الجمهوريـة العربيــة الســورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسـم الاتصالات
العام الدراسي 2025/2024

HIAST
Communication
Department

بحث

أُعدّ لنيل درجة الماجستير في نظم الاتصالات الراديوية والنقالة

استخدام تقنيات التعلّم المعزّز لتحقيق النفاذ الديناميكي للطيف في شبكات الراديو الإدراكي

Dynamic Spectrum Access based on Reinforcement Learning technics in Cognitive Radio Networks

تقديم الطّالب
ميّار منصور

إشراف
د. وسام التبان

2025/11/14

Syrian Arab Republic

Higher Institute for Applied Sciences and Technology

Department of Telecommunication

2024-2025

Dynamic Spectrum Access based on Reinforcement Learning technics in Cognitive Radio Networks

This thesis has been prepared for obtaining a master's degree in communication systems

Prepared by:

Mayyar Mansour

Supervised by:

Wissam Altabban

14/11/2025

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology – HIAST

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدثت بموجب المرسوم التشريعي رقم /24/ لعام 1983 وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي لدراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم الروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

إلى جانب النشاط التعليمي، يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات أطره العلمية ومختبراته.

واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشبكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب 31983

Higher Institute for Applied Sciences and Technology – HIAST, P. O. Box 31983, Damascus, Syrian Arab Republic.

هاتف 00963115123819 - فاكس 00963115140761

بريد إلكتروني contact@hiast.edu.sy

موقع إلكتروني https://hiast.edu.sy/

تصريح

أنا الموقع أدناه **ميار منذر منصور** معدّ أطروحة الماجستير التي تحمل العنوان:

" استخدام تقنيات التعلّم المعزّز لتحقيق النفاذ الديناميكي للطيف في شبكات الراديو الإدراكي "

أصرّح بأن:

- الأعمال والنتائج المعروضة في هذه الأطروحة هي نتيجة جهودي الشخصية وبتوجيه من المشرف، وأن ما عدا ذلك من معلومات ونتائج قد نُسبت إلى مصادرها ومؤلفيها، وأُشير إلى ذلك في متن النص وفي قائمة المراجع.
- المعطيات والمعلومات المستخدمة في هذه الأطروحة جرى تحصيلها بطرائق سليمة ومشروعة ونُسبت إلى مصادرها في المواضع الملائمة.
- كلّ مكوّن من مكونات هذه الأطروحة (مقطع نصي، صورة، مخطط، ...) مقتبس من عمل آخر جرى تمييزه بوضوح ونُسب إلى مصدره.
- الأعمال والنتائج المعروضة في هذه الأطروحة لم تُستخدم سابقاً وليست قيد الاستخدام للحصول على أي شهادة أكاديمية أخرى.

التوقيع

دمشق 2025/11/14

كلمة شكر

أتوجه بجزيل الشكر والامتنان إلى الدكتورة وسام التّبّان المشرفة على إنجاز هذا البحث والتي لم تبخل في تقديم النصائح والملاحظات العلميّة القيّمة والمتابعة والتوجيه المستمر أثناء مراحل سير البحث، وأخصّ المهنيّة العالية التي لطالما عهدتُها منها.

كما أشكر إدارة المعهد العالي وكامل أعضاء الهيئة التدريسيّة في قسم الاتصالات على الدعم والمساعدة لإنجاز هذا العمل.

ملخّص

يُعدّ الراديو الإدراكي CR التقنية التمكينية الرئيسية لحل مشكلة ندرة الطيف الترددي الناتجة عن سياسات التخصيص الثابت. تركّز هذه الأطروحة على النفاذ الديناميكي للطيف DSA من نمط Underlay، الذي يسمح للمستخدمين الثانويين SUs بالنفاذ للطيف بالتزامن مع المستخدمين الأساسيين PUs بشرط الالتزام الصارم بقيود التداخل. ويشكّل تحدي تخصيص الموارد بكفاءة لضمان الأداء الأفضل في ظلّ هذه القيود المعقّدة المشكلةَ الأساسية التي يعالجها هذا البحث.

تتمثل المسألة الأساسية في تجاوز مقاييس الأداء التقليدية والانتقال إلى تعظيم جودة التجربة QoE للمستخدم النهائي، والتي تُقاس كمّياً بـ متوسط درجة الرأي MOS. لمواجهة الطبيعة المتغيّرة للبيئة الراديوية، تم اعتماد منهجية التعلم المعزز RL لقدرتها الفائقة على التعلّم واتخاذ القرار. تمّ في هذا البحث تطبيق ومقارنة خوارزميتين محوريّتين: خوارزمية Q-Learning (QL) التقليدية، وخوارزمية Deep Q-Learning (DQL) التي تدمج الشبكات العصبونية العميقة للتعامل مع حالات تتطلّب سرعة في الوصول إلى الحلّ.

تمّ في هذه الأطروحة اقتراح نموذج تحكّم مركزي تكون فيه المحطة القاعدية الثانوية SBS بمثابة عميل ذكي وحيد Single Agent يتّخذ قرارات التحكّم. هذا العميل مسؤول عن تخصيص عتبات نسبة الإشارة إلى التداخل والضجيج SINR لجميع المستخدمين الثانويين بهدف تعظيم قيمة MOS الإجمالية مع ضمان احترام قيود التداخل والتشوه. أظهرت نتائج المحاكاة والمقارنة مع الأدبيات التفوق الواضح لنموذج DQL المقترح، حيث حقق قيم MOS أعلى (تراوحت بين 4.8 و3.86) وسرعة تقارب أكبر (بين 13.5 و30 تكراراً). وفي المقابل أظهرت خوارزمية QL المقترحة معدّل ازدحام Congestion Rate أقلّ في السيناريوهات ذات الأعداد الكبيرة من المستخدمين (أكثر من 5)، مما يوضّح أن نموذج DQL هو الأنسب لتعظيم جودة التجربة وسرعة الاستجابة.

Abstract

Cognitive Radio (CR) is the main enabling technology to solve the problem of spectrum scarcity resulting from static allocation policies. This thesis focuses on Dynamic Spectrum Access (DSA) of the Underlay type, which allows Secondary Users (SUs) to access the spectrum concurrently with Primary Users (PUs) under the condition of strict adherence to interference constraints. The challenge of efficiently allocating resources to ensure optimal performance under these complex constraints constitutes the primary problem addressed by this research.

The core issue is to move beyond traditional performance metrics and shift towards maximizing the Quality of Experience (QoE) for the end-user, which is quantitatively measured by the Mean Opinion Score (MOS). To cope with the changing nature of the radio environment, the Reinforcement Learning (RL) methodology was adopted for its superior ability to learn and make decisions. In this research, two pivotal algorithms were applied and compared: the traditional Q-Learning (QL) algorithm, and the Deep Q-Learning (DQL) algorithm, which integrates deep neural networks to handle situations requiring speed in reaching the solution.

In this thesis, a central control model was proposed in which the Secondary Base Station (SBS) acts as a single intelligent agent (Single Agent) that makes control decisions. This agent is responsible for allocating Signal to Interference and Noise Ratio (SINR) thresholds for all secondary users with the aim of maximizing the total MOS value while ensuring respect for interference and distortion constraints. The simulation results and comparison with literature showed the clear superiority of the proposed DQL model, as it achieved higher MOS values (ranging between 4.8 and 3.86) and faster convergence speed (between 13.5 and 30 iterations). In contrast, the proposed QL algorithm showed a lower Congestion Rate in scenarios with a large number of users (more than 5), which clarifies that the DQL model is the most suitable for maximizing Quality of Experience and response speed.

فهرس المحتويات

فهرس الأشكال i

فهرس الاختصارات والمصطلحات ii

الفصل الأوّل **مقدّمة** 1

1-1 الرّاديو الإدراكي Cognitive Radio 1

2-1 النفاذ الديناميكي للطيف DSA 1

1-2-1 نموذج الاستخدام الحصري الديناميكي Dynamic Exclusive Use Model 1

2-2-1 نموذج المشاركة المفتوحة Open Sharing Model 2

3-2-1 نموذج الوصول الهرمي Hierarchical Access Model 2

3-1 معايير خاصّة بالراديو الإدراكي 3

4-1 الهدف من البحث 3

5-1 المساهمة العلميّة 4

6-1 تنظيم الأطروحة 4

الفصل الثّاني **الدّراسة المرجعيّة** 5

1-2 مراجعة الأدبيّات 5

2-2 خاتمة 11

الفصل الثالث **الدراسة النظريّة** 13

1-3 مقدّمة (علاقة الراديو الإدراكي بتعلّم الآلة) 13

2-3 تعلّم الآلة 13

3-3 التعلّم المعزّز 14

1-3-3 بنية التعلّم المعزّز 15

2-3-3 تصنيفات خاصّة بالتعلّم المعزّز 16

3-3-3 عمليات ماركوف المحدودة لاتخاذ القرار 17

4-3-3 خوارزميّة Q-Learning 20

5-3-3 خوارزميّة Deep Q-Learning 21

4-3 معيار MOS 22

5-3 خاتمة 23

الفصل الرّابع **النمذجة** 25

1-4 نموذج العمل 25
2-4 معايير الأداء 27
3-4 سيناريو العمل 27
4-4 خوارزميّة Q-Learning 28
1-4-4 العميل Agent 28
2-4-4 فضاء الأفعال Action Space 28
3-4-4 فضاء الحالات State Spaces 29
4-4-4 المكافأة Reward 29
5-4-4 السياسة Policy 30
5-4 خوارزميّة Deep Q-Learning 30
1-5-4 بارامترات التعلّم في DQN 30
2-5-4 تخزين سلاسل الانتقال 31
6-4 خاتمة 31
الفصل الخامس **النتائج** 33
1-5 نتائج استخدام خوارزميّة Q-Learning 33
1-1-5 معيار MOS 33
2-1-5 معدّل الازدحام Congestion Rate 34
3-1-5 عدد التكرارات اللّازمة للوصول إلى التقارب 34
2-5 دراسة تماسك النموذج 35
1-2-5 أثر تغيّر عدد المستخدمين الثانويين SUs 35
2-2-5 أثر تغيّر المساحة الجغرافيّة المخصّصة للمستخدمين الثانويين SUs 37
3-5 نتائج استخدام خوارزميّة Deep Q-learning 39
1-3-5 معيار MOS 39
2-3-5 معدّل الازدحام Congestion Rate 39
3-3-5 عدد التكرارات اللازمة للوصول إلى التقارب 40
4-5 مقارنة النتائج 41
1-4-5 معيار MOS 41
2-4-5 معدّل الازدحام Congestion Rate 41
3-4-5 عدد التكرارات اللازمة للوصول إلى التقارب 42

5-5 خاتمة 42
الخاتمة والآفاق المستقبليّة 43
المراجع 45

فهرس الأشكال

الشكل 1: يبيّن مخطّط لتصنيف نماذج النفاذ الديناميكي للطيف.[11] 2
الشكل 2: مقارنة بين نتائج البحثين من حيث قيمة MOS. 8
الشكل 3: مقارنة بين نتائج البحثين من حيث معدّل ازدحام الشبكة. 9
الشكل 4: مقارنة بين نتائج البحثين من حيث عدد التكرارات للوصول إلى الحلّ. 9
الشكل 5: يبيّن دورة الإدراك في الراديو الإدراكي.[12] 14
الشكل 6: عناصر التعلّم المعزّز.[8] 17
الشكل 7: يوضّح بنية خوارزميّةDQL.[9] 21
الشكل 8: يوضّح الفرق بين QL وDQL من حيث البنية الأساسيّة.[18] 22
الشكل 9: يوضّح نموذج العمل. 25
الشكل 10: تغيّر منحني قيم MOS بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL. 33
الشكل 11: تغيّر منحني معدّل الازدحام بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL. 34
الشكل 12: تغيّر منحني عدد التكرارات اللازمة للتقارب بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL. 34
الشكل 13: تماسك النموذج من حيث معدّل الازدحام في حالة عدد SUs كبير. 35
الشكل 14: تماسك النموذج من حيث قيم MOS في حالة عدد SUs كبير. 36
الشكل 15: تماسك النموذج من حيث عدد التكرارات اللازمة للتقارب في حالة عدد SUs كبير. 36
الشكل 16: تماسك النموذج من حيث قيم MOS في حالة مساحة جغرافيّة أكبر. 37
الشكل 17: تماسك النموذج من حيث معدّل الازدحام في حالة مساحة جغرافيّة أكبر. 38
الشكل 18: تماسك النموذج من حيث عدد التكرارات اللازمة للتقارب في حالة مساحة جغرافيّة أكبر. 38
الشكل 19: تغيّر منحني قيم MOS بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL. 39
الشكل 20: تغيّر منحني معدّل الازدحام بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL. 40
الشكل 21: تغيّر منحني عدد التكرارات اللازمة للتقارب بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL. 40
الشكل 22: مقارنة النتائج من حيث معيار MOS. 41
الشكل 23: مقارنة النتائج من حيث معدّل الازدحام. 41
الشكل 24: مقارنة بين QL وDQL من حيث عدد التكرارات للوصول إلى التقارب. 42

فهرس الاختصارات والمصطلحات

الاختصار	المصطلح
CR	Cognitive Radio
QoE	Quality of Experience
SU	Secondary Users
PU	Primary User
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
DDPG	Deep Deterministic Policy Gradient
RADDPG	Resource Allocation DDPG
QL	Q-Learning
DQN	Deep Q Network
DQL	Deep Q Learning
MOS	Mean Opinion Score
CSI	Channel State Information
CE	Cognitive Engine
DSA	Dynamic Spectrum Access
SINR	Signal to Interference and Noise Ratio
V2V	Vehicle to Vehicle
TVWS	TV White Spaces
NB-IOT	Narrowband Internet of Things
SNR	Signal to Noise Ratio
SN	Secondary Network
CBS	Cognitive Base Station
TDL	Temporal-Difference Learning
PN	Primary Network

الفصل الأوّل

مقدّمة

في هذ الفصل تُشرح المفاهيم الأساسيّة المستخدمة في البحث والمبدأ النظري للخوارزميّات المستخدمة وتُذكر بعض المعايير الدوليّة والغاية من البحث وتنظيم الأطروحة.

يواجه قطاع الاتصالات اللاسلكية تحدياً يتمثّل في الاعتقاد الشائع بأن الترددات الراديوية القابلة للاستخدام أوشكت على النفاد، وقد تعزّز هذا الاعتقاد بفعل الازدحام الشديد في مخططات تخصيص الترددات والأسعار الباهظة التي سُجلت في مزادات الطيف. لكن المفارقة تكمن في أن القياسات الفعلية لاستخدام الطيف تشير إلى أن أجزاء كبيرة من هذا الطيف تبقى غير مُستغلّة (خاملة) في أي لحظة زمنية وموقع جغرافي معيّن. تكشف هذه المفارقة أن الإحساس بالندرة لا يعود إلى شحّ فيزيائي في الموارد الترددّية بل ينبع بشكل أساسي من سياسات إدارة الطيف المتّبعة. إنّ وجود نطاقات ترددّية خاملة هو أمر حتميّ في ظل وجود سياسة تخصيص الطّيف الثابتة Static Spectrum Allotment التي تمنح الاستخدام الحصري للمستخدمين المرخص لهم. من هذا المنطلق برزت الحاجة الماسّة لتقنيّات وسياسات جديدة لإدارة الطيف بكفاءة أعلى، ومنها تقنيّة الراديو الإدراكي (CR: Cognitive Radio) الذي أتاح استخدام نموذج النفاذ الديناميكي للطيف (DSA: Dynamic Spectrum Access).

1-1 الرّاديو الإدراكي Cognitive Radio

تُعد تقنية الراديو الإدراكي Cognitive Radio التقنية التمكينيّة الرئيسية للنفاذ الديناميكي للطيف DSA، وهي من بين التقنيات الأساسيّة الفعّالة في أنظمة الاتصالات اللّاسلكية. على عكس الراديو التقليدي المقتصر على العمل فقط في نطاقات طيفية محدّدة يمتلك الراديو الإدراكي CR القدرة على العمل في نطاقات طيفيّة مختلفة بفضل قدرته على استشعار بيئته اللاسلكية وفهمها والتعلّم من التجارب السابقة. تتوفّر هذه الميزات للراديو الإدراكي CR بواسطة حزمة برمجيات ذكية تسمى المحرّك الإدراكي (CE: Cognitive Engine). يدير المحرك الإدراكي CE موارد الراديو لإنجاز الوظائف الإدراكية، ويقوم بتخصيص ومواءمة موارد الراديو لتحسين أداء الشبكة. يزداد الاهتمام بالراديو الإدراكي بشكل كبير، وتتطوّر هذه التقنيّة بقفزات هائلة نتيجةً لكونه حلاً لمشكلة الاستثمار غير الفعّال للطيف الترّددي.

2-1 النفاذ الديناميكي للطيف DSA

النفاذ الديناميكي للطيف أحد تطبيقات الرّاديو الإدراكي CR، ويتمّ من خلاله مشاركة الطّيف من قبل المستخدمين غير المرخّص لهم (المستخدمين الثانويّين SUs) مع المستخدمين المرخّص لهم (المستخدمين الأساسيّين PUs) وفق قواعد وشروط يحدّدها نموذج النفاذ الديناميكي للطيف. إنّ قدرة الراديو الإدراكي على الاستشعار والتعلّم والتكيّف هي التي تجعل تطبيق النفاذ الديناميكي للطيف ممكناً، مما يسمح باستغلال الطيف بكفاءة أعلى. يمكن تصنيف هذه النماذج ضمن ثلاثة نماذج أساسية:

1-2-1 نموذج الاستخدام الحصري الديناميكي Dynamic Exclusive Use Model

يحافظ هذا النموذج على الهيكل الأساسي للسياسة التنظيمية الحالية للطيف، حيث يتم ترخيص حزم الطيف لخدمات معينة للاستخدام الحصري. الفكرة الرئيسية هنا هي إدخال المرونة لتحسين كفاءة الطيف. ويشمل هذا النموذج مقاربتين:

- **حقوق ملكية الطيف (Spectrum Property Rights):** تسمح هذه المقاربة للمرخص لهم ببيع الطيف والمتاجرة به، واختيار التقنية التي يريدون استخدامها بحرية. بهذا، تلعب قوى الاقتصاد والسوق دوراً أهم في الدفع نحو الاستخدام الأكثر ربحية لهذا المورد المحدود.
- **التخصيص الديناميكي للطيف (Dynamic Spectrum Allocation):** تهدف هذه المقاربة إلى تحسين كفاءة الطيف من خلال تخصيص الطيف بشكل ديناميكي، وذلك عبر استغلال الإحصائيات المكانية والزمانية للحركة المرورية traffic statistics للخدمات المختلفة. أي أنه في منطقة معينة ووقت معين، يتم تخصيص الطيف للخدمات بشكل حصري، لكن هذا التخصيص يتغير بمعدل أسرع بكثير من السياسة الحالية.[11]

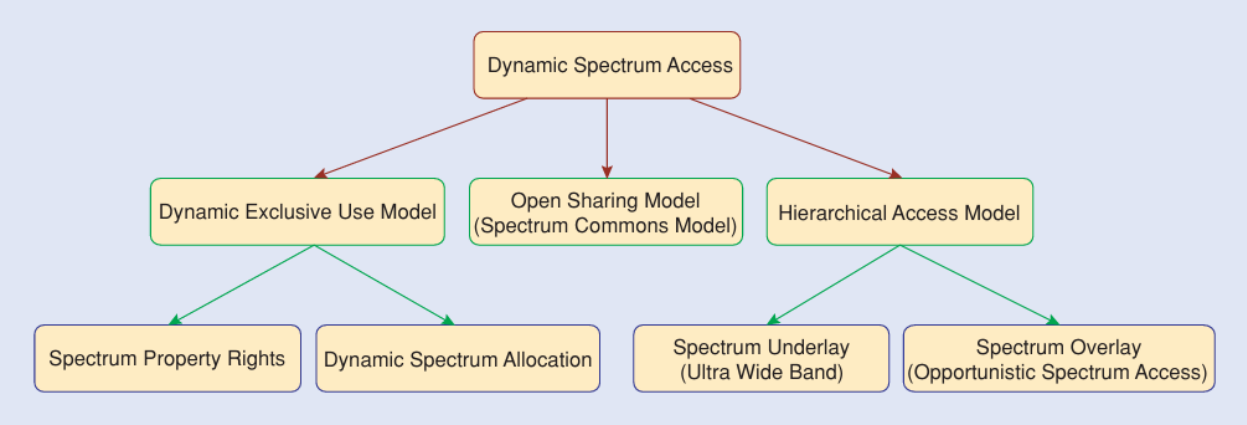
1-2-2 نموذج المشاركة المفتوحة Open Sharing Model

يُشار إلى هذا النموذج أيضاً باسم مشاعات الطيف Spectrum Commons. يعتمد هذا النموذج على المشاركة المفتوحة بين المستخدمين الأقران peer users كأساس لإدارة منطقة طيفيّة معيّنة. يستمد مؤيّدو هذا النموذج دعمهم من النجاح الهائل للخدمات اللاسلكية العاملة في النطاقات الراديوية الصناعية والعلمية والطبية غير المرخصة، مثل شبكات WIFI.[11]

1-2-3 نموذج الوصول الهرمي Hierarchical Access Model

يتبنّى هيكليّة نفاذ هرمي يتضمن وجود نوعين من المستخدمين: المستخدمون الأساسيّون (PU: Primary Users) وهُم المستخدمون المرخّص لهم (ترجع ملكيّة الطّيف لهم)، والمستخدمون الثانويّون (SU: Secondary Users) وهُم مستخدمون لا يمتلكون ترخيص لاستخدام الطيف. يُعد هذا النموذج الهرمي هو النموذج الأكثر توافقاً مع سياسات إدارة الطيف الحالية، ويهدف إلى إمكانيّة مشاركة الطّيف المرخّص من قبل المستخدمين الثانويين، وهناك مقاربتان رئيسيّتان ضمن هذا النموذج:

- **Spectrum Underlay:** تفرض هذه المقاربة قيوداً شديدة على استطاعة إرسال المستخدمين الثانويين بحيث تبقى SINR الخاصّة بكل مستخدم ثانوي SU أقلّ من حدّ معيّن، وذلك لمنع التداخل مع المستخدم الأوّلي PU ومع المستخدمين الثانويّين SUs.
- **Spectrum Overlay:** تستهدف هذه المقاربة بشكل مباشر الفجوات المكانية والزمانية في الطيف. هي لا تفرض بالضرورة قيوداً شديدة على استطاعة إرسال المستخدمين الثانويين SUs بل تفرض شروط على مكان وزمان الإرسال.[11]



الشكل 1: يبيّن مخطّط لتصنيف نماذج النفاذ الديناميكي للطيف.[11]

3-1 معايير خاصّة بالراديو الإدراكي

نظراً لأهميّة تقنية الراديو الإدراكي كحلّ واعد لمشكلة ندرة الطيف وتحسين كفاءة استخدامه عملت العديد من الهيئات الدولية على وضع معايير لتنظيم وتطوير هذه التقنية وضمان التشغيل البينيّ Interoperability والتعايش Coexistence بين الأنظمة اللّاسلكيّة المختلفة. من أبرز هذه الهيئات معهد مهندسي الكهرباء والإلكترونيات IEEE والمعهد الأوروبي لمعايير الاتصالات ETSI وقطاع الاتصالات الراديوية في الاتحاد الدولي للاتصالات ITU-R.

يُعدّ معيار IEEE 802.22 الخاص بشبكات المنطقة الإقليمية اللاسلكية WRAN من أبرز المعايير التطبيقية للراديو الإدراكي. يهدف هذا المعيار إلى توفير وصول لاسلكي عريض الحزمة الطيفيّة في المناطق الريفية عبر استغلال الفراغات الطيفية (TVWS: TV White Spaces) في حزم البث التلفزيوني VHF وUHF. يحدد المعيار متطلبات صارمة لتحسّس الطيف لضمان حماية المستخدمين الأوليين (مثل محطات التلفاز والميكروفونات اللاسلكية)، حيث يشترط احتمال كشف لا يقل عن 0.9 واحتمال إنذار خاطئ لا يزيد عن 0.1 لإشارات التلفاز.[13]

بالإضافة إلى ذلك تعمل لجنة معايير النفاذ الديناميكي للطيف في IEEE (DySPAN-SC، التي كانت تُعرف سابقاً بـ SCC41) على تطوير سلسلة المعايير IEEE 1900 التي تضع الأسس النظرية والتطبيقية للراديو الإدراكي والنفاذ الديناميكي للطيف DSA. تتضمن هذه السلسلة معايير هامة مثل:

- معيار IEEE 1900.1: يوحد المصطلحات والمفاهيم الأساسية.
- معيار IEEE 1900.2: يقدم ممارسات موصى بها لتحليل التداخل والتعايش بين الأنظمة الراديوية المختلفة.
- معايير أخرى في السلسلة تغطي موضوعات حيوية مثل واجهات تحسّس الطيف (IEEE 1900.6)، وهيكلية أنظمة الراديو القابلة لإعادة التشكيل (IEEE 1900.7).[14]

كما تمّ تكييف بعض معايير IEEE 802 الأخرى لدمج ودعم جوانب من الراديو الإدراكي مثل معيار IEEE 802.11af: يُمكن شبكات WIFI من العمل في الفراغات الطيفية التلفزيونية TVWS.[15]

على الصعيد الأوروبي، ينشط المعهد الأوروبي لمعايير الاتصالات ETSI من خلال لجنته التقنية للأنظمة الراديوية القابلة لإعادة التشكيل (TCRRS: Technical Committee for Reconfigurable Radio Systems). يركز ETSI على تطوير معايير للراديو الإدراكي تتوافق مع البيئة التنظيمية الأوروبية، بما في ذلك معايير تشغيل الأجهزة في TVWS (مثل EN 301 598) وسياسات إدارة الطيف مثل الوصول المرخص المشترك (LSA: Licensed Shared Access).[17]

أما قطاع الاتصالات الراديوية في الاتحاد الدولي للاتصالات ITU-R فيساهم من خلال نشر تقارير وتوصيات فنية حول أنظمة الراديو الإدراكي والراديو المعرف برمجياً SDR، مما يساعد في تنسيق الجهود العالمية وتوفير إرشادات للدول الأعضاء في هذا المجال.[16]

تهدف هذه المعايير بمجملها إلى تحقيق التشغيل البيني بين أجهزة ومكوّنات الراديو الإدراكي المختلفة، وضمان الاستخدام الفعال والمشترك للطيف الترددي المحدود، وتقليل التداخل المحتمل بين الأنظمة اللاسلكية المتنوعة، مما يمهد الطريق لانتشار أوسع وأكثر تنظيماً لتقنيات الراديو الإدراكي وتحقيق فوائدها المرجوّة.

4-1 الهدف من البحث

يكمن الهدف من هذا البحث في دراسة وتوظيف تقنيات التعلم المعزز لتحسين إدارة الموارد وتخصيصها في سياق النفاذ الديناميكي للطيف DSA ضمن شبكات الراديو الإدراكي CR. يركّز البحث بشكل خاص على معالجة مشكلة تخصيص الموارد

للمستخدمين الثانويين SUs العاملين وفق نمط المشاركة Underlay، وذلك بهدف تعظيم جودة التجربة QoE الإجمالية للشبكة والتي تُقاس باستخدام معيار متوسط درجة الرأي MOS مع الالتزام الصّارم بقيود التداخل المفروضة لحماية المستخدم الأولي PU وضمان عدم حدوث تشوّه. يهدف البحث إلى إثبات فعالية النموذج المقترح وتحقيق تحسين في مقاييس الأداء مقارنة بالدراسات المرجعية، مما يقدّم حلاً عملياً وفعّالاً لإدارة الطّيف الديناميكية في شبكات الراديو الإدراكي الحديثة.

5-1 المساهمة العلميّة

قدّم هذا البحث المساهمات العلميّة التّالية:

- دراسة النفاذ الديناميكي للطيف في شبكات الراديو الإدراكي من نمط Underlay دراسة مرجعيّة تشمل المتطلّبات والصعوبات والتقنّيات المستخدمة. انتهت هذه الدراسة بتصنيف لنتائج مختلف خوارزميّات التعلّم المعزّز التي ذُكرت في الأدبيّات.
- اقتراح نموذج تحكّم مركزي يعتمد على تعلّم معزّز بعميل وحيد متمثّلاً بالمحطّة القاعديّة الثانويّة. هذه المقاربة تسعى لمنع التعقيدات المرتبطة بتضارب السّياسات الذي قد يحدث عند تعلّم عدّة عملاء بشكل متزامن.
- تحسين أداء خوارزميّة Q-Learning المقترحة من حيث قيمة MOS ومعدّل الازدحام وعدد التكرارات الوسطي اللازمة للتقارب.
- تحسين أداء خوارزميّة Deep Q-Learning المقترحة من حيث قيمة MOS وعدد التكرارات الوسطي اللازمة للتقارب.

6-1 تنظيم الأطروحة

تُنظّم هذه الأطروحة في خمسة فصول. يتناول الفصل الأوّل المفاهيم الأساسيّة المستخدمة في البحث والمبدأ النظري للخوارزميات المستخدمة، كما تُذكر بعض المعايير الدولية والهدف من البحث. ثم في الفصل الثاني دراسة مرجعية للعديد من الأبحاث التي تطرّقت لمشكلة البحث أو التي استخدمت خوارزميات مشابهة، وتُستعرض الحلول المقدمة في كل بحث. في الفصل الثالث يتمّ توضيح الدراسة النظرية وتعريف تعلّم الآلة وتصنيفاته مع شرح نظري للخوارزميّات المستخدمة في البحث، وتحديداً التعلّم المعزّز وخوارزميّتي Q-Learning وDeep Q-Learning. يلي ذلك الفصل الرابع الذي يقدّم نمذجة المسألة، ويتضمّن شرح نموذج العمل ومعايير الأداء بالإضافة إلى سيناريو العمل، وتطبيق خوارزميّات التعلّم المعزّز المستخدمة وعناصرها. أخيراً في الفصل الخامس يتمّ عرض النتائج التي تمّ الحصول عليها ومناقشتها، والمقارنة بينها ونتائج الخوارزميات في الأدبيّات.

7-1 خاتمة

عرض هذا الفصل مقدّمة عامّة عن مشكلة ندرة الطيف الترددي الناجمة عن سياسات التخصيص الثابت، ودور تقنية الراديو الإدراكي CR كحلٍّ فعّال لرفع كفاءة استغلال الطيف. ثم استعرض المفهوم الأساسي للنفاذ الديناميكي للطيف DSA وصنّف نماذجه الرئيسية، بما في ذلك نموذج الوصول الهرمي Hierarchical Access Model ومقاربتيه Spectrum Underlay وSpectrum Overlay. إضافة إلى ذلك تطرّق إلى أبرز المعايير الدولية التي تضع أسس عمل الراديو الإدراكي مثل معايير IEEE وETSI. وأخيراً حدد الهدف الرئيسي من هذه الأطروحة مع المساهمات العلمية المقدّمة، وبيّن تنظيم فصول الأطروحة.

الفصل الثّاني
الدّراسة المرجعيّة

يقدّم الفصل دراسة مرجعيّة للعديد من الأبحاث التي تطرّقت لنفس مشكلة البحث أو التي استخدمت خوارزميّات البحث، ويستعرض الحلول المقدّمة لكلّ بحث.

1-2 مراجعة الأدبيّات

يقدّم [6] مفهوم التعلّم التعاوني Cooperative Learning لتعظيم جودة التجربة QoE المقاسة بـ MOS لحركة مرور غير متجانسة (بيانات وفيديو) مع الالتزام بقيود التّداخل المفروضة، حيثُ يهدف إلى تطوير خوارزمية لتخصيص الموارد تعتمد على التعلم المعزز في شبكات الراديو الإدراكي من نمط Underlay. اعتمد البحث على خوارزميات Q, DQL في بيئة متعددة العملاء (multi-agent)، وقدّم مفهوم التعلم التعاوني Cooperative Learning كآلية لتسريع عملية تعلّم العملاء الجدد (المستخدمين الثانويين SUs) الذين ينضمّون إلى الشبكة بالاستفادة من خبرة العملاء الموجودين مسبقاً (مشابه لفكرة Transfer Learning). كل مستخدم ثانوي SU يعمل كعميل تعلم معزز حيث يحاول اختيار أفضل عتبة لنسبة الإشارة إلى التداخل والضجيج SINR من مجموعة خيارات منفصلة لتعظيم مكافأته. آلية التعلم التعاوني تسمح للعميل الجديد (المستخدم الثانوي المضاف حديثاً) ببدء عملية التعلم من نقطة متقدمة بناءً على سياسات العملاء الآخرين (المستخدمين الثانويين). أظهرت النتائج أن استخدام التعلم التعاوني يقلل بشكل كبير من عدد التكرارات اللازمة للوصول إلى سياسة مستقرة مقارنة بالتّعلم الفردي دون التأثير سلباً على متوسط MOS المحقّق في الشبكة. كما أظهرت تفوق DQL على Q-learning من حيث سرعة التقارب. تكمن المساهمة الرئيسية لهذا البحث في إدخال آلية التعلم التعاوني الفعال لتسريع أداء خوارزميات التعلم المعزز الموزّعة في شبكات الرّاديو الإدراكي وخاصّة في السّيناريوهات الديناميكية.

عمل البحث [7] على معالجة مشكلة قصور الطيف الترددي المخصص لتقنية إنترنت الأشياء ضيقة النطاق Narrowband Internet-of-Things - NB-IoT المحدود (180-200 KHz) في استيعاب الزيادة الهائلة المتوقعة في عدد أجهزة NB-IoT. تمّ اقتراح آلية لتخصيص الموارد بكفاءة عبر السماح لأجهزة NB-IoT بالوصول الانتهازي للطيف المرخّص الشّاغر مع التركيز على تقليل عدد الإرسالات المتكررة واستيعاب عدد أكبر من الأجهزة. اقترح الباحثون تقنية أطلقوا عليها اسم NB-Cognitive Radio-IoT (NB-CR-IoT)، والتي تدمج قدرات الراديو الإدراكي CR في تشغيل شبكات NB-IoT التقليدية. يتم صياغة مشكلة تخصيص الموارد كعملية قرار ماركوف (MDP: Markov Decision Process) وحلها باستخدام خوارزمية DQL. تفترض بنية NB-CR-IoT المقترحة أن أجهزة NB-IoT لديها قدرات إدراكية لاستشعار الطيف وتحديد القنوات المرخّصة الشاغرة للوصول إليها بشكل انتهازي، ويتم استخدام DQN لاتخاذ قرارات تخصيص الموارد مثل اختيار القناة أو تحديد الحاجة للإرسال المتكرر. تُعتبر خوارزميّة DQL مناسبة لأنها تستطيع التعامل مع فضاء الحالات والإجراءات كبير الأبعاد بشكل أفضل من خوارزميّة QL التقليدية. أظهرت نتائج المحاكاة أن خوارزمية DQL تتفوق بشكل واضح على خوارزمية QL التقليدية في سياق NB-CR-IoT المقترح، حيثُ حققت DQN درجة رضا satisfaction degree أعلى للمستخدمين ومعدلات بت قابلة للتحقيق achievable bit rates أفضل بتحسّن قدره 7.41% مقارنة بـ Q-learning مع زيادة عدد أجهزة NB-CR-IoT في الشبكة.

في [3] يوجد مشكلة النفاذ الديناميكي للطيف DSA في شبكة موزعة متعددة الخلايا multi-cell ومتعددة المستخدمين MU-MIMO حيث لا يوجد متحكم مركزي لإدارة الطيف. يهدف البحث إلى تحقيق هدف مزدوج: تعظيم جودة الخدمة لمستخدمي الشبكة المرخصة Licensee Network (المستخدمين الأوليين)، وفي نفس الوقت تجنب التداخل على الشبكة الأساسية Incumbent Network. لتحقيق ذلك، يقترح الباحثون إطار عمل يعتمد على التعلّم المعزّز متعدد العملاء Multi-Agent RL حيث تعمل كل خلية في الشبكة المرخّصة كعميل مستقل. يقوم البحث بتحليل ومقارنة أداء أربع خوارزميات تعلم معزز مختلفة هي Q-Learning, DQN, DDPG, TD3 (Twin Delayed Deep Deterministic). أظهرت النتائج أنه في الشبكات

الصغيرة ذات فضاء الحالات والأفعال المحدود تقدم خوارزمية Q-learning أداءً شبه أمثلي حيث تحقّق معدل نقل بيانات مرتفع نسبياً وتحافظ على التداخل ضمن الحدود الآمنة. بينما في الشبكات الأكبر والأكثر تعقيداً يتفوق أداء خوارزميات التعلم العميق مثل DDPG التي تمكنت من تحقيق أعلى متوسط لمعدل نقل البيانات. ومع ذلك، أظهرت خوارزميات DDPG وTD3 سلوكاً غير مستقراً في تخصيص الاستطاعة في المراحل الأولى من التعلّم.

استخدم [4] خوارزميّة DDPG الموزّعة في بيئة اتصالات من مركبة إلى مركبة (V2V) متّبعاً نهجاً لا مركزيّاً متعدّد العملاء (Distributed Multi-agent) حيث تتّخذ كل مركبة قراراتها بنفسها، وهذا ما يتيح قابليّة التوسّع Scalability. يوضح البحث كيف يمكن لخوارزمية DDPG تحسين كفاءة أخذ العينات وسرعة الوصول للحل في الأنظمة التي تحتوي على عدد كبير من المستخدمين.

يعالج [5] هذا البحث مشكلة استشعار الطيف Spectrum Sensing للحصول على معلومات حالة القناة CSI، وتعتبر هذه العملية عمليّة مكلفة وتستهلك الموارد. لذلك يقترح البحث إطار عمل لتحسين عمليتي الاستشعار وتخصيص الموارد بشكل مشترك jointly بدلاً من التعامل مع كل منهما على حدة. الهدف المزدوج هو تعظيم أداء شبكة الراديو الإدراكي وفي نفس الوقت حماية المستخدم الأساسي (الأوّلي) عبر شرط للتداخل. إن قابلية المراقبة الجزئية Partial observability لحالة القناة (بسبب المعلومات المشوشة والمتقادمة) تستدعي استخدام مقدّرات تتابعية لتتبع مكاسب قناة التداخل، بالإضافة إلى استخدام أدوات البرمجة الديناميكية لتصميم المخططات المُثلى. وإلى جانب المخططات المُثلى. يُبرز هذا البحث أهميّة تكامل عملية الاستشعار مع عملية اتخاذ القرار. في نماذج التعلم المعزز يمكن تمثيل عدم دقة الاستشعار كجزء من حالة البيئة State أو دالة المكافأة Reward لجعل النموذج أكثر واقعيّة.[5]

يعالج [2] مشكلة تخصيص الموارد في شبكة راديو إدراكي من نمط Underlay لتعظيم جودة التجربة QoE المقاسة باستخدام مقياس MOS. يركّز بشكل أساسي على استخدام خوارزميّة DQN في بيئة متعدّدة العملاء Multi-Agent، حيثُ يمثّل كل مستخدم ثانوي SU عميلاً مستقلاً. الإضافة الأكثر أهميّة في هذا البحث هي فكرة التعلّم بالنقل Transfer Learning التي تم استخدامها لتسريع عمليّة تعلّم المستخدمين الجدد الذين ينضمون إلى الشبكة.

اقترح البحث المنهجيّة التالية:

- **النموذج الأساسي (Multi-Agent)**
  يعتمد هذا النموذج نظاماً لا مركزيّاً حيث كل مستخدم ثانوي SU لديه شبكة DQN خاصة به بدلاً من وجود شبكة DQN مركزيّة لجميع المستخدمين. يتعلّم كل مستخدم ثانوي SU بشكل مستقل عن بقية المستخدمين السّياسة المثلى لاختيار أفضل عتبة SINR من فضاء عيّنة متقطّع من أجل تعظيم قيمة MOS. يعتبر كل SU باقي المستخدمين كجزء من بيئة التعلّم المعزّز التي يتفاعل معها، وهو تبسيط شائع في الأنظمة متعددة العملاء.
- **تقنيّة التعلّم بالنقل (Transfer Learning)**
  تتيح تقنية التعلّم بالنقل في هذا النموذج مشاركة الخبرة بين المستخدمين الثانويين، فبما أنّ المستخدمين الثانويين المتقاربين جغرافياً يواجهون بيئة لاسلكية متشابهة فإن سياساتهم المثلى ستكون متشابهة. عندما ينضم مستخدم ثانوي جديد إلى الشبكة يقوم بنسخ أوزان شبكة DQN من أقرب مستخدم ثانوي قديم (جار له) كنقطة بداية بدلاً من أن يبدأ عملية التعلم من الصفر (بأوزان عشوائية لشبكته). هذه العملية تسمى نقل الخبرة أو التعلّم بالنقل، وهي تقلل بشكل كبير من الوقت الذي يحتاجه المستخدم الجديد للوصول إلى سياسة فعالة.

ساهم هذا البحث في تطبيق التعلّم بالنقل في سياق شبكات الراديو الإدراكي، ويوضّح البحث أنّ الخبرة التي اكتسبها عملاء DRL (في هذه الحالة المستخدمون الثانويون) يمكن نقلها بفعاليّة لتسريع أداء النظام بشكل كبير، وخاصّة في البيئات الديناميكية التي ينضم إليها ويغادرها المستخدمون باستمرار.

أظهرت النتائج أن استخدام التعلّم بالنقل قلّل من عدد التكرارات اللازمة للوصول إلى الحلّ الأمثل بنسبة تصل إلى 72% مقارنة بخوارزمية Q-Learning التقليدية وبنسبة 25% مقارنة بخوارزمية DQN التي لا تستخدم التعلم بالنقل ومن دون أي تضحية في أداء متوسط قيم MOS النهائي.[2]

يقدّم [1] حلّاً لمشكلة الولوج الديناميكي للطيف (DSA: Dynamic Spectrum Access) في شبكات الراديو الإدراكي Cognitive Radio من نمط Underlay على الوصلة الصاعدة Uplink. الهدف الأساسي هو تعظيم جودة التجربة QoE لمجموعة من مستخدمي بيانات الوسائط المتعدّدة Multimedia (المستخدمين الثانويين SUs) مع ضمان إبقاء التداخل مع المستخدم الأوّلي (الأساسي) ضمن الحدّ المسموح. لتحقيق ذلك، اقترح الباحثون خوارزميّة (DDPG: Deep Deterministic Policy Gradient) وهي إحدى خوارزميّات التعلّم المعزّز العميق DRL، وأطلقوا عليها (RADDPG: Resource Allocation DDPG). في شبكات الراديو الإدراكي من نمط Underlay يقوم المستخدمون الثانويون SUs بمشاركة الطيف الترددي مع المستخدم الأوّلي PU بشكل متزامن بشرط ألّا يتجاوز التداخل حدّاً معيّناً. المشكلة تكمن في تخصيص الموارد (مثل استطاعة الإرسال) بشكل ديناميكي لكلّ SU وتحقيق شروط التداخل مع المستخدم الأوّلي PU وبين المستخدمين الثانويين SUs أيضاً لضمان أفضل جودة تجربة، والتي تقاس بمعيار (MOS: Mean Opinion Score).

الخوارزميّة المقترحة (RADDPG: Resource Allocation Deep Deterministic Policy Gradient) وهي عبارة عن إطار عمل هجين من مرحلتين يدمج بين شبكة Q العميقة (DQN: Deep Q Learning) وإطار عمل Actor-Critic وفق الآتي:

- **المرحلة الأولى: تحديد عتبة التداخل (Interference Threshold Assignment)**
  في هذه المرحلة يتم استخدام شبكة DQN لاختيار فعل من فضاء الأفعال، وهذا الإجراء يحدّد قيمة عتبة نسبة الإشارة إلى التداخل والضجيج SINR لكل مستخدم ثانوي. تعتبر خوارزميّة DQL مناسبة لأنّ مجموعة العتبات المتاحة للمستخدمين الثانويين منفصلة ومحدودة أي أنّ فضاء الأفعال متقطّع Discrete Action Space.
- **المرحلة الثانية: تخصيص استطاعة الإرسال (Power Assignment)**
  بناءً على العتبة التي تم اختيارها في المرحلة الأولى يتم استخدام خوارزميّة DDPG لتحديد قيمة استطاعة الإرسال لكل مستخدم ثانوي. خوارزميّة DDPG من خوارزميّات Actor-Critic وهي مناسبة للتعامل مع فضاء الأفعال المستمرّ، ممّا يُجنّب أخطاء التكميم Quantization Error التي قد تنتج عن تقسيم الاستطاعة إلى مستويات منفصلة.

تكمن المساهمة الأساسية في اقتراح بنية RADDPG الهجينة التي تتعامل بفعالية مع فضاء الأفعال المستمرّ والمتقطّع في مشكلة تخصيص الموارد. هذا النهج يجمع بين أفضل ما في الخوارزميّتين: قدرة خوارزميّة DQL على التعامل مع فضاء الأفعال المتقطّع بسرعة معالجة أكبر من خوارزميّة QL، وقدرة خوارزميّة DDPG على التعامل الدقيق مع فضاء الأفعال المستمرّ، مما يؤدي إلى تحسين سرعة التعلم واستقراره ودقّته مقارنة بالخوارزميات التي تعتمد على نهج واحد فقط.

أظهرت نتائج المحاكاة تفوّق خوارزميّات RADDPG وDQL على خوارزميّة QL من حيث ثلاث مقاييس أداء رئيسيّة:

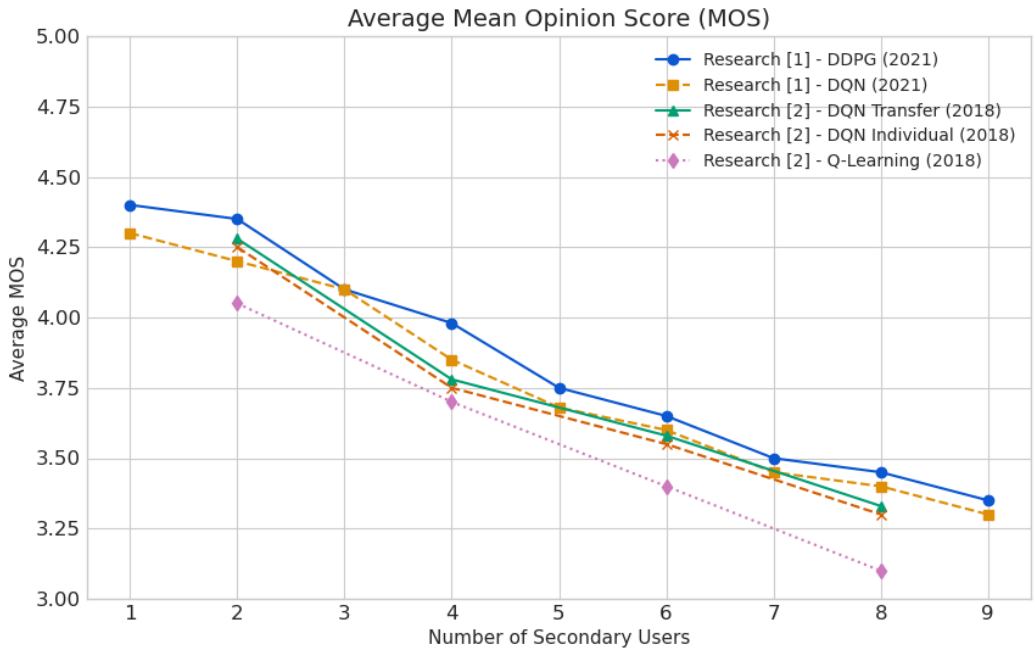
1. MOS (Mean Opinion Score): حقّقت خوارزميّة RADDPG أعلى قيم.
2. معدّل ازدحام الشبكة Network Congestion Rate: أظهرت أداء أفضل مع زيادة عدد المستخدمين الثانويين.
3. عدد التكرارات اللّازمة للوصول إلى الحلّ الأمثل Iteration number till convergence: حقّقت خوارزميّة DDPG أسرع وصول إلى سياسة مستقرّة.

أشار الباحثون إلى أنه على الرغم من أداء RDDPG الجيّد إلا أنّ سرعة تعلمها قد لا تكون كافية للتطبيقات العملية التي تتطلب استجابة سريعة جداً .واقترحوا كعمل مستقبلي دمج Meta-Learning مع خوارزميات Actor -Critic لتسريع عملية التعلّم بشكل أكبر وتحسين قدرة الخوارزميّة على التكيّف مع المهام الجديدة بسرعة.[1]

كلا البحثين [1] و[2] يعالجان نفس المشكلة وهي تخصيص الموارد لتعظيم جودة التجربة QoE في شبكات الراديو الإدراكي من نمط Underlay، ويستخدمان التعلّم المعزّز العميق كأداة أساسية للحلّ، ويعتمدان على مقياس MOS كمؤشّر لجودة التجربة. يكمن الاختلاف الرئيسي بين البحثين في المنهجيّة، حيثُ أنّ البحث [1] يركّز على دقّة التحكّم من خلال التعامل مع استطاعة الإرسال كقيمة مستمرّة ممّا يجعله أكثر دقّة من الناحية النظريّة. أمّا البحث [2] فيركّز على كفاءة التعلّم في البيئات الديناميكية، حيث يقدّم حلّاً عمليّاً لمشكلة انضمام المستخدمين الجدد إلى الشبكة وتسريع تكيُّفهم.

فيما يلي مقارنة بين نتائج البحثين [1] و[2] توضُّحها المنحنيات البيانيّة الموجودة في كل منهما، وذلك وفق مقاييس الأداء المعتمدة من قبل كلي البحثين (نفس المقاييس). تتضمن هذه المقاييس كلّاً من MOS، ومعدّل ازدحام الشبكة Network Congestion Rate، وعدد التكرارات للوصول إلى الحلّ Number of iteration till convergence.

- **MOS**

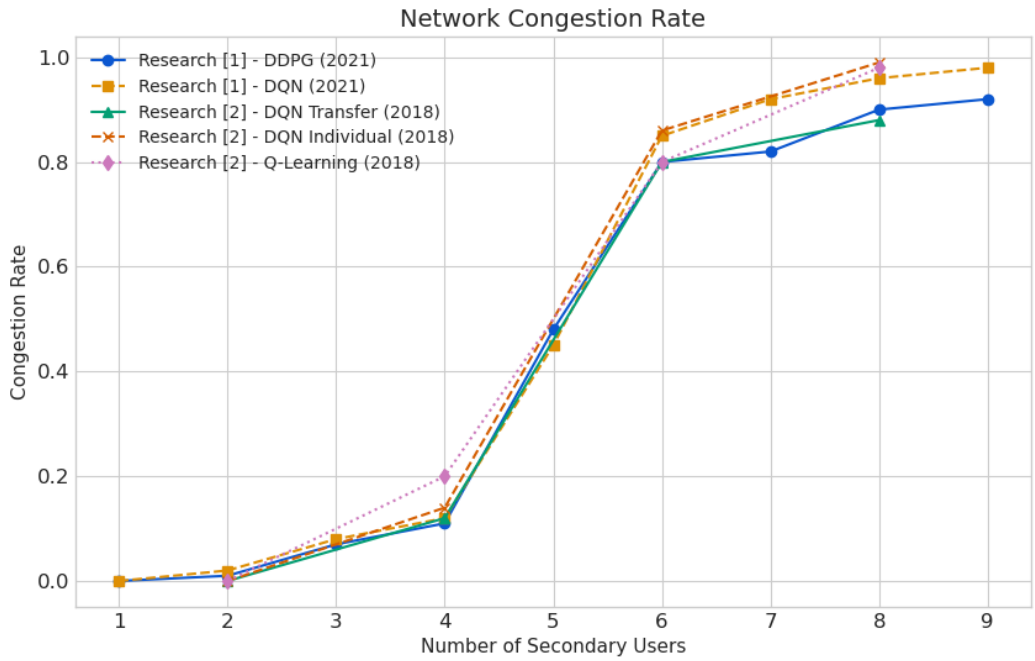


الشكل 2: مقارنة بين نتائج البحثين من حيث قيمة MOS.

تُظهر كلا الخوارزميّتان سلوكاً متوقعاً حيث تنخفض قيمة MOS مع زيادة عدد المستخدمين الثانويين بسبب زيادة التداخل. خوارزميّة البحث [1] RADDPG تحقّق أداءً أعلى بشكل طفيف ممّا يعكس فائدة التحكّم الدقيق والمستمرّ في استطاعة الإرسال. خوارزميّة DQN في كلي البحثين تظهر أداءً متقارباً جداً وتتفوّق على خوارزميّة Q-Learning.

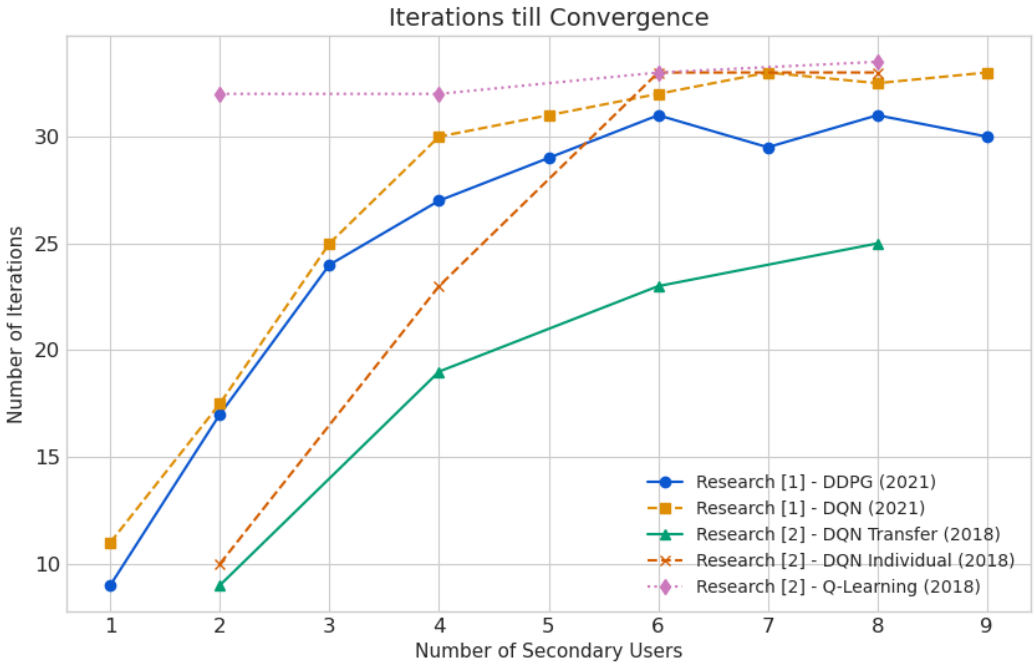
- **معدّل ازدحام الشبكة**

في الشكل التالي تُظهر خوارزميّة DQN Transfer أقلّ معدّل ازدحام عند وجود عدد أكبير من المستخدمين ممّا يشير إلى أنّ تقنية نقل الخبرة تساعد شبكة الرّاديو الإدراكي على استيعاب عدد أكبر من المستخدمين بفعاليّة. خوارزميّة DDPG وكذلك خوارزميّة DQN تُظهر أداءً قويّاً جدّاً في الحفاظ على معدّل ازدحام منخفض مقارنةً بخوارزميّات Q-Learning.



الشكل 3: مقارنة بين نتائج البحثين من حيث معدّل ازدحام الشبكة.

- **عدد التكرارات للوصول إلى الحلّ**



الشكل 4: مقارنة بين نتائج البحثين من حيث عدد التكرارات للوصول إلى الحلّ.

هنا يظهر التفوّق الواضح لخوارزميّة DQN مع تقنيّة التعلّم بالنقل (نقل الخبرة)، وهذا يؤكّد أنّ البدء مع وجود خبرة سابقة يقلّل بشكل كبير من وقت التعلّم. خوارزميّة DDPG تُظهر سرعة تقارب ممتازة خاصّةً عند وجود عدد قليل من المستخدمين الثانويين. من الواضح أنّ خوارزميّة Q-Learning هي الأبطأ بفارق كبير حيث تتطلّب عدداً ثابتاً وكبيراً نسبيّاً من التكرارات بغض النظر عن عدد المستخدمين الثانويين.

من خلال دمج النتائج يمكن استنتاج أنّ كلّاً من التحكّم الدقيق في خوارزميّة RADDPG وتسريع التعلّم عبر نقل الخبرة في خوارزميّة DQL Transfer يكمّلان بعضهما البعض لتحسين أداء نظام الراديو الإدراكي، حيثُ توفّر RADDPG دقّة أعلى في الأداء النهائي بينما يوفّر التعلّم بالنّقل باستخدام DQL سرعة أكبر في الوصول إلى هذا الأداء وهو أمر حاسم في البيئات اللاسلكيّة الديناميكيّة. بالإضافة إلى المرونة في بيئات العمل ممّا يجعلها خياراً معتمَداً لكثير من الأدبيات.

المرجع	مقاربة البحث	مقاييس الأداء	النتائج
[1]	خوارزميّة RADDPG متعددة العملاء تتكوّن من مرحلتين: - تحديد عتبة التداخل باستخدام DQN. - تخصيص استطاعة الإرسال باستخدام DDPG.	- MOS - Congestion Rate - Iteration number till convergence	تتفوّق خوارزميّة RADDPG على كل من QL وDQL
[2]	خوارزميّة DQL استخدمت لكل SU بشكل مستقلّ مع تقنيّة التعلم بالنقل (TL: Transfer Learning)	- MOS - Congestion Rate - Iteration number till convergence	التعلم بالنقل يقلّل عدد تكرارات التقارب بنسبة تصل إلى 72% مقارنة بخوارزمية QL وبنسبة 25% مقارنة بخوارزمية DQL التي لا تستخدم TL
[3]	إطار عمل يعتمد على التعلّم المعزّز متعدد العملاء Multi-Agent RL حيث تعمل كل خلية في الشبكة المرخّصة كعميل مستقل	- معدّل نقل البيانات - مستويات التداخل	في الشبكات الصغيرة تقدم خوارزمية QL أداءً شبه أمثلي حيث تحقّق معدل نقل بيانات مرتفع نسبياً وتحافظ على التداخل ضمن الحدود الآمنة. بينما في الشبكات الأكبر تتفوّق DDPG التي تمكنت من تحقيق أعلى متوسط لمعدّل نقل البيانات
[4]	نهج لا مركزيّاً متعدّد العملاء حيث تتّخذ كل مركبة قراراتها بنفسها ممّا يتيح قابليّة التوسّع	كفاءة أخذ العينات وسرعة التقارب	حسّنت خوارزمية DDPG كفاءة أخذ العينات وسرعة الوصول للحل في الأنظمة التي تحتوي على عدد كبير من المستخدمين
[5]	استخدام مقدّرات تتابعية لتتبع مكاسب قناة التداخل، بالإضافة إلى استخدام أدوات البرمجة الديناميكية لتصميم المخططات المُثلى	الفعالية الطيفية الوسطيّة	يُبرز هذا البحث أهميّة تكامل عملية الاستشعار مع عملية اتخاذ القرار
[6]	خوارزميّة DQL استخدمت في بيئة متعدّدة العملاء مع تقنيّة التعلم التعاوني	- MOS - عدد التكرارات اللازمة للتقارب - سرعة التقارب	التعلم التعاوني يقلل بشكل كبير عدد تكرارات التقارب بالنسبة للتعلّم الفردي، حيثُ أظهرت DQL سرعة أكبر من QL
[7]	خوارزميّة DQL مع تقنية NB-CR-IoT	- درجة رضا المستخدمين - معدّلات بت قابلة للتحقّق	حققت DQL درجة رضا أعلى للمستخدمين ومعدلات بت قابلة للتحقيق أفضل بنسبة 7.41% مقارنة بـ QL.

جدول 1: خلاصة الدراسة المرجعية.

يبيّن الجدول السّابق (جدول 1) خلاصة الدراسة المرجعيّة حيث تُستعرض فيها مقاربة كل بحث ومقاييس الأداء والنتائج.

2-2 خاتمة

يتّضح من خلال مراجعة الأدبيات أن التوجه الحديث لحلّ مشكلة تخصيص الموارد في شبكات الراديو الإدراكي من نمط underlay يرتكز بشكل متزايد على تقنيات الذكاء الاصطناعي وتحديداً التعلم المعزز العميق DRL متعدّد العملاء. إن الهدف الأساسي المشترك بين هذه الأبحاث هو الانتقال من مقاييس الأداء التقليدية إلى مقياس أكثر تمثيلاً لرضا المستخدم مثل جودة التجربة QoE والذي يُقيّم باستخدام معيار MOS، وقد نجحت خوارزميّات QL وDQL في الوصول إلى نتائج تُرضي المستخدم.

الفصل الثالث

الدراسة النظريّة

في هذا الفصل سيتمّ توضيح سبب اتّخاذ تعلّم الآلة كمنهجيّة مقاربة، والتعرف على تعلّم الآلة وتصنيفاتها بالإضافة إلى شرح نظري للخوارزميات المستخدمة في البحث.

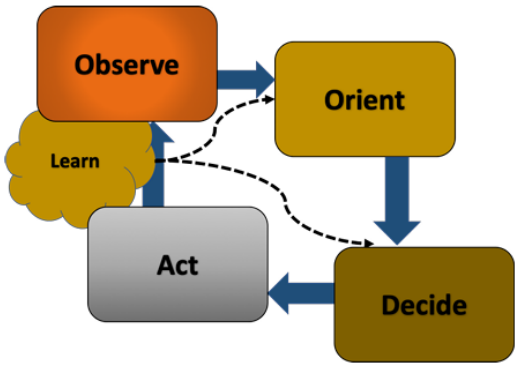
3-1 مقدّمة (علاقة الراديو الإدراكي بتعلّم الآلة)

على عكس الراديو التقليدي الذي يمكنه العمل فقط على نطاق طيفي محدد مسبقاً بناءً على القيود التنظيمية، فإن الراديو الإدراكي قادر على العمل في نطاقات طيفيّة مختلفة، ومن أجل ذلك يجب أن يمتلك القدرة على استشعار وفهم بيئته اللاسلكية وتغيير نمط تشغيله بشكل استباقي حسب الحاجة. تُنفّذ هذه المجموعة من الأنشطة الأساسية والتي تسمى دورة الإدراك Cognition Cycle بواسطة الراديو الإدراكي من أجل تلبية احتياجات المستخدم النهائي، وتتضمّن هذه المجموع مراقبة البيئة واكتساب المعلومات والتوجيه والتعلم من التجارب السابقة ومن ثمّ اتخاذ القرار، وأخيراً تنفيذ الإجراءات. في البداية تقوم مرحلة اتخاذ الفعل Action phase بتهيئة الراديو الإدراكي لتوفير جودة اتصال محسّنة فيما يتعلق بأهداف المستخدم النهائي، ويمكن أن تكون هذه التهيئة على سبيل المثال اختيار واجهة الراديو اللاسلكية التي سيتم استخدامها للاتصال أو ضبط معاملات نظام الاتصالات. في مرحلة المراقبة Observation phase يتمّ جمع إحصائيات لتمييز البيئة الخارجية مثل أنماط حمل الحركة المرورية؛ قياسات نسبة الإشارة إلى الضجيج SNR؛ معدل خطأ الرزم؛ زمن التأخير Latency. تتكوّن مرحلة التوجيه Orientation phase من فهم تأثير البيئة الخارجية على أداء الاتصال، ويتمّ ذلك عن طريق تحديد علاقة وظيفيّة بين القياسات ومتطلّبات الاتصال مثل الإنتاجية؛ التأخير؛ الموثوقيّة. بعد ذلك تأتي مرحلة اتخاذ القرار Decision phase التي يتمّ فيها إيجاد الحلول التي تحقّق الأهداف المحدّدة من قبل المستخدم، والتي يتمّ التعبير عنها بمقاييس أداء عالية المستوى التي يتطلّبها التطبيق من معدّل نقل البيانات المناسب والتأخير والموثوقية وكذلك التكلفة واستهلاك الطاقة. وأخيراً مرحلة التّعلّم Learning phase لتقييم نتائج القرارات التي تم اتخاذها، وبالتالي جمع المعرفة لاستغلالها في مراحل التوجيه المستقبلية بهدف أن يكون أكثر فعالية في مرحلة اتخاذ القرار. يوضّح الشكل (5) دورة الإدراك التي يتفاعل من خلالها الراديو الإدراكي مع البيئة. يُطلق على دماغ الراديو الإدراكي حيث يتم تنفيذ مجموعة الأنشطة المكوّنة لدورة الإدراك اسم المحرّك الإدراكي (CE: Cognitive Engine). يمكن استخدام تعلم الآلة (ML: Machine Learning) على نطاق واسع في تنفيذ مراحل التعلم والتوجيه واتخاذ القرار في أجهزة الراديو الإدراكي وتحديداً المحرّك الإدراكي CE.[12] تحتاج شبكات الراديو الإدراكي من نمط Underlay إلى تصميم خوارزمية تخصيص موارد للمستخدمين الثانويين SUs بحيث تضمن التعايش بين جميع المستخدمين، وبشكل خاصّ تُستخدم تقنيات تعلّم الآلة لتخصيص موارد الراديو للمستخدمين الثانويين بحيث يتمّ تعظيم مقياس أداء الشبكة ضمن الشبكة الثانويّة SN.[11]

3-2 تعلّم الآلة

أدّى التطور السريع للأجهزة الذكية وتطوّر الحوسبة السّحابية والمحاكاة الافتراضيّة للشبكات إلى زيادة أسيّة في حركة البيانات، حيثُ أصبحت الشبكات أكثر تعقيدًا لأنها تتضمن عددًا كبيرًا من الأجهزة وتدعم تطبيقات متنوعة، ممّا خلق تحدّيات أكبر في إدارة موارد الشبكة واستثمارها بفعالية. يعدّ استخدام الذكاء الصّنعيّ حلاً ممكناً لهذه التحديات من خلال تطبيقاته مثل تعلّم الآلة Machine Learning التي تعتمد خوارزميّاتها على جمع وتحليل البيانات من أجل إيجاد الأنماط والعلاقات بينها وإنتاج نموذج يمكنه ربط المدخلات بالمخرجات.[12] يمكن تصنيف خوارزميّات تعلّم الآلة إلى ثلاثة أنواع:

- **تعلّم خاضع للإشراف Supervised Machine Learning:** خوارزميّات هذا النوع تتدرّب على مجموعة بيانات Data Set محدّدة مسبقاً أو موسومة (labeled) بتصنيفات معيّنة. أي أنّ النموذج لا يحتاج أن يصنّف البيانات لكي يستطيع التعرف على البيانات الجديدة. مهمّة خوارزمية تعلم الآلة الخاضع للإشراف هي إيجاد الأنماط patterns وبناء نماذج رياضية، ليتمّ بعد ذلك تقييم هذه النماذج بناءً على القدرة التنبؤية predictive capacity فيما يتعلق بمقاييس التباين في البيانات نفسها. تعدّ نماذج التصنيف classification models ونماذج الانحدار regression models من النماذج الشائعة في هذا النوع من الخوارزميّات.
- **تعلّم غير خاضع للأشراف Unsupervised Machine Learning:** المهمة الرئيسية للتعلّم غير الخاضع للإشراف هي تطوير وسوم التصنيف classifications labels لمجموعة البيانات Data Set تلقائياً. تبحث هذه الخوارزميات عن التشابه similarity بين أجزاء البيانات لتحديد ما إذا كان يمكن تصنيفها وإنشاء مجموعة، حيثُ تُسمّى هذه المجموعات بالعناقيد clusters، وهي تمثّل عائلة كاملة من تقنيّات تعلم الآلة غير الخاضعة للإشراف.
- **تعلّم معزّز (RL: Reinforcement Learning):** سيتم شرحه في الفقرة القادمة.



الشكل 5: يبيّن دورة الإدراك في الراديو الإدراكي.[12]

3-3 التعلّم المعزّز

التعلم المعزّز Reinforcement Learning هو أحد أفرع تعلّم الآلة Machine Learning الذي يعتمد على تعلم الأنظمة كيفية اتخاذ قرارات من خلال التفاعل مع البيئة. يتعلّم النظام من خلال التجربة والخطأ حيثُ يختبر الأفعال المختلفة ويقيّم نتائجها بناءً على مكافآت يتلقّاها من البيئة دون الحاجة لإشراف مباشر أو أمثلة موجهة (Data Set) مثل التعلّم الخاضع للإشراف أو التعلّم غير الخاضع للإشراف. يتعامل التعلّم المعزّز مع مشكلة تفاعلية يكون فيها توليد أمثلة للسلوك المرغوب صعباً أو حتى غير عمليّ التحقيق، ويعتبر مناسباً لتعلّم النماذج الديناميكيّة وتحديداً في المواقف التفاعلية أو غير المعروفة حيث يتعيّن على العميل Agent أن يتعلّم من تجربته الخاصة بالتفاعل مع البيئة بهدف تعظيم المكافآت التي يحصل عليها العميل عبر اتخاذ سلسلة من الأفعال. يتفاعل العميل مع البيئة عن طريق اتخاذ الأفعال Actions وتلقّي مكافآت Rewards تُظهر تأثير الفعل على البيئة لينتقل إلى حالة جديدة State متمّماً ما يسمّى دورة التعلّم. في بداية كل دورة تعلم يتلقّى العميل ملاحظة كاملة أو جزئية للحالة الحالية بالإضافة إلى المكافأة المتراكمة، فيقوم باستخدام ملاحظة الحالة وقيمة المكافأة بتحديث سياسته Policy خلال مرحلة التعلّم.[12][8]

تتراوح تطبيقات التعلّم المعزّز RL من الألعاب إلى الروبوتات حيث يمكن تدريب العملاء على بيئات متعدّدة الاحتمالات مثل الشطرنج، حيث يمكن للعملاء تعلّم استراتيجيّات معقّدة لزيادة فرص الفوز. كما يستخدم في تدريب الروبوتات على مهام كالتّنقل في أماكن متنوّعة.[8]

1-3-3 بنية التعلّم المعزّز

تتكوّن بنية التعلّم المعزّز من أربعة عناصر رئيسيّة، وهي: السّياسة Policy والمكافأة Reward وتابع القيمة Value Function والنموذج Model. سيتمّ شرح كل منها:

السياسة Policy: ويرمز لها ب π. هي المكوّن الذي يحدد استراتيجية اتخاذ القرار للعميل Agent. تُمثّل السياسة عملية ربط mapping بين فضاء الحالات State space وفضاء الأفعال Action space. يمكن أن تكون السّياسة حتمية deterministic حيث تحدّد فعلاً واحداً دقيقاً لكل حالة، أو سياسة احتماليّة stochastic تعرّف توزيعاً احتماليّاً على مجموعة الأفعال الممكنة عند التواجد في حالة ما. الهدف الجوهري لخوارزميّة التعلم المعزز هو البحث التكراري عن السّياسة المثلى التي تعظّم العائد التراكمي المتوقّع (المكافأة التراكميّة الإجماليّة).

المكافأة Reward: ويرمز لها ب R_t. هي قيمة عدديّة فوريّة يتلقّاها العميل من البيئة في كل خطوة زمنية t. تُعدّ المكافأة الهدف القصير المدى للعميل، فهي تقيّم الفعل المتخذ في الحالة الحالية. لا يتحكّم العميل بقيمة المكافأة بل هي ناتجة عن ديناميكيات البيئة وتُستخدم لتقييم أفعال العميل وتوجيه السياسة نحو السلوك المرغوب الذي يؤدّي إلى تعظيم المكافأة التراكميّة الإجماليّة التي تمثّل مجموع المكافآت على المدى الطويل.[8]

تابع القيمة Value Function: هو المكوّن الذي يقيّم الجدوى طويلة المدى لحالة ما أو لفعل ما أو لثنائيّة معيّنة مكوّنة منهما. على عكس المكافأة الفورية R، يتنبّأ تابع القيمة بالمكافأة المستقبلية المتوقعة الذي يمكن للعميل تجميعه بدءاً من تلك النقطة. يوجد نوعان لتوابع القيمة:

- تابع قيمة الحالة State-Value Function $V_\pi(s)$: يحسب المكافأة المتوقّعة عند البدء من الحالة s باتّباع السياسة π بعد ذلك. يمثّل رياضيّاً بالشكل التالي: $V_\pi(s) = \mathbb{E}[G_t|S_t = s]$.
- تابع قيمة الفعل Action-Value Function $Q_\pi(s,a)$: يحسب المكافأة المتوقّعة عند البدء من الحالة s واتّخاذ الفعل a باتّباع السياسة π بعد ذلك. يمثّل رياضيّاً بالشكل التالي: $Q_\pi(s,a) = \mathbb{E}[G_t|S_t = s, A_t = a]$.

تعدّ توابع القيمة ضروريّة لتقييم السياسات والمفاضلة بين الأفعال المختلفة والمتاحة التي قد تؤدّي إلى مكافآت فورية منخفضة ولكن مكافآت مستقبلية تراكمية عالية.

النموذج Model: هو تمثيل رياضي لسلوك البيئة يهدف إلى التنبّؤ بكيفية استجابتها لأفعال العميل. يصنّف هذا العنصر التعلّم المعزّز إلى نوعين:

- التعلّم القائم على نموذج Model-Based Learning: يُعرَّف بأنه الأساليب التي يتعلم فيها الوكيل نموذجاً لديناميكيات البيئة من خلاله يمكن للوكيل استخدامه ل محاكاة التفاعلات مع البيئة داخلياً لاتخاذ قرارات أفضل وتحديث سياسته أو تابع القيمة الخاصة به.
- التّعلم غير القائم على نموذج Model-Free Learning: يُعرَّف بأنه الأساليب التي يتمّ تعلّم السياسات policies أو توابع القيمة value functions مباشرة من تجارب التفاعل مع البيئة (من خلال التجربة والخطأ) دون بناء أو استخدام نموذج صريح لكيفية عمل البيئة. تُعد خوارزميات مثل Q-Learning وDDPG أمثلة شهيرة على هذا النوع.[8]

طريقة اختيار الفعل ϵ-greedy: تعدّ طريقة اختيار الفعل من أساسيّات بنية التعلّم المعزّز، ذلك ولأنّ العميل يواجه أثناء التعلّم معضلة اختيار الفعل من بين مجموعة الأفعال المتاحة عند حالة ما نظراً لعدم معرفته بالسياسة المثلى في البداية، فيتصرّف بشكل دون الأمثل sub-optimally من خلال اللجوء إلى آليّة ϵ-greedy التي تنصّ على وجود طريقتين لاختيار الفعل. الطريقة الأولى وهي الاستغلال exploitation وتتمثّل في اختيار الفعل الذي ينتج عنه القيمة العظمى المعروفة لدى العميل. الطريقة الثانية هي الاستكشاف exploration وتتمثّل في الاختيار العشوائي للفعل. يختار العميل طريقة الاستكشاف باحتمال

$(p = \epsilon)$ بينما يختار طريقة الاستغلال باحتمال $(p = 1 - \epsilon)$، بحيث تكون قيمة ϵ أعلى ما يمكن في بداية عمليّة الملاحة navigation (التعلّم) حيثُ يتم اتباع طريقة الاستكشاف في اختيار الأفعال في الغالبية الساحقة من مرات الاختيار، وتتناقص قيمة ϵ بعدها بحيث تصبح طريقة الاستغلال هي المتّبعة بالغالبية الساحقة في نهاية التعلّم. إنّ استخدام هذه السياسة يُظهر موازنة trade-off جيّدة من حيث الاستكشاف والاستغلال الاستكشاف مطلوب لاكتشاف المزيد عن الاستراتيجية المثلى، ولكن الاستغلال مطلوب أيضاً لمعرفة المزيد عن الحالات التي تمت زيارتها مسبقاً والمعروفة جزئياً في البيئة.

2-3-3 تصنيفات خاصّة بالتعلّم المعزّز

في الحقيقة يوجد عدّة تصنيفات لخوارزميّات التعلّم المعزّز. في الفقرة السابقة عند الحديث عن النموذج في بنية التعلّم المعزّز صُنّفت خوارزميّات التعلّم المعزّز من حيث النموذج إلى تعلّم قائم على نموذج Model-Based Learning وتعلّم غير قائم على نموذج Model-Free Learning. علاوةً على ذلك، يمكن تقسيم خوارزميات التعلم المعزز RL إلى ثلاث فئات رئيسية كما يلي:

- البرمجة الديناميكية (DP: Dynamic Programming): يكون لدى العميل نموذج مثالي للبيئة معطى كعملية قرار ماركوف (MDP: Markov Decision Process)، والهدف هو تعلّم السياسة المُثلى لاختيار أفضل الأفعال.
- طرق مونت كارلو (MCM: Monte Carlo Methods): في هذه الحالة لا يُفترض وجود معرفة كاملة بالبيئة، وبالتالي يجب على العميل أن يتعلّم إما من خلال تجربة البيئة أو من خلال تجارب مُحاكاة حيث يتم تمثيل البيئة بنموذج بسيط جدًا.
- تعلم الفارق المؤقّت (TDL: Temporal-Difference Learning): والذي يمكن تعريفه كمزيج من طرق MCM وDP. حيثُ يمكن لعملاء تعلّم TD التعلم مباشرة من تجربتهم مع البيئة، دون الحاجة إلى الديناميكيات الكاملة للبيئة، بالإضافة إلى تحديث التقديرات (إما سياسة أو تابع قيمة) بناءً على تقديرات أخرى مكتسبة.[12]

يُعد تصنيف خوارزميات التعلم المعزّز بناءً على السياسة أحد أهم الفروقات الجوهرية في كيفية عمل هذه الخوارزميات. يقسم هذا التصنيف الخوارزميّات إلى فئتين رئيسيتين: التعلم وفقًا للسياسة On-Policy والتعلم خارج السياسة Off-Policy. يكمن الاختلاف الأساسي بينهما في السياسة التي تتعلم الخوارزمية قيمتها والتي تُعرف بالسياسة المستهدفة Target Policy مقارنة بالسياسة التي تستخدمها الخوارزمية فعلياً لاختيار الأفعال والتفاعل مع البيئة وتعرف بالسياسة السّلوكية Behavior Policy.

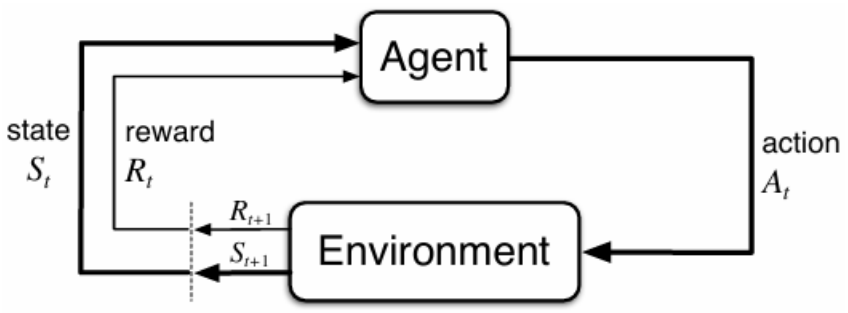
- التعلم وفقًا للسياسة On-Policy Learning: في هذا النوع من الخوارزميات تتعلّم الخوارزمية قيمة السياسة التي تتبعها حالياً، أي أن السياسة المستهدفة والسياسة السلوكية هما نفس الشيء. يقوم العميل باتّخاذ فعل ما a في الحالة s بناءً على سياسته الحاليّة π، وبعد مراقبة المكافأة r والحالة التالية s' يقوم بتحديث تابع القيمة بناءً على الإجراء a' الذي سيتّخذه فعليّاً في الحالة s' وفقاً لنفس السياسة π.
  يمكن تشبيه هذه الخوارزميّة بخوارزميّة تعلّم قيادة السيارة. يتعلّم العميل كيفية تحسين قيادته بناءً على الأخطاء والنجاحات التي يرتكبها بنفسه أثناء التجربة، فهو يحسّن السياسة التي يتّبعها. تعدّ خوارزميّة SARSA من الأمثلة الشائعة.
- التعلم خارج السياسة Off-Policy Learning: في هذا النوع تتعلم الخوارزميّة قيمة سياسة مُثلى النظر عن السّياسة التي تتبعها حالياً لاختيار الأفعال. يقوم العميل بالتفاعل مع البيئة باستخدام سياسة سلوكيّة استكشافيّة مثل اختيار أفعال عشوائية في كثير من الأحيان، ولكن عند تحديث تابع القيمة فإنه سيتخذ أفضل إجراء ممكن في الحالة التالية بغض النظر عمّا إذا كان سيختار هذا الفعل أم لا، وهذا يسمح بفصل عملية الاستكشاف عن عملية تعلم السّياسة المثلى.[12]
  يمكن تشبيه هذه الخوارزميّة بخوارزميّة تعلّم القيادة المثاليّة من خلال مشاهدة سائق محترف وتعبّر عن السياسة المثلى لهذه الحالة، حتى لو كان الشخص المتعلّم (العميل في حالتنا) نفسه لا يزال يقود بحذر شديد أو بشكل عشوائي وتعني السياسة السلوكيّة، حيثُ يتعلّم قيمة أفضل سياسة حتى لو لم يكن يطبقها. من الأمثلة الشهيرة خوارزميّة Q-Learning.

3-3-3 عمليات ماركوف المحدودة لاتخاذ القرار

تمثل عمليات ماركوف لاتخاذ القرار عملية اتخاذ القرار بشكل متسلسل، حيث تؤثر الأفعال على المكافآت الفورية والحالات اللاحقة، ومن خلالها على المكافآت المستقبلية. تحتاج عمليات ماركوف لاتخاذ القرار إلى حساب المكافأة المتأخرة بالإضافة إلى المقايضة بين المكافأة الفورية والمتأخرة. إن عمليات ماركوف لاتخاذ القرار هي شكل رياضي مثالي لمشكلة التعلم المعزز، وتتألف بنيتها الرياضية من عناصر أساسية هي المكافأة وتوابع القيمة ومعادلات بيلمان Bellman equations.[8]

التفاعل بين العميل والبيئة

كما ذكرنا فإنّ عمليات ماركوف لاتخاذ القرار تشكل إطار عمل لعملية التعلّم من خلال التفاعل مع البيئة، ويسمى المتعلّم وصانع القرار بالعميل، في حين يسمى كل شيء يتفاعل معه هذا العميل بالبيئة. تتفاعل البيئة باستمرار مع العميل وتستجيب للأفعال التي يتّخذها وهذا ما قد ينقله إلى حالات جديدة. تقدّم البيئة المكافآت للعميل، وهي قيم عددية يسعى العميل إلى تعظيمها بمرور الوقت من خلال اختياره للأفعال التي تحقق له أكبر قدر من المكافأة.



الشكل 6: عناصر التعلّم المعزّز.[8]

في كل لحظة زمنيّة ينتقل العميل إلى حالة ما state، ونرمز لها في اللحظة t بالرمز $S_t \in \mathcal{S}$ وبناءً على الحالة التي يتواجد فيها العميل يمكن أن يتّخذ الفعل $A_t \in \mathcal{A}(S)$ من مجموعة الأفعال actions المتاحة ضمن هذه الحالة، أو يختار بشكل عشوائي كما ذُكر سابقاً وفق آليّة ϵ-greedy. نتيجة لهذا الإجراء يتلقّى العميل مكافأة عددية $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ وقد ينتقل إلى حالة جديدة S_{t+1}. يؤدي هذا التسلسل إلى ظهور مسار Trajectory الذي يمكن أن نعبر عنه بالشكل التالي:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, A_3, \dots$$

في عمليات ماركوف المحدودة لاتخاذ القرار تكون البيئة ذات عدد منتهي من الحالات $\mathcal{S}$، ويكون عدد الأفعال التي يمكن أن يتخذها العميل في كل حالة يتواجد فيها $\mathcal{A}(S)$ منتهي، وبالتالي فإن مجموعة المكافآت $\mathcal{R}$ التي يحصل عليها العميل من البيئة منتهية. تمثل كلاً من الحالة S_t التي يمكن التي يتواجد فيها العميل والمكافأة R_{t+1} التي يحصل عليها متحولين عشوائيين متقطعين يعتمدان فقط على الحالة والفعل السّابقين. ويعرّف التابع: $p: \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ ديناميكيات عمليّات ماركوف لاتخاذ القرار بالشكل:

$$p(s', r|s, a) = Pr(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a) \qquad (\mathbf{1})$$

وبما أن p يمثل توزيعاً احتماليّاً، عندئذ أيّاً كان $s', s \in \mathcal{S}$ و $R_{t+1} \in \mathcal{R}$ و $a \in \mathcal{A}(s)$، فإنّ

$$\sum_{s' \in S} \sum_{r \in \mathbb{R}} p(s', r|s, a) = 1 \qquad (\mathbf{2})$$

يُعرف التابع p البيئة بشكل كامل ونلاحظ أن الحالة S_t والمكافأة R_t تعتمد فقط على الحالة S_{t-1} والفعل A_{t-1} السابقين. يمكن أن ننظر إلى ذلك بأنه تقييد للحالة، أي أن الحالة يجب أن تتضمن معلومات عن كل جوانب التفاعل السابق بين العميل والبيئة، وعندئذ نقول إن الحالة ماركوفية. يمكن من خلال تابع ديناميكيات عمليات ماركوف المحدودة لاتخاذ القرار حساب أي شيء آخر نريد معرفته عن البيئة، مثل احتمالات الانتقال والتي يمثلها التابع $p: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ وتعرف بالشكل:

$$p(s'|s, a) = Pr(S_t = s'|S_{t-1} = s, A_{t-1} = a) = \Sigma_{r \in \mathbb{R}} p(s', r|s, a) \qquad (\mathbf{3})$$

كما يمكن أيضاً حساب المكافأة المتوقعة Expected Reward للزوج حالة - فعل، حيث تعرف المكافأة المتوقعة عند اتخاذ الفعل a في الحالة s بالتابع $r: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ الذي له بارامترين بالشكل:

$$r(s,a) = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathbb{R}} r \sum_{s' \in S} p\,(s', r | s, a) \qquad (\mathbf{4})$$

وتعرف المكافأة المتوقعة للثلاثية حالة - فعل - حالة أي المكافأة المتوقعة عند الانتقال إلى الحالة s' انطلاقاً من الحالة s بعد اتخاذ الفعل a بالتابع $r: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ الذي له ثلاث بارامترات بالشكل:

$$r(s', a, s) = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathbb{R}} r \frac{p(s', r | s, a)}{p(s' | s, a)} \qquad (\mathbf{5})$$

إن إطار عمليات ماركوف لاتخاذ القرار هو إطار عمل مرن، فمثلاً ليس بالضرورة أن تعبّر الخطوات الزمنية عن الزمن الحقيقي، ويكفي أن تعبر عن سلسلة متتالية من عمليات اتخاذ القرارات والأفعال. يمكن أن تكون هذه الأفعال على مستوى منخفض مثل عملية التحكم بالجهد (voltage) أو الاستطاعة، أو قرار على مستوى عالي مثل الانتقال من مكان ما والذهاب إلى آخر. يمكن أيضاً التعبير عن الحالات بأشكال مختلفة قد تكون مرتبطة بمستوى منخفض من الإدراك الناتج عن قراءة حسّاس ما، أو مستوى عالي كإدراك الأوصاف المحددة لأغراض موجودة في الغرفة. للتمييز بين البيئة والعميل يعتبر أي شيء لا يمكن للعميل تغييره جزء من بيئته، وغالباً ما يعرف العميل القليل عن بيئته مما يسمح له بحساب مكافأته وبالتالي ليس كل شيء في البيئة معروف بالنسبة للعميل.[8]

المكافآت والعائد (المكافأة التراكميّة الإجماليّة)

المكافأة هي إشارة خاصة تعبّر بشكل رياضي عن غرض أو هدف العميل وتُعطى من البيئة إلى العميل في كل خطوة. تسمح المكافأة بإيصال ما نريد تحقيقه إلى العميل وليس الطريقة التي نريد تحقيقه بها، وعلى الرغم من أن صياغة الأهداف كإشارة خاصّة تسمى المكافأة قد تبدو مقيدة إلّا أنها أثبتت من الناحية العملية أنها مرنة وقابلة للتطبيق على نطاق واسع. ويهدف العميل إلى زيادة إجمالي المكافأة التي يتلقاها، أي أنه يهدف إلى تعظيم المكافأة التراكميّة على المدى الطويل. يمكن التعبير عن المكافأة التراكميّة على المدى الطويل (الإجماليّة) في أبسط أشكالها عبر تابع العائد المتوقع Expected Return الذي يكون عبارة عن مجموع المكافآت وهو يعرُف بالشكل:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \qquad (\mathbf{6})$$

حيث تعبّر G_t عن قيمة العائد و R_t عن قيمة المكافأة، و T عن اللحظة الزمنية الأخيرة. في شكل آخر يُعرف العائد على أنّه مجموع المكافآت المتناقصة ووفقاً لهذا التعريف فإن العميل يسعى إلى تعظيم مجموع المكافآت المستقبلية المتناقصة، أي أنه سيختار الإجراء $A_t = a$ الذي يسعى إلى تعظيم العائد المتناقص المتوقع الذي يعرف بالشكل:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \qquad (\mathbf{7})$$

حيث $0 \le \gamma \le 1$ معدل التناقص، وهو يحدّد القيمة الحاليّة للمكافآت المستقبلية، أي أنّ المكافأة التي يتلقّاها العميل في الخطوة الزمنيّة k تساوي قيمتها مخفّضة بمقدار γ^{k-1} مرّة في حال تلقّاها في الخطوة الحاليّة. عندما يكون $\gamma < 1$ فإن المجموع اللانهائي للعائد G_t له قيمة محددة على اعتبار أن قيمة المكافأة R_t محدودة. أمّا عندما يكون $\gamma = 0$ يكون العميل قصير المدى ويهتم فقط بالمكافآت المباشرة فقط ويتم تعظيم العائد من خلال اختيار الفعل الذي يعظّم المكافأة R_{t+1}. يرتبط العائد (المكافأة التراكميّة الإجماليّة) ببعضه البعض خلال سلسلة متتالية من الأزمنة بطريقة مهمّة لخوارزميات التعلم المعزّز كالتالي:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$

$$G_t = R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots)$$

$$G_t = R_{t+1} + \gamma G_{t+1} \qquad (\mathbf{8})$$

يمكن النظر أيضاً إلى معدل التناقص γ على أنه مدى تأثر المكافآت المستقبلية بالفعل الحالي الذي يتّخذه العميل. ونلاحظ أنه عندما يكون $0 \leq \gamma < 1$ حتى لو كان الزمن $T \to \infty$ فإن G_t لا يزال محدوداً طالما أنّ R_t محدودة، أمّا في الحالة التي يكون فيها $\gamma = 1$ يجب أن يكون الزمن T محدود، أي أن الفعل المتّخذ يؤثّر على عدد منتهٍ من المكافآت المستقبليّة.[8]

توابع السّياسة والقيمة

يحتاج العميل إلى تقييم الفعل الذي يتّخذه في حالة ما ليتحقّق من جودة اختيار الفعل الذي اتّخذه، ويتمّ ذلك عن طريق حساب تابع القيمة. تتحدّد جودة الفعل هنا بتقييم العائد الذي يمكن توقُّعه، فالمكافآت التي يتوقّع حصولها العميل في المستقبل تعتمد على الأفعال التي سيتّخذُها، وبناءً على ذلك يتم تعريف تابع القيمة بالاعتماد على طرق معيّنة لاتخاذ الأفعال تسمّى السّياسات. تعرّف السياسة على أنها عملية مقابلة من الحالات إلى احتمالات الأفعال، فإذا كان الوكيل يتبع السّياسة π في اللحظة t عندئذ يعبّر $\pi(a|s)$ عن احتمال اتخاذ الفعل $A_t = a$ في الحالة $S_t = s$ وبالتالي فإن $\pi(a|s)$ هو توزيع احتمالي للأفعال $a \in \mathcal{A}$ من أجل جميع الحالات $s \in \mathcal{S}$. خوارزميّات التعلّم المعزّز تحدّد كيفيّة تغيّر سياسة العميل كنتيجة لتجاربه.

ويُعرّف تابع القيمة للحالة $\forall\, s \in \mathcal{S}$ عند اتّباع السّياسة π بأنه العائد المتوقّع (المكافأة التراكميّة الإجماليّة المتوقّعة) عندما يبدأ العميل التعلّم في الحالة s ويتّبع السّياسة π بعد ذلك ويعرف بالشكل:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right] \qquad (\mathbf{9})$$

وبنفس الطريقة تعرف قيمة اتخاذ الفعل a عند اتباع السياسة π بأنه العائد المتوقّع من الحالة s واتخاذ الفعل a وبعد ذلك اتباع السياسة π:

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right] \qquad (\mathbf{10})$$

من أهم الخصائص الأساسية لتابع القيمة التي يتم الاستفادة منها في التعلّم المعزّز هي العوديّة recursive، فمهما كانت السياسة π والحالة s يكون:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']]$$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \ , \forall s \in \mathcal{S} \qquad (\mathbf{11})$$

تسمّى المعادلة الأخيرة السابقة بمعادلة بيلمان وهي تعبّر عن دالّة القيمة بطريقة تكراريّة ويكون v_π حلاً وحيداً لمعادلة بيلمان. إن السّياسة التي تحقّق القيمة العظمى لتابع القيمة تسمّى بالسياسة المثلى، ويمكن أن يوجد أكثر من سياسة مثلى وجميعها تعطي نفس القيمة وترمز لها $v_*(s)$ وقيمتها تساوي:

$$v_*(s) = \max_\pi v_\pi(s) \ , \forall s \in \mathcal{S} \qquad (\mathbf{12})$$

وكذلك جميع السّياسات المُثلى تحقّق نفس القيمة لتابع قيمة الفعل أياً كان $s \in \mathcal{S}$ و $a \in \mathcal{A}(s)$ وهي:

$$q_*(s,a) = \max_\pi q_\pi(s,a) \qquad (\mathbf{13})$$

ومن أجل الزوج (s,a) يكون العائد المتوقّع لاتخاذ الفعل a في الحالة s، وبعد ذلك اتّباع السّياسة المثلى كما هو موضّح في العلاقة التّالية:

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \qquad (\mathbf{14})$$

عندها يمكن كتابة معادلة بيلمان الأمثلية بالصيغة:

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a)$$

$$v_*(s) = \max_a \mathbb{E}_{\pi_*}[G_t | S_t = s, A_t = a]$$

$$v_*(s) = \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

$$v_*(s) = \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

$$v_*(s) = \max_a \Sigma_{s',r} p(s',r|s,a)[r + \gamma v_*(s')] \qquad (\mathbf{15})$$

تعبّر العلاقتين 14 و15 عن شكلين مختلفين لمعادلة بيلمان الأمثلية لدالة v_* وتكون معادلة بيلمان الأمثلية للتابع q_* من الشكل:

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a]$$

$$q_*(s,a) = \Sigma_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')] \qquad (\mathbf{16})$$

في عمليات ماركوف المحدودة لاتخاذ القرار يكون لمعادلة بيلمان المُثلى حلّاً وحيداً، حيثُ تعبّر معادلة بيلمان عن مجموعة من المعادلات التي تنتج عن الحالات فإذا كان لدينا n حالة فيكون لدينا عندئذ n معادلة وn^* مجهول. عندما يكون تابع ديناميكيات البيئة p معروفاً تكون المعادلات قابلة للحل، فعند الحصول على القيمة المثلى للحالة v_* يمكن إيجاد السياسة المثلى للعميل.[8]

(جميع المعادلات الموجودة في الفقرة 3-3-3 من المرجع [8]).

3-3-4 خوارزميّة Q-Learning

تُعد خوارزمية Q-Learning واحدة من أشهر وأهم خوارزميات التعلم المعزز RL. تندرج هذه الخوارزمية ضمن فئة الخوارزميات التي تتبع طريقة التعلّم بالفارق المؤقّت TDL، وغير القائمة على نموذج Model-Free ممّا يعني أنها لا تحتاج إلى بناء نموذج كامل لديناميكيّات البيئة حيث تتعلم الخوارزميّة السياسة المثلى مباشرة من خلال التجربة والخطأ والتفاعل مع البيئة. تُصنّف Q-Learning أيضاً بأنها خوارزمية خارج السياسة Off-Policy، ويعني هذا أنها تتعلّم قيمة السّياسة المثلى بغض النظر عن السياسة التي يتبعها العميل فعلياً لاتخاذ الأفعال أثناء مرحلة التعلم. تعتمد هذه الخوارزميّة على آليّة ϵ-greedy لاختيار الفعل التي توازن بين طريقتي الاستكشاف Exploration والاستغلال Exploitation، فتبدأ باختيار الأفعال بطريقة الاستكشاف حيث تكون قيمة $\epsilon = 1$ وتتناقص فيما بعد لتنعدم في نهاية عملية التعلّم فتقوم بالاستغلال باحتمال $\epsilon = 1$. تخزّن خوارزميّة Q-Learning قيم تابع القيمة في بنية بيانات جدوليّة Q-Table، وهو فعّال فقط في حال كان فضاء الحالات والأفعال صغيرين ومنتهيين ويُعتبر هذا القيد تحدّي لهذه الخوارزميّة.[12]

تهدف Q-Learning إلى إيجاد السياسة المثلى عن طريق تعظيم قيمة التابع $Q(s,a)$ الذي يمثّل القيمة المتوقّعة لعميل ما في حالة معينة s واتخاذه لفعل معين a. تختار Q-Learning عند كل خطوة زمنيّة t في حالة ما s_t فعل ما a يعظّم تابع القيمة

الخاصة بها، وبشكل متكرّر تحصل في النهاية على السياسة المثلى π. تستخدم هذه الخوارزميّة معادلة بيلمان لتحديث قيمة Q بشكل تكراري وفق المعادلة:

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha\left[R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a)\right] \qquad (17)$$

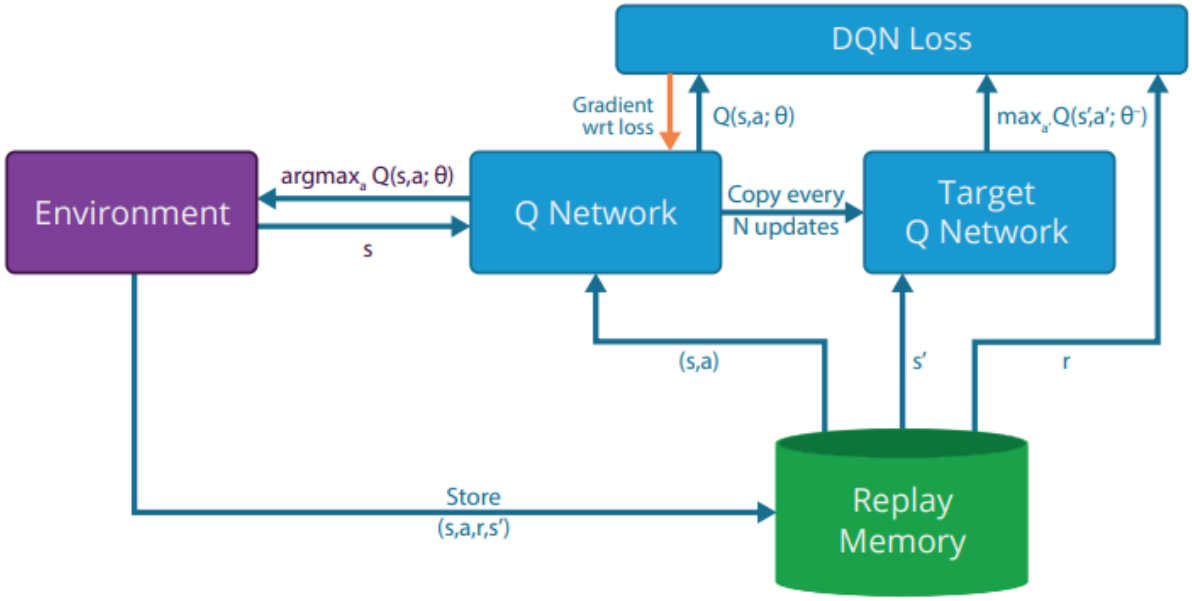
يمثّل $Q_t(s_t, a_t)$ تابع القيمة (حالة - فعل)، و α هو معدّل التعلّم Learning Rate حيث $0 < \alpha \leq 1$، و r_t هي المكافأة الفوريّة، و γ معامل التناقص Discount Factor حيثُ $0 < \gamma < 1$. تمثّل $\max_a Q_t(s_{t+1}, a)$ قيمة تابع القيمة-الفعل المثلى في الخطوة الزمنية التالية على جميع الأفعال الممكنة a.[8]

5-3-3 خوارزميّة Deep Q-Learning

تُعد خوارزمية (DQL: Deep Q-Learning) امتداداً مباشراً لخوارزمية Q-Learning، وهي مصمّمة للتغلّب على التحدّي الأساسي المتمثّل في الفضاءات المستمرّة. في خوارزمية Q-Learning يتم تخزين تابع القيمة $Q(s, a)$ في بنية بيانات جدولية Q-Table، وكما ذُكر سابقاً هذه المقاربة المعروفة بالطرق الجدولية تكون فعالة فقط عندما يكون فضاء الحالات State Space وفضاء الأفعال Action Space صغيرين ومنتهيين. أمّا في المسائل المعقدة والواقعيّة مثل تخصيص الموارد في شبكات الراديو الإدراكي أو البيئات ذات المدخلات الحسيّة عالية الأبعاد يصبح من المستحيل حسابياً أو من ناحية الذاكرة إنشاء وإدارة هذا الجدول. لحل هذه المشكلة، تستبدل خوارزمية DQN الجدول بطريقة أخرى لمقاربة تابع قيمة غير خطي باستخدام الشبكة العصبونية العميقة (DNN: Deep Neural Network). بدلاً من البحث عن قيمة $Q(s, a)$ في جدول يتم تغذية الحالة s كمدخل للشبكة العصبونية، وتقوم الشبكة بتقدير قيم Q لجميع الأفعال الممكنة a في تلك الحالة كمخرجات للشبكة.

لتحسين أداء التعلّم تُضمّن فكرة Experience Replay في خوارزميّة DQL وهي عبارة عن Buffer يحتوي على سلاسل الانتقالات للعميل $e(t)$، بحيث تخزّن كل تجربة مع البيئة كصفّ Tuple في ذاكرة تسمّى Replay Memory بالشكل التالي:

$$e(t) = a(t), s(t), r(t), s(t+1)$$

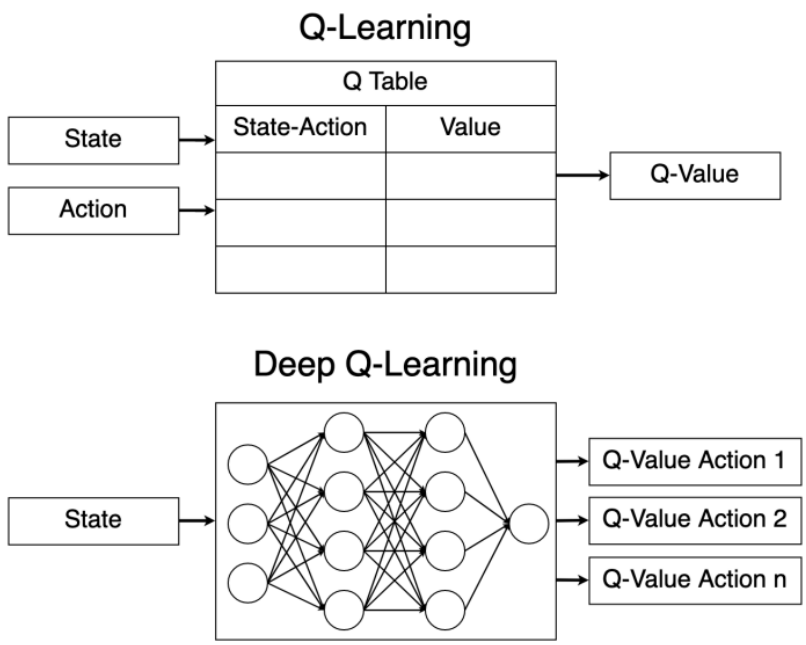


الشكل 7: يوضّح بنية خوارزميّةDQL.[9]

يتم تدريب الشبكة العصبونية على تقليل الخطأ بين تقديرها الحالي وبين القيمة المستهدفة Target Value المستمدّة من معادلة بيلمان الأمثلية Bellman Optimality Equation. يتم تكييف معادلة التحديث الخاصة بـ Q-Learning لتصبح دالة خسارة Loss Function للشبكة، وغالباً ما تكون متوسّط الخطأ التربيعي MSE بين القيمة المتوقّعة والقيمة المستهدفة (المعادلة 18).

$$L(\theta) = \mathbb{E}[\left(\mathrm{r}(\mathrm{s}, a) + \gamma \max_{a'}(q_*(S_{t+1}, a', \theta^-)) - \mathrm{q}(\mathrm{s}, a; \theta)\right)^2] \qquad (\mathbf{18})$$

تمثّل كلّاً من θ^-, θ البارامترات الحاليّة والقديمة على الترتيب. عند كلّ خطوة زمنيّة يتم اختيار دفعة مصغّرة عشوائيّة من المدخلات Random mini-Batch في الذاكرة لتدريب معاملات الشبكة العصبونيّة، وهذا ذو اهميّة كبيرة في تقليل التباين بين العيّنات لأنّ عشوائيّة العيّنات تكسر الارتباطات القويّة التي تنشأ نتيجة تتالي العيّنات.[12] تحتاج DQL إلى شبكتين، الأولى شبكة سياسة Policy Network لتقدير $\mathrm{Q}(s_t, a_t)$، والثانية شبكة هدف Target Network تقوم بحساب القيمة المستهدفة $r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a)$. يتمّ تحديث أوزان شبكة الهدف بسرعة أقلّ أي أنّ جميع بارامترات شبكة السّياسة يتمّ نسخها ممّا يمنع القيمة المستهدفة من التذبذب السريع ويجعل عمليّة التدريب أكثر استقراراً.[12]



الشكل 8: يوضّح الفرق بين QL وDQL من حيث البنية الأساسيّة.[18]

4-3 معيار MOS

يُعد متوسط درجة الرأي (MOS: Mean Opinion Score) المقياس الشّخصي subjective metric الأكثر شيوعاً واستخداماً لقياس جودة التجربة QoE كما يُدركها المستخدم النهائي. يُعرَّف MOS بأنه القيمة المتوسطة للتقييمات التي يقدمها مجموعة من الأشخاص المستخدمين في بيئة اختبار خاضعة للرقابة. يتم تجميع هذه التقييمات ضمن مقياس معياري خماسي النقاط five point scale يتراوح من 1 وهي قيمة سيّئة إلى 5 وتعتبر قيمة ممتازة، وذلك لوضع قياس كمّي لمستوى الرضا عن جودة الخدمة. تعتبر قيمة MOS شديدة الاعتماد على التطبيق application-dependent حيثُ أنّ العوامل التي تؤثر

على جودة التجربة تختلف جذرياً بين خدمات مثل نقل الصوت VoIP أو بث الفيديو video streaming أو خدمات البيانات. نظراً لأنّ إجراء الاختبارات الذاتية لتقييم MOS في الزمن الحقيقي غير عملي لإدارة الشبكة يتمّ اللجوء إلى النّماذج الموضوعيّة objective models. تقوم هذه النماذج بمقابلة mapping مقاييس أداء الشبكة الموضوعية مثل التأخير أو معدّل فقدان الرزم أو معدل ذروة الإشارة إلى الضجيج (PSNR) بقيمة MOS المتوقّعة.[20][19]

نمذجة MOS لحركة البيانات:

بالنسبة لخدمات البيانات Data Services يتمّ غالباً نمذجة MOS كتابع لوغاريتمي مرتبط مباشرة بمعدّل الإرسال R الذي يختبره المستخدم النهائي. تُستخدم صيغ مثل:

$$MOS_D = alog_{10}(bR) \qquad (\mathbf{19})$$

حيث a, b هي معاملات تعتمد على نوع التطبيق.

نمذجة MOS لحركة الفيديو:

تُعد نمذجة MOS للفيديو أكثر تعقيداً، فغالباً ما يُستخدم مقياس معدل ذروة الإشارة إلى الضجيج (PSNR: Peak SNR) كمقياس جودة موضوعي لأداء الترميز. ومع ذلك من المعروف أن PSNR لا يعكس دائماً الإدراك البشري الذاتي بدقّة، لذلك يتمّ استخدام تابع لربط PSNR بـ MOS. يمكن صياغة MOS بالصيغة التالية:

$$MOS_V = \frac{a}{1 + \exp\left(b(PSNR - c)\right)} \qquad (\mathbf{20})$$

حيثُ تُعطى PSNR بالعلاقة التالي:

$$PSNR = klog_{10}(R) + p \qquad (\mathbf{21})$$

تربط هذه العلاقة قيمة PSNR بقيمة معدّل الإرسال R، ومن ثمّ تُستخدم العلاقة السابقة (21) لترجمة قيمة PSNR إلى قيمة MOS النهائية. في أبحاث تخصيص الموارد يُستخدم MOS سواء MOS_D أو MOS_v كمقياس الأداء الذي تسعى خوارزميات التعلم المعزّز إلى تعظيمه لضمان تلبية متطلبات المستخدم النهائي.

يبيّن الجدول التالي دلالات قيم MOS:

MOS	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

جدول 2: دلالات قيم MOS. [19]

5-3 خاتمة

قدّم هذا الفصل تعريف مفصّل بالتعلّم المعزّز وتصنيفاته، ودراسة واسعة عن بنية التعلّم المعزّز كسيرورة ماركوفية. بالإضافة إلى شرح نظري لخوارزميّات العمل ومعيار MOS.

الفصل الرّابع

النمذجة

يقدّم هذا الفصل نمذجة المسألة وشرح نموذج العمل من مكوّنات وفرضيّات وقيود، وتطبيق خوارزميّات التعلّم المعزّز المستخدمة وعناصرها.

1-4 نموذج العمل

يتكون نموذج العمل من شبكتين لا سلكيّتين تتشاركان حزمة طيف راديوية. الشبكة الأوّلية PN: Primary Network هي المسؤولة عن حزمة الطيف، وشبكة ثانويّة SN: Secondary Network تستخدم هذه الحزمة بحسب ما تسمح به الشبكة الأوّليّة، وذلك باتّباع أحد أنماط النفاذ الديناميكي للطيف DSA: Dynamic Spectrum Access وهو نمط Underlay. في الشبكة الأوليّة مستخدم أوّليPU: Primary User ومحطة قاعدية أوّلية PBS: Primary Base station، وفي الشبكة الثانويّة مجموعة مستخدمين ثانويين SUs: Secondary Users بعدد N تربطهم محطة قاعدية ثانويّة SBS: Secondary Base Station يتشاركون قناة المستخدم الأوّلي لاستخدامها في نقل بيانات وسائط متعددة Multimedia بنمطين (Video, Data). تبعد المحطتان عن بعضهما مسافة 2.5 km، ويتوزّع كل من المستخدم الأوّلي والمستخدمين الثانويين حول محطاتهم القاعدية ضمن دائرة نصف قطرها 300 m.

الشكل 9: يوضّح نموذج العمل.

نعتبر أنّ القناة شبه ثابتة (غير متغيرة زمنيّاً)، وتخضع لضجيج أبيض غوصي باستطاعة σ^2. يُرمز لاستطاعة إرسال المستخدم الأوّلي P_0 واستطاعة إرسال المستخدم الثانوي رقم i يُرمز لها ب P_i. تُعطى نسبة استطاعة الإشارة المرسلة إلى استطاعة إشارات التداخل بالإضافة لاستطاعة الضجيج $SINR$ للمستخدم الأوّلي بالعلاقة التالية:

$$SINR_0^P = \frac{G_0^P P_0}{\sum_{i=1}^{N} G_i^P P_i + \sigma^2} \qquad (22)$$

- G_0^P : ربح القناة من المستخدم الأولي وصولاً للمحطة القاعديّة الأوليّة.
- G_i^P : ربح القناة من المستخدم الثانوي i وصولاً للمحطة القاعدية الأوّليّة.

وبشكل مماثل نسبة $SINR$ للمستخدم الثانوي:

$$SINR_i^S = \frac{G_i^S P_i}{G_0^S P_0 + \sum_{j=1, j\neq i}^{N} G_j^S P_j + \sigma^2} \qquad (\mathbf{23})$$

- G_i^S : ربح القناة من المستخدم الثانوي i وصولاً للمحطة القاعديّة الثانويّة.
- G_j^P : ربح القناة من المستخدم الثانوي j وصولاً للمحطة القاعدية الأوّليّة.
- G_0^S : ربح القناة من المستخدم الأوّلي وصولاً للمحطة القاعدية الثانويّة.

يفرض النفاذ الديناميكي للطيف من نمط underlay شروط على المستخدمين الثانويين لتجنّب حدوث تداخل مع إشارات المستخدم الأوّلي أو فيما بينهم حيث يوجد حدّ للتداخل لكل مستخدم، ويمكن التعبير عنها بالشكل التّالي:

$$SINR_0^P \geq K_0 \; ; \; SINR_i^S \geq K_i \qquad (\mathbf{24})$$

حيث K_0 عتبة التداخل الخاصة بالمستخدم الأوّلي، و K_i عتبة التداخل الخاصة بالمستخدم الثانوي i. من العلاقتين (23) و(24) نجد أنّ الحدّ الأدنى ل P_i تكتب بالشكل التالي:

$$P_i = \frac{\beta_i(\sigma^2 + G_0^S P_0)}{G_i^S(1 - \sum_{i=1}^{N} \beta_i)} \; ; \quad \beta_i = \left(1 + \frac{1}{K_i}\right)^{-1} \qquad (\mathbf{25})$$

لتحقيق شرط تخصيص الاستطاعة (لتجنّب التشوّه) يجب أن تتحقّق المتراجحة:

$$1 - \sum_{i=1}^{N} \beta_i > 0 \; ; \quad \beta_i = \left(1 + \frac{1}{K_i}\right)^{-1} \qquad (\mathbf{26})$$

ويمكن أن يكتب شرط التداخل في (24) بعد تعويض العلاقة (22) في (25) بالشكل التالي:

$$\sum_{i=1}^{N} \alpha_i \beta_i \leq 1; \quad \alpha_i = \frac{G_i^P(\sigma^2 + G_0^S P_0)}{G_i^S\left(\frac{G_0^P P_0}{K_0} - \sigma^2\right)} + 1 \qquad (\mathbf{27})$$

نفترض أنّ الوصلة الأوّلية والوصلات الثانويّة تُرسِل مستخدمةً تعديل وترميز تكيّفي (AMC)، بحيث يتكيّف نوع التعديل ومعدّل ترميز القناة وفقاً لحالة وصلة الإرسال وعادةً ما تقاس بدلالة SINR.[1] ضمن هذه الفرضية تقوم المحطة القاعدية الثانوية بالحفاظ على تطبيق الشروط في (26) و(27).

تعطى العلاقة بين معدّل نقل بيانات الأعظمي (سعة القناة) للمستخدم الثانوي i وعتبة التداخل الخاصّة به بالشكل التالي:

$$R_i^S = W log_2(1 + k.SINR_i^S) \qquad (\mathbf{28})$$

W هي عرض طيف القناة التردّدي. $(1 + k.SINR_i^S)$ تمثّل عدد البتات في رمز التعديل، حيث $k = \frac{1.5}{-ln(5.BER)}$ ثابت يتعلّق بمتطلّبات معدّل خطأ بت الإرسال BER الأعظمي المطلوب. بالنسبة للنفاذ الديناميكي للطيف من النمط underlay يقوم المستخدمون الثانويّون بتخصيص الموارد عبر اختيار قيم $SINR$ المناسبة من قبل المحطّة القاعدية الإدراكيّة CBS، بما يحقق الشروط المفروضة في (24).[1]

2-4 معايير الأداء

إن تقييم الأداء المستند إلى جودة تجربة المستخدم QoE يكتسب اهتماماً كبيراً، وبناءً على ذلك اسُتخدمت جودة تجربة المستخدم كمقياس لأداء الشبكة من حيث جودة نقل البيانات. من بين المقاييس المستخدمة لتقييم جودة تجربة المستخدم تمّ اختيار مقياس MOS بسبب شعبيّته واستخدامه الواسع. نظرًا لوجود نمطين من البيانات (Data & Video) مع اختلاف متطلّبات كلّ منها، فإنّ لحساب قيمة MOS تابعين:

$$Q_D = alog_{10}(bR^s)\,;a = 1.8719, b = 0.795 \qquad (\mathbf{29})$$

- Q_D: قيمة MOS لحركة البيانات العاديّة.
- p_{e2e}: احتمال فقد رزمة البيانات من طرف إلى طرف.
- a و b ثوابت.

$$Q_V = \frac{c}{1 + \exp\left(d(PSNR - f)\right)};\ PSNR = k.log(R^s) + p \qquad (\mathbf{30})$$
$$c = 6.5433, d = -0.1433, f = 31.4266, k = 10.4, p = -28.7221$$

- Q_V: قيمة MOS لحركة نقل بيانات الفيديو.
- $PSNR$: نسبة الإشارة إلى الضجيج العليا.
- c, d, f, k, p: ثوابت.

قيم الثوابت الموجودة في المعادلة (26) والثوابت c, d, f الموجودة في المعادلة (30) بالإضافة إلى المعادلات مأخوذة من المرجع [1]، أمّا الثوابت k, p مأخوذة من المرجع [6]. بالإضافة إلى معيار MOS تم استخدام معدّل الازدحام Congestion Rate وعدد التكرارات حتى التقارب Iteration number till convergence، وهذه هي المعايير المستخدمة في بعض الأدبيّات أيضاً ممّا يتيح المقارنة بسهولة.

3-4 سيناريو العمل

أُجريت محاكاة نموذج العمل على برمجية MATLAB. في البداية تمّ تعريف مصفوفة قيم SINR التي تتكوّن من 11 قيمة حيثُ أنّ كلّاً منها تمثّل عتبة تداخل، ومصفوفة أعداد المستخدمين الثانويّين SUs، وأيضاً معادلات الحصول على قيم MOS لكل نمط من البيانات Data & Video. تتكرّر عمليّة التعلّم لكل عدد مستخدمين ثانويّين عدد محدّد من المرّات Episodes ليتمّ حساب متوسّط كل نتيجة. في كل Episode تُولّد قيم جديدة لمواقع كل من المستخدم الأوّلي والمحطة القاعدية للمستخدمين الثانويين ومجموعة المستخدمين الثانويين بتوزّع منتظم ضمن المناطق المخصّصة لهم بحسب ما تم شرحُه سابقاً في نموذج العمل من خلال تابع خاصّ، بالإضافة إلى تحديد نمط البيانات بشكل عشوائي بتوزّع منتظم لكل مستخدم ثانوي. بعد ذلك يتم تطبيق خوارزميّة العمل، ومن ثمّ حساب القيم الوسطيّة لكل من MOS ومعدّل الازدحام وعدد التكرارات اللازم للوصول إلى التقارب، في نهاية المحاكاة تُستعرض النتائج بالنسبة لعدد المستخدمين الثانويين لكل من المعايير السابقة.

تم اختيار خوارزميّات QL وDQL لمقاربة الحلّ وذلك لاعتمادهما في كثير من الأدبيّات وقدرتهما على تحقيق نتائج جيّدة جدّاً في جودة التجربة، بالإضافة إلى المرونة في العمل باستخدام أي تقنيّة والتكيّف مع البيئة بسلاسة. إنّ إجراء المحاكاة تطلّب

إنشاء البيئة الخاصّة بالتعلّم المعزّز التي تتكوّن من مجموعة حالات ومجموعة أفعال والمكافأة وتابع القيمة. الجدول التّالي يعرض قيم البارامترات الخاصّة بالمسألة:

البارامیتر	القيمة
مجموعة عتبات تداخل المستخدمين الثانويّين	[-5:2:15] dB
مجموعة أعداد المستخدمين الثانويّين	1:1:9
استطاعة الضجيج	1 nW
استطاعة إرسال المستخدم الأوّلي	10 mW
المعامل الأسّي لفقد المسار	2.9
عرض الحزمة	10 MHz
عتبة تداخل المستخدم الأوّلي	1.2 dB
معدّل الخطأ BER	10^{-5}
Episodes number	2000
Iterations number	100

جدول 3: يعرض قيم بارامترات النموذج.

4-4 خوارزميّة Q-Learning

تعتبر خوارزميّة QL البيئة كنظام ديناميكي عشوائي ذي حالات منتهية وزمن متقطّع. يلاحظ العميل Agent حالته الراهنة $s \in S$ وبناءً عليها يتّخذ فعل $a \in A$، هذا ما يؤدّي إلى حصوله على مكافأة فوريّة قياسيّة R_t. تصبح المسألة عبارة عن إيجاد سياسة ما تعظّم المكافأة التراكميّة الإجماليّة. فيما يلي بنية التعلّم المعزّز لهذه الخوارزميّة:

1-4-4 العميل Agent

تقوم المحطّة القاعديّة الثانويّة SBS أو ما تعرف بالمحرّك الإدراكي Cognitive Engine بدور العميل Agent في نموذج التعلّم المعزّز لتمنح كل مستخدم ثانوي عتبة تداخل لا يمكن أن يتجاوزها بما يتناسب مع شروط التداخل المفروضة في (21). إنّ اختيار المحطّة القاعديّة الثانويّة كعميل في التعلّم المعزّز هو قرار تصميمي يفرض نموذج تحكّم مركزي بناءً على الأسباب التالية:

- المحطّة القاعديّة الثانويّة هي الكيان الوحيد الذي يمتلك النظرة الشاملة اللازمة لتقييم هذه المنفعة الكليّة، فهي التي تستطيع الموازنة بين احتياجات المستخدمين الثانويّين وتخصيص الموارد لتحقيق جودة تجربة أفضل.
- يعمل النظام وفق نمط underlay الذي يفرض قيوداً صارمة على عتبات تداخل المستخدمين الثانويين، فالمستخدم الثانوي الفردي SU لا يمتلك أي معلومات عن غيره من المستخدمين الثانويّين لذلك من الصعب على عميل لامركزي أن يضمن احترام هذه الشروط بالشكل الأمثلي.
- على الرغم من أنّ نموذج التحكّم اللّامركزي مفيد لتقليل الحمل overhead، لكنّ اختيار عميل واحد يمنع التعقيدات المرتبطة بتضارب السّياسات الذي قد يحدث عند تعلّم عدّة عملاء بشكل متزامن، ويغيّر كلّاً منهم سلوك الآخر.

2-4-4 فضاء الأفعال Action Space

تتّخذ المحطّة القاعديّة الثانويّة SBS (العميل) الأفعال من فضاء منتهي متقطّع $A(i) = \{a_1, a_2, \ldots\ldots, a_{11}\}$، وهي قيم SINR التي تمثّل عتبات التداخل للمستخدمين الثانوي وعددها 11 فعل. في كل تكرار iteration يتم اتّخاذ مجموعة من الأفعال يتغيّر حجمها بحسب عدد المستخدمين الثانويّين SU بهدف منح كل مستخدم ثانوي قيمة SINR مناسبة التي لا تلبّي الشروط المفروضة

فحسب بل تعطي أفضل جودة تجربة ممكنة. تتّبع عمليّة اتخاذ الفعل من قبل العميل ألية ϵ-greed التي تتضمّن طريقتين في الاختيار الأولى استكشاف Exploration والثانية استغلال Exploitation. يختار العميل طريقة الاستكشاف باحتمال $p = \epsilon$ بينما يختار طريقة الاستغلال باحتمال $p = 1 - \epsilon$، بحيث تكون قيمة ϵ أعلى ما يمكن في بداية عمليّة الملاحة navigation (التعلّم) حيثُ يتم اتّباع طريقة الاستكشاف في اختيار الأفعال في الغالبية الساحقة من مرات الاختيار، وتتناقص قيمة ϵ بعدها بحيث تصبح طريقة الاستغلال هي المتّبعة بالغالبية الساحقة في نهاية التعلّم. إنّ استخدام هذه السياسة يُظهر موازنة trade-off جيّدة من حيث الاستكشاف والاستغلال الاستكشاف مطلوب لاكتشاف المزيد عن الاستراتيجية المثلى، ولكن الاستغلال مطلوب أيضاً لاستثمار الحالات المعروفة جزئياً في البيئة.

3-4-4 فضاء الحالات State Spaces

يتكوّن فضاء الحالات من أربع حالات تم تعريفها بحسب قيود نموذج العمل. العلاقة (23) تعبّر عن شرط تجنّب التشوّه (تخصيص الاستطاعة) والعلاقة (24) تعبّر عن شرط التداخل بين المستخدمين.

$$I = \begin{cases} 0; & 1 - \sum_{i=1}^{N} \beta_i > 0 \\ 1; & otherwise \end{cases} \qquad (\mathbf{31})$$

$$L = \begin{cases} 0; & \sum_{i=1}^{N} \alpha_i \beta_i \leq 1 \\ 1; & otherwise \end{cases} \qquad (\mathbf{32})$$

State	I	L
S_1	0	0
S_2	0	1
S_3	1	0
S_4	1	1

جدول 4: يعرض فضاء الحالات في نموذج العمل.

الحالة S_1 هي الحالة الوحيدة التي تضمن احترام الشروط (قيود النموذج)، لأنّ التداخل بين المستخدمين يكون ضمن الحدّ المسموح به ولا يوجد تشوّه لدى أي مستخدم. كل من الحالات الأخرى تُخلّ بإحدى الشروط المفروضة S_2, S_3 أو تُخلّ بكليهما S_4.

4-4-4 المكافأة Reward

يُعرّف تابع المكافأة كتابع للحالة والفعل بالشكل التالي:

$$R(s,a) = \begin{cases} MOS; & s = S_1 \\ cte; & otherwise \end{cases} \qquad (\mathbf{33})$$

طالما أنّ الهدف الرئيسي هو ضمان عدم التداخل وتجنّب التشوّه، فإنّ أي حالة من الحالات الأربعة باستثناء الحالة S_1 تُعد انتهاك للشروط المفروضة وبالتالي يتم تخصيص مكافأة سلبيّة قيمتها تساوي $cte = -5$. لكن في حال الانتقال الأوّل فقط (بعد اتّخاذ أوّل فعل) قيمة المكافأة السلبيّة (العقوبة) فتكون $cte = -0.03$. أمّا في حالة S_1 فتكون المكافأة هي قيمة MOS نفسها بحسب نوع بيانات كل مستخدم سواءً كانت Data أو Video.

5-4-4 السياسة Policy

يتمّ تحديد سياسة π لعمليّة الربط بين الحالة والفعل state to action mapping. تقوم السّياسة المُثلى بتوجيه العميل باتجاه تعظيم جودة التجربة مع الموازنة في نفس الوقت مع القيود المفروضة. يعرّف تابع القيمة الذي يتّبع سياسة π بدلالة المكافأة التراكميّة الإجماليّة $r_t(s_t, a_t)$ والحالة s_t والفعل a_t بالشّكل التالي:

$$Q_\pi(s_t, a_t) = \mathbb{E}_\pi[r_t|s_t, a_t] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \, r_{t+k+1}|s_t, a_t\right]$$

$$Q_\pi(s_t, a_t) = \mathbb{E}[r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})|s_t, a_t] \qquad (\mathbf{34})$$

يعرّف تابع القيمة المُثلى التي تتبع السياسة المُثلى باستخدام معادلة بيلمان العوديّة recursive Bellman equation كما ورد في العلاقة (17):

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha\left[R_{t+1} + \gamma \max_a Q_t(s_{t+1}, a)\right] \qquad (\mathbf{35})$$

حيثُ γ معامل التناقص ويساوي $\gamma = 0.95$، ومعدّل التعلّم وقيمتهُ $\alpha = 0.0001$. تقوم خوارزميّة Q بتحديث قيم تابع القيمة التي تخزّن في جدول Q-Table حتى يستقرّ وتصبح جميع الأفعال التي تؤدّي إلى الهدف معروفة، وبالتالي يكون العميل قد أنهى مرحلة التعلّم وتكوّنت الخبرة اللازمة للوصول إلى الهدف المطلوب. بما أنّ المحطّة القاعديّة الثانويّة هي المتحكّم المركزي، فهي التي تقوم بإسناد قيمة عتبة التداخل المناسبة لكل مستخدم ثانوي، وهذا يتطلّب أن يكون لكلّ مستخدم ثانوي جدول Q خاصّ به حيثُ كل فعل يكافئ إسناد قيم عتبات تداخل مناسبة لكل مستخدم ثانوي. لكنّ الحالة في كل انتقال تكون موحّدة لكل المستخدمين الثانويين وهذا أمر طبيعي لأنّ النموذج ليس متعدّد العملاء بل مركزي.

5-4 خوارزميّة Deep Q-Learning

تعتمد خوارزمية DQL على المكوّنات الأساسيّة في عملية ماركوف لاتخاذ القرار، وتمّ استخدام نفس عناصر التعلّم المعزّز في خوارزميّة QL من حيث العميل وفضاء الحالات وفضاء الأفعال وتابع المكافأة. لكن لمقاربة تابع القيمة تمّ استخدام شبكة عصبونيّة أماميّة التغذية Feed-Forward Neural Network، وتتألف من:

- طبقة الدخل Input Layer: مكوّنة من 4 وحدات تمثّل الحالات الأربعة المختلفة.
- الطبقة المخفيّة الأولى First Hidden Layer: تتألف من 16 عصبون Neuron.
- تابع تنشيط عصبونات الطبقة الأولى: استُخدم تابع (RELU: Rectified Linear Unit).
- الطبقة المخفيّة الثانية Second Hidden Layer: تتألف من 16 عصبون Neuron.
- تابع تنشيط عصبونات الطبقة الثانية: استُخدم تابع RELU.
- طبقة الخرج Output Layer: مكوّن من 11 وحدة تمثّل فضاء الأفعال.

لضمان استقرار التعلّم تم استخدام شبكة هدف Target Network منفصلة، وهي نسخة من شبكة السياسة Policy Network (أي الشبكة الأساسيّة)، حيثُ يتم تحديث أوزان شبكة الهدف ببطء باستخدام تابع التحديث الناعم Soft Update.

1-5-4 بارامترات التعلّم في DQN

يوضّح الجدول التالي البارامترات الخاصّة بعملية التدريب:

البارامیتر	القیمة
episodes	2000
gamma	0.95
lr (Learning Rate)	0.001
batchSize	100
bufferSize	2000
targetTau	0.005
eps0	1.0
epsMin	0.05
epsDecay	30
t_max	100

جدول 5: يعرض قيم بارامترات DQN.

2-5-4 تخزين سلاسل الانتقال

هو عبارة عن مصفوفة متعدّدة الأنماط Tuple تحتوي على معلومات كل انتقال (الحالة؛ الفعل؛ المكافأة؛ الحالة الجديدة) ليتمّ تخزينها في ذاكرة كبيرة تسمّى Replay Buffer. يتمّ سحب دفعة عشوائيّة منها لتدريب الشبكة العصبونيّة، وهذا يؤدّي إلى:

- زيادة استقرار التدريب بحيث يمنع الشبكة من الانحياز نحو مجموعة محدّدة من الخبرات أي سلوك محدّد.
- كفاءة استخدام البيانات عن طريق إعادة استخدام الخبرات السابقة عدة مرات في تحديث بارامترات الشبكة.

6-4 خاتمة

تمّ هذا الفصل استعراض نموذج العمل والمعادلات المستخدمة ومعايير الأداء، وتعريف سيناريو العمل مع تحديد بارامترات البيئة بالإضافة إلى إسناد عناصر التعلّم المعزّز والبارامترات الخاصّة بكلّ خوارزميّة.

الفصل الخامس

النتائج

يعرُض هذا الفصل نتائج تطبيق الخوارزميّات مع مناقشتها، ويقارن بينها وبين نتائج الخوارزميّات في الأدبيّات.

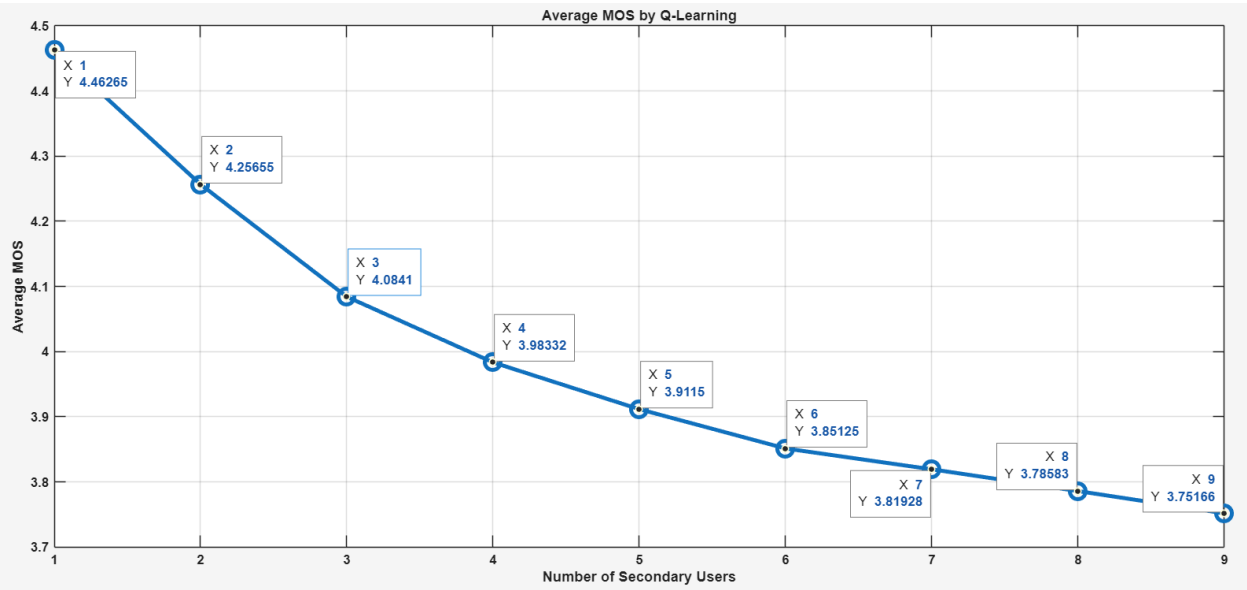
يستعرض هذا الفصل نتائج المحاكاة التي أُجريت لتقييم أداء الخوارزميات المقترحة لتنظيم عمل المحطّة القاعديّة الثانويّة العاملة وفق النمط underlay باعتبارها المحرّك الإدراكي Cognitive Engine. تعملُ المحطّة القاعديّة الثانوية كعميل وحيد مسؤول عن ضمان عدم حدوث تشوّه لدى أي مستخدم ثانوي، أو تداخل بين المستخدمين عن طريق تخصيص عتبات تداخل لجميع المستخدمين الثانويين SUs التابعين لها. تمّ تقييم النتائج باستخدام المعايير المتّبعة في الأدبيّات وهي معيار MOS ومعدّل الازدحام Congestion Rate وعدد التكرارات اللازمة للوصول إلى التّقارب Number of Iteration till Convergence.

في البداية سيتم مناقشة نتائج خوارزميّة Q-Learning، ومن ثمّ دراسة تماسك النموذج باستخدام خوارزميّة Q-Learning. بعد ذلك ستُعرض نتائج خوارزميّة Deep Q-Learning، وفي النهاية سيتمّ مقارنة نتائج الخوارزميّتين مع بعضهما ومع النتائج في الأدبيّات. تمّ اختيار عدد مستخدمين ثانويين يساوي 9 أسوةً بالأدبيّات من أجل تسهيل المقارنة وإظهار التحسينات.

5-1 نتائج استخدام خوارزميّة Q-Learning

فيما يلي نتائج تطبيق خوارزميّة QL وفق المعايير المذكورة سابقاً بدلالة عدد المستخدمين الثانويّين:

5-1-1 معيار MOS

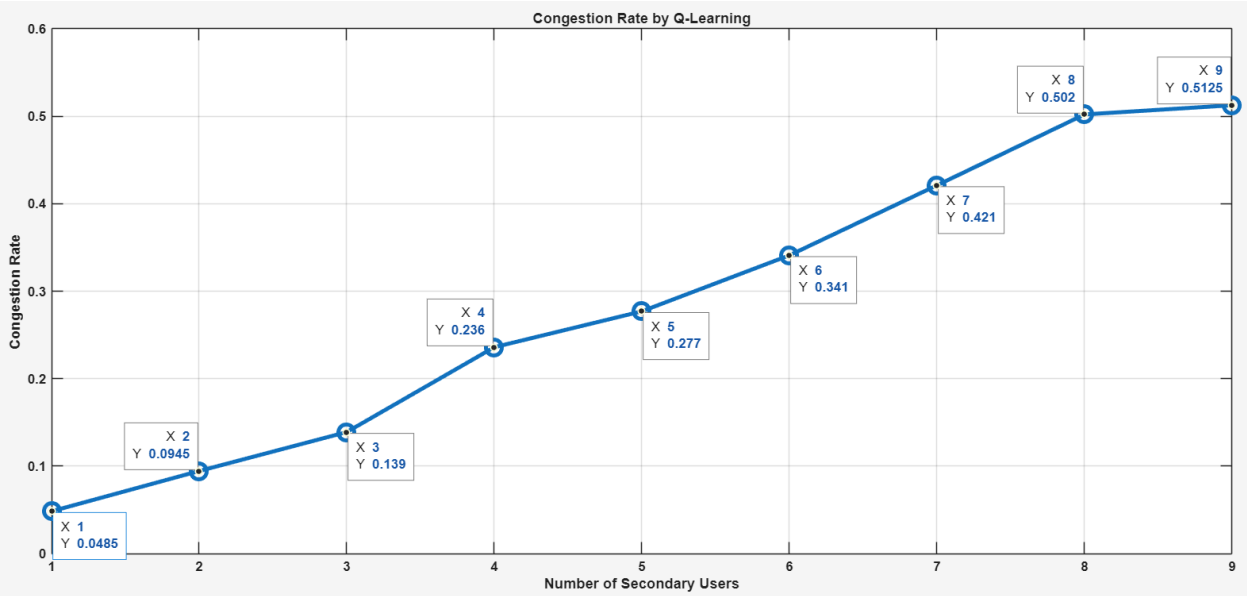


الشكل 10: تغيّر منحني قيم MOS بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL.

يتدرّج المنحني بسلاسة وبشكل متناقص مع تزايد أعداد المستخدمين الثانويّين بدءاً من القيمة 4.463 عند مستخدم ثانوي وحيد

حتّى يصل إلى قيمة 3.752 عند 9 مستخدمين ثانويّين. يحدث هذا التناقص نتيجة تقليل عتبة SINR لكل مستخدم ثانوي من أجل احترام الشّروط المفروضة، وبالتالي تقلّ جودة الخدمة. تعتبر هذه القيم ضمن الحدّ المقبول (MOS > 3).

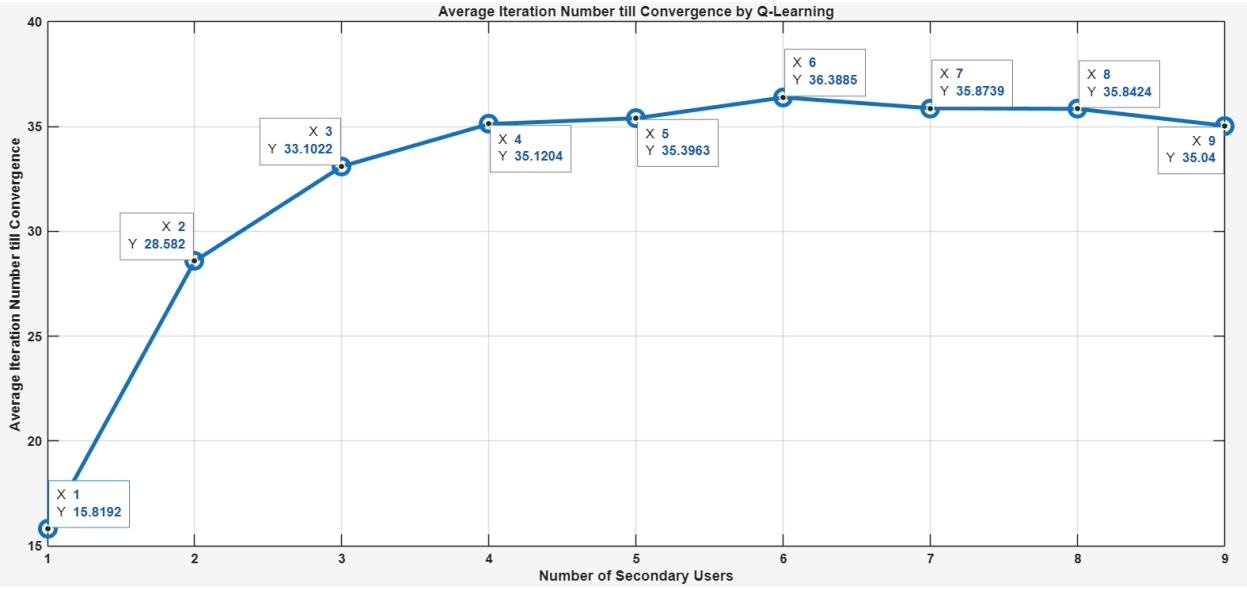
2-1-5 معدّل الازدحام Congestion Rate



الشكل 11: تغيّر منحني معدّل الازدحام بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL.

يعبّر معدّل الازدحام عن النسبة المئويّة لعدد مرات فشل محاولات التقارب. نلاحظ من الشكل (11) أنّ متوسّط معدّل الازدحام وصل إلى 51.25% عند 9 مستخدمين ثانويّين وهذه النّسبة جيّدة جدّاً نسبيّاً، أي نصف عدد محاولات التقارب ناجحة تقريباً.

3-1-5 عدد التكرارات اللّازمة للوصول إلى التقارب



الشكل 12: تغيّر منحني عدد التكرارات اللازمة للتقارب بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة QL.

يمثّل عدد التكرارات كفاءة الخوارزميّة في سرعة وصول العميل إلى الحلّ. من الشكل (12) نجد أنّ متوسّط عدد التكرارات بدأ ب 15.82 بالنسبة لمستخدم وحيد، ثمّ استقرّ على 35 تكرار تقريباً بعد 4 مستخدمين ثانويين. يعتبر عدد التكرارات الذي احتاجه العميل مقبولاً، ولكن بعض الأنظمة قد تحتاج إلى عدد أقلّ من أجل تحقيق سرعة أكبر في التّقارب.

إنّ نتائج تطبيق خوارزميّة Q-Learning أعطى نتائج جيّدة من حيث المعايير الثلاثة. في فقرة لاحقة سيتمّ مقارنة نتائج QL مع نتائج من الأدبيّات لتبيان التحسّن في النتائج.

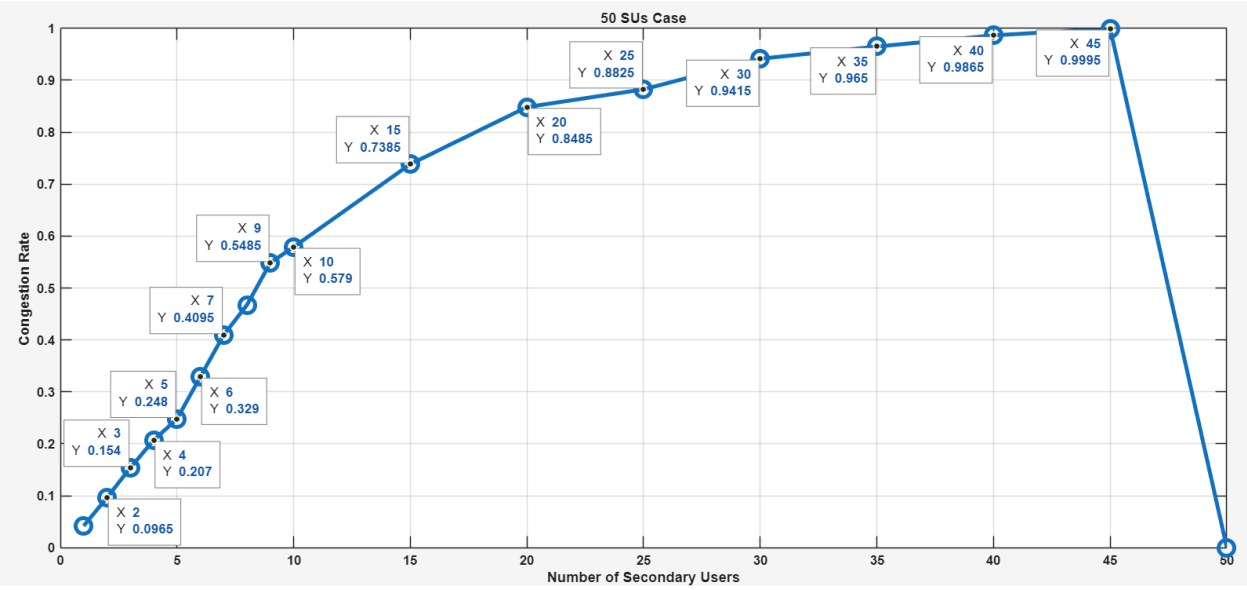
2-5 دراسة تماسك النموذج

بعد إنشاء بيئة التعلم المعزّز وتطبيق خوارزمية Q-Learning تم التأكد من صحّة المحاكاة عن طريق دراسة تماسك النموذج من خلال دراسة أثر عدد مستخدمين ثانويين أكبر وأثر مساحة جغرافيّة أكبر للمستخدمين الثانويين.

1-2-5 أثر تغيّر عدد المستخدمين الثانويين SUs

تمّت المحاكاة على مجموعة أعداد مستخدمين ثانويّين تبدأ بمستخدم وحيد لتصل إلى 50 مستخدم ثانوي وفق معايير التقييم الثّلاث بدلالة عدد المستخدمين الثانويّين.

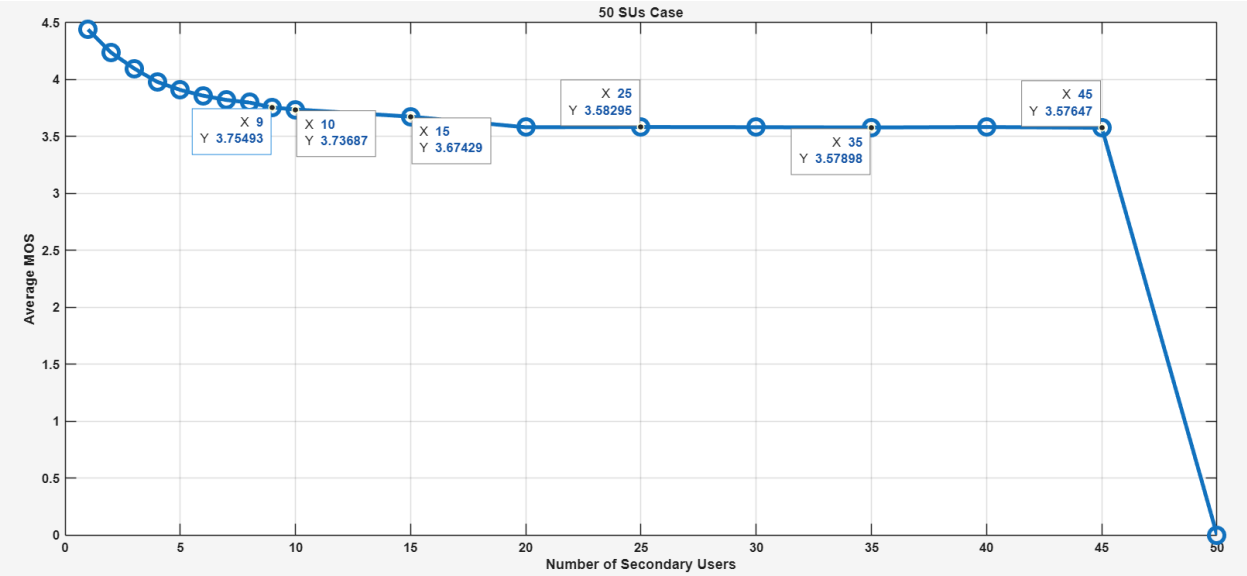
- **معدّل الازدحام Congestion Rate**



الشكل 13: تماسك النموذج من حيث معدّل الازدحام في حالة عدد SUs كبير.

يبيّن الشكل (13) أنّ متوسّط معدّل الازدحام وصل إلى 0.9995 عند 45 مستخدم ثانوي، وهذا يعني أنّ محاولة واحدة من 2000 محاولة تقاربت إلى تحقيق توزيع عتبات تداخل للمستخدمين الثانويّين بحيث ترضي الشروط المفروضة. انهار الأداء عند 50 مستخدم ثانوي. إنّ نتيجة معدّل الازدحام مهمّة لمعرفة محدوديّة النموذج بالنسبة لعدد المستخدمين الثانويّين. نستطيع القول أنّ المحطّة القاعديّة الثانويّة في هذا النموذج تُنجز أداءً مقبولاً حتّى 15 مستخدم ثانوي باحتمال تقارب أكبر من 25%. أمّا عند 9 مستخدمين ثانويّين فبلغ معدّل الازدحام 0.5485 أي احتمال تقارب 45.15%.

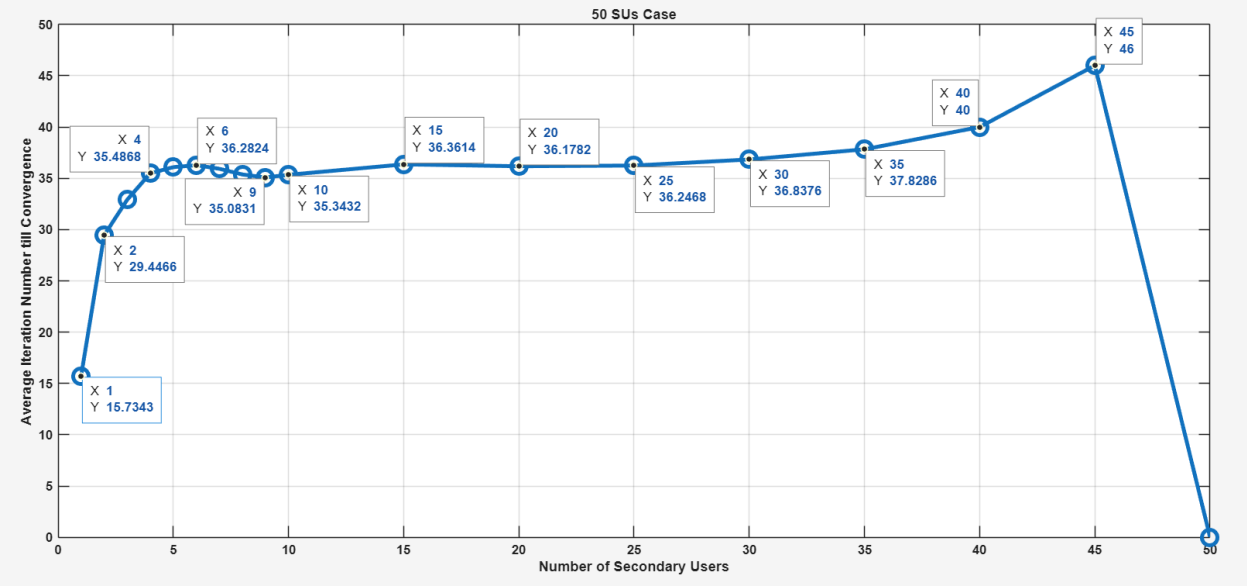
- **معيار MOS**



الشكل 14: تماسك النموذج من حيث قيم MOS في حالة عدد SUs كبير.

نلاحظ في الشكل (14) استقرار متوسّط قيم MOS على القيمة 3.58 ابتداءً من 20 مستخدم ثانوي لينهار النظام عند 50 مستخدم ثانوي. إنّ تلك القيم منطقيّة طالما فرص التقارب نادرة، وهذا يدلّ على أنّ حفاظ النموذج على جودة التجربة هو الذي يتحكّم بمعدّل الازدحام، فالنموذج يعطي عتبات تداخل بحيث تحافظ على الحدّ المقبول من جودة التجربة على حساب احتمال التقارب كما في الشكل (13).

- **عدد التكرارات اللازمة للوصول إلى التقارب**



الشكل 15: تماسك النموذج من حيث عدد التكرارات اللازمة للتقارب في حالة عدد SUs كبير.

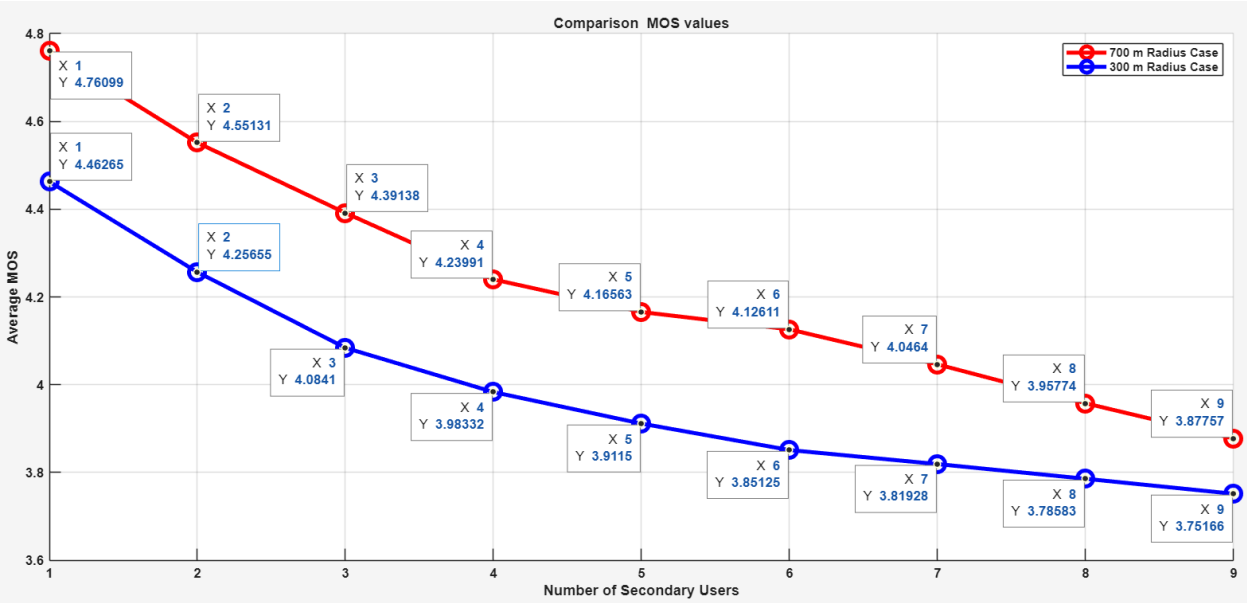
يبيّن الشكل السابق (15) أنّ متوسّط عدد التكرارات اللازمة للتقارب استقرّ عند القيمة 36 بشكل تقريبي حتى 30 مستخدم ليصل إلى 46 تكرار عند 45 مستخدم ثانوي على الرغم من ندرة حدوث التقارب في هذه الحالة. في حالة 50 مستخدم ثانوي لم يحصل أي تقارب.

نستنتج من خلال دراسة تماسك النموذج في حال زيادة عدد المستخدمين الثانويين أنّ النموذج يعمل على تحقيق الشروط المفروضة مع الحفاظ على جودة التجربة بالحدّ المقبول (MOS > 3)، وهذا ما يجعل معدّل التّزاحم كبير نسبيّاً عند عدد مستخدمين ثانويّين أكبر من 15 مستخدم. يُفسّر هذا الأداء بتقيّد نموذج العمل بقيم محدّدة من عتبات التداخل بحيث تضمن الأداء المقبول على حساب احتمال التّقارب.

2-2-5 أثر تغيّر المساحة الجغرافيّة المخصّصة للمستخدمين الثانويين SUs

تمّت المحاكاة على 9 مستخدمين ثانويّين يتمّ تموضعهم ضمن دائرة نصف قطرها 700 متر وفق معايير التقييم الثّلاث بدلالة عدد المستخدمين الثانويّين. وبما أنّ عدد المستخدمين الثانويين مساوٍ لعددهم في فقرة نتائج استخدام QL، فتمّ دمج المنحنيين المتقابلين في شكل واحد من أجل كل معيار وذلك لتسهيل المقارنة.

- **معيار MOS**



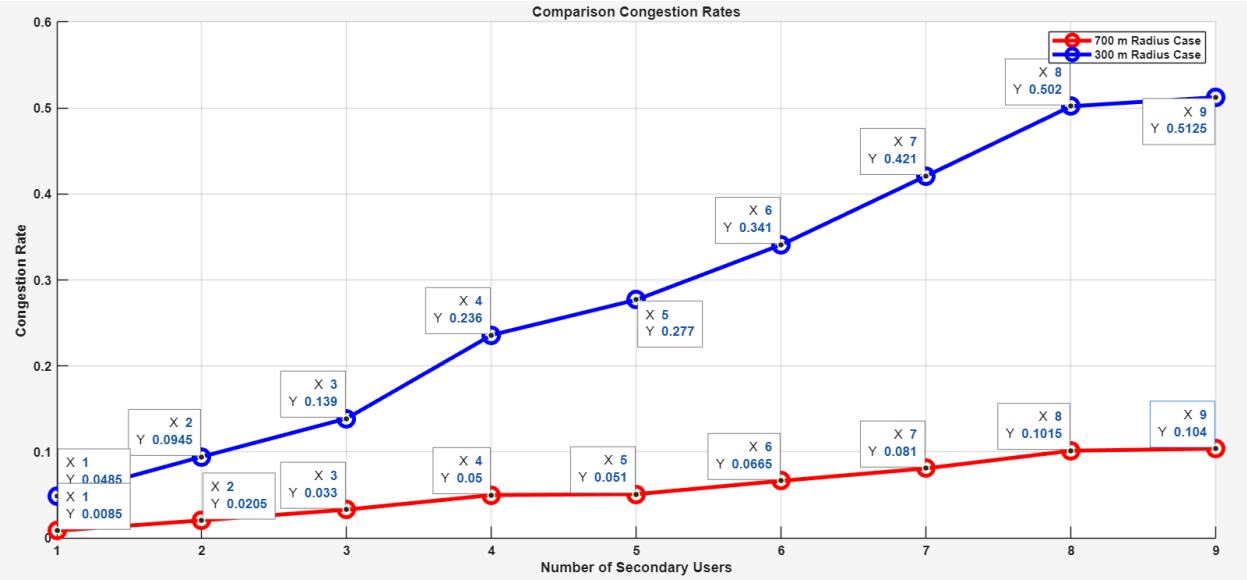
الشكل 16: تماسك النموذج من حيث قيم MOS في حالة مساحة جغرافيّة أكبر.

من الشكل (15) نلاحظ الفرق الكبير بين الحالتين، ويتراوح من 0.3 إلى 0.12. تعتبر هذه النتيجة طبيعيّة جداً، فمن الطبيعي أن يقلّ التداخل بين المستخدمين الثانويين والتداخل مع المستخدم الأوّلي نتيجةً لتوزعهم على مساحة كبيرة، وبالتالي رفع عتبات التداخل للمستخدمين الثانويين أي زيادة SINR لكل منهم.

- **معدّل الازدحام Congestion Rate**

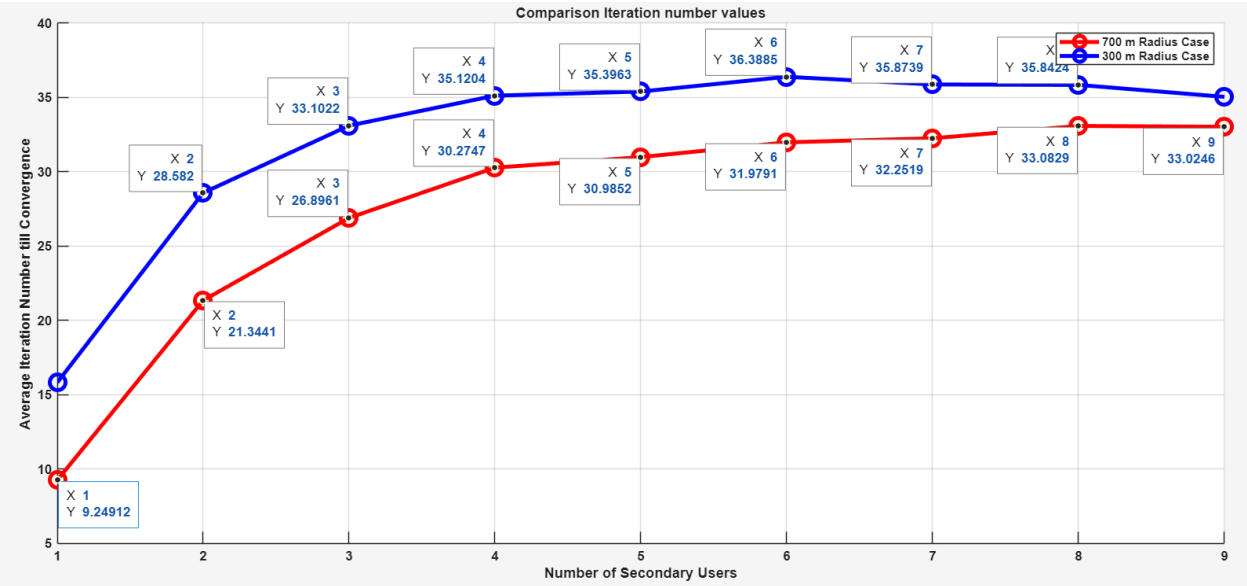
يبيّن الشكل (17) أنّ معدّل الازدحام تناقص بشكل ملحوظ جدّاً في حالة نصف قطر يساوي 700 متر، وبلغ في أعلى قيمة عند

9 مستخدمين ثانويين 10.4%، أي أنّ 1792 محاولة من أصل 2000 محاولة تقاربت إلى إرضاء الشروط المفروضة. تُعتبر النتيجة منطقيّة جدّاً لأنّ ابتعاد المستخدمين الثانويين عن بعضهم يؤدّي إلى تداخل أقلّ بينهم، وبالتالي احتمال تحقيق الشروط المطلوبة أكبر.



الشكل 17: تماسك النموذج من حيث معدّل الازدحام في حالة مساحة جغرافيّة أكبر.

- **عدد التكرارات اللازمة للوصول إلى التقارب**



الشكل 18: تماسك النموذج من حيث عدد التكرارات اللازمة للتقارب في حالة مساحة جغرافيّة أكبر.

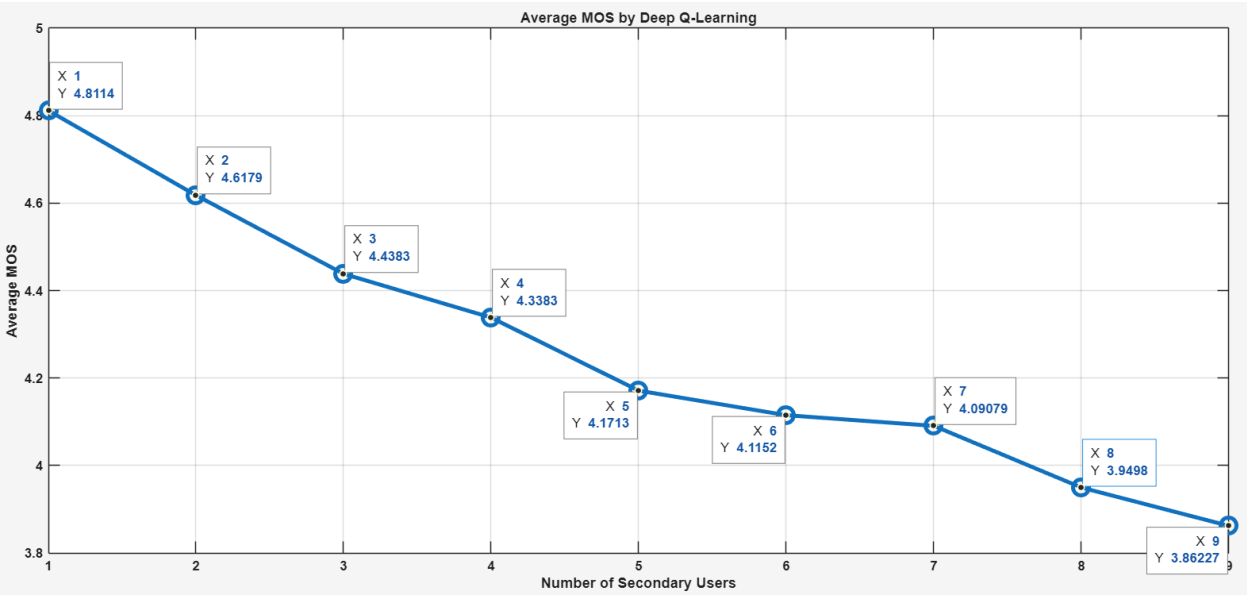
نلاحظ من الشكل (18) انخفاض عدد التكرارات اللازمة للتقارب بشكل ملحوظ، ولكن بفارق صغير نسبيّاً ممّا يدلُّ على ازدياد سرعة التّقارب في حالة نصف قطر يساوي 700 متر، وهذا نتيجة منطقيّة أيضاً بسبب ابتعاد المستخدمين الثانويين عن بعضهم ما يؤدّي إلى تداخل أقلّ بينهم، وبالتالي سرعة أكبر في الوصول إلى التقارب.

من خلال دراسة تماسك النموذج من حيث أثر تغيّر عدد المستخدمين الثانويين وأثر تغيّر المساحة الجغرافيّة المخصّصة للمستخدمين الثانويين يمكن تبيان الاستجابة الطبيعيّة المتوقّعة لهذه التغيرات مما يدلّ على صحة المحاكاة وبالتالي تماسك النموذج.

3-5 نتائج استخدام خوارزميّة Deep Q-learning

فيما يلي نتائج تطبيق خوارزميّة DQL وفق المعايير المذكورة سابقاً بدلالة عدد المستخدمين الثانويّين:

1-3-5 معيار MOS

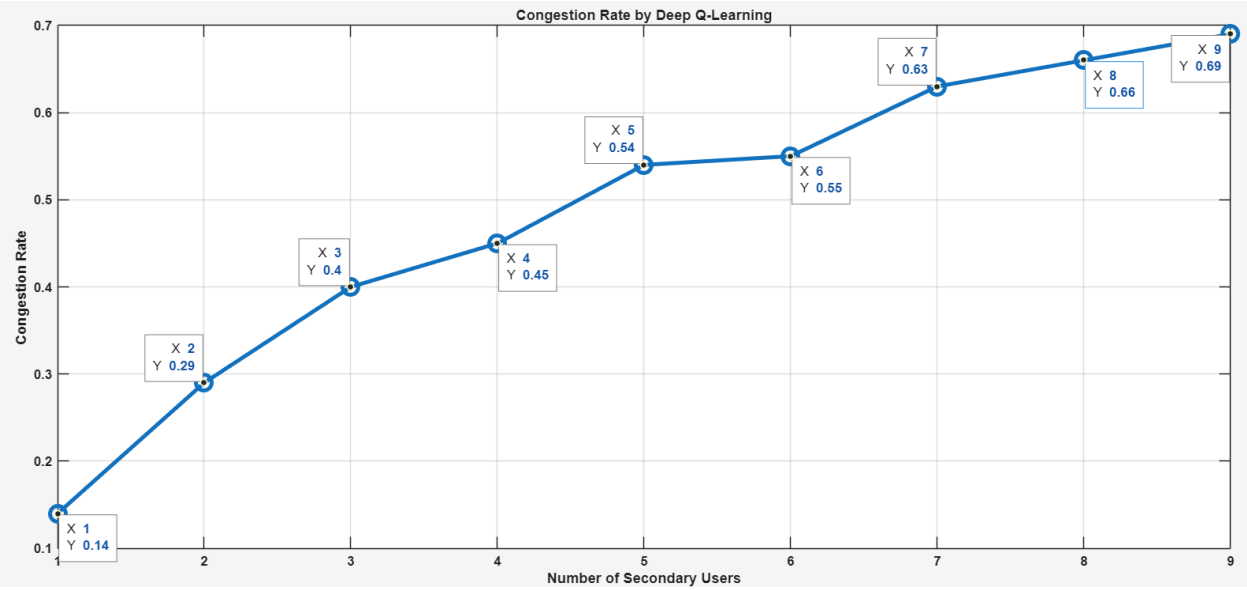


الشكل 19: تغيّر منحني قيم MOS بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL.

يبيّن الشكل 19 تدرّج المنحني بشكل متناقص مع تزايد أعداد المستخدمين الثانويّين بدءاً من القيمة 4.81 عند مستخدم ثانوي وحيد حتّى يصل إلى قيمة 3.86 عند 9 مستخدمين ثانويّين. جميع قيم MOS أكبر من 4 باستثناء آخر قيمتين، أي أنّ جودة الخدمة تبقى بالحدّ الجيد.

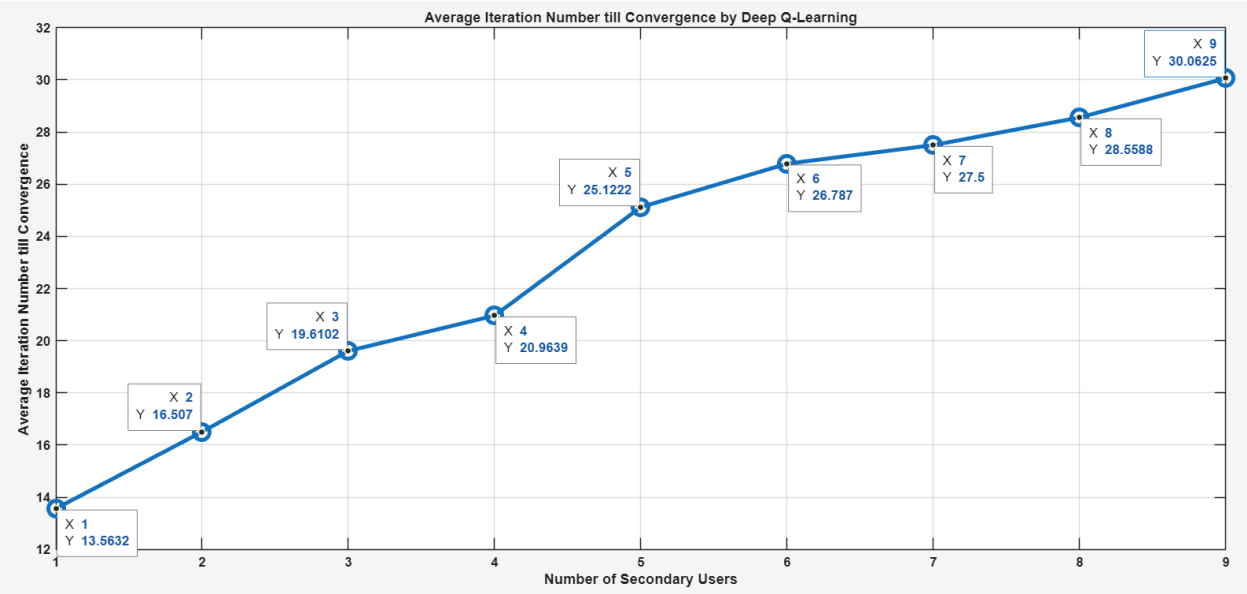
2-3-5 معدّل الازدحام Congestion Rate

نلاحظ من الشكل (20) أنّ متوسّط معدّل الازدحام بدأ بنسبة 14% في حالة مستخدم وحيد، وَوصل إلى 69% عند 9 مستخدمين ثانويّين وهذه النّسبة جيّدة نسبيّاً.



الشكل 20: تغيّر منحني معدّل الازدحام بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL.

3-3-5 عدد التكرارات اللازمة للوصول إلى التقارب



الشكل 21: تغيّر منحني عدد التكرارات اللازمة للتقارب بدلالة عدد المستخدمين الثانويين باستخدام خوارزميّة DQL.

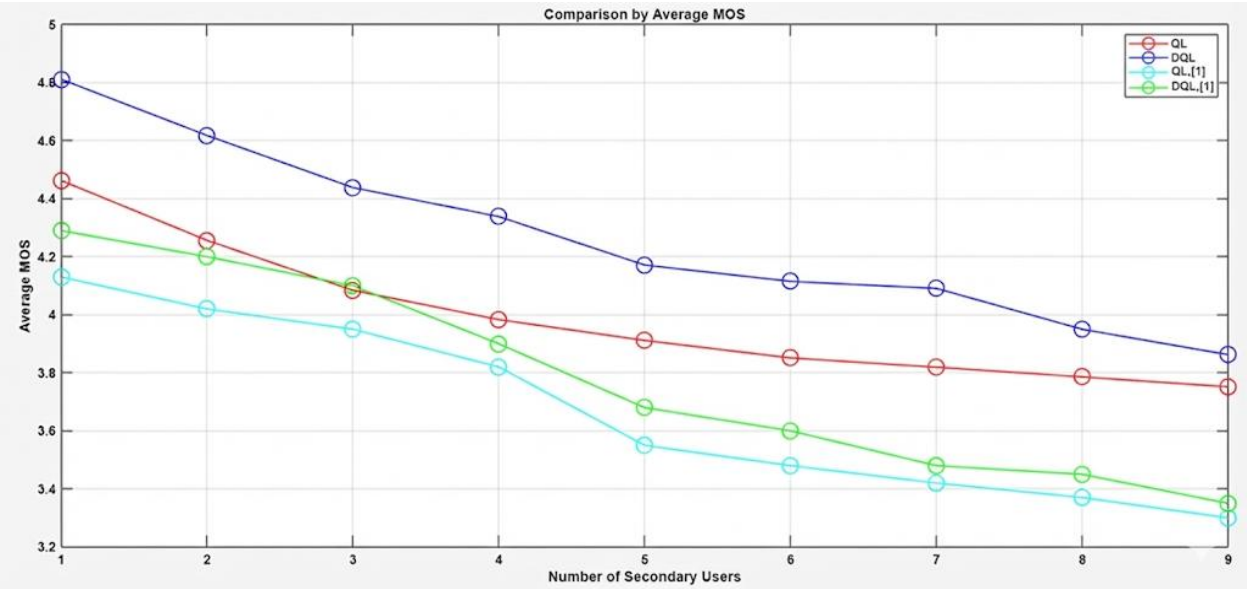
من الشكل (21) نجد أنّ متوسّط عدد التكرارات بدأ ب 13.56 بالنسبة لمستخدم وحيد، ثمّ تابع متزايداً حتّى 30.06 تكرار عند 9 مستخدمين ثانويّين. يعتبر عدد التكرارات الذي احتاجه العميل جيّداً، ويدلّ على سرعة الوصول إلى الحلّ المطلوب.

نتائج تطبيق خوارزميّة Deep Q-Learning تعتبر جيّدة جداً من حيث المعايير الثلاثة. في الفقرة التالية سيتمّ مقارنة هذه النتائج مع نتائج من الأدبيّات لتبيان التحسّن في النتائج.

4-5 مقارنة النتائج

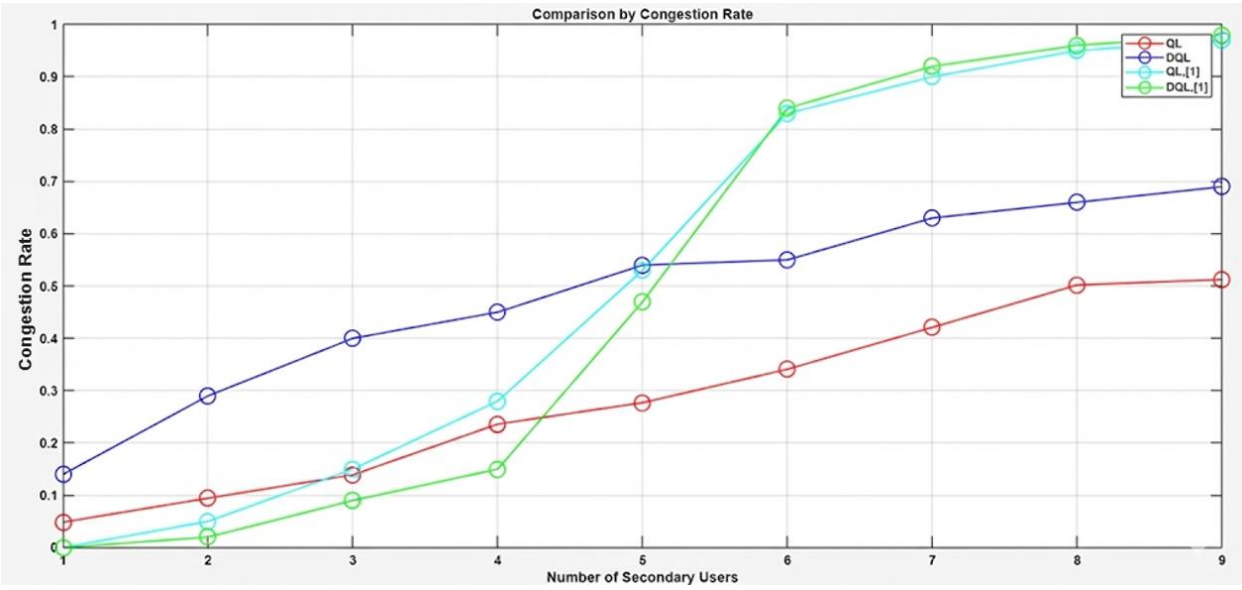
في هذه الفقرة ستتم المقارنة بين نتائج كلّاً من Q-Learning وDeep Q-Learning في هذا البحث وبين النتائج في [1] وفق المعايير الثلاثة المتّبعة:

1-4-5 معيار MOS



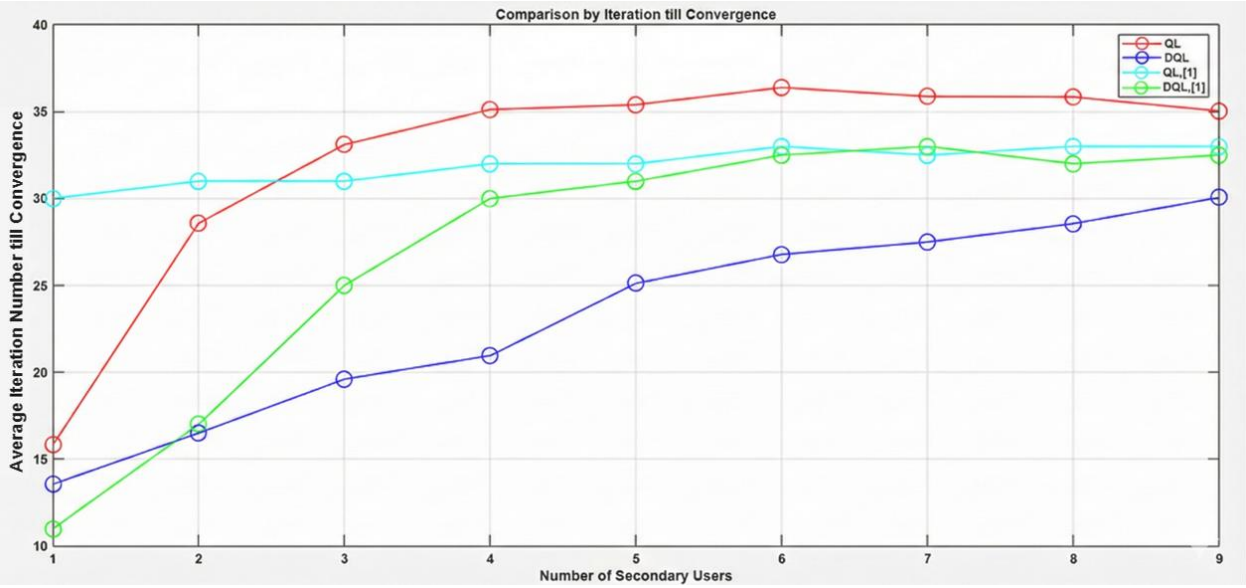
الشكل 22: مقارنة النتائج من حيث معيار MOS.

2-4-5 معدّل الازدحام Congestion Rate



الشكل 23: مقارنة النتائج من حيث معدّل الازدحام.

3-4-5 عدد التكرارات اللازمة للوصول إلى التقارب



الشكل 24: مقارنة بين QL وDQL من حيث عدد التكرارات للوصول إلى التقارب.

نلاحظ من الأشكال السابقة (22, 23, 24) التفوّق الواضح لخوارزميّة DQL في النموذج المقترح بفارق كبير من حيث قيم MOS التي تتراوح بين 3.86 و4.8، وعدد التكرارات اللازمة للوصول إلى التقارب التي تتراوح بين 13.5 و30 تكرار. وهذا منطقي لأنّ استخدام الشبكة العصبونيّة يزيد من سرعة الوصول إلى التقارب وإمكانيّة اختيار الأفعال الأكثر كفاءةً. أمّا من ناحية معدّل الازدحام أبدت الخوارزميّات في النموذج المرجعي أداءً أفضل من المقترحة عند أعداد مستخدمين ثانويين أصغر من 5، ولكن عند عدد مستخدمين ثانويين أكبر من 5 فتفوّقت خوارزميّة QL في النموذج المقترح.

في النهاية يمكن القول بأنّ نموذج Deep Q-learning وحيد العميل أبدى نجاحاً أكبر من النماذج الأخرى الموجودة في الأدبيّات. إنّ الاختيار بين الخوارزميّتين يعتمد على شروط المسألة، فإذا كان الهدف هو قيمة MOS وعدد التكرارات فخوارزميّة DQL هي التي ستعطي أفضل أداء، أمّا إذا كان الهدف هو معدّل الازدحام فخوارزميّة QL هي الحلّ. وهذا ما يحدّد أي الخوارزميّتين سيتم اختيارهما.

5-5 خاتمة

عرض هذا الفصل نتائج تطبيق خوارزميّة Q-Learning ودرس تماسك النموذج، ونتائج تطبيق خوارزميّة Deep Q-Learning، ومقارنة النتائج مع النتائج الموجودة في المرجع الذي استخدم نفس البارامترات وفق معيار MOS ومعدّل الازدحام Congestion Rate وعدد التكرارات اللازمة للوصول إلى التّقارب Number of Iteration till Convergence.

الخاتمة والآفاق المستقبليّة

قدّم هذا البحث نموذج تحكم مركزي يعتمد على التعلم المعزز RL لحل مشكلة النفاذ الديناميكي للطيف DSA في شبكات الراديو الإدراكي من نمط Underlay. استُخدمت المحطّة القاعدية الثانوية SBS كعميل وحيد (محرك إدراكي) لاتخاذ قرارات تخصيص عتبات SINR. تم تطبيق ومقارنة خوارزميتي Q-Learning وDeep Q-Learning بهدف تعظيم جودة التجربة وفق مقياس MOS للمستخدمين الثانويين مع ضمان شروط التداخل. أظهرت نتائج المحاكاة أن النموذج المقترح DQL حقّق تفوّقاً واضحاً مقارنةً بخوارزمية QL التقليدية والنماذج المرجعية في [1] خاصةً في قيم MOS المحققة وفي سرعة التقارب، حيث قلّص عدد التكرارات اللازمة للوصول للحل بشكل ملحوظ. كما أبدت خوارزمية QL المقترحة تحسّناً في معدل الازدحام مقارنة بالمرجع [1] عند ازدياد عدد المستخدمين. يثبت هذا النموذج المركزي فعاليّته في تبسيط المشكلة وتجنّب التعقيدات المرتبطة بتضارب السياسات التي قد تحدث في النماذج متعددة العملاء. اقتصر البحث الحالي على استخدام فضاء أفعال متقطع (عتبات SINR المحددة مسبقاً).

كآفاق مستقبلية يُؤمل تطوير هذا العمل ليتعامل مع فضاء أفعال مستمر، مثل تخصيص استطاعة الإرسال بشكل دقيق. على سبيل المثال يمكن استخدام خوارزميات Actor-Critic المتقدمة مثل DDPG، والتي تتيح التعامل مع قيم الاستطاعة المستمرّة ممّا قد يؤدّي إلى تحسين إضافي في الأداء. كما يمكن تحسين النموذج للبيئات عالية الحركية High Mobility، فالنموذج الحالي افترض قناة شبه ثابتة، ولكن يمكن أن يركّز العمل المستقبلي على نمذجة مستخدمين ثانويين SUs ذوي حركيّة عالية وقنوات سريعة التغير Fast-Fading. سيتطلّب هذا من الخوارزميات التي ستُستخدم أن تكون قادرة على اتخاذ قرارات فعّالة في بيئة شديدة الديناميكية.

المراجع

[1]: Mishra, Nikita, Sumit Srivastava, and Shivendra Nath Sharan. "**Raddpg: Resource allocation in cognitive radio with deep reinforcement learning**". 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS). IEEE, 2021.

[2]: Shah-Mohammadi, Fatemeh, and Andres Kwasinski. "**Deep reinforcement learning approach to QoE-driven resource allocation for spectrum underlay in cognitive radio networks**". 2018 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2018.

[3]: Albinsaid, Hasan, et al. "**Multi-agent reinforcement learning-based distributed dynamic spectrum access**". IEEE Transactions on Cognitive Communications and Networking 8.2 (2021): 1174-1185.

[4]: Nguyen, Khoi Khac, et al. "**Distributed deep deterministic policy gradient for power allocation control in D2D-based V2V communications**". IEEE Access 7 (2019): 164533-164543.

[5]: Lopez-Ramos, Luis M., Antonio G. Marques, and Javier Ramos. "**Joint sensing and resource allocation for underlay cognitive radios**". 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

[6]: Mohammadi, Fatemeh Shah, and Andres Kwasinski. "**QoE-driven integrated heterogeneous traffic resource allocation based on cooperative learning for 5G cognitive radio networks**". 2018 IEEE 5G World Forum (5GWF). IEEE, 2018.

[7]: Muteba, K. F., Karim Djouani, and Thomas O. Olwal. "**Deep reinforcement learning based resource allocation for narrowband cognitive radio-IoT systems**". Procedia Computer Science 175 (2020): 315-324.

[8]: Richard S. Sutton and Andrew G. Barto. "**Reinforcement Learning An Introduction**". 2018.

[9]: Nair, Arun, et al. "**Massively parallel methods for deep reinforcement learning**". arXiv preprint arXiv:1507.04296 (2015).

[10]: Kwasinski, Andres, Wenbo Wang, and Fatemeh Shah Mohammadi. "**Reinforcement learning for resource allocation in cognitive radio networks**". Machine Learning for Future Wireless Communications (2020):27-44.

[11]: Zhao, Qing, and Brian M. Sadler. "**A survey of dynamic spectrum access**". IEEE signal processing magazine 24.3 (2007): 79-89.

[12]: Mohammadi, Fatemeh Shah. “**Machine Learning-enabled Resource Allocation for Underlay Cognitive Radio Networks**”. Rochester Institute of Technology, 2020.

[13]: Stevenson, Carl R., et al. "**IEEE 802.22: The first cognitive radio wireless regional area network standard**". IEEE communications magazine 47.1 (2009): 130-138.

[14]: Granelli, Fabrizio, et al. "**Standardization and research in cognitive and dynamic spectrum access networks: IEEE SCC41 efforts and other activities**". IEEE Communications Magazine 48.1 (2010): 71-79.

[15]: Flores, Adriana B., et al. "**IEEE 802.11 af: A standard for TV white space spectrum sharing**". IEEE Communications Magazine 51.10 (2013): 92-100.

[16]: Yoshino, Hitoshi. "**ITU-R standardization activities on cognitive radio**". IEICE transactions on communications 95.4 (2012): 1036-1043.

[17]: Mueck, Markus Dominik, Srikathyayani Srikanteswara, and Biljana Badic. "**Spectrum sharing: Licensed shared access (LSA) and spectrum access system (SAS)**". Intel White Paper (2015): 1-26.

[18]: Sebastianelli, Alessandro, et al. "**A Deep Q-Learning based approach applied to the Snake game**" 2021 29th Mediterranean Conference on Control and Automation (MED). IEEE, 2021.

[19]: Piamrat, Kandaraj, et al. "**Quality of experience measurements for video streaming over wireless networks**" 2009 Sixth International Conference on Information Technology: New Generations. IEEE, 2009.

[20]: Khan, Shoaib, et al. "**MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication**" Advances in Multimedia 2007.1 (2007): 094918.

ملخّص

يُعدّ الراديو الإدراكي CR التقنية التمكينية الرئيسية لحل مشكلة ندرة الطيف الترددي الناتجة عن سياسات التخصيص الثابت. تركّز هذه الأطروحة على النفاذ الديناميكي للطيف DSA من نمط Underlay، الذي يسمح للمستخدمين الثانويين SUs بالنفاذ للطيف بالتزامن مع المستخدمين الأساسيين PUs بشرط الالتزام الصارم بقيود التداخل. ويشكّل تحدي تخصيص الموارد بكفاءة لضمان الأداء الأفضل في ظلّ هذه القيود المعقّدة المشكلةَ الأساسية التي يعالجها هذا البحث.

تتمثل المسألة الأساسية في تجاوز مقاييس الأداء التقليدية والانتقال إلى تعظيم جودة التجربة QoE للمستخدم النهائي، والتي تُقاس كمّياً بـ متوسط درجة الرأي MOS. لمواجهة الطبيعة المتغيّرة للبيئة الراديوية، تم اعتماد منهجية التعلم المعزز RL لقدرتها الفائقة على التعلّم واتخاذ القرار. تمّ في هذا البحث تطبيق ومقارنة خوارزميتين محوريّتين: خوارزمية Q-Learning (QL) التقليدية، وخوارزمية Deep Q-Learning (DQL) التي تدمج الشبكات العصبونية العميقة للتعامل مع حالات تتطلّب سرعة في الوصول إلى الحلّ.

تمّ في هذه الأطروحة اقتراح نموذج تحكّم مركزي تكون فيه المحطة القاعدية الثانوية SBS بمثابة عميل ذكي وحيد Single Agent يتّخذ قرارات التحكّم. هذا العميل مسؤول عن تخصيص عتبات نسبة الإشارة إلى التداخل والضجيج SINR لجميع المستخدمين الثانويين بهدف تعظيم قيمة MOS الإجمالية مع ضمان احترام قيود التداخل والتشوه. أظهرت نتائج المحاكاة والمقارنة مع الأدبيات التفوق الواضح لنموذج DQL المقترح، حيث حقق قيم MOS أعلى (تراوحت بين 4.8 و3.86) وسرعة تقارب أكبر (بين 13.5 و30 تكراراً). وفي المقابل أظهرت خوارزمية QL المقترحة معدّل ازدحام Congestion Rate أقلّ في السيناريوهات ذات الأعداد الكبيرة من المستخدمين (أكثر من 5)، مما يوضّح أن نموذج DQL هو الأنسب لتعظيم جودة التجربة وسرعة الاستجابة.

Abstract

Cognitive Radio (CR) is the main enabling technology to solve the problem of spectrum scarcity resulting from static allocation policies. This thesis focuses on Dynamic Spectrum Access (DSA) of the Underlay type, which allows Secondary Users (SUs) to access the spectrum concurrently with Primary Users (PUs) under the condition of strict adherence to interference constraints. The challenge of efficiently allocating resources to ensure optimal performance under these complex constraints constitutes the primary problem addressed by this research.

The core issue is to move beyond traditional performance metrics and shift towards maximizing the Quality of Experience (QoE) for the end-user, which is quantitatively measured by the Mean Opinion Score (MOS). To cope with the changing nature of the radio environment, the Reinforcement Learning (RL) methodology was adopted for its superior ability to learn and make decisions. In this research, two pivotal algorithms were applied and compared: the traditional Q-Learning (QL) algorithm, and the Deep Q-Learning (DQL) algorithm, which integrates deep neural networks to handle situations requiring speed in reaching the solution.

In this thesis, a central control model was proposed in which the Secondary Base Station (SBS) acts as a single intelligent agent (Single Agent) that makes control decisions. This agent is responsible for allocating Signal to Interference and Noise Ratio (SINR) thresholds for all secondary users with the aim of maximizing the total MOS value while ensuring respect for interference and distortion constraints. The simulation results and comparison with literature showed the clear superiority of the proposed DQL model, as it achieved higher MOS values (ranging between 4.8 and 3.86) and faster convergence speed (between 13.5 and 30 iterations). In contrast, the proposed QL algorithm showed a lower Congestion Rate in scenarios with a large number of users (more than 5), which clarifies that the DQL model is the most suitable for maximizing Quality of Experience and response speed.