission based on the coalitions' data fusion rules (Khaf *et al.*, 2022). All SUs then perform the actions by transmitting with the selected $p_t^{s_i}$ over the selected channels, and achieve a certain throughput $\rho_t^{s_i}(o_t, a_t)$, which, together with any interference they create, the constraints mentioned in (3.10), and feedback they receive from the environment, is used to determine their reward. They also receive the next observation o' and terminal state done flag d from the environment for the channels they chose to sense. This information (o_t, a_t, r_t, o', d) is stored in each agent's memory $\mathcal{D}$ for future reference.

CriticNet is the network responsible for approximating the action value and evaluates the quality of $ActorNet$. $ActorNet$ uses o_t as its input and a_t as the output. *CriticNet* uses (o_t, a_t) as its input and the estimated action value $Q_\phi(o_t, a_t)$ as its output, which is used to measure the quality of training. *CriticNet*'s objective is to maximize this value in order to maximize $ActorNet$'s quality.

The inputs and outputs to networks are normalized. The *ActorNet* uses an input layer with 3 neurons, 2 hidden layers with 64 neurons and ReLU activation, and an output layer with 3 neurons and $tanh$ activation. The *CriticNet* uses input layer with 6 neurons, 2 hidden layers with 64 neurons and ReLU activation, and output layer with 1 neuron and linear activation. *TargetActorNet* and *TargetCriticNet* are copies of *ActorNet* and *CriticNet*, respectively, that have the same input and output structure but are used in a soft update manner to stabilize the latter.

During training, a random batch of transitions $\mathcal{B} = (o, a, r, o', d)$ is sampled from memory $\mathcal{D}$ and used to compute the target y as

$$y(r, o', d) = r + \gamma(1 - d)Q_\phi^{\text{targ}}(o', \mu_\theta^{\text{targ}}(o')), \tag{3.12}$$

where $Q_{\phi}^{\text{targ}}(.)$ represents the output of *TargetCriticNet* and $\mu_{\theta}^{\text{targ}}(.)$ represents the output of *TargetActorNet*. Gradient descent is performed to update *CriticNet* as:

$$\nabla_{\phi}\frac{1}{\mathcal{B}}\sum_{\mathcal{B}}(Q_{\phi}(o,a) - y(r,o',d))^2, \tag{3.13}$$

where $\mathcal{B}$ is a mini-batch of samples drawn from memory $\mathcal{D}$ without replacement. The mini-batch size is 64, which is selected through hyper-parameter turning. The actor network is updated using gradient ascent as:

$$\nabla_{\theta}\frac{1}{\mathcal{B}}\sum_{\mathcal{B}}Q_{\phi}(o,\mu_{\theta}(o)), \tag{3.14}$$

where $\mu_{\theta}(.)$ represents the output of $ActorNet$. *TargetActorNet*'s and *TargetCriticNet*'s weights are updated at each step in a soft update manner as:

$$\phi^{\text{targ}} \leftarrow \alpha\phi^{\text{targ}} + (1-\alpha)\phi \tag{3.15}$$

$$\theta^{\text{targ}} \leftarrow \alpha\theta^{\text{targ}} + (1-\alpha)\theta. \tag{3.16}$$

To facilitate the exploration of different actions in a continuous action space, we add uncorrelated Gaussian noise to the continuous parts of the actions as:

$$\psi_t^{s_i} \leftarrow \psi_t^{s_i} + \psi_n \sim \mathcal{N}(\mu_{\text{noise}}, \sigma_{\text{noise}}^2) \tag{3.17}$$

$$p_t^{s_i} \leftarrow p_t^{s_i} + p_n \sim \mathcal{N}(\mu_{\text{noise}}, \sigma_{\text{noise}}^2). \tag{3.18}$$

The model is labeled as the current best model in the beginning of training after at least $\mathcal{W}_n$ episodes, and is updated periodically as the training progresses whenever the average throughput and energy efficiency of the model over $\mathcal{W}_n$ recent episodes exceeds the current best performance. This approach prevents unnecessary storage. The complexity of a training pass of IMPACT is $O(\mathcal{B}.(2f_{\theta} + 2f_{\phi}))$, where f_{θ} and f_{ϕ} represent the number of neurons in $ActorNet$ and $TargetNet$, respectively. Complexity of decision making by a trained network is $O(f_{\theta})$.

3.4.2 Algorithm 2

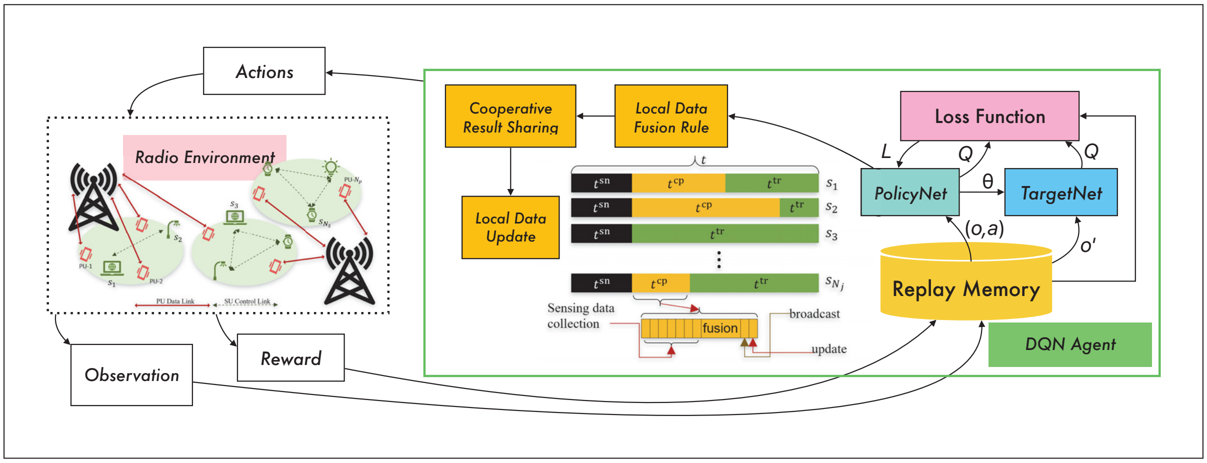

Figure 3.5 Flowchart of the proposed discretized algorithm

Algorithm 3.2 Discretized intelligent multi-user participatory algorithm for cooperative transmission

Input: $\delta_\psi, \delta_p, \alpha, \gamma, N_{ep}, \mathcal{F}$;
Create $\mathcal{A}_d$ from $\mathcal{A}$ such that $f : \mathcal{A} \to \mathcal{A}_d$ as: $f_\psi : \mathbb{R}_\psi \to \mathbb{N}^{\delta_\psi}$, and $f_p : \mathbb{R}_p \to \mathbb{N}^{\delta_p}$;
Initialize replay memory $\mathcal{D}$;
Initialize $PolicyNet$ with random weights θ ;
Initialize $TargetNet$ with weights $\theta^{\text{targ}} = \theta$;
for $episode = 1$ *to* N_{ep} **do**
  Reset Environment;
  for $t = 1$ *to* T **do**
    Choose a_t with an epsilon-greedy policy as $a_t = argmaxPolicyNet(o_t, a_t; \theta)$;
    Cooperative agents transmit their inferences, passive listeners get the inferred results from other;
    Perform local data fusion, decide on ζ, p ;
    Execute action a_t, obtain reward r_t and the next observation o' ;
    Store tuple (o_t, a_t, r_t, o', d) in replay memory $\mathcal{D}$;
    Sample mini-batch of transitions (o, a, r, o', d) from memory $\mathcal{D}$;
    Set $y_j = r_j + \gamma \max_{a'} \hat{Q}(o_{j+1}, a'; \theta^{\text{targ}})$;
    Update the $PolicyNet$ by minimizing the loss according to equations 3.20 and 3.21;
  end for
  Every $\mathcal{F}$ perform a hard update as $\theta^{\text{targ}} \leftarrow \theta$
end for

IMPACT is a powerful algorithm that can deal with hybrid action spaces. To analyze its performance in comparison to that of a closely matched discretized algorithm we propose IMPACT-D, which uses the discretized action space $\mathcal{A}_d$ instead of $\mathcal{A}$ and tries to find the optimal actions while utilizing half as many networks as IMPACT. IMPACT-D uses the DQN architecture and is shown as a flowchart in Figure 3.5. It utilizes two identical neural networks, called $PolicyNet$ and $TargetNet$, to approximate the Q-values of the SU's actions. $PolicyNet$ serves to estimate the Q-value $Q(o_t, a_t; \theta)$ for all (o_t, a_t) pairs, where θ denotes $PolicyNet$'s parameters. $TargetNet$ is identical to $PolicyNet$ except for the fact that it is updated only every $\mathcal{F}$ steps. Using $PolicyNet$ and $TargetNet$ prevents network oscillations and instability so that the network is not chasing a constantly moving target. Unlike IMPACT, IMPACT-D cannot work with continuous or hybrid action spaces and does not strive to learn a deterministic policy. Instead, its [IMPACT-D's] only objective is to find the (discrete) actions that maximize the estimated action values.

At each time step t each SU observes o_t and chooses its action a_t using an epsilon-greedy policy. The value of epsilon decreases exponentially as training progresses. After the SUs interact with the environment, they receive a reward r_t and the tuple (o_t, a_t, r_t, o', d) is stored in memory $\mathcal{D}$. Each episode ends in a terminal state, called terminal o_t, after a fixed number of steps. After the SUs reach the terminal state, the environment is reset and a new episode begins. A mini-batch is sampled from the memory in each episode, and the Q-values are calculated as:

$$y_j = \begin{cases} r_j & \text{for terminal } o_t \\ r_j + \gamma \max_{a'} \hat{Q}(o', a'; \theta) & \text{for non-terminal } o_t \end{cases} \tag{3.19}$$

Afterward, $PolicyNet$ is updated using the following loss function:

$$L(x, y) = \frac{1}{n} \sum_i z_i, \tag{3.20}$$

$$z_i = \begin{cases} 0.5(x_i - y_i)^2 & \text{if}|x_i - y_i| < 1 \\ |x_i - y_i| - 0.5 & \text{otherwise,} \end{cases} \tag{3.21}$$

where (3.21) defines *SmoothL1Loss*, which has been proven to be less sensitive to outliers than *MSELoss*.

Finally, gradient descent is used to minimize the loss function and update $PolicyNet$'s weights, and training continues until a certain level of accuracy is achieved or the maximum number of episodes has passed.

The SUs explore the discrete action space using an epsilon-greedy policy during training. They also periodically store the trained model and update it whenever the model performance over the recent $\mathcal{W}_n$ episodes exceeds the previous best performance. IMPACT-D's main steps are highlighted in Algorithm 3.2. IMPACT-D's training and testing complexity is $O(\mathcal{B}.(f_\theta + f_\phi))$ and $O(f_\theta)$, respectively.

3.5 Performance evaluation

In this section, we present simulation results and compare the two proposed algorithms' performance with that of random agents (RAs), greedy bandit agents, and exhaustive search agents (ESAs). The greedy bandit algorithm is an iterative algorithm that solves multi-agent CRNs as multi-armed bandit problems in which the agents always select the arm with the highest estimated reward at each step. Each SU's maximum sensing capacity is assumed to be 10, which is a realistic assumption for most low-powered IoT devices. We evaluate the algorithms with $N_p = [2, 3, 5, 7, 10]$ and $N_s = [2, 3, 5, 7, 10]$. The simulations were carried out for $1,000$ episodes, with $T = 20$ steps per episode. We use $E_{sn} = 0.1$ mJ – 5 mJ and $E_{cp} = 0.05$ mJ – 1.5 mJ for different SUs based on (Guo & Zhao, 2022; Lauridsen *et al.*, 2018; Davies, 2021).

Figure 3.6 shows the training performance of the various agents considered with different types of CSI. IMPACT outperforms all the other agents with perfect CSI and achieves a near-optimal

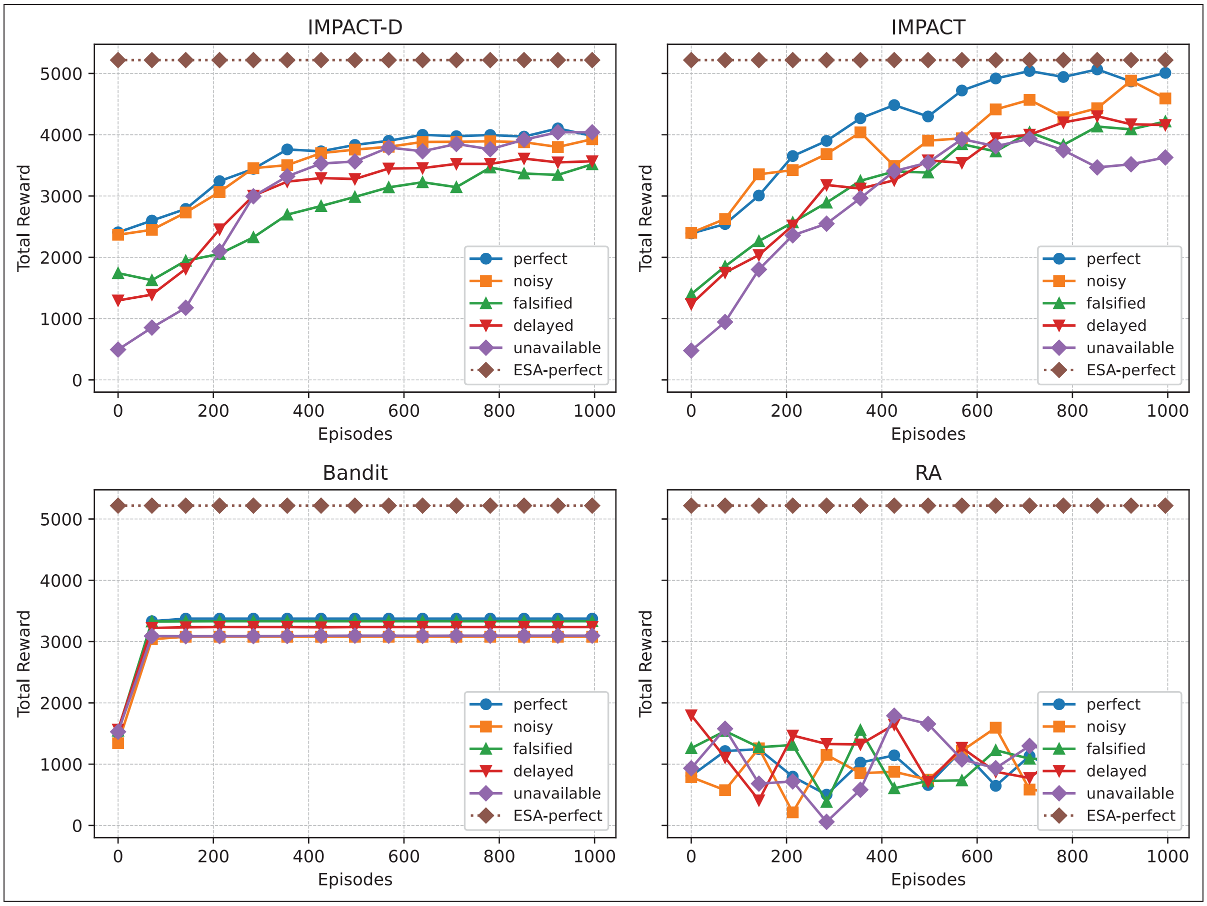

Figure 3.6 Training performance of different agents under varying channel conditions

reward when compared with that of the ESAs, as shown. IMPACT-D doesn't perform optimally due to the discretization of the action space limiting the actions the agents can take. A comparison with different values of δ_ψ and δ_p is not provided for the sake of brevity, but IMPACT-D's performance degrades as $\mathcal{A}_d$ becomes smaller. Both IMPACT and IMPACT-D learn to adapt their actions with noisy CSI and closely match the performance achieved under perfect CSI, although IMPACT is more affected by the ICSI than IMPACT-D is. IMPACT using more neural networks and its continuous parts being more sensitive to noisy CSI. All the agents' training performance is degraded more under delayed and falsified CSI conditions, with falsified having a greater impact. This is expected, and the impact of said data falsification on cooperative agents and the proposed solutions can be found in our previous work (Khaf *et al.*, 2022). When CSI is unavailable, IMPACT-D seems to outperform IMPACT in training at least. We suspect this is also

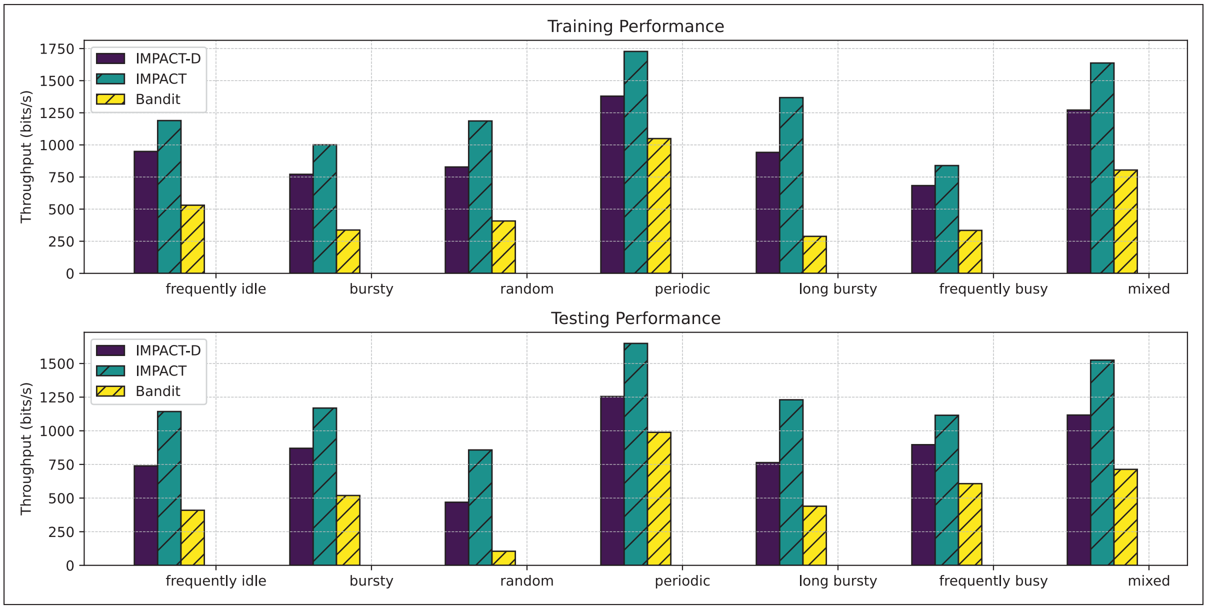

Figure 3.7 Average throughput under different traffic patterns explored for training and testing

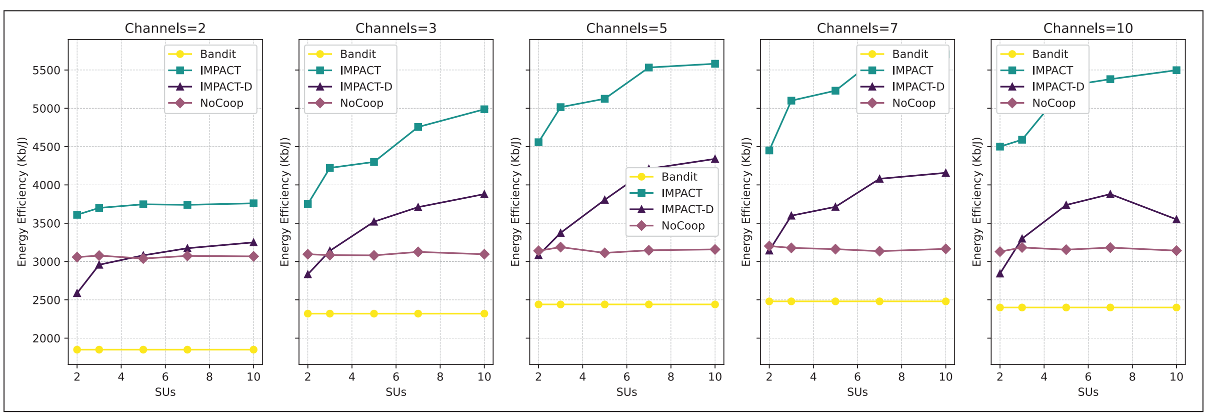

Figure 3.8 Average energy efficiency for different number of secondary users and channels

due to IMPACT-D having fewer neural networks involved and its discrete-only decision-making being less complex, but further research that is beyond the scope of this paper is needed to draw a more definitive conclusion. The greedy bandit agents' performance is severely limited due to the size of the action space and the fact that the bandits easily get stuck in the local optimums. As expected, the RAs perform the worst and are not affected by any variation in CSI.

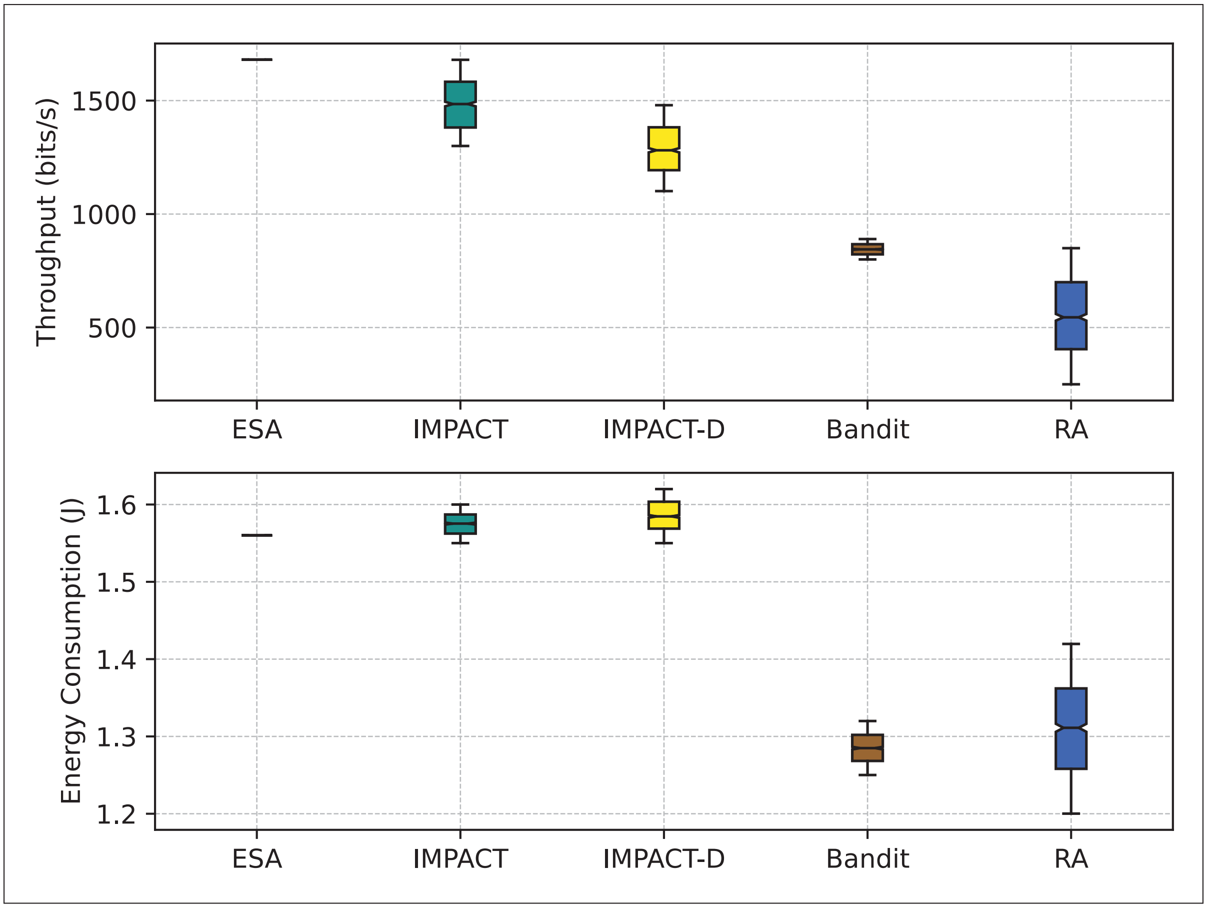

Figure 3.9 Average throughput and energy consumption of different algorithms

Figure 3.7 shows the training and testing throughput of various agents under different traffic patterns explored in this work. Agents that are trained on a periodic traffic pattern outperform the other agents during training but generalize relatively poorly when they are tested on all types of traffic patterns, whereas the agents that are trained in a hybrid environment comprising multiple traffic patterns generalize far better than the other agents. As expected, random traffic patterns do not provide any meaningful information for agents to learn. The simpler greedy bandit agents perform relatively well in a periodic traffic pattern but lose to others by a significant margin for all the other traffic types.

Figure 3.8 shows the energy efficiency of SUs as the number of SUs per coalition varies from 2 to 10. All the SUs in a coalition sense and transmit over the same subset of channels. The

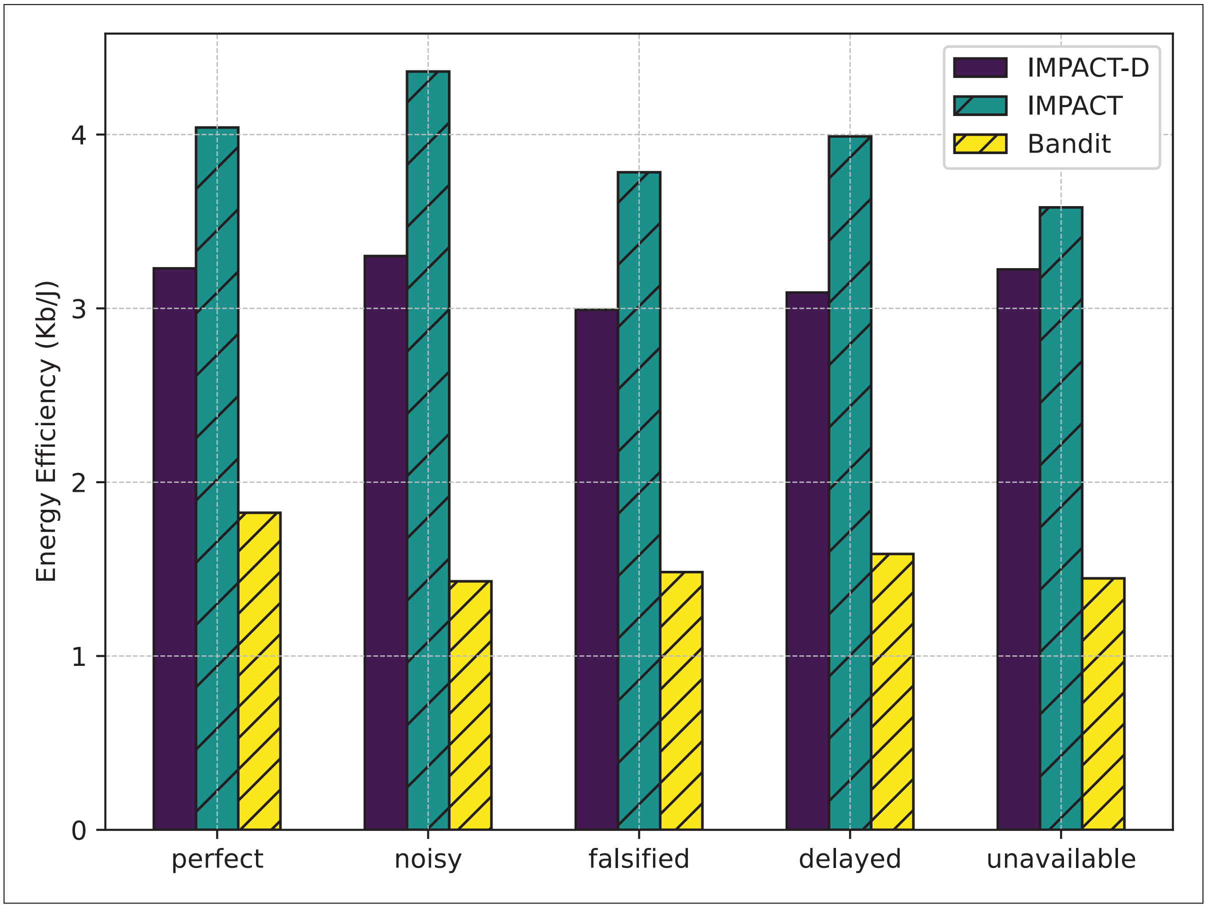

Figure 3.10 Energy efficiency of different agents under varying channel conditions

number of channels is set at $N_p = [2, 3, 5, 7, 10]$ from one subplot to the next going from left to right. When $N_s = 2$ there is no significant difference in average agent's energy efficiency as the number of SUs participating in sensing and cooperation increases. When $N_s = [3, 5, 7]$, both IMPACT and IMPACT-D benefit from having more SUs involved in sensing and cooperation, and receive overall higher efficiency. IMPACT-D starts running into scalability challenges at $N_s = 10$, where effectively exploring the action space becomes a challenge. IMPACT, on the other hand, remains scalable and continues to benefit from having more SUs participating in sensing and cooperation as the number of channels grows. The non-cooperative algorithm (NoCoop) has a considerably lower energy efficiency due to frequent sensing, SFA and SDF attacks, and demonstrates the benefits of cooperation.

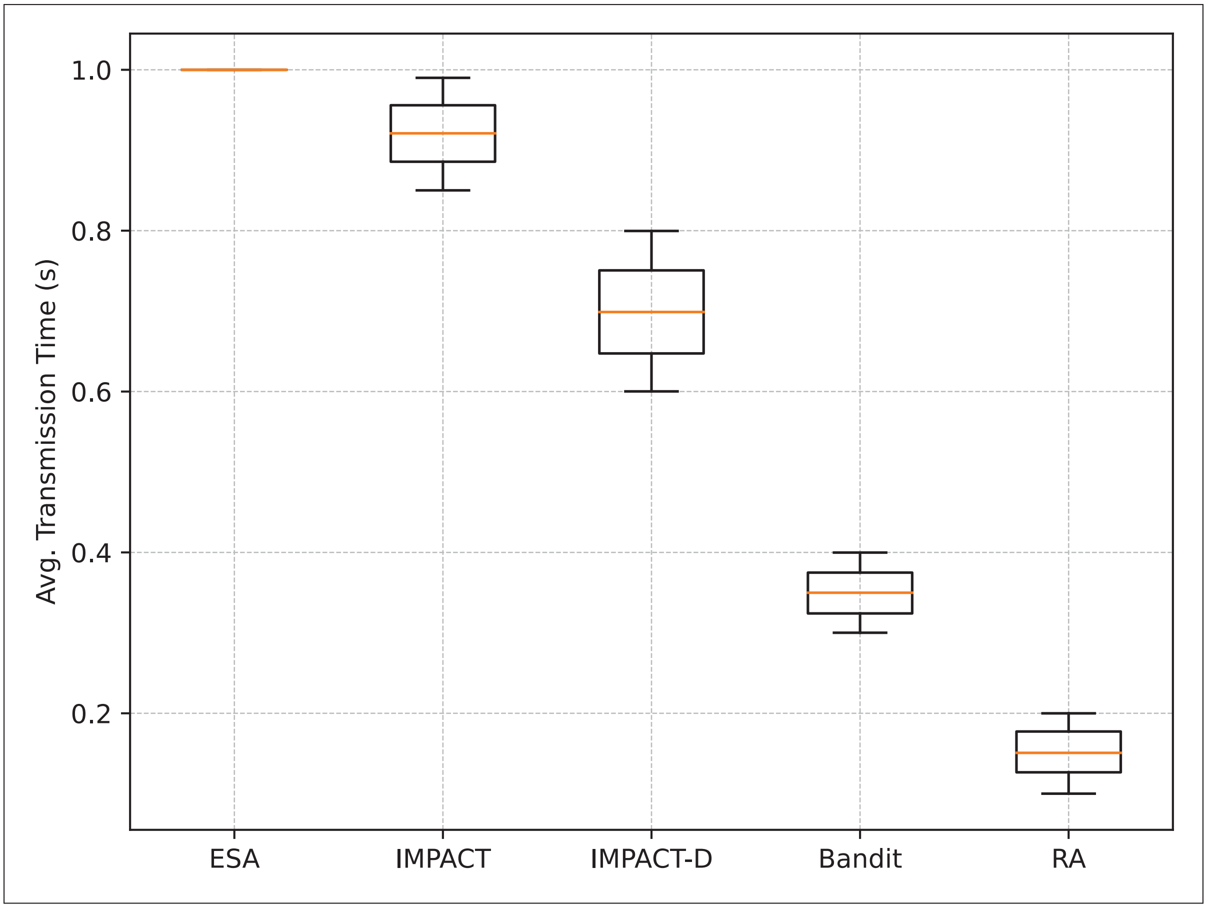

Figure 3.11 Average transmission time for different agents

Figure 3.9 shows the distribution of the average throughput and energy consumption of the approaches considered for $N_p = N_s = 10$. IMPACT achieves near-optimal throughput while keeping its energy consumption low. IMPACT-D achieves less throughput and consumes slightly less energy than IMPACT. IMPACT's energy consumption is expected to be lower in smaller environments where N_p and N_s are relatively small. The greedy bandit agents did not consume a lot of energy or achieve a high throughput and were stuck in local optimal solutions, and the RAs perform the worst of all. For the ESAs, it is important to note that the amount of energy consumed when searching through all the actions is not taken into account here; the plot shows only the energy consumed for the optimal actions. It can be observed from this figure that IMPACT achieves near-optimal performance on both fronts.

Figure 3.10 generalizes the energy efficiency of the various agents considered under different CSI conditions. The agents that are trained under noisy CSI conditions outperform all the other agents when tested in all types of CSI conditions. The CSI conditions have a greater impact on IMPACT than IMPACT-D, whereas the greedy bandit agents generalize poorly due to the simplistic nature of their algorithm. IMPACT's heightened sensitivity to training conditions is due to its underlying policy gradient methods' inherent sensitivity to environmental condition, and serves as an advantage when it is trained under noisy CSI conditions. As expected, training under falsified CSI conditions results in worse generalization than any other type of CSI conditions due to the highly irregular and non-deterministic nature of this falsification.

Figure 3.11 shows different agents' distributions of average transmission time. It is important to note that IMPACT and IMPACT-D cannot and should not use the full time slot for transmission because they are always relying on cooperative mechanisms to achieve high throughput without always having to sense. When that is taken into account, both IMPACT and IMPACT-D manage to utilize most of each time slot for transmission and use only a small portion of time overall for sensing and cooperation.

3.6 Conclusion

In this paper, we proposed two cooperative spectrum sensing and transmission algorithms, IMPACT and IMPACT-D, that maximize the energy efficiency of CRNs. The former offers the ability to incorporate hybrid action spaces comprising discrete and continuous actions, achieves higher performance in challenging traffic conditions, and generalizes better overall under different CSI conditions, while the latter offers a low-complexity alternative for relatively simpler environments. Both algorithms outperform both the greedy bandit agents and the random agents, and approach the performance upper bound of the exhaustive search agents. The fact that these algorithms can achieve the same throughput and energy efficiency as exhaustive search is a testament to their performance, since the computational complexity of exhaustive search makes it impractical for implementation on low-powered IoT devices. These hybrid and scalable algorithms are monumental for the rapid growth of distributed heterogeneous IoT networks.

Future works may consider more granular optimization of the cooperation process by optimizing participation based on IoT device characteristics. Other possible directions include networks comprising both terrestrial and non-terrestrial components, on-the-fly network reconfiguration, along with exploring more algorithms e.g., proximal policy optimization (PPO), advantage actor critic (A2C), trust region policy optimization (TRPO) , and so forth.

CHAPTER 4

FEDERATED HIERARCHICAL REINFORCEMENT LEARNING FOR RESILIENT SPECTRUM SHARING IN 6G NON-TERRESTRIAL NETWORKS

Sadia Maverick Khaf[1] , Georges Kaddoum[1]

[1]Department of Electrical Engineering, École de technologie supérieure, Montreal, QC H3C 1K3, Canada

Published in the IEEE Open Journal of the Communications Society, May 2026.

Abstract

High-altitude platform stations (HAPS) offer a pragmatic backbone for extending 6G non-terrestrial networks (NTNs) into dense urban cores, sparsely connected remote regions, and disaster-stricken areas. In these environments, cognitive radio networks (CRNs) must arbitrate spectrum access without trusting all agents or relying on intact terrestrial backhaul. Accordingly, in the present study, we propose a federated hierarchical reinforcement learning (F-HRL) framework where secondary users (SUs) cooperatively access primary users' (PUs) spectrum, ground next-generation Node Bs (gNBs) and uncrewed aerial vehicles (UAVs) enforce trust-aware scheduling, while HAPS coordinators perform Byzantine-resilient aggregation using satellite backhaul. Area-specific reward shaping allows the same policy stack to prioritize spectral efficiency in urban deployments, coverage in remote regions, and rapid recovery after infrastructure loss. We benchmark F-HRL against non-cognitive baselines, simple heuristics, and the hierarchical deep reinforcement learning framework across the following three key scenarios over 5 000 decision steps: nominal spectrum sharing, disaster recovery with a gNB outage, and adversarial environments with malicious SUs. Experimental results demonstrate that F-HRL achieves a nearly 8-fold higher spectrum efficiency than the non-cognitive operation, maintains 86% SU quality-of-service (QoS) satisfaction with a Jain fairness index of 0.950, and sustains over 95% of nominal throughput during infrastructure failures. Under attack conditions with 25% compromised agents, trust-weighted aggregation successfully isolates malicious updates

while preserving legitimate throughput. These findings demonstrate that HAPS-centric F-HRL provides adaptive, secure, and resilient spectrum management for future NTN deployments.

4.1 Introduction

The evolution towards 6G wireless networks requires unprecedented capabilities such as ultra-low latency on the order of microseconds (Lotfi *et al.*, 2021), ultra-high-speed and massive connectivity, as well as major gains in energy efficiency as compared to 5G (Tehrani *et al.*, 2021). To achieve these goals, it is essential to ensure seamless integration between terrestrial and non-terrestrial segments, forming multi-layered non-terrestrial networks (NTNs) that enhance service continuity, resilience, and resource utilization (Umer *et al.*, 2025; Das *et al.*, 2024).

Key performance improvements are measured using key performance indicators (KPIs), while broader societal benefits are quantified through key value indicators (KVIs). Meeting these metrics requires addressing the following three interdependent challenges: (i) sustaining licensed services in dense urban cells; (ii) extending coverage to remote or under-served regions; and (iii) restoring connectivity when terrestrial infrastructure is compromised (Manchón *et al.*, 2025).

However, these requirements are difficult to meet with conventional cognitive radio approaches that assume always-on terrestrial backhaul and complete channel state information. When next-generation Node Bs (gNBs) are disabled, sensing reports are withheld and malicious secondary users (SUs) can falsify observations, which may lead to centralized coordination collapses (Getu *et al.*, 2019b,a; Quach *et al.*, 2019). High-altitude platform stations (HAPS) operating in the lower stratosphere offer a compelling backbone for resilient cognition (Umer *et al.*, 2025; Abbasi *et al.*, 2025; Qureshi *et al.*, 2023). A single HAPS can supervise multiple clusters of gNBs, uncrewed aerial vehicles (UAVs), and SUs while coordinating with neighbouring clusters through satellite backhaul (Khaf *et al.*, 2024). However, disaster-struck urban cores, sparse remote cells, and mixed civilian–public-safety traffic inflict significant heterogeneity, thus requiring adaptive, privacy-preserving decision making beyond static policies.

Furthermore, deterministic coordination becomes brittle under infrastructure loss, privacy constraints, or adversarial interference, with centralized cognitive radio networks (CRNs) struggling to maintain reliable service when base stations fail or when sensing is falsified (Ning *et al.*, 2020; Silvirianti & Kaddoum, 2025). This suggests that first responders and mission-critical internet-of-things (IoT) installations require autonomous, decentralized, and learning-driven spectrum management where both decision making and coordination dynamically evolve under uncertainty (Ngomane *et al.*, 2018; Khaf *et al.*, 2022).

Adaptive resource and spectrum management in dynamic NTNs increasingly relies on machine learning, especially the combination of federated learning (FL) and reinforcement learning (RL) (Umer *et al.*, 2025; Das *et al.*, 2024; Ahmadzadeh *et al.*, 2024; Li *et al.*, 2022). Recent hierarchical multi-agent federated DRL was proposed for TN-NTN resource allocation and task offloading with auction-based incentives (Seid *et al.*, 2026), while unified distributed ML architectures (FSTL/GFSTL) target split/federated/transfer learning pipelines for 6G intelligent transportation data, rather than control policies (Naseh *et al.*, 2025). Furthermore, GAIL-powered federated inverse RL was previously applied to multi-satellite beamforming and spectrum allocation using a matching-game formulation with expert demonstrations (Hassan *et al.*, 2024). Accordingly, in this paper, we propose a federated hierarchical reinforcement learning (F-HRL) framework designed to address the non-stationary, heterogeneous, and security-critical nature of 6G NTNs. This framework extends capabilities of traditional FL–RL models by integrating multi-tier coordination, trust-aware learning, and transfer-based adaptation.

Critically, F-HRL differs from existing hierarchical DRL and federated RL approaches in the following three respects: (i) it decomposes spectrum management across four distinct tiers (SU, gNB, UAV, and HAPS) with tier-specific RL formulations and area-specific reward shaping, rather than a flat or two-tier hierarchy; (ii) it integrates sensing-layer majority-vote fusion with federated divergence-based trust weighting, whereas prior FL-based methods assumed honest participants or treat trust in isolation; and (iii) it incorporates context-aware transfer learning for zero-shot UAV deployment during disasters, which is lacking in the previously proposed federated or hierarchical RL frameworks.

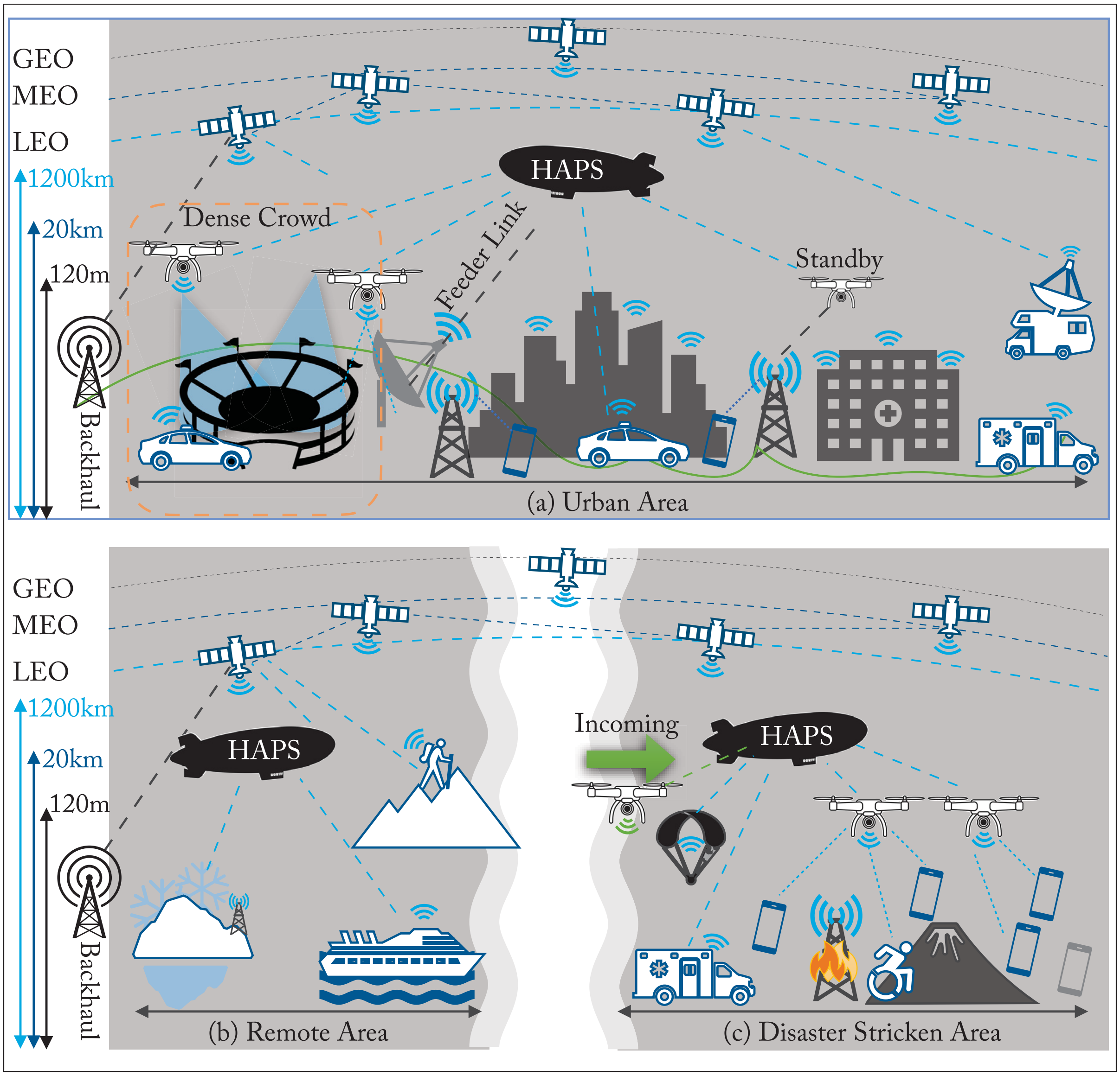

Figure 4.1 The proposed architecture in three representative scenarios: (a) urban, (b) remote, and (c) disaster-stricken areas

To overcome these limitations, in this study, we propose an F-HRL system that combines multi-agent reinforcement learning (MARL) for distributed policy learning, federated learning (FL) for privacy-preserving coordination, and transfer learning (TL) for rapid adaptation in new or emergency deployments. The proposed system is shown in Fig. 4.1. Section 4.4 details how each tier executes tailored learning policies coordinated through low-Earth orbit (LEO) satellite backhaul.

The primary contributions of this work are as follows:

- **Cross-tier cognitive spectrum management**: We developed a four-tier cognitive F-HRL framework spanning SUs, gNBs, UAVs, and HAPS that adapts to heterogeneous quality-of-service (QoS) objectives and constraints. SUs learn opportunistic access while gNBs enforce primary-user (PU) priority, unlike prior hierarchical approaches that treated PUs and SUs equivalently.
- **Federated coordination via HAPS aggregation**: We integrated periodic model aggregation at the HAPS tier to stabilize multi-agent learning across regions, addressing non-stationarity in MARL.
- **Disaster-resilient UAV failover**: We developed a disaster recovery framework that leveraged context-aware similarity-based selective model sharing for UAV deployment upon gNB outage to support emergency relief scenarios.
- **Trust-aware adversarial detection**: We introduced a Byzantine-robust aggregation and trust-weighted cooperative sensing that combines majority-vote sensing consistency with federated weight-divergence filtering to withstand falsified reports, model poisoning, and malicious agents attacks without sharing raw data.

Collectively, these contributions address persistent gaps where existing FL- and RL-based solutions confine learning to a single tier, ignore disaster-induced outages, or lack defenses against adversaries. By contrast, the proposed approach enables rapid deployment in emergency scenarios while maintaining security, privacy, and performance guarantees essential for critical communication infrastructure.

The remainder of this paper is organized as follows. Section 4.2 reviews related research in CRNs, FL, and HAPS-assisted communications. In Section 4.3, we present the system model, network architecture, and problem formulation. Section 4.4 details the proposed F-HRL framework, including security and TL mechanisms. Section 4.5 discusses experimental validation and performance analysis. Finally, Section 4.6 concludes the paper with an outline of future research directions.

4.2 Related work

To motivate F-HRL, we group previous research efforts into three key themes related to the critical gaps addressed by our unified architecture.

4.2.1 Multi-tier hierarchical coordination

Recent studies demonstrated the benefits of combining FL and DRL for adaptive spectrum access, resource management, and joint optimization across complex wireless architectures. Integrating massive antenna arrays (e.g., in HAPS) and managing large, heterogeneous networks frequently leads to a "curse of dimensionality" for optimization algorithms (Khoshkbari *et al.*, 2025; Abbasi *et al.*, 2025). HRL approaches tackle this complexity by introducing temporal abstraction via options and macros, modeling large problems as Semi-Markov Decision Processes (SMDPs). For resource scheduling in HAPS-integrated networks, HDRL reduces complexity by delegating tasks, enabling fast execution (e.g., 50× faster than exhaustive search), while maintaining up to 95% of the spectral efficiency achieved by exhaustive search. HDRL was also used for resource scheduling, UE association, and channel assignment in HAPS systems.

Previous architectural surveys and design studies argued that NTN deployments must explicitly balance cross-tier coordination, communication overhead, and energy objectives (Naous *et al.*, 2023; Farajzadeh *et al.*, 2025). Recent FL frameworks for satellite and LEO scenarios demonstrated viability of federated coordination under constrained backhaul (Qureshi *et al.*, 2023) and energy-aware client participation (Hlophe *et al.*, 2021). In terrestrial IoT settings, reinforcement learning was paired with FL to improve routing fairness and convergence (Malik *et al.*, 2023), thus reinforcing the need for multi-agent formulations embraced in the present study.

Furthermore, techniques such as Federated Deep Deterministic Policy Gradient (F-DDPG) were previously proposed for satellite Reconfigurable Intelligent Surface (RIS)-enhanced Integrated Sensing and Communication (ISAC) systems (Pala *et al.*, 2025), yielding significant gains (e.g., up to 76.8% improvement in spectral efficiency) as compared to non-federated baselines.

Federated DRL algorithms were reported to perform comparably to centralized learning while achieving faster execution times (Ahmadzadeh *et al.*, 2024; Xu *et al.*, 2020).

While cognitive radio techniques have been extensively studied under practical imperfections (Sharma *et al.*, 2015) and hardware constraints (Hamza *et al.*, 2019), applying DRL approaches such as Deep Q-Networks (DQNs) to HAPS systems for dynamic resource allocation remains limited by the "curse of dimensionality," arising from large coverage areas and massive antenna arrays. To mitigate this challenge, complexity-reduction strategies such as HRL and HDRL were explored (Umer *et al.*, 2025; Lee, 2021), decomposing large action spaces into smaller sub-tasks, or *options* (Khoshkbari *et al.*, 2025; Li *et al.*, 2022).

Recent HDRL frameworks employing PPO-based policies across multiple tiers have demonstrated effective hierarchical coordination for joint UE association, channel assignment, and resource scheduling in HAPS-integrated networks (Umer *et al.*, 2025). Similarly, hierarchical multi-agent federated DRL was previously proposed for TN-NTN resource allocation and task offloading with auction-based incentives (Seid *et al.*, 2026), while unified distributed ML architectures (FSTL/GFSTL) focus on split/federated/transfer learning pipelines for 6G intelligent transportation data rather than control policies (Naseh *et al.*, 2025). Despite these advances, existing HDRL approaches do not explicitly address adversarial robustness, distributed trust, or coordinated disaster recovery across heterogeneous NTN layers.

It is worth contrasting F-HRL with two additional categories of approaches. First, multi-agent PPO (MA-PPO) methods employ independent PPO agents with shared centralized critics, but lack federated aggregation for cross-region knowledge transfer and trust-weighted Byzantine resilience. Second, optimization-based methods such as the Whale Optimization Algorithm (WOA) can effectively solve per-step allocation problems, but are inherently unable to adapt to non-stationary PU traffic or infrastructure failures without re-optimization. Both categories are evaluated as baselines in Section 4.5.

4.2.2 Resilience and rapid deployment

Mission-critical applications require the network to maintain connectivity despite mobility, extreme congestion, or infrastructure loss. DRL and FL are critical for managing dynamic and mobile environments, particularly in multi-tiered NTNs involving UAVs and ground stations.

Providing flexible coverage extension and serving as temporary gNBs, UAVs are essential in 5G. Due to the high mobility of UAVs, fast handover management is crucial. Dynamic Spectrum Sharing (DSS) based on FL and DRL was previously proposed for emergency CRNs to deal with spectrum shortages after disasters (Yang *et al.*, 2023). Federated UAV coordination was also explored for dense IoT deployments, combining hypergraph models with FRL to stabilize throughput (Yang *et al.*, 2024), while over-the-air transfer learning across UAV swarms was reported to improve modulation recognition yet assumes trusted agents (Zhou *et al.*, 2024).

While DRL/FL is applied to emergency scenarios, previous studies generally assumed static infrastructure or reliable ad-hoc relaying, thus failing to investigate the challenge of sequential infrastructure failure such as multiple gNB outages—and the subsequent need for rapid system reconfiguration. Furthermore, implementing DL algorithms in real-time CRNs faces the challenge that most modern DL systems are trained offline, which results in the lack of verified generalization ability in complicated and dynamic physical channels.

Furthermore, previous HDRL frameworks addressed cross-tier coordination by imposing constraints–i.e., higher tiers make decisions that constrain local tiers, ensuring coordinated spectrum use and managing dynamic movement (Hu *et al.*, 2022).

Current federated and hierarchical frameworks do not integrate explicit mechanisms for rapid redeployment and knowledge transfer, i.e., transfer learning, across non-stationary topologies. The lack of reliable terrestrial infrastructure in disaster zones prevents the use of traditional centralized retraining. F-HRL addresses this concern by using pre-trained base models across representative conditions (urban, remote, disaster) and using context-similarity scoring to rapidly fine-tune the model upon deployment or failure detection. This mechanism allows

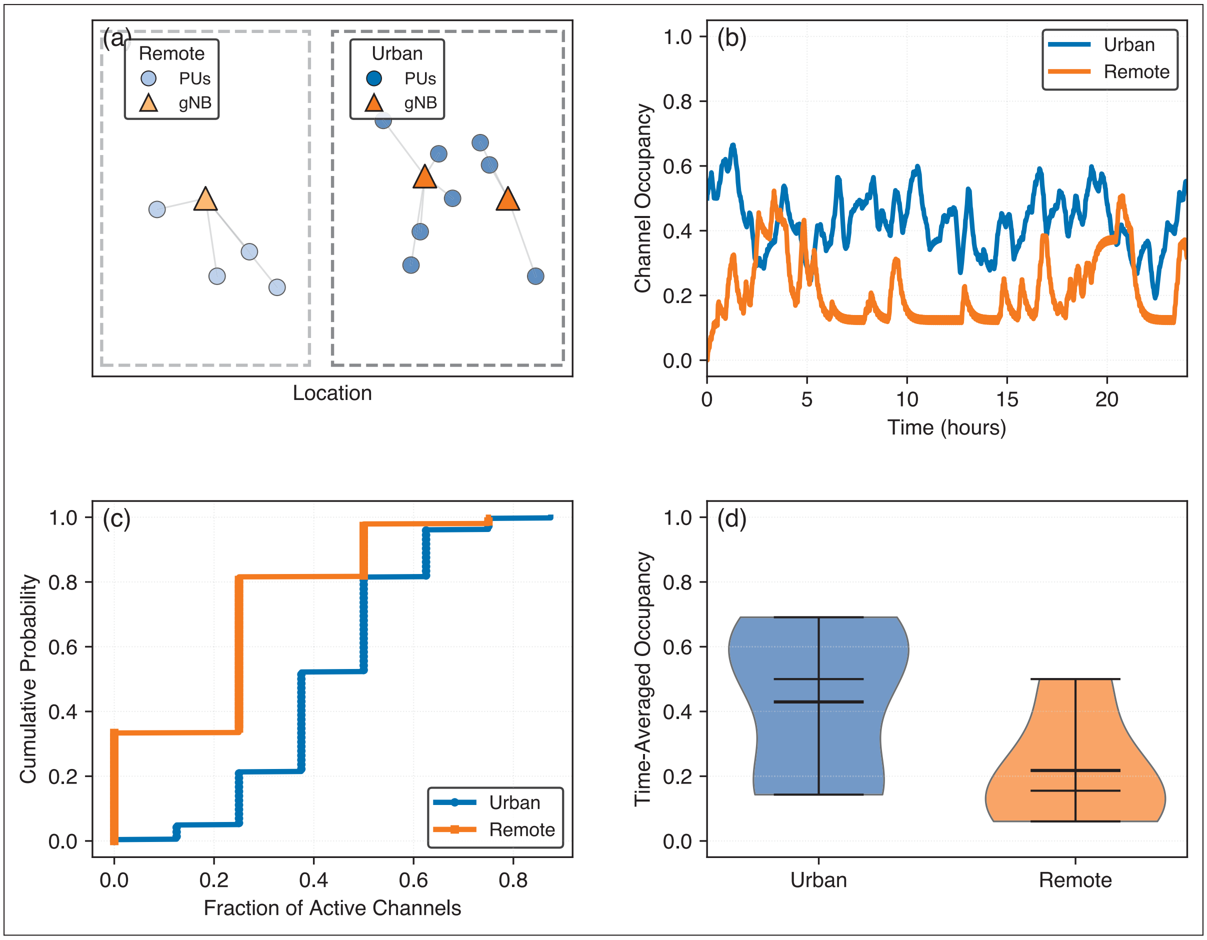

Figure 4.2 Primary user traffic across urban and remote areas: (a) topology, (b) diurnal occupancy, (c) distribution of active-channel, and (d) time-averaged occupancy

UAV-assisted coordination to quickly restore over 85% of pre-failure throughput after sequential gNB outages (Yang *et al.*, 2023; Le & Kaddoum, 2021).

4.2.3 Trust, robustness, and security in federated coordination

Distributed learning in critical infrastructure requires resilience against adversarial behavior–a concern magnified in CRNs where agents opportunistically access shared spectrum. Unless tensor-based detectors and trust scoring are incorporated, cooperative sensing remains vulnerable to falsified observations and hidden terminal effects (Getu *et al.*, 2019b,a). While vertical FL reduces data leakage by partitioning features across participants (Zhang *et al.*, 2020b) and deep recurrent methods handle partial observability in heterogeneous networks (Khoshkbari & Kaddoum, 2023), these approaches typically assume honest participants.

Although recent efforts deployed FRL to mitigate jamming and noisy gradients (Sharma *et al.*, 2023), larger-scale surveys stressed that existing countermeasures treat isolated threat vectors (Das *et al.*, 2024). Byzantine-resilient aggregation and communication-efficient FRL were proposed to tolerate poisoned updates and data heterogeneity (Zuo *et al.*, 2025; Krouka *et al.*, 2022), while NTN security studies advocated for coupling spectrum control with anomaly detection (Alharbi *et al.*, 2025). These works rarely integrated sensing trust with hierarchical recovery, thus reinforcing the need for the unified trust-weighted aggregation that underpins our F-HRL stack.

Based on a literature review conducted through the joint lens of coordination depth, adaptation strategy, and resilience features, in Table 4.1, we compare representative studies across these facets. The catalogue spans contributions published in 2020–2025 covering satellite RIS, UAV swarms, vehicular CRNs, and HAPS-assisted deployments. This perspective highlights that most prior efforts optimized a single tier or a single threat model, thus leaving cross-tier failover and trust governance largely unexplored.

Table 4.1 Representative studies

Study	**Coordination scope**	**Adaptation strategy**	**Resilience / trust handling**
(Pala *et al.*, 2025)	Satellite–RIS link (single tier)	Federated DDPG for RIS beamforming; spectral-efficiency maximization	No explicit failover or security model; assumes benign participants and intact infrastructure.
(Ahmadzadeh *et al.*, 2024)	RIS-aided CRN (single tier)	Over-the-air FL with DRL for robust RIS-aided cognitive radio	Addresses channel distortion during aggregation, but omits adversarial, trust, or disaster considerations.

Continued on next page

Table 4.1 Representative studies (continued)

Study	**Coordination scope**	**Adaptation strategy**	**Resilience / trust handling**
(Yang *et al.*, 2023)	UAV-assisted emergency CRN (dual tier)	Federated multi-agent Q-learning for post-disaster spectrum sharing	Restores traffic after a single outage, but lacks trust modelling and sequential failure analysis.
(Hu *et al.*, 2022)	UAV relay + gNB hierarchy	Hierarchical actor–critic with cache-aware scheduling	Ensures QoS under mobility, yet depends on centralized coordination, without federated robustness.
(Tehrani *et al.*, 2021)	Distributed gNB cluster	Federated DRL for coordinated NextG control loops	Boosts distributed control, yet assumes stable backhaul and trusted participants.
(Ning *et al.*, 2020)	Cooperative CRN sensing	Centralized RL with cooperative feedback loops	Handles sensing uncertainty, but no federation or hierarchical fallback paths.
(Farajzadeh *et al.*, 2025)	NTN architectural survey	Blueprint for federated orchestration across satellite and HAPS tiers	Highlights scalability/energy trade-offs, but leaves disaster adaptation open.
(Naous *et al.*, 2023)	NTN survey (multi-tier)	Categorizes RL enablers for satellite/HAPS integration	Frames open problems, but does not deliver a deployed stack.
(Sharma *et al.*, 2023)	gNB with cooperative SUs	Federated D3QN for anti-jamming power control	Mitigates jamming, but omits disaster recovery and multi-tier orchestration.
(Zhou *et al.*, 2024)	UAV swarm for V2X monitoring	Over-the-air federated transfer learning for modulation recognition across mobile relays	Tackles mobility-induced drift, yet assumes trusted agents and no spectrum failover plan.

Continued on next page

Table 4.1 Representative studies (continued)

Study	Coordination scope	Adaptation strategy	Resilience / trust handling
(Zhang *et al.*, 2020b)	Cooperative CRN sensors (single tier)	Vertical FL coordinating heterogeneous feature sets	Preserves sensing privacy, but leaves Byzantine behaviour and cross-tier recovery unaddressed.
(Chen *et al.*, 2022)	Distributed SU cluster	Lightweight FL-based cooperative sensing with reliability weighting	Improves detection accuracy; resilience is limited to statistical weighting without dynamic trust feedback.
(Khaf *et al.*, 2024)	HAPS-assisted terrestrial/ UAV users	Capability-aware cooperation with partial hierarchy	Enhances coverage after single outages; lacks federated training and sustained multi-tier trust governance.
(Le & Kaddoum, 2021)	Vehicular CRN (multi-agent SUs)	LSTM-based channel access with multi-agent RL	Addresses vehicular mobility, but independent agents, no federated aggregation or coordinated failover.
(Yang *et al.*, 2024)	UAV-enabled IoT (dense)	FRL with hypergraph scheduling for dense UAV clusters	Improves throughput, but assumes reliable backhaul and honest agents.
(Malik *et al.*, 2023)	CR-enabled IoT mesh	RL-based routing for fairness and load balancing	Focuses on terrestrial IoT without NTN layers or trust governance.
(Getu *et al.*, 2019b)	MIMO cognitive sensing	Tensor-based detection to mitigate hidden terminals	Enhances sensing fidelity, yet lacks federated coordination.

Continued on next page

Table 4.1 Representative studies (continued)

Study	**Coordination scope**	**Adaptation strategy**	**Resilience / trust handling**
(Zuo *et al.*, 2025)	FL aggregation layer	Byzantine-robust aggregation for heterogeneous clients	Addresses poisoning, but not spectrum context or multi-tier failover.
(Krouka *et al.*, 2022)	Multi-agent FRL cluster	Communication-efficient consensus via ADMM	Improves communication cost, yet assumes trustworthy sensing inputs.
(Das *et al.*, 2024)	Survey across FRL deployments	Taxonomy of privacy- and security-aware FL primitives	Identifies need for trust-weighted aggregation and cross-layer resilience, but offers no integrated implementation.
(Seid *et al.*, 2026)	TN-NTN multi-tier	FeDRL/MAFDRL with auction-based incentives for resource allocation and task offloading	Targets incentive-driven offloading; no PU/SU spectrum sharing, trust weighting, or disaster recovery.
(Naseh *et al.*, 2025)	TN-NTN ITS hierarchy	FSTL/GFSTL split–federated–transfer learning architecture for ITS data	Learning framework only; no RL spectrum policy, PU/SU coordination, or adversarial resilience.
(Hassan *et al.*, 2024)	6G multi-satellite spectrum	GAIL-based federated inverse RL for beamforming and spectrum allocation with matching-game association	Focuses on satellite beamforming/SE, not CRN PU/SU coordination or disaster recovery.

In Table 4.2, we summarize the principal literature gaps and highlight how the proposed approach addresses each dimension.

Table 4.2 Summary of literature gaps and main contributions

Theme	Representative gaps	F-HRL Contributions
NTN FL (Complexity & Cross-Tier)	Single-agent DRL methods struggle with the exponentially large action spaces and state spaces in massive MIMO HAPS/NTNs. Centralized approaches incur significant computational bottlenecks and communication overhead.	Employs HAPS as a coordinating layer (Tier 4) for cross-tier adaptation (SUs, gNBs, UAVs), thus mitigating the curse of dimensionality via policy decomposition.
FL CRN (Resilience & Topology)	Available FL assumes static infrastructure or relies on constant cooperation. Current frameworks lack specific mechanisms for managing sequential infrastructure failures (disaster recovery) and adapting policies in non-stationary topologies.	Manages multiple sequential gNB outages. Achieves rapid adaptation and self-configuration using context-aware TL to fine-tune pre-trained models for emergency deployment.
RL CRN (Trust & Security)	Available DRL solutions handle threats (jamming or poisoning) in isolation. Secure model sharing remains an open challenge. Trust mechanisms are minimally integrated with sensing capabilities.	Implements robust trust-weighted aggregation and observation-consistency checks to suppress outlier updates, thus ensuring resilience against Byzantine nodes and sensing data falsification (SDF).

Continued on next page

Table 4.2 Summary of literature gaps and main contributions (continued)

Theme	Representative gaps	F-HRL Contributions
Scalability & Overhead	Previous FL-RL methods do not analyze communication overhead or transfer learning search cost in dense NTN deployments. Scalability to hundreds of agents remains unquantified.	Provides quantitative complexity analysis with concrete deployment numbers (Section IV-D), showing quadratic context cost is negligible for $N_h < 1000$ and identifying gradient compression for denser deployments.

In contrast to the above literature, our F-HRL framework unifies cross-tier coordination, trust-aware aggregation, and disaster recovery within a single NTN-compatible architecture.

4.3 System model

We consider a heterogeneous wireless system integrating HAPSs, terrestrial gNBs, UAVs, and clustered user equipments (UEs) (see Fig. 4.1). HAPSs are deployed at altitudes of 20 km–50 km and act as regional spectrum coordinators and FL aggregators. Inter-HAPS coordination relies on low-Earth orbit (LEO) satellite backhaul, while medium Earth orbit and geostationary satellite layers are excluded due to latency exceeding the slot-level decision framework discussed in this paper.

UEs are divided into the following two sets: the set of PUs denoted by $\mathcal{U}_P$ and the set of SUs denoted by $\mathcal{U}_S$. Thus, the total UE set is $\mathcal{U} = \mathcal{U}_P \cup \mathcal{U}_S$ with $\mathcal{U}_P \cap \mathcal{U}_S = \emptyset$. PU activity is exogenous and bursty, generated by a mix of Markov-modulated Poisson process (MMPP) traffic, heavy-tailed Pareto on/off processes, and fixed traces. Fig. 4.2 summarizes PU traffic across the urban and remote areas: (a) the PU–gNB topology, (b) diurnal occupancy with urban maintaining 40–50% utilization and remote 15–25% with evening peaks, (c) a right-shifted CDF indicating consistently higher urban utilization relative to sparser remote activity, and (d) per-PU heterogeneity captured via a violin plot where the thin line denotes the median, the bold line the mean, and the width the density. Urban PUs show a broader occupancy variability, while remote PUs are more concentrated at low-to-moderate usage, highlighting the challenge for SUs to adapt to spatially heterogeneous traffic. PUs consume their allocations without adaptation, while SUs operate as cognitive users that sense occupancy and opportunistically transmit when no PU activity is detected.

4.3.1 Spectrum and channel model

Let $\mathcal{U}_{S,b}$ denote the SUs associated with gNB b, $\mathcal{K}_b$ the channel set at gNB b, $\mathcal{U}_d$ the UEs served by UAV d, $\mathcal{B}_h$ the gNBs coordinated by HAPS h, and $\mathcal{D}_h$ the UAVs managed by HAPS h. The global gNB and UAV sets are $\mathcal{B} = \cup_h \mathcal{B}_h$ and $\mathcal{D} = \cup_h \mathcal{D}_h$, respectively.

The network operates over 6 channels spanning 3GPP bands n77, n78, n79, and n258. Channel resource block capacities range from 50 to 113 RBs depending on bandwidth allocation (20–100 MHz); gNB scheduling capacities range from 106 to 273 RBs. Propagation follows 3GPP TR 38.901/38.811:

- Urban links: UMi LoS pathloss,

$$\begin{aligned} \mathrm{PL}_{\mathrm{UMi}} = {} & 32.4 + 21 \log_{10}(f_c[\mathrm{GHz}]) \\ & + 20 \log_{10}(d_{3\mathrm{D}}[\mathrm{m}]) \end{aligned} \tag{4.1}$$

- Remote links: RMa pathloss,

$$\begin{aligned} \mathrm{PL}_{\mathrm{RMa}} = {} & 32.4 + 20 \log_{10}(f_c[\mathrm{GHz}]) \\ & + 23.3 \log_{10}(d_{3\mathrm{D}}[\mathrm{m}]) \end{aligned} \tag{4.2}$$

Small-scale fading follows Rician distribution for HAPS–UE links with elevation-dependent K-factors, and Rayleigh distribution for terrestrial links under non-line-of-sight conditions. Shadowing is log-normal with standard deviation 4 dB for HAPS–UE links and 8 dB for terrestrial links. UAVs fly at the altitudes of 100 m–300 m and use the same channel model as gNBs.

The channel gain from UE u to its serving node on channel k is

$$g_{u,k} = \mathrm{PL}(d_{3\mathrm{D}})^{-1} \cdot |h_{u,k}|^2 \cdot 10^{-\chi/10}, \tag{4.3}$$

where $h_{u,k}$ captures small-scale fading (Rician or Rayleigh as above) and χ is log-normal shadowing with standard deviation 4 dB for HAPS links and 8 dB for terrestrial links. The SINR at UE u on channel k is then

$$\gamma_{u,k} = \frac{p_u \, g_{u,k}}{N_0 B_k + \sum_{j \neq u} p_j \, g_{j,k}}, \tag{4.4}$$

where N_0 is the noise power spectral density and the sum captures co-channel interference from other active transmitters. The achievable rate follows as $R_{u,k} = B_k \log_2(1 + \gamma_{u,k})$.

Each SU performs energy detection to sense channel occupancy, and collaborative SUs improve their accuracy using majority vote (Khaf *et al.*, 2022).

4.3.1.0.1 Modeling assumptions and limitations

The present model adopts several simplifying assumptions that we explicitly acknowledge below. First, the HAPS is treated as quasi-stationary, since stratospheric station-keeping maintains its position within a 400 m radius, which is negligible relative to its 20 km altitude. Second, while UAV positions update per decision step, intra-step Doppler effects are not modeled; this is justified for low-altitude UAVs at 10–30 m/s in sub-6 GHz bands where the coherence time (in the order of milliseconds) significantly exceeds the slot duration. Third, signaling overhead for handover is not explicitly modeled, but is quantified in terms of communication cost in Section 4.4. Extending the framework to incorporate detailed channel dynamics, including Doppler effects and mobility-induced handover overhead, is left for future work.

4.3.2 Problem formulation

We model the network as a partially observable Markov decision process (POMDP). The global state is as follows:

$$\mathbf{s} \;=\; \left\{ (\bar{\lambda}_b,\; \bar{\gamma}_b,\; I_b)_{b\in\mathcal{B}},\;\; (E_d)_{d\in\mathcal{D}} \right\}, \tag{4.5}$$

where $\bar{\lambda}_b$ is the aggregate spectrum demand at gNB b, $\bar{\gamma}_b$ is the average SINR at gNB b, I_b is the local interference indicator at gNB b, and E_d is the residual energy of UAV d.

Each tier $i \in \{h, b, d, u\}$ (HAPS, gNB, UAV, and SU, respectively) observes only a partial, and thus, imperfect view of $\mathbf{s}$:

$$\mathbf{o}_i^t \;=\; \mathbf{s}_i^t + \boldsymbol{\epsilon}_i^t, \tag{4.6}$$

where $\mathbf{s}_i^t \subset \mathbf{s}^t$ denotes the state components visible to tier i, and $\boldsymbol{\epsilon}_i^t \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$ captures measurement and sensing errors (e.g., imperfect CSI).

We decompose the overall spectrum management problem into four tier-specific subproblems as outlined below.

4.3.2.0.1 (S1) SU opportunistic access and power control

Each SU u chooses transmission channel k_u and power p_u to maximize its achievable rate $R_{u,k_u} = R_{u,k}\big|_{k=k_u}$ as shown below.

$$\max_{k_u, p_u} \; R_{u,k_u} = B_{k_u} \log_2(1 + \gamma_{u,k_u}) \tag{4.7}$$

subject to k_u being vacant, $0 \le p_u \le P^{\max}$, and $\gamma_{u,k_u} \ge \Gamma_{\min}$.

4.3.2.0.2 (S2) gNB local scheduling

Each gNB b schedules SUs on residual capacity after satisfying PU demand:

$$\max_{x_{u,k}} \sum_{u \in \mathcal{U}_{S,b}} \sum_{k \in \mathcal{K}_b} x_{u,k} R_{u,k} \tag{4.8}$$

subject to PU-priority allocation, $\sum_u x_{u,k} \leq 1$, $I_k^{\mathrm{PU}} \leq I_k^{\max}$, and $x_{u,k} \in \{0, 1\}$.

4.3.2.0.3 (S3) UAV placement and coverage

Each UAV d determines position $\mathbf{p}_d$ to maximize coverage of displaced UEs:

$$\max_{\mathbf{p}_d} \sum_{u \in \mathcal{U}_d} w_u \mathbb{I}\{\mathrm{cov}_d(u)\} R_{u,k_u} \tag{4.9}$$

subject to activation $a_d \in \{0, 1\}$, $E_d \geq E_{\min}$, and backhaul constraints $C_d^{\mathrm{BH}} \geq C_{\min}$.

4.3.2.0.4 (S4) HAPS backhaul routing and recovery coordination

Each HAPS routes admitted loads from gNBs and UAVs subject to backhaul capacity:

$$\max_{x_{b,h}, x_{d,h}, a_d} \sum_{b \in \mathcal{B}_h} x_{b,h} \bar{\lambda}_b + \sum_{d \in \mathcal{D}_h} x_{d,h} \Lambda_d \tag{4.10}$$

subject to

$$\sum_{b \in \mathcal{B}_h} x_{b,h} \bar{\lambda}_b + \sum_{d \in \mathcal{D}_h} x_{d,h} \Lambda_d \leq C_h^{\mathrm{BH}}, \tag{4.11}$$

$$a_d \in \{0, 1\}, \quad x_{d,h} \leq a_d, \quad x_{b,h}, x_{d,h} \in \{0, 1\}. \tag{4.12}$$

4.3.2.0.5 Coupling conditions

Cross-tier consistency requires (i) each SU associated with exactly one node; (ii) admitted SU and UAV loads respecting HAPS backhaul; (iii) interference $I_{u,k}$ including co-channel transmissions across gNBs and UAVs, and (iv) PU protection $I_k^{\mathrm{PU}} \leq I_k^{\max}$.

4.3.2.0.6 Constraint enforcement in the RL framework

The coupling constraints are enforced within the RL framework as follows. The HAPS backhaul capacity constraint (S4) is enforced via action masking: the HAPS-tier RL agent's discrete action space is restricted so that only routing decisions satisfying the backhaul capacity limit are presented to the policy. The binary routing and activation variables in the same subproblem are inherently satisfied by the discrete action space. For PU-priority constraints (S2), the gNB scheduler allocates PU demand first and offers only residual capacity to SUs. For S1, the SU action space enforces channel vacancy and power bounds via masking/clipping. For S3, UAV activation and routing binaries are enforced by discrete actions, with backhaul feasibility enforced by the HAPS routing mask. As a safety net, post-hoc constraint violation detection triggers a reward penalty when any constraint is violated during execution, reinforcing thus feasibility through learning.

4.3.2.0.7 Reward shaping

Since S1–S4 are intractable to solve jointly, each tier optimizes a shaped reward that combines throughput, QoS satisfaction, collision penalties, and regional fairness. Urban regions emphasize fairness, remote regions prioritize coverage, and disaster regions increase the PU-protection weight. The penalty weights are held fixed throughout training. Table 4.4 summarizes the key simulation parameters.

In summary, the hierarchical structure ensures that PUs remain protected, SUs opportunistically exploit spectrum under constraints, UAVs provide resilience under failures, and, finally, HAPS orchestrates regional backhaul feasibility.

Table 4.3 Key notations

Symbol	**Meaning**
Network Sets	
$\mathcal{U}_P, \mathcal{U}_S$	Sets of PUs and SUs
$\mathcal{U}_{S,b}$	SUs associated with gNB b
$\mathcal{K}_b$	Channel set at gNB b
$\mathcal{U}_d$	UEs served by UAV d
$\mathcal{B}, \mathcal{D}$	Global sets of gNBs and UAVs
$\mathcal{B}_h, \mathcal{D}_h$	gNBs and UAVs coordinated by HAPS h
$\mathcal{R}_h$	Agent set participating in HAPS h federation round
N_h	Number of agents in HAPS region h
Decision Variables	
k_u, p_u	Channel and power for SU u
$x_{u,k}$	Scheduling indicator for SU u on channel k
$\mathbf{p}_d$	UAV position
$x_{b,h}, x_{d,h}$	Routing indicators via HAPS h
a_d	UAV activation binary
Channel and Signal	
$g_{u,k}$	Channel gain of UE u on channel k
N_0	Noise power spectral density
B_k	Bandwidth of channel k
$\gamma_{u,k}$	SINR of UE u on channel k
$R_{u,k}$	Achievable rate of UE u on channel k
System Parameters	
$\bar{\lambda}_b$	Spectrum demand at gNB b
Λ_d	Traffic load from UAV d
w_u	Priority weight for UE u

Continued on next page

Table 4.3 Key notations (continued)

Symbol	Meaning
E_d	Residual energy of UAV d
C_d^{BH}	UAV backhaul capacity
C_h^{BH}	HAPS backhaul capacity
$\mathbf{o}_i^t$	Noisy observation received by tier-i agent at step t
$P^{\max}$	Maximum SU transmit power
I_b	Interference indicator at gNB b
I_k^{PU}	Interference observed at PU channel k
$I_{u,k}$	Aggregate co-channel interference at UE u on channel k
$E_{\min}$	Minimum energy threshold for UAVs
$\Gamma_{\min}$	Minimum SINR threshold
$I_k^{\max}$	Maximum interference threshold on channel k
$\bar{\gamma}_b$	Average SINR at gNB b
σ_e^2	Estimation error variance
$C_{\min}$	Minimum backhaul capacity
$\mathrm{cov}_d(u)$	Coverage indicator: UAV d covers UE u
τ	Federated aggregation period in steps

4.4 Proposed solution

We design an F-HRL framework that coordinates spectrum access for UEs, terrestrial gNBs, UAV relays, and HAPS coordinators. The key objective is to maximize spectrum utilization and service resilience without centralizing raw observations, while concurrently maintaining explicit protection for licensed PUs.

4.4.1 Federated hierarchical reinforcement learning

Each tier executes a learning component tailored to its sensing capability and control scope as follows:

- **SUs** run lightweight models that perform channel selection and power control using local measurements such as occupancy and SINR estimates.
- **gNBs** host RL-based schedulers that allocate RBs among associated SUs subject to PU-protection and congestion constraints, and maintain a centralized critic for CTDE training of local SUs.
- **UAVs** employ RL-based policies to choose positions and activation status, trading off coverage and energy consumption.
- **HAPS** nodes act as federated aggregators: aggregating regional deltas, they apply trust-aware weighting and dispatch adapted policies for redeployment or disaster response.

During operation, each agent i maps its partial observation $\mathbf{o}_i^t$ to action a_i^t through policy π_i, while the HAPS tier maintains global parameters Θ_h shared across regions.

Fig. 4.3 illustrates the architecture.

4.4.2 Two-timescale convergence

To mitigate non-stationarity in MARL, F-HRL employs two-timescale learning: fast local policy updates and slower HAPS-level aggregation every τ steps, yielding a quasi-stationary reference model (Borkar, 2008; Leslie & Collins, 2005).

Each agent $i \in \mathcal{R}_h$ performs local updates

$$\Theta_i^{t+1} = \Theta_i^t - \alpha_t \nabla_{\Theta_i} \mathcal{L}_i(\Theta_i^t), \tag{4.13}$$

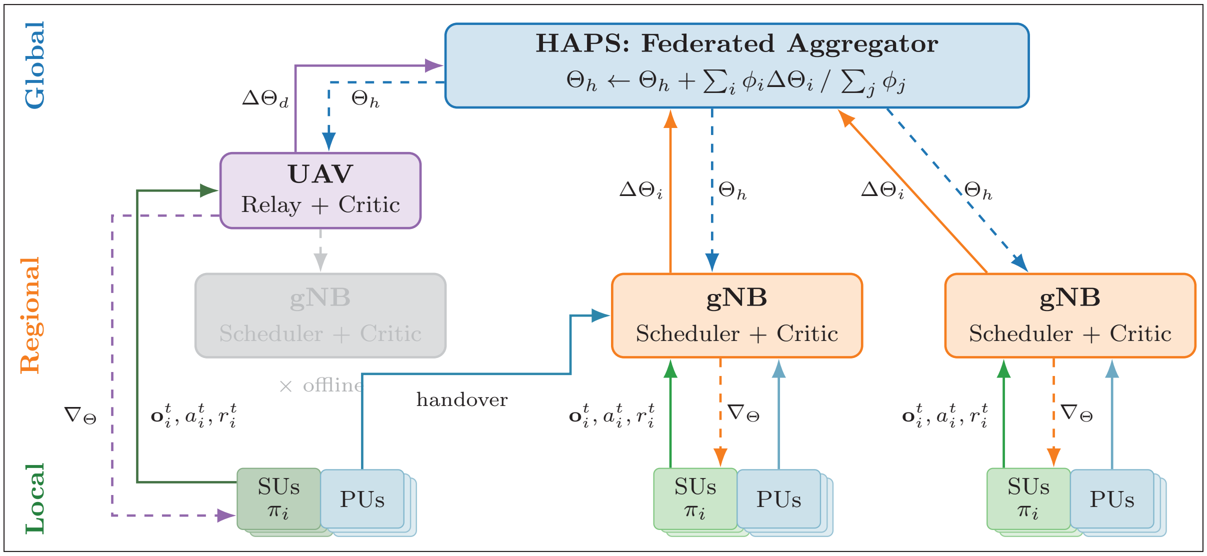

Figure 4.3 The disaster recovery protocol: gNB_0 is offline; users are served by nearest available ground and aerial supports

while the HAPS aggregator updates every τ steps as

$$\Theta_h^{r+1} = \Theta_h^r + \beta_r \frac{\sum_{i\in\mathcal{R}_h} \phi_i^{(r+1)} (\Theta_i^{(r)} - \Theta_h^{(r)})}{\sum_{j\in\mathcal{R}_h} \phi_j^{(r+1)}}. \tag{4.14}$$

Assume diminishing step sizes with $\sum_t \alpha_t = \sum_r \beta_r = \infty$, $\sum_t \alpha_t^2, \sum_r \beta_r^2 < \infty$, and $\beta_r/\alpha_{r\tau} \to 0$, bounded iterates, and Lipschitz losses $\mathcal{L}_i$.

Under these conditions, Θ_h^r evolves on a slower timescale and is quasi-static over each federated interval. The local iterates track the ODE $\dot{\Theta}_i = -\nabla_{\Theta_i} \mathcal{L}_i(\Theta_i; \Theta_h^r)$ and converge to $\Theta_i^\star(\Theta_h^r)$. Substitution into the slow update yields a reduced ODE for Θ_h, whose convergence follows from standard two-timescale stochastic approximation; the detailed proof can be found in Borkar (Borkar, 2008) and Leslie (Leslie & Collins, 2005).

Remark: In practice, F-HRL uses fixed step sizes, rather than diminishing schedules. Then, the iterates converge to an ϵ-neighborhood of the equilibrium, with radius scaling as $\epsilon = O(\alpha_{\text{loc}} + \beta_r^2/\alpha_{\text{loc}})$; equivalently, $\limsup_{r\to\infty} \mathbb{E}[\|\Theta_h^r - \Theta_h^\star\|^2] \le C_1\alpha_{\text{loc}} + C_2\beta_r^2/\alpha_{\text{loc}}$ for some constants $C_1, C_2 > 0$. This means that smaller local steps and stronger timescale separation yield tighter neighborhoods.

4.4.3 Federated aggregation and adaptation

Every τ steps, HAPS h aggregates updates from $\mathcal{R}_h$ using trust weights (Algorithm 4.1). Each agent's delta $\Delta\Theta_i^{(r)} = \Theta_i^{(r)} - \Theta_h^{(r)}$ is weighted by

$$\phi_i^{(r+1)} = \rho\phi_i^{(r)} + (1-\rho)\exp\left(-\frac{\|\Delta\Theta_i^{(r)} - \tilde{\Delta}^{(r)}\|^2}{2\sigma_{\text{trust}}^2}\right), \tag{4.15}$$

where $\tilde{\Delta}^{(r)}$ is the coordinate-wise median, ρ is trust momentum, and σ_{trust}^2 penalizes divergence from consensus.

To enable rapid redeployment without full retraining, HAPS construct context vectors over the preceding τ steps as follows:

$$\mathbf{c}_i = \left[\bar{\mathbf{o}}_i,\, \bar{r}_i\right], \tag{4.16}$$

where $\bar{\mathbf{o}}_i$ and $\bar{r}_i$ are the temporal means of agent i's local observations and rewards, respectively. HAPS computes cosine similarities $\mathrm{sim}(\mathbf{c}_i, \mathbf{c}_j)$. If the best match exceeds threshold η, the stored model $\Theta_{j^\star}$ is transferred; otherwise, the base model is adapted via

$$\Theta_{\text{adapted}} = \Theta_{\text{base}} - \alpha_{\text{adapt}}\nabla_\Theta\mathcal{L}_i(\Theta_{\text{base}}), \tag{4.17}$$

$$\alpha_{\text{adapt}} = \alpha_0 \max_j \mathrm{sim}(\mathbf{c}_i, \mathbf{c}_j), \tag{4.18}$$

where $\mathcal{L}_i$ is the local RL training loss for agent i.

4.4.4 Complexity analysis

Let $N_h = |\mathcal{R}_h|$ denote the number of agents in region h, and let $|\mathcal{A}_i|$ and $|\Theta_i|$ denote the action-space cardinality and parameter dimension of agent $i \in \mathcal{R}_h$. A flat controller requires searching

Algorithm 4.1 Federated hierarchical reinforcement learning

Input: Base model Θ_{base}, trust parameters $(\rho, \sigma_{\text{trust}})$, aggregation period τ
Output: Updated policies $\{\Theta_i\}$ and Θ_h
for $r \leftarrow 1$ **to** R **do**
  foreach *agent i* **do**
    $\Theta_i^{(r)} \leftarrow \Theta_h^{(r)} - \alpha_{\text{loc}} \nabla_\Theta \mathcal{L}_i(\Theta_h^{(r)})$
    $\Delta\Theta_i^{(r)} \leftarrow \Theta_i^{(r)} - \Theta_h^{(r)}$
  end foreach
  $\tilde{\Delta}^{(r)} \leftarrow \text{median}\Big(\{\Delta\Theta_i^{(r)}\}_{i \in \mathcal{R}_h}\Big)$
  $\phi_i^{(r+1)} \leftarrow \rho\phi_i^{(r)} + (1-\rho)\exp\Big(-\|\Delta\Theta_i^{(r)} - \tilde{\Delta}^{(r)}\|^2 / 2\sigma_{\text{trust}}^2\Big)$
  $\Theta_h^{(r+1)} \leftarrow \Theta_h^{(r)} + \dfrac{\sum_i \phi_i^{(r+1)} \Delta\Theta_i^{(r)}}{\sum_j \phi_j^{(r+1)}}$
  Broadcast $\Theta_h^{(r+1)}$ to all agents
end for

$\prod_i |\mathcal{A}_i|$, which grows exponentially with N_h. In contrast, the hierarchical decomposition reduces this complexity to $\sum_i |\mathcal{A}_i|$, scaling linearly with the number of agents.

Per-step inference: Each agent performs a forward pass with complexity $O(|\Theta_i|)$, yielding an aggregate regional cost of $O\big(\sum_{i \in \mathcal{R}_h} |\Theta_i|\big)$ across the HAPS, $|\mathcal{B}_h|$ gNBs, $|\mathcal{D}_h|$ UAVs, and SUs. Under CTDE training, each gNB additionally evaluates a centralized critic of size $|\Theta_{\text{C}}|$, incurring an extra $O(|\mathcal{B}_h|\,|\Theta_{\text{C}}|)$.

Federated round (Algorithm 4.1): Every τ steps, each agent computes a gradient update on $\mathcal{L}_i(\Theta_i)$ using a mini-batch of size M, with cost $O(M|\Theta_i|)$. The HAPS aggregator computes the coordinate-wise median $\tilde{\Delta}^{(r)}$ and the trust-weighted aggregation, each at $O(N_h|\Theta|)$, and evaluates pairwise context similarities $\text{sim}(\mathbf{c}_i, \mathbf{c}_j)$ at $O(N_h^2 d_{\text{ctx}})$. Letting $|\Theta| = \max_i |\Theta_i|$, the per-round complexity is

$$O\Big(N_h M |\Theta| + N_h^2 d_{\text{ctx}}\Big). \tag{4.19}$$

If a HAPS manages hundreds of agents, this can become expensive. In such cases, we propose using locality-sensitive hashing to reduce similarity search to $O(N_h \log N_h d_{\text{ctx}})$. Since $d_{\text{ctx}} \ll |\Theta|$ (typically $d_{\text{ctx}} \approx 10$ vs. $|\Theta| \approx 10^4$), the gradient computation dominates.

Communication: Each federated round exchanges $O(N_h|\Theta|)$ parameters for uplink model deltas $\Delta\Theta_i^{(r)}$ and downlink broadcasts of $\Theta_h^{(r+1)}$. Inter-HAPS synchronization over LEO backhaul every T_{sync} rounds introduces an amortized overhead of $O((H-1)S_{\text{sync}}/T_{\text{sync}})$, where H denotes the number of HAPS regions and S_{sync} the compressed summary size. For the deployment in Section 4.5 ($N_h = 18$, $|\Theta| = 10^4$), the uplink+downlink exchange is about $2N_h|\Theta|$ parameters, i.e., approximately 3.6×10^5 values (~1.4 MB with 32-bit floats) per round; with $\tau = 50$ this is ~0.12 s on a 100 Mbps backhaul. For a denser region ($N_h = 200$, $|\Theta| = 10^4$), the exchange is about 4×10^6 values (~16 MB), motivating gradient compression.

4.4.5 Security and trust management

The proposed F-HRL framework achieves resilience against adversarial behavior through complementary mechanisms operating at the sensing and federated layers.

Sensing layer: Cooperative SUs mitigate spectrum-sensing data falsification (SDF) by sharing local occupancy decisions and applying majority voting prior to channel access (Khaf *et al.*, 2022). This mechanism tolerates up to $\lfloor(|\mathcal{U}_{S,b}| - 1)/2\rfloor$ falsified reports per gNB, preventing any single malicious SU from dominating local decisions.

Federated layer: Trust-weighted aggregation (4.15) provides Byzantine resilience during model fusion. The coordinate-wise median $\tilde{\Delta}^{(r)}$ and Gaussian trust kernel in ϕ_i suppress updates that deviate from regional consensus. Furthermore, HAPS tracks each agent's spectrum-access behavior (request patterns, grant efficiency, and adaptivity) as an auxiliary diagnostic metric for anomaly monitoring and post-hoc analysis; this behavioral metric is not part of the active aggregation weight in Eq. (4.15). Empirically, malicious SUs converge to lower scores than benign SUs (Section 4.5).

4.5 Performance evaluation

We benchmarked the proposed F-HRL stack against a non-cognitive baseline, a simpler cognitive baseline, a standard 5G heuristic baseline, and the HDRL implementation of (Umer *et al.*, 2025) across the following key NTN operating scenarios: nominal spectrum sharing, disaster recovery, and malicious SU attacks. Unless otherwise stated, the results were averaged over 5000 decision steps with synchronized random seeds.

Network configuration

Unless stated otherwise, we simulated a single HAPS at 20 km coordinating three terrestrial gNBs and three standby UAVs. A total of 12 PUs and 12 SUs were distributed across urban and remote regions, contending over six channels (20–100 MHz bandwidth) allocated across the gNBs. Propagation, channel models, and transmit power levels followed Section 4.3. Table 4.4 summarized the common configuration used in all experiments.

Table 4.4 Simulation settings and hyperparameters

Parameter	**Value**
Network topology	
HAPS altitude	20 km
Terrestrial gNBs / UAV relays	3 / 3 standby
PUs / SUs	12 / 12
Spectrum & Physical layer	
Active channels	6 (20–100 MHz each)
Channel / gNB capacity	50–113 / 106–273 RBs
PU / SU transmit power	2 W / 0.5–1.0 W
Noise PSD (N_0)	−174 dBm/Hz
RL hyperparameters	
SU DQN: α / γ / ε-decay	0.003 / 0.99 / 0.995
SU DQN: hidden / batch / buffer	64 / 32 / 200
gNB/HAPS DQN: α / buffer	0.001 / 10000
Learning & Federation	
Horizon	5000 steps
FL aggregation (τ) / Trust ρ	50 steps / 0.9
Disaster window	Steps 2500–3500

SU agents employed a range of RL algorithms with state vectors capturing sensed PU occupancy, and RB allocation trends among collaborative SUs. gNB and UAV policies enforced PU priority and protected against malicious SUs. HAPS performed federated aggregation with robust weighting every 50 steps and redistributed channels among active UAVs and gNBs as needed.

Attack model

Malicious SUs falsified sensing with probability 0.8 and always required maximum RBs at full transmit power. 25% of SUs were modeled as malicious agents for the trust evaluation experiments.

Fig. 4.4 provided a comprehensive multi-metric comparison across all evaluated policies. Six normalized KPIs were displayed: SU throughput, Jain fairness index, QoS satisfaction ratio, spectral efficiency, energy efficiency (i.e., throughput per unit transmit energy), and interference footprint (defined as the inverse of collision rate weighted by transmit power). As can be seen in the results, F-HRL attained the highest throughput, the highest fairness score (0.950), and the lowest interference among learning-based approaches. HDRL achieved 97.9% QoS satisfaction, but lower throughput at 94.7 Mbps, while the CRN baseline exhibited the highest interference due to its fixed full-power transmissions. The 5G Standard policy, which excluded SU access entirely, achieved trivially zero interference, but forfeited all secondary throughput. These results confirmed that federated coordination enabled SUs to more effectively balance throughput, fairness, and interference than single-tier or non-cognitive alternatives. Importantly, the F-HRL architecture is algorithm-agnostic: the hierarchical coordination and federated aggregation layers are independent of the per-tier RL algorithm, so DQN, PPO, actor–critic, or other methods can be substituted at any tier without altering the overall framework.

KPIs and KVIs

Table 4.5 showed the mean indicators across the 5000-step evaluation window. Spectrum efficiency reported aggregate RB utilization as a fraction of available capacity. Disaster and

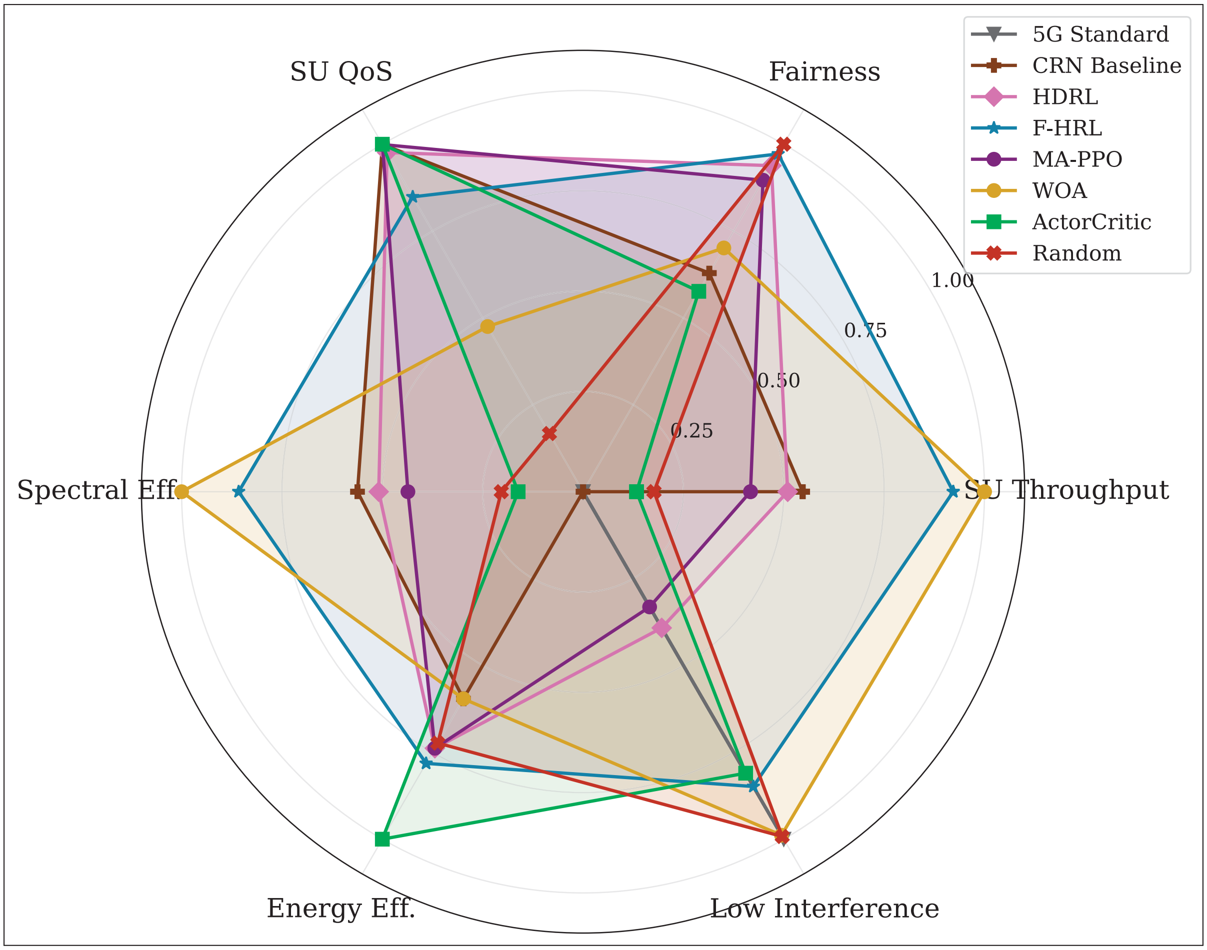

Figure 4.4 Multi-metric policy comparison

attack scenarios additionally logged retention and recovery dynamics, referenced herein in dedicated subsections.

Table 4.5 Key performance indicators across scenarios

Scenario	**Spec. Eff.**	**SU Tput (Mbps)**	**Col. (/step)**	**Fair. (Jain)**
Non-cognitive	0.111	0.0	0.00	1.000
F-HRL (nominal)	0.871	167.5	5.13	0.950
F-HRL (attack)	0.748	128.9	8.46	0.887
F-HRL (disaster)	0.964	164.4	4.76	0.954

Fig. 4.5 presented an ablation study comparing four hierarchy configurations: (i) a heuristic baseline with fixed small RB requests (5 RBs), random channel selection, and no learning at any tier, (ii) SU-only DQN where only SUs learned while gNB and HAPS used fixed policies, (iii) SU+gNB DQN adding gNB-level learning, and (iv) F-HRL with learning at all tiers. The heuristic baseline remained the lowest (0.279 spectral efficiency, 38.0 Mbps SU throughput). Relative to SU-only learning (0.418 spectral efficiency, 64 Mbps), adding the gNB tier (SU+gNB: 0.416, 63 Mbps) showed no statistically significant degradation across runs. Furthermore, adding the HAPS tier (full F-HRL: 0.423, 67.4 Mbps) preserved efficiency/throughput and slightly improved both metrics, indicating that hierarchy scales coordination without sacrificing core performance. Fig. 4.5 includes error bars across 10 random seeds to show the run-to-run variance. All configurations maintained perfect PU protection (1.0). Multi-seed runs across 10 random seeds exhibited negligible variance.

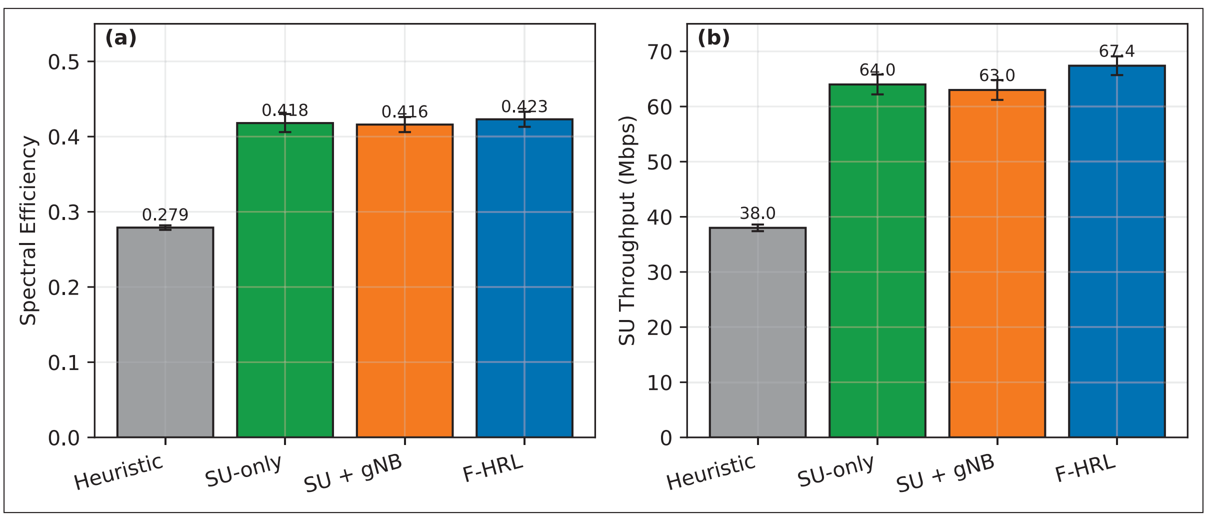

Figure 4.5 Hierarchy ablation study (a) spectral efficiency, (b) secondary user throughput

4.5.1 Secondary user coverage and grant fulfillment

Fig. 4.6 further examined SU access performance through two complementary metrics: coverage (i.e., the percentage of SUs receiving grants per step), and fulfillment (i.e., the ratio of granted to requested RBs). F-HRL achieved 74% coverage and 92% fulfillment, outperforming HDRL at

66% and 87%, respectively, by learning adaptive request strategies. The CRN baseline attained perfect scores on both metrics because its fixed, conservative requests were always satisfiable; however, this came at the cost of reduced throughput (see Fig. 4.4). Random access yielded only 17% on both metrics, confirming that uncoordinated spectrum contention severely limited SU performance.

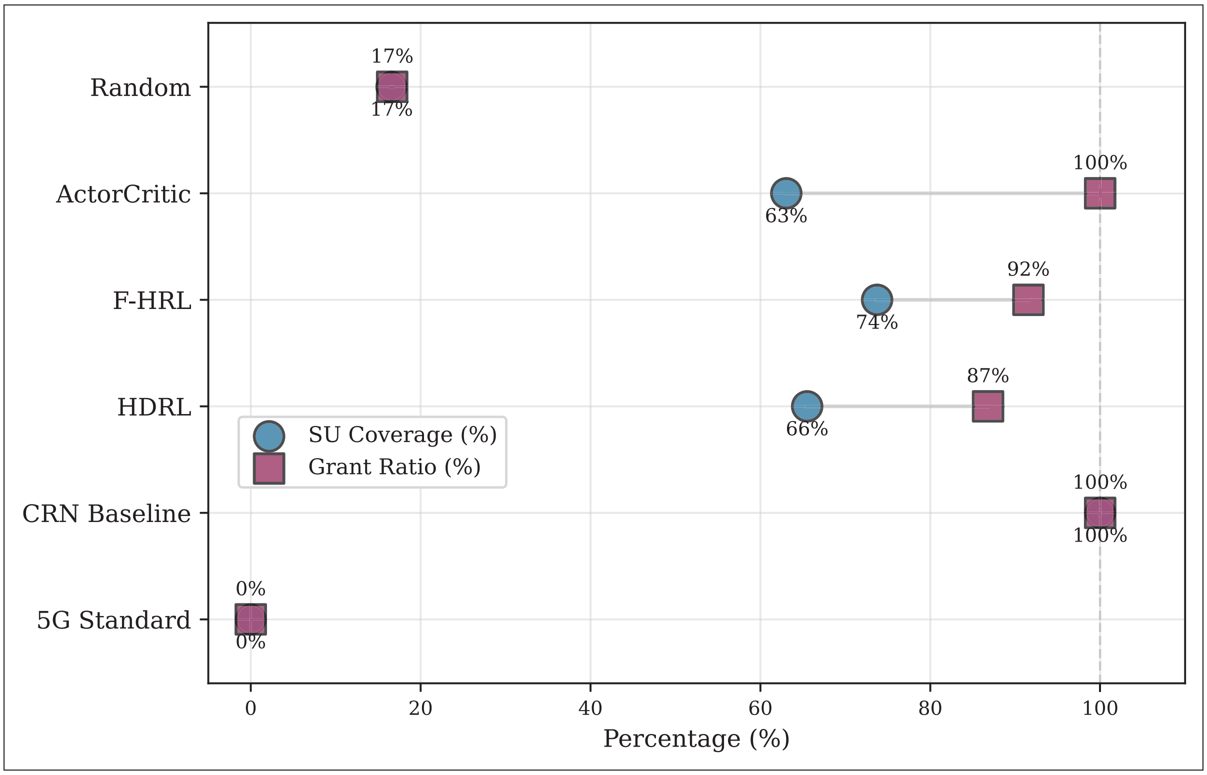

Figure 4.6 Secondary user coverage and grant fulfillment

The convergence curve showed that learning remained stable across all tiers, with variance across Monte Carlo running below 0.01. This stability was a direct result of federated aggregation and trust-weighted updates, which dampened the effect of outlier agents. Importantly, the convergence rate was faster than in single-tier RL baselines reported in previous research (Umer *et al.*, 2025; Hu *et al.*, 2022), thus validating the benefit of hierarchical learning. However, the persistent contention post-convergence suggested that further improvements in cross-tier reward shaping or adaptive aggregation frequency may be needed.

4.5.2 Resilience to infrastructure outages

Fig. 4.7 showed disaster recovery performance when gNB0 failed at step 2 500. Fig. 4.7(a) depicted UE reassociation dynamics: prior to failure, gNB0 served the majority of UEs; upon failure, PUs and SUs tried to connect to the next available node and service degraded. HAPS recognized gNB0 failure and activated a UAV, while UEs continued trying to get service from next available nodes while the UAV was in transit. gNB1 experienced a high load and kept HAPS informed of the UE demand. HAPS detected the overload and activated another UAV to meet the demand and redistributed the available channels. Once gNB0 was restored, both UAVs returned to standby. Fig. 4.7(b) decomposed throughput into PU and SU contributions, demonstrating that PU traffic was prioritized throughout the recovery phase. During the disaster window, PU throughput retention reached 99.1% while SU retention amounted to 95.1%, confirming that the scheduling policy enforced PU protection even under constrained relay capacity. The results validated the system's resilience: HAPS-coordinated UAV deployment restored PU service within 50 steps, matching best-case recovery times in recent literature.

4.5.3 Adversarial robustness and trust management

Three of the twelve SUs, i.e., 25%, were configured as greedy attackers that falsified spectrum sensing reports and required maximum RBs at full transmit power. Fig. 4.8 presents the resulting score distribution after 100 episodes of HAPS-coordinated federated aggregation. Benign SUs converged to a median score of 0.75 (IQR 0.70–0.85), while malicious SUs were suppressed to a median of 0.34 (IQR 0.28–0.39), yielding a separation of 0.40. The upper whisker of malicious scores reached 0.44, which is below the minimum benign score of 0.455, giving the optimal threshold 0.45, with 100% detection accuracy (and 0 false positives and 0 false negatives). At a lower threshold $\tau = 0.30$, one malicious agent remained undetected, allowing that attacker to influence aggregation, yielding 91.7% accuracy. At a higher threshold $\tau = 0.50$, one benign agent was misclassified. These results confirm that the F-HRL trust-weighted aggregation mechanism reliably distinguishes adversarial from benign agents.

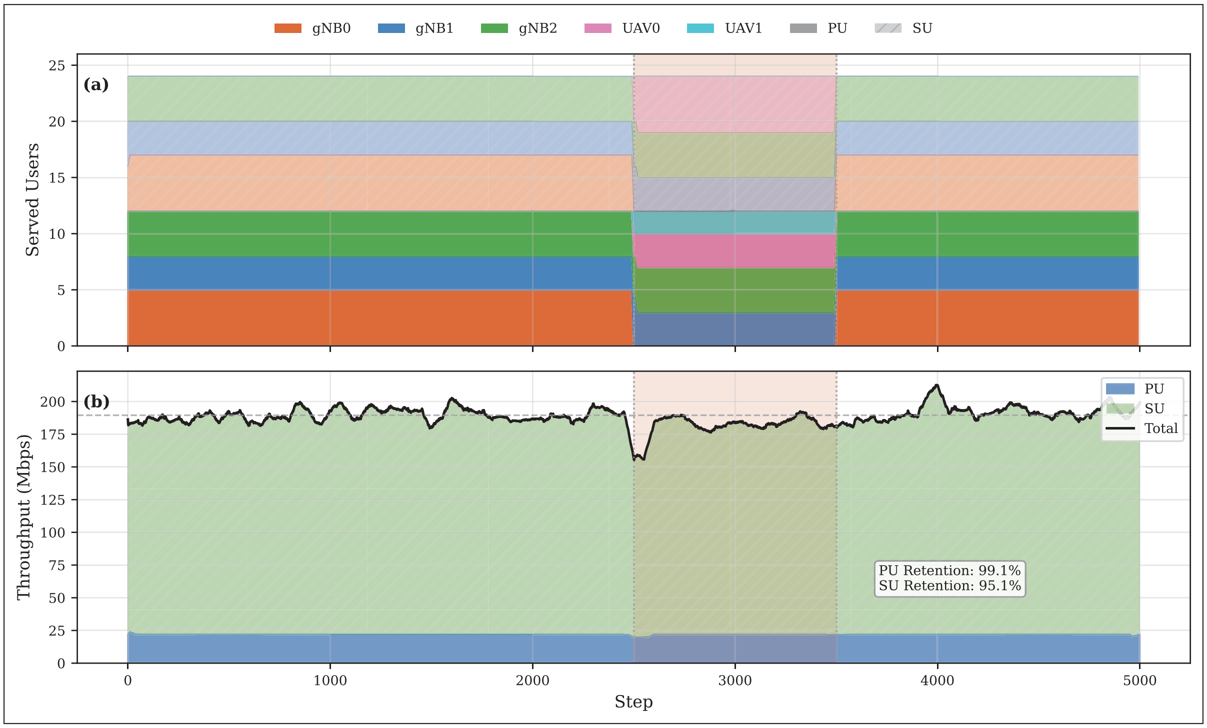

Figure 4.7 Disaster recovery: (a) user reassociation during outage, (b) throughput retention

4.5.4 Policy benchmarking and fairness

To contextualize the hierarchical ablation in Fig. 4.5, we compared F-HRL against a broader range of RL algorithms and baseline policies. Table 4.6 and Fig. 4.9 provide comprehensive policy benchmarking. As can be seen in the results, F-HRL achieved the highest SU throughput among learning-based methods (167.5 Mbps) and outperformed other RL alternatives, while HDRL attained 94.7 Mbps and the CRN Baseline demonstrated conservative yet reliable performance at 97.1 Mbps. Two additional baselines were evaluated: MA-PPO achieved 73.5 Mbps, while WOA achieved the highest SU throughput at 177.5 Mbps, but with a lower SU QoS (47.5%) and a higher delay (0.4000), indicating a poorer QoS compliance despite centralized optimization. Delay reports average SU request-to-grant backlog-fulfillment latency (in scheduler steps); zero values indicate no delayed backlog fulfillment events in the logged runs.

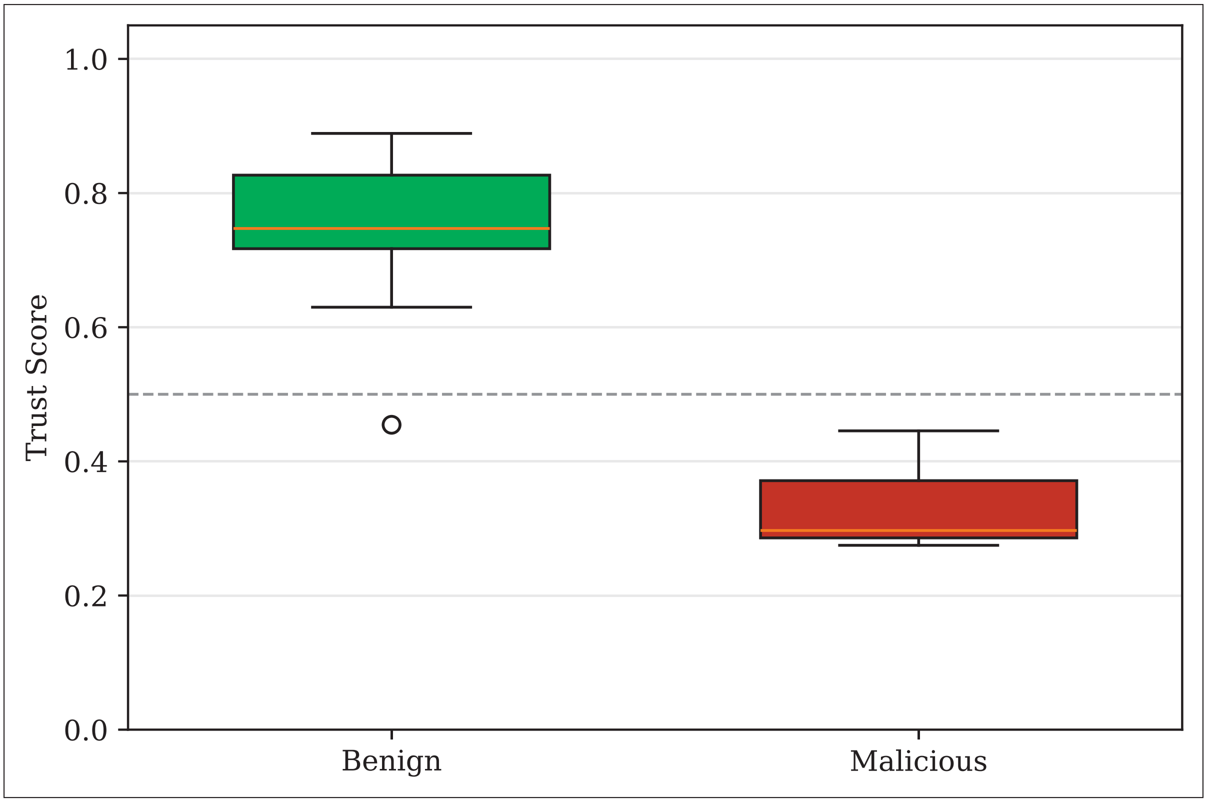

Figure 4.8 Trust score distribution

Table 4.6 Policy comparison across key metrics

Policy	**SU Throughput (Mbps)**	**SU QoS (%)**	**Delay (steps)**
5G Standard	0.0	0.0	0.0000
CRN Baseline	97.1	100.0	0.0000
HDRL	94.7	97.9	0.0000
F-HRL	167.5	86.2	0.6000
MA-PPO	73.5	99.9	0.0000
WOA	177.5	47.5	0.4000
ActorCritic	29.9	100.0	0.0000
Random	31.1	17.0	0.0000

Table 4.7 Regional performance breakdown

Region	**gNB**	**RBs**	**SUs**	**PU QoS (%)**	**SU QoS (%)**
Urban-1	gNB0	273	5	100.0	91.0
Urban-2	gNB1	200	3	96.9	95.7
Remote	gNB2	106	4	100.0	89.5

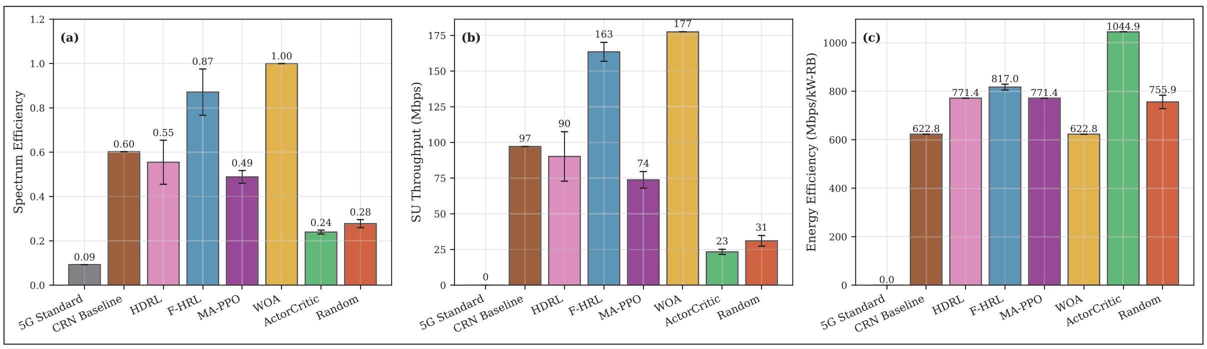

Figure 4.9 Policy comparison: (a) spectral efficiency, (b) secondary user throughput, and (c) energy efficiency

4.5.5 Discussion

Table 4.7 provided a per-region breakdown of F-HRL performance. Urban-1, which was found to have the highest capacity of 273 RBs and served five SUs, achieved 100% PU QoS and 91.0% SU QoS. Urban-2 attained the best SU QoS at 95.7%, which can be attributed to its lower contention ratio of 3 SUs sharing 200 RBs. The remote region, despite its limited 106 RB budget and four SUs, still reached 89.5% SU QoS, thus confirming that F-HRL adapted resource allocation to heterogeneous regional capacities.

We verified the convergence neighborhood prediction from Proposition 1 by running the F-HRL framework under three learning rate configurations ($\alpha \in \{0.003, 0.001, 0.0003\}$), as shown in Fig. 4.10. As α decreased from 0.003 to 0.0003, the steady-state parameter drift $\|\Theta_h^r - \Theta_h^{r-1}\|_2$ decreased from 0.125 to 0.116, confirming that in line with our expectation, smaller step sizes produce tighter convergence neighborhoods.

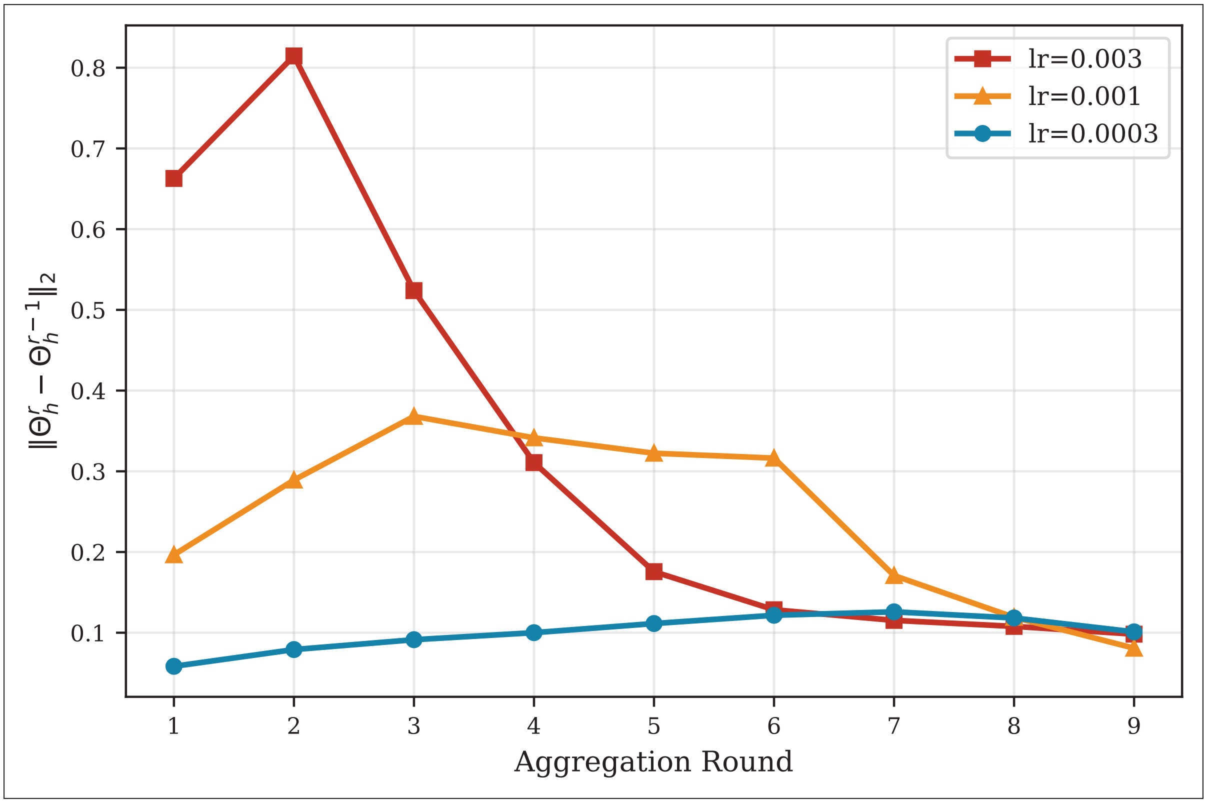

Figure 4.10 Convergence neighborhood: parameter drift per aggregation round for three learning rate configurations

Overall, F-HRL maximized spectrum utilization and per-SU throughput, and traded off energy efficiency and QoS fulfillment according to regional priorities. Disaster recovery rapidly restored PU service and efficiently allowed SU access, while maintaining PU protection and priority. These insights demonstrated that F-HRL is highly adaptable to heterogeneous environments and provided a solid foundation for 6G NTN CRNs.

4.6 Conclusion

In this paper, we presented a federated hierarchical reinforcement learning framework for HAPS-assisted CRNs operating across urban, remote, and disaster-stricken environments. The proposed architecture couples lightweight SU-level cooperative learning, trust-aware gNB/UAV coordination, and Byzantine-resilient aggregation at the HAPS tier, thereby enabling spectrum access decisions without centralizing raw observations. Simulation results demonstrated that the proposed framework consistently improved QoS across heterogeneous environments while maintaining PU protection and rapid post-failure recovery.

Major findings are briefly summarized below:

- **High spectrum utilization with hierarchical learning:** F-HRL achieved 87.1% spectrum efficiency as compared to 11.1% for the non-cognitive baseline, demonstrating a nearly 8-fold improvement while maintaining sustained per-SU throughput and high fairness.
- **Robust disaster recovery:** When a gNB failed, UAV relays quickly restored PU service with 99.1% retention and maintained 95.1% of original SU throughput, demonstrating effective resilience despite constrained relay capacity.
- **Adversarial robustness:** Trust-aware aggregation successfully isolated compromised agents and maintained legitimate throughput near nominal levels during attacks with 25% malicious SUs.
- **Policy comparison validates federated approach:** F-HRL achieved 74% SU coverage with 92% grant fulfillment, thus outperforming HDRL and maintaining the highest fairness score among learning-based approaches.

A brief agenda for future work is outlined below. First, context-aware selective model transfer should be explored to analyze the impact of regional transferability. Second, broader adversarial test suites such as coordinated power jamming should be incorporated to stress-test the trust pipeline. Third, extending the framework to account for detailed channel dynamics, including Doppler effects from UAV mobility and handover signaling overhead, would improve fidelity of the physical layer model. Fourth, investigating gradient compression techniques and hierarchical aggregation trees for deployments exceeding 1000 agents would address scalability requirements of dense satellite constellations.

Despite these open directions, the present results demonstrate that the proposed HAPS-centric F-HRL pipeline provides a practical template for resilient, privacy-aware spectrum coordination in emerging non-terrestrial networks.

CONCLUSION AND RECOMMENDATIONS

This thesis examined spectrum utilization as an end-to-end coordination problem involving both spectrum-consuming nodes and spectrum-granting infrastructure. The first two articles laid down the secondary-user (SU)-side foundation: namely, reliable cooperative sensing and joint sensing-access optimization. The third article expanded this foundation into a multi-tier architecture coordinating SUs, next-generation Node Bs (gNBs), uncrewed aerial vehicles (UAVs), and high-altitude platform stations (HAPS) under heterogeneous demand, infrastructure failures, privacy constraints, and adversarial behavior.

Synthesis of the Research

The first two articles addressed the decisions available to SUs before infrastructure-side coordination was introduced. First, partially cooperative multi-agent reinforcement learning (PCMARL) established reliable cooperative sensing under sensing uncertainty, energy constraints, and sensing data falsification (SDF) attacks. Second, the Intelligent Multi-User Participatory Algorithm for Cooperative Transmission (IMPACT) then extended the SU-side decision space from sensing to access by jointly optimizing sensing participation, cooperation time, channel selection, and transmission power under imperfect channel state information (CSI).

Taken together, these results revealed that SUs learned when to sense, when to cooperate, which channel to access, and how much power to use. We also found why SU-side optimization alone remained incomplete: while SUs improved local sensing and transmission behavior, they did not provide access authorization, grant allocation, regional coordination, trust management, or recovery from infrastructure failures.

Federated hierarchical reinforcement learning (F-HRL) addressed this broader end-to-end problem. To this end, spectrum management was decomposed across four tiers with distinct roles: SUs performed local channel selection and power control; gNBs allocated resources while enforcing primary user (PU) priority and congestion constraints; UAVs provided coverage and relay support when terrestrial infrastructure was impaired; and, finally, HAPS nodes coordinated regional aggregation, policy adaptation, and recovery. This decomposition extended the work beyond SU-only optimization to coordinated spectrum management across consuming and granting nodes.

The F-HRL framework also addressed additional risks introduced by infrastructure-side control. While gNBs improved coordination, they could fail, become overloaded, or lose reliable backhaul. UAVs provided backup support during these events, while HAPS nodes provided a higher-level coordination layer supporting federation across regions without centralizing raw observations. Trust-weighted aggregation further allowed the system to suppress malicious or unreliable participants during model fusion.

Findings and Contributions

The results showed that selective cooperation improved sensing reliability without requiring full-network sensing participation. PCMARL was found to improve sensing accuracy and energy efficiency under SDF attacks by rewarding useful sensing participation and reducing unnecessary sensing activity.

SU-side spectrum access also benefited from joint sensing and transmission optimization. IMPACT and IMPACT-D showed that access decisions under imperfect CSI require both discrete and continuous control. The proposed hybrid-action formulation improved throughput and energy efficiency while accounting for PU protection and SU-to-SU interference.

Infrastructure-side coordination changed the nature of the spectrum-management task. In F-HRL, gNBs did not simply add another network tier; instead, they introduced the granting side of the system by scheduling access, enforcing PU priority, and coordinating local SU decisions. This closed a gap left by SU-only approaches.

Resilience required explicit support beyond gNB-side coordination. The F-HRL disaster-recovery scenario showed that UAV activation and HAPS coordination maintained PU service and preserved SU throughput during gNB outage. Therefore, end-to-end spectrum management had to account not only for nominal allocation, but also for continuity under infrastructure disruption.

Federation and trust management were central to multi-tier spectrum coordination. HAPS-based aggregation enabled regional coordination without centralizing raw observations, while trust-weighted model fusion suppressed malicious updates and unreliable sensing behavior. These mechanisms connected privacy preservation, adversarial robustness, and scalable coordination within the same framework.

The main contribution of the present thesis was the development of an end-to-end view of spectrum management that began with SU-side learning and culminated in cross-tier coordination. The SU-side articles provided the sensing and access mechanisms needed by spectrum-consuming nodes. The subsequent F-HRL article extended the work to the spectrum-granting and coordination side, where scheduling, PU priority, federation, resilience, and trust were handled jointly.

The resulting architecture assigned a distinct function to each technical layer. Specifically, sensing provided spectrum awareness. While access optimization determined how SUs used detected opportunities under power and interference constraints, gNB coordination introduced grant decisions and policy enforcement. Furthermore, UAV support addressed impaired terrestrial infrastructure, while HAPS federation enabled wider-area coordination, privacy-preserving aggregation, and trust-aware adaptation. The combined result was an integrated framework for resilient beyond 5G (B5G) spectrum management.

Scope of the Findings

The proposed methods were evaluated through simulation. This choice enabled controlled comparison across sensing uncertainty, imperfect CSI, adversarial behavior, and infrastructure disruption. However, this choice also limited the conclusions to the assumptions represented in the simulation models. Practical deployment would require additional validation of processing cost, signalling overhead, synchronization, latency, and hardware constraints.

The SU-side studies focused on the sensing and access decisions available to SUs. PCMARL considered partially cooperative sensing under bounded SDF behavior, while IMPACT considered joint sensing and access under imperfect CSI. Together, these models capture important sources of uncertainty; however, they do not include all device-level constraints, MAC-layer signalling costs, or very large state spaces potentially arising in dense low-power deployments.

The end-to-end F-HRL framework was evaluated under representative urban, remote, disaster-recovery, and adversarial scenarios. The evaluated failure and attack models included gNB outage, malicious SUs, and trust-weighted aggregation. However, further research is needed on broader coordinated attacks, detailed UAV mobility effects, handover signalling, inter-HAPS synchronization, and large-scale aggregation across many regions.

Future Work

Future work could first validate the proposed methods on network-emulation or software-defined radio testbeds. Such validation would clarify computational cost, signalling overhead, latency, and hardware feasibility under realistic sensing, access, and coordination conditions.

In addition, the SU-side methods could then be extended through function approximation, device-aware cooperation, and more detailed modelling of CSI uncertainty and traffic variation.

Furthermore, F-HRL framework could also be evaluated in larger multi-region deployments with multi-HAPS coordination, selective model transfer, realistic backhaul constraints, and stronger adversarial models.

Finally, the proposed coordination mechanisms could be aligned with practical network interfaces and emerging B5G and 6G standardization requirements.

In summary, the results of the present thesis demonstrated that improved spectrum utilization required SU-side sensing and access optimization together with granting-side coordination, infrastructure resilience, federated aggregation, and trust-aware control. This work contributes to that direction by integrating SU-side learning with gNB, UAV, and HAPS coordination for resilient spectrum management in future wireless networks.

BIBLIOGRAPHY

5G Americas. (2023). *Update on 5G non-terrestrial networks*. Bellevue, WA.

Abbasi, O., Kaddoum, G. & Yanikomeroglu, H. (2025). An ML-Assisted OFDM-Based Hemispherical Array Antenna with Hybrid Beamforming for HAPS. *IEEE Transactions on Wireless Communications*, 24(10), 8908–8922. doi: 10.1109/TWC.2025.3569831.

Ahmadfard, A., Jamshidi, A. & Keshavarz-Haddad, A. (2017). Probabilistic spectrum sensing data falsification attack in cognitive radio networks. *Signal Processing*, 137, 1–9. doi: 10.1016/j.sigpro.2017.01.033.

Ahmadzadeh, M., Pakravan, S., Hodtani, G., Zeng, M. & Chouinard, J.-Y. (2024). Deep Reinforcement Learning for Robust RIS-Aided Over-the-Air Federated Learning in Cognitive Radio. *IEEE Middle East Conference on Communications and Networking (MECOM)*.

Ahmed, A., Mishra, D., Prasad, G. & Baishnab, K. L. (2021). Cognitive radio timing protocol for interference-constrained throughput maximization. *IEEE Transactions on Cognitive Communications and Networking*, 1-1. doi: 10.1109/TCCN.2021.3137487.

Alfa, A. S., Maharaj, B. T., Lall, S. & Pal, S. (2016). Mixed-Integer Programming based Techniques for Resource Allocation in Underlay Cognitive Radio Networks : A Survey. *Journal of Communications and Networks*, 18(5), 744–761.

Alghorani, Y., Kaddoum, G., Muhaidat, S. & Pierre, S. (2015). On the Approximate Analysis of Energy Detection over n Rayleigh Fading Channels Through Cooperative Spectrum Sensing. *IEEE Wireless Communications Letters*, 4(4), 413–416. doi: 10.1109/LWC.2015.2427283.

Alharbi, F., Alhuzali, A., Alalwany, E. & Alfahaid, A. (2025). Dynamic Orbital Resource Management (DORM) for 6G Networks: Enhancing Cybersecurity in Non-Terrestrial Systems. *IEEE Access*, 13, 137033–137047.

Ali, A. & Hamouda, W. (2017). Advances on Spectrum Sensing for Cognitive Radio Networks: Theory and Applications. *IEEE Communications Surveys and Tutorials*, 19(2), 1277–1304. doi: 10.1109/COMST.2016.2631080.

Ali, Z., Jiao, L., Baker, T., Abbas, G., Abbas, Z. H. & Khaf, S. (2019). A deep learning approach for energy efficient computational offloading in mobile edge computing. *IEEE Access*, 7, 149623–149633.

Ali, Z., Khaf, S., Abbas, Z., Abbas, G., Jiao, L., Irshad, A., Kwak, K. & Bilal, M. (2020a). A Comprehensive Utility Function for Resource Allocation in Mobile Edge Computing. *Cmc -Tech Science Press-*, 66, 1461-1477. doi: 10.32604/cmc.2020.013743.

Ali, Z., Khaf, S., Abbas, Z. H., Abbas, G., Muhammad, F. & Kim, S. (2020b). A Deep Learning Approach for Mobility-Aware and Energy-Efficient Resource Allocation in MEC. *IEEE Access*, 8, 179530–179546.

Bendouda, D., Rachedi, A. & Haffaf, H. (2018). Programmable architecture based on Software Defined Network for Internet of Things: Connected Dominated Sets approach. *Future Generation Computer Systems*, 80, 188–197. doi: 10.1016/j.future.2017.09.070.

Bkassiny, M., Li, Y. & Jayaweera, S. K. (2013). A Survey on Machine-Learning Techniques in Cognitive Radios. *IEEE Communications Surveys and Tutorials Tutorials*, 15(3), 1136–1159. doi: 10.5120/ijca2016907355.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press.

Chang, H.-h., Song, H., Yi, Y., Zhang, J., He, H. & Liu, L. (2019). Distributive Dynamic Spectrum Access Through Deep Reinforcement Learning: A Reservoir Computing-Based Approach. *IEEE Internet of Things Journal*, 6(2), 1938–1948. doi: 10.1007/978-981-13-8285-7.

Chen, H., Zhou, M., Wang, K. & Li, J. (2016). Joint Spectrum Sensing and Resource Allocation Scheme in Cognitive Radio Networks with Spectrum Sensing Data Falsification Attack. *IEEE Transactions on Vehicular Technology*, 65(11), 9181–9191. doi: 10.1109/TCOMM.2018.2807432.

Chen, Z., Xu, Y.-Q., Wang, H. & Guo, D. (2022). Federated Learning-Based Cooperative Spectrum Sensing in Cognitive Radio. *IEEE Communications Letters*, 26(2), 330–334.

Cheng, P., Chen, Z., Ding, M., Li, Y., Vucetic, B. & Niyato, D. (2020). Spectrum Intelligent Radio: Technology, Development, and Future Trends. *IEEE Communications Magazine*, 58(1), 12–18. doi: 10.1109/MCOM.001.1900200.

Dai, Z., Wang, Z. & Wong, V. W. (2016). An Overlapping Coalitional Game for Cooperative Spectrum Sensing and Access in Cognitive Radio Networks. *IEEE Transactions on Vehicular Technology*, 65(10), 8400–8413. doi: 10.1109/TVT.2015.2505711.

Das, S. K., Mudi, R., Rahman, M. S., Rabie, K. M. & Li, X. (2024). Federated Reinforcement Learning for Wireless Networks: Fundamentals, Challenges and Future Research Trends. *IEEE Open Journal of Vehicular Technology*, 5, 1400–1440.

Davies, H. (2021). *Making energy harvesting work for edge IoT devices*.

Farajzadeh, A., Yadav, A. & Yanikomeroglu, H. (2025). Federated Learning in NTN: Design, Architecture, and Challenges. *IEEE Communications Magazine*, 63(6), 26–33.

Furtado, A., Irio, L., Oliveira, R., Bernardo, L. & Dinis, R. (2016). Spectrum sensing performance in cognitive radio networks with multiple primary Users. *IEEE Transactions on Vehicular Technology*, 65(3), 1564-1574. doi: 10.1109/TVT.2015.2406254.

Getu, T. M., Ajib, W., Landry, R. & Kaddoum, G. (2019a). Toward overcoming a hidden terminal problem arising in MIMO cognitive radio networks: A tensor-based spectrum sensing algorithm. *IEEE Transactions on Vehicular Technology*, 68(10), 9833–9847. doi: 10.1109/TVT.2019.2933857.

Getu, T. M., Ajib, W., Landry, R. & Kaddoum, G. (2019b). A Tensor-Based Spectrum Sensing Technique for MIMO Cognitive Radio Networks. *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5. doi: 10.1109/GlobalSIP45357.2019.8969440.

Guo, S. & Zhao, X. (2022). Deep reinforcement learning optimal transmission algorithm for cognitive internet of things with RF energy harvesting. *IEEE Transactions on Cognitive Communications and Networking*, 8(2), 1216-1227. doi: 10.1109/TCCN.2022.3142727.

Haddadou, N., Rachedi, A. & Ghamri-Doudane, Y. (2015). A Job Market Signaling Scheme for Incentive and Trust Management in Vehicular Ad Hoc Networks. *IEEE Transactions on Vehicular Technology*, 64(8), 3657–3674. doi: 10.1109/TVT.2014.2360883.

Hamza, E. K., Mahdy, L. S. & Thabit, A. A. (2019). Design and implementation of cognitive radio (CR) based on Xilinx FPGA. *ARPN Journal of Engineering and Applied Sciences*, 14(4), 892–897.

Hassan, S. S., Park, Y. M., Tun, Y. K., Saad, W., Han, Z. & Hong, C. S. (2024). Enhancing Spectrum Efficiency in 6G Satellite Networks: A GAIL-Powered Policy Learning via Asynchronous Federated Inverse Reinforcement Learning. Retrieved from: https://arxiv.org/abs/2409.18718.

He, H. & Jiang, H. (2019). Deep Learning Based Energy Efficiency Optimization for Distributed Cooperative Spectrum Sensing. *IEEE Wireless Communications*, 26(3), 32–39.

Hlophe, M. C., Maharaj, B. & Sande, M. M. (2021). Energy-Efficient Transmissions in Federated Learning-Assisted Cognitive Radio Networks. *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, pp. 216–222.

Hlophe, M. C. & Maharaj, S. B. (2019). Spectrum Occupancy Reconstruction in Distributed Cognitive Radio Networks Using Deep Learning. *IEEE Access*, 7, 14294–14307.

Hu, F., Deng, Y. & Aghvami, H. (2022). Cooperative Multigroup Broadcast 360° Video Delivery Network: A Hierarchical Federated Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications*, 21(6), 4009–4024.

Hussain, F., Hassan, S. A., Hussain, R. & Hossain, E. (2020). Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges. *IEEE Communications Surveys and Tutorials*, 22(2), 1251–1275. doi: 10.1109/COMST.2020.2964534.

Jagannathan, K., Menache, I., Modiano, E. & Zussman, G. (2012). Non-cooperative spectrum access the dedicated vs. free spectrum choice. *IEEE Journal on Selected Areas in Communications*, 30(11), 2251–2261. doi: 10.1109/JSAC.2012.121217.

Jin, C., Allen-Zhu, Z., Bubeck, S. & Jordan, M. I. (2018). Is Q-learning provably efficient? *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS), 4863–4873.

Joseph, K. (2018). *Analysis of Canadian Wireless Spectrum Auctions: Licence Ownership and Deployment in the 700 MHz, 2500 MHz and 3500 MHz Frequency Ranges*. Retrieved from: https://crtc.gc.ca/eng/acrtc/prx/2018joseph.htm.

Kaur, A. & Kumar, K. (2022). Imperfect CSI based intelligent dynamic spectrum management using cooperative reinforcement learning framework in cognitive radio networks. *IEEE Transactions on Mobile Computing*, 21(5), 1672-1683. doi: 10.1109/TMC.2020.3026415.

Kaur, A., Kumar, K., Prakash, A. & Tripathi, R. (2023). Imperfect CSI-based resource management in cognitive IoT networks: A deep recurrent reinforcement learning framework. *IEEE Transactions on Cognitive Communications and Networking*, 9(5), 1271-1281. doi: 10.1109/TCCN.2023.3271144.

Kearns, M. & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2), 209–232.

Khaf, S., Alkhodary, M. T. & Kaddoum, G. (2022). Partially cooperative scalable spectrum sensing in cognitive radio networks under SDF attacks. *IEEE Internet of Things Journal*, 9(11), 8901–8912. doi: 10.1109/JIOT.2021.3116928.

Khaf, S., Kaddoum, G. & Altamimi, M. (2024). HAPS-Assisted Cognitive Radio with UE Capability-Centered Cooperation. *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, pp. 1317–1322.

Khalid, W. & Yu, H. (2019). Spatial–temporal sensing and utilization in full duplex spectrum-heterogeneous cognitive radio networks for the internet of things. *Sensors*, 19(6), 1441. doi: 10.3390/s19061441.

Khan, A. A., Rehmani, M. H. & Rachedi, A. (2016). When Cognitive Radio meets the Internet of Things? *2016 International Wireless Communications and Mobile Computing Conference, IWCMC*, 469–474. doi: 10.1109/IWCMC.2016.7577103.

Khan, A. A., Rehmani, M. H. & Rachedi, A. (2017). Cognitive-Radio-Based Internet of Things: Applications, architectures, spectrum related functionalities, and future research directions. *IEEE Wireless Communications*, 24(3), 17–25.

Khoshkbari, H. & Kaddoum, G. (2023). Deep Recurrent Reinforcement Learning for Partially Observable User Association in a Vertical Heterogenous Network. *IEEE Communications Letters*, 27(12), 3235-3239. doi: 10.1109/LCOMM.2023.3331216.

Khoshkbari, H., Sharifi, S. & Kaddoum, G. (2023). User association in a VHetNet with delayed CSI: A deep reinforcement learning approach. *IEEE Communications Letters*, 27(8), 2257-2261. doi: 10.1109/LCOMM.2023.3291613.

Khoshkbari, H., Kaddoum, G. & Altamimi, M. (2025). A Generalized Deep Hierarchical Reinforcement Learning for HAPS Resource Scheduler. *2025 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 91–96. doi: 10.1109/IWCMC65282.2025.11059542.

Kim, M. & Ko, I.-Y. (2015). An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments. *2015 IEEE International Conference on Web Services*, pp. 543–550.

Koçkaya, K. & Develi, I. (2020). Spectrum sensing in cognitive radio networks: Threshold optimization and analysis. *EURASIP Journal on Wireless Communications and Networking*, 2020(1), 255.

Koike, R., Ariizumi, R. & Matsuno, F. (2023). Simultaneous optimization of discrete and continuous parameters defining a robot morphology and controller. *IEEE Transactions on Neural Networks and Learning Systems*, 1-14. doi: 10.1109/TNNLS.2023.3272068.

Krouka, M., Elgabli, A., Issaid, C. B. & Bennis, M. (2022). Communication-Efficient and Federated Multi-Agent Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking*, 8(1), 311–320.

Lauridsen, M., Krigslund, R., Rohr, M. & Madueno, G. (2018). An empirical NB-IoT power consumption model for battery lifetime estimation. *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1-5. doi: 10.1109/VTCSpring.2018.8417653.

Le, T.-D. & Kaddoum, G. (2021). LSTM-Based Channel Access Scheme for Vehicles in Cognitive Vehicular Networks With Multi-Agent Settings. *IEEE Transactions on Vehicular Technology*, 70(9), 9132–9143.

Lee, D. J. (2015). Adaptive random access for cooperative spectrum sensing in cognitive radio networks. *IEEE Transactions on Wireless Communications*, 14(2), 831–840. doi: 10.1109/TWC.2014.2360857.

Lee, K. (2021). Low-complexity transmit power control for secure communications in wireless-powered cognitive radio networks. *Sensors*, 21(23), 7837. doi: 10.3390/s21237837.

Leslie, D. S. & Collins, E. J. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2), 495–514.

Li, F., Shen, B., Guo, J., Lam, K.-Y., Wei, G. & Wang, L. (2022). Dynamic Spectrum Access for Internet-of-Things Based on Federated Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology*, 71(7), 7952–7956.

Li, H., Xing, X., Zhu, J., Cheng, X., Li, K., Bie, R. & Jing, T. (2017). Utility-Based Cooperative Spectrum Sensing Scheduling in Cognitive Radio Networks. *IEEE Transactions on Vehicular Technology*, 66(1), 645–655. doi: 10.1109/TVT.2016.2532886.

Li, X., Dong, L., Xue, L. & Sun, C. (2023). Hybrid reinforcement learning for optimal control of non-linear switching system. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11), 9161-9170. doi: 10.1109/TNNLS.2022.3156287.

Li, Y., Zhang, W., Wang, C.-X., Sun, J. & Liu, Y. (2020). Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks. *IEEE Transactions on Cognitive Communications and Networking*, 6(2), 464-475. doi: 10.1109/TCCN.2020.2982895.

Liu, X., Sun, C., Zhou, M., Wu, C., Peng, B. & Li, P. (2020). Reinforcement Learning-based Multislot Double-threshold Spectrum Sensing with Bayesian fusion for Industrial Big Spectrum Data. *IEEE Transactions on Industrial Informatics*, 3203(c), 1–1. doi: 10.1109/tii.2020.2987421.

Lo, B. F. (2011). A survey of common control channel design in cognitive radio networks. *Physical Communication*, 4(1), 26-39. doi: https://doi.org/10.1016/j.phycom.2010.12.004.

Lorincz, J., Ramljak, I. & Begušić, D. (2022). Analysis of the impact of detection threshold adjustments and noise uncertainty on energy detection performance in MIMO-OFDM cognitive radio systems. *Sensors*, 22(2), 631. doi: 10.3390/s22020631.

Lotfi, F., Semiari, O. & Saad, W. (2021). Semantic-Aware Collaborative Deep Reinforcement Learning Over Wireless Cellular Networks. *ICC 2022 - IEEE International Conference on Communications*.

Lunden, J., Koivunen, V., Kulkarni, S. R. & Poor, H. V. (2011). Exploiting spatial diversity in multiagent reinforcement learning based spectrum sensing. *2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP 2011*, 325–328. doi: 10.1109/CAMSAP.2011.6136016.

Lunden, J., Kulkarni, S. R., Koivunen, V. & Poor, H. V. (2013). *Journal article n°5. Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks*. IEEE Journal on Selected Topics in Signal Processing.

MacGillivray, C. & Reinsel, D. (2019). Worldwide Global DataSphere IoT Device and Data Forecast, 2019-2023. Retrieved from: Availableathttps://www.idc.com/getdoc.jsp?containerId=prUS45213219(2019/06/18).

Malik, T. S., Malik, K. R., Afzal, A., Ibrar, M., Wang, L., Song, H. & Shah, N. (2023). RL-IoT: Reinforcement Learning-Based Routing Approach for Cognitive Radio-Enabled IoT Communications. *IEEE Internet of Things Journal*, 10, 1836–1847.

Manchón, C. N., Maaz, I., Zecchin, M., Simeone, O., Rita, N. D., Tan, A. S., Yatribi, A., Mestrah, Y., Jagyasi, D. G., Jadoon, M. A., Saini, A., Tehrani, A., Chiskis, A., Verbin, R., Katranaras, E., Adeogun, R., Wang, X., Hakimi, S., Cammerer, S., Valcarce, A., Liu, B., Chiaramello, E., Parazzini, M., Pipaoy, A. & Sarbu, S. (2025). *Towards an AI-native, user-centric air interface for 6G networks: KPIs/KVIs, Testing Methodologies and Benchmarking*.

Mastronarde, N. & Van Der Schaar, M. (2011). Fast reinforcement learning for energy-efficient wireless communication. *IEEE Transactions on Signal Processing*, 59(12), 6262–6266. doi: 10.1109/TSP.2011.2165211.

Mitola, J. & Maguire, G. (1999). Cognitive radio: making software radios more personal. *IEEE Personal Communications*, 6(4), 13-18. doi: 10.1109/98.788210.

Mosleh, S., Ma, Y., Rezac, J. D. & Coder, J. B. (2020). Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and beyond. *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020-May(Ml), 1–7. doi: 10.1109/VTC2020-Spring48590.2020.9129381.

Naous, T., Itani, M., Awad, M. & Sharafeddine, S. (2023). Reinforcement Learning in the Sky: A Survey on Enabling Intelligence in NTN-Based Communications. *IEEE Access*, 11, 19941–19968.

Naseh, D. et al. (2025). Unified Distributed Machine Learning for 6G Intelligent Transportation Systems: A Hierarchical Approach for Terrestrial and Non-Terrestrial Networks. *Network*, 5(3), 41. doi: 10.3390/network5030041.

Ngomane, I., Velempini, M. & Dlamini, S. V. (2018). The detection of the spectrum sensing data falsification attack in cognitive radio ad hoc networks. *2018 Conference on Information Communications Technology and Society, ICTAS 2018 - Proceedings*, 1–5. doi: 10.1109/ICTAS.2018.8368742.

Ni, Q., Zhu, R., Wu, Z., Sun, Y., Zhou, L. & Zhou, B. (2013). Spectrum Allocation Based on Game Theory in Cognitive Radio Networks. *JNW*, 8(3), 712–722.

Ning, W., Huang, X., Yang, K., Wu, F. & Leng, S. (2020). Reinforcement learning enabled cooperative spectrum sensing in cognitive radio networks. *Journal of Communications and Networks*, 22(1), 12–22.

Oksanen, J., Lundén, J. & Koivunen, V. (2012). Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks. *Neurocomputing*, 80, 102–110. doi: 10.1016/j.neucom.2011.07.027.

Oueis, J., Strinati, E. C. & Barbarossa, S. (2015). The Fog Balancing: Load Distribution for Small Cell Cloud Computing. *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1-6. doi: 10.1109/VTCSpring.2015.7146129.

Pala, S., Singh, K., Li, C.-P. & Dobre, O. (2025). Empowering ISAC Systems With Federated Learning: A Focus on Satellite and RIS-Enhanced Terrestrial Integrated Networks. *IEEE Transactions on Wireless Communications*, 24(1), 810–824.

Panayiotou, T., Manousakis, K., Chatzis, S. P. & Ellinas, G. (2019). A Data-Driven Bandwidth Allocation Framework with QoS Considerations for EONs. *Journal of Lightwave Technology*, 37(9), 1853–1864. doi: 10.1109/JLT.2019.2894179.

Pham, C., Tran, N. H., Do, C. T., Moon, S. I. & Hong, C. S. (2014). Spectrum handoff model based on Hidden Markov model in Cognitive Radio Networks. *The International Conference on Information Networking 2014 (ICOIN2014)*, pp. 406-411. doi: 10.1109/ICOIN.2014.6799714.

Quach, T. X., Tran, H., Uhlemann, E., Kaddoum, G. & Tran, Q. A. (2019). Power allocation policy and performance analysis of secure and reliable communication in cognitive radio networks. *Wireless Networks*, 25(4), 1477–1489.

Quan, Z., Cui, S. & Sayed, A. H. (2008). Optimal linear cooperation for spectrum sensing in cognitive radio networks. *IEEE Journal of selected topics in signal processing*, 2(1), 28–40.

Qureshi, M. A., Lagunas, E. & Kaddoum, G. (2023). Reinforcement Learning for Link Adaptation and Channel Selection in LEO Satellite Cognitive Communications. *IEEE Communications Letters*, 27, 951–955.

Rahman, M., Hassan, Z., Reed, J. H. & Liu, L. (2023). Non-terrestrial network (NTN): A novel alternate fractional programming for the downlink channels power allocation. *MILCOM 2023 - 2023 IEEE Military Communications Conference: Communications Supporting Military Operations in a Contested Environment*, pp. 963–968. doi: 10.1109/MILCOM58377.2023.10356339.

Saad, M. M., Tariq, M. A., Khan, M. T. R. & Kim, D. (2024). Non-terrestrial networks: An overview of 3GPP release 17 & 18. *IEEE Internet of Things Magazine*, 7(1), 20–26. doi: 10.1109/iotm.001.2300154.

Sajid, A., Khalid, B., Ali, M., Mumtaz, S., Masud, U. & Qamar, F. (2020). Securing Cognitive Radio Networks using blockchains. *Future Generation Computer Systems*, 108, 816–826. doi: 10.1016/j.future.2020.03.020.

Sami, H., Mourad, A., Otrok, H. & Bentahar, J. (2020). FScaler: Automatic Resource Scaling of Containers in Fog Clusters Using Reinforcement Learning. *2020 International Wireless Communications and Mobile Computing, IWCMC 2020*, 1824–1829. doi: 10.1109/IWCMC48107.2020.9148401.

Sami, H., Mourad, A., Otrok, H. & Bentahar, J. (2021a). Demand-Driven Deep Reinforcement Learning for Scalable Fog and Service Placement. *IEEE Transactions on Services Computing*, 1–14. doi: 10.1109/TSC.2021.3075988.

Sami, H., Otrok, H., Bentahar, J. & Mourad, A. (2021b). AI-based Resource Provisioning of IoE Services in 6G: A Deep Reinforcement Learning Approach. *IEEE Transactions on Network and Service Management*, 1–14. doi: 10.1109/TNSM.2021.3066625.

Sarikhani, R. & Keynia, F. (2020). Cooperative Spectrum Sensing Meets Machine Learning: Deep Reinforcement Learning Approach. *IEEE Communications Letters*, 24(7), 1459–1462. doi: 10.1109/LCOMM.2020.2984430.

Scutari, G. & Pang, J. S. (2013). Joint sensing and power allocation in nonconvex cognitive radio games: Nash equilibria and distributed algorithms. *IEEE Transactions on Information Theory*, 59(7), 4626–4661. doi: 10.1109/TIT.2013.2239354.

Seid, A. M. et al. (2026). A Hierarchical MAFDRL-Based Resource Allocation and Incentive Mechanism for TN-NTN in 6G Networks. *IEEE Transactions on Mobile Computing*, 25(2), 2426–2443. doi: 10.1109/TMC.2025.3608291.

Sharma, H., Kumar, N. & Tekchandani, R. (2023). Mitigating Jamming Attack in 5G Heterogeneous Networks: A Federated Deep Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology*, 72(2), 2439–2452.

Sharma, S. K., Bogale, T. E., Chatzinotas, S., Ottersten, B., Le, L. B. & Wang, X. (2015). Cognitive radio techniques under practical imperfections: A survey. *IEEE Communications Surveys & Tutorials*, 17(4), 1858–1884.

Shi, Z., Gao, W., Zhang, S., Liu, J. & Kato, N. (2020). AI-Enhanced Cooperative Spectrum Sensing for Non-Orthogonal Multiple Access. *IEEE Wireless Communications*, 27(2), 173-179.

Shin, K. S., Hwang, G. H. & Jo, O. (2020). Distributed reinforcement learning scheme for environmentally adaptive IoT network selection. *Electronics Letters*, 56(9), 441–444. doi: 10.1049/el.2019.3891.

Silvirianti & Kaddoum, G. (2025). Quantum Adaptive Learning for Coverage Optimization in LEO Satellite Network. *ICC 2025 - IEEE International Conference on Communications*, pp. 2382-2387. doi: 10.1109/ICC52391.2025.11161401.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J. & Littman, M. L. (2006). PAC Model-Free Reinforcement Learning. *Proceedings of the 23rd International Conference on Machine Learning*, 881–888. doi: 10.1145/1143844.1143955.

Sun, H., Dong, Y., Zhang, Y., Li, X., Wang, J., Zhao, N. & Pan, M. (2022). A cost-efficient skipping based spectrum sensing scheme via reinforcement learning. *IEEE Transactions on Vehicular Technology*, 71(2), 2220–2224. doi: 10.1109/TVT.2021.3136197.

Sun, M., Wang, X., Zhao, C., Li, B., Liang, Y. C., Goussetis, G. & Salous, S. (2018). Adaptive sensing schedule for dynamic spectrum sharing in time-varying channel. *IEEE Transactions on Vehicular Technology*, 67(6), 5520–5524. doi: 10.1109/TVT.2018.2797318.

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning:An Introduction* (ed. 2). The MIT Press.

Tan, C. H., Tan, K. C. & Tay, A. (2011). Dynamic Game Difficulty Scaling Using Adaptive Behavior-Based AI. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(4), 289–301.

Tan, X., Zhou, L., Wang, H., Sun, Y., Zhao, H., Seet, B.-C., Wei, J. & Leung, V. C. M. (2022). Cooperative multi-agent reinforcement-learning-based distributed dynamic spectrum access in cognitive radio networks. *IEEE Internet of Things Journal*, 9(19), 19477-19488. doi: 10.1109/JIOT.2022.3168296.

Tehrani, P., Restuccia, F. & Levorato, M. (2021). Federated Deep Reinforcement Learning for the Distributed Control of NextG Wireless Networks. *International Symposium on Dynamic Spectrum Access Networks (DySPAN)*.

Thilina, K. M., Choi, K. W., Saquib, N. & Nazmus, E. (2013). Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks. *IEEE Journal on Selected Areas in Communications*, 31(11), 2209–2221. doi: 10.1109/JSAC.2013.131120.

Tsitsiklis, J. N. (1994). *Asynchronous Stochastic Approximation and Q-Learning*. doi: 10.1002/art.21867.

Umer, M., Mohsin, M. A., Nasir, A. A., Abou-Zeid, H. & Hassan, S. A. (2025). Intelligent Spectrum Sharing in Integrated TN-NTNs: A Hierarchical Deep Reinforcement Learning Approach. *IEEE Wireless Communications*, 32(3), 64–71.

Verma, S., Kawamoto, Y., Fadlullah, Z. M., Nishiyama, H. & Kato, N. (2017). A Survey on Network Methodologies for Real-Time Analytics of Massive IoT Data and Open Research Issues. *IEEE Communications Surveys and Tutorials*, 19(3), 1457–1477. doi: 10.1109/COMST.2017.2694469.

Vu-Van, H. & Koo, I. (2011). Cooperative spectrum sensing with collaborative users using individual sensing credibility for cognitive radio network. *IEEE Transactions on Consumer Electronics*, 57(2), 320-326. doi: 10.1109/TCE.2011.5955162.

Wu, J., Wang, C., Yu, Y., Song, T. & Hu, J. (2020). Sequential fusion to defend against sensing data falsification attack for cognitive Internet of Things. *ETRI Journal*, 42(August 2019), 976–986. doi: 10.4218/etrij.2019-0388.

Xu, Y., Yu, J., Headley, W. C. & Buehrer, R. M. (2019). Deep Reinforcement Learning for Dynamic Spectrum Access in Wireless Networks. *Proceedings - IEEE Military Communications Conference MILCOM*, 2019-Octob, 207–212. doi: 10.1109/MILCOM.2018.8599723.

Xu, Y., Yang, Y., Liu, Q. & Li, Z. (2020). Joint energy-efficient resource allocation and transmission duration for cognitive HetNets under imperfect CSI. *Signal Processing*, 167, 107309. doi: 10.1016/j.sigpro.2019.107309.

Yang, F., Zhao, Z., Huang, J., Liu, P., Tolba, A., Yu, K. & Guizani, M. (2024). A Federated Reinforcement Learning Approach for Optimizing Wireless Communication in UAV-Enabled IoT Network With Dense Deployments. *IEEE Internet of Things Journal*, 11(20), 33953–33966.

Yang, T., Zhang, W., Bo, Y., Sun, J. & Wang, C.-X. (2023). Dynamic Spectrum Sharing Based on Federated Learning and Multi-Agent Actor-Critic Reinforcement Learning. *International Conference on Wireless Communications and Mobile Computing (IWCMC)*.

Zhang, H., Nie, Y., Cheng, J., Leung, V. C. & Nallanathan, A. (2017, 2). Sensing time optimization and power control for energy efficient cognitive small cell with imperfect hybrid spectrum sensing. *IEEE Transactions on Wireless Communications*, 16(2), 730–743. doi: 10.1109/TWC.2016.2628821.

Zhang, H., Yang, N., Huangfu, W., Long, K. & Leung, V. C. (2020a). Power Control Based on Deep Reinforcement Learning for Spectrum Sharing. *IEEE Transactions on Wireless Communications*, 19(6), 4209–4219. doi: 10.1109/TWC.2020.2981320.

Zhang, W., Yang, Y. & Yeo, C. K. (2015). Cluster-based cooperative spectrum sensing assignment strategy for heterogeneous cognitive radio network. *IEEE Transactions on Vehicular Technology*, 64(6), 2637-2647. doi: 10.1109/TVT.2014.2342216.

Zhang, X. & Su, H. (2011). CREAM-MAC: Cognitive Radio-EnAbled Multi-Channel MAC Protocol Over Dynamic Spectrum Access Networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(1), 110-123. doi: 10.1109/JSTSP.2010.2091941.

Zhang, Y., Wu, Q. & Shikh-Bahaei, M. (2020b). Vertical Federated Learning Based Privacy-Preserving Cooperative Sensing in Cognitive Radio Networks. *2020 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6.

Zhang, Y., Cai, P., Pan, C. & Zhang, S. (2019). Multi-Agent Deep Reinforcement Learning-Based Cooperative Spectrum Sensing With Upper Confidence Bound Exploration. *IEEE Access*, 7, 118898–118906. doi: 10.1109/access.2019.2937108.

Zhang, Z., Wen, X., Xu, H. & Yuan, L. (2018). Sensing nodes selective fusion scheme of spectrum sensing in spectrum-heterogeneous cognitive wireless sensor networks. *IEEE Sensors Journal*, 18(1), 436-445. doi: 10.1109/JSEN.2017.2760925.

Zhou, Q., Wu, S., Jiang, C., Zhang, R. & Jing, X. (2024). Over-the-Air Federated Transfer Learning Over UAV Swarm for Automatic Modulation Recognition in V2X Radio Monitoring. *IEEE Transactions on Vehicular Technology*, 73, 3597–3607.

Zia, K., Javed, N., Sial, M. N., Ahmed, S., Pirzada, A. A. & Pervez, F. (2019). A Distributed Multi-Agent RL-Based Autonomous Spectrum Allocation Scheme in D2D Enabled Multi-Tier HetNets. *IEEE Access*, 7, 6733–6745. doi: 10.1109/ACCESS.2018.2890210.

Zou, Y., Zhu, J., Yang, L., Liang, Y. C. & Yao, Y. D. (2015). Securing physical-layer communications for cognitive radio networks. *IEEE Communications Magazine*, 53(9), 48–54. doi: 10.1109/MCOM.2015.7263345.

Zuo, S., Yan, X., Fan, R., Hu, H., Shan, H., Quek, T. Q. S. & Zhao, P. (2025). Federated Learning Resilient to Byzantine Attacks and Data Heterogeneity. *IEEE Transactions on Mobile Computing*, 24(10), 10729–10742.